PQStat Software
Statistical Computation Software

User Manual - PQStat

Barbara Więckowska

Copyright © 2010-2026 U-smile Stats

All rights reserved

To the version 2.0.0

Publisher: U-smile Stats, Poznań
Publication date: 8 October 2026
ISBN 978-83-983915-1-1

www.pqstat.pl

Spis treści

1 SYSTEM REQUIREMENTS 8

2 INSTALLATION 8

3 SETTINGS 9

4 AI TOOLS 13

5 AI: RESULTS INTERPRETER 13

6 DATA QUALITY CHECK 17

7 PROGRAM OPERATION 20

7.1 WORKING WITH DATASHEETS . . . 22

7.1.1 ADDING, REMOVING AND EXPORTING DATASHEETS . . . 22

7.1.2 INPUTTING DATA INTO A DATASHEET . . . 22

7.1.3 DATASHEET WINDOW . . . 24

7.1.4 VARIABLE PROPERTIES . . . 25

7.1.5 EDITING DATA . . . 27

7.1.6 SORTING DATA . . . 29

7.1.7 CONVERTING RAW DATA INTO A CONTINGENCY TABLE . . . 30

7.1.8 CONVERTING A CONTINGENCY TABLE INTO RAW DATA . . . 30

7.1.9 RESHAPING DATA BETWEEN THE WIDE AND THE LONG FORMAT . . . 31

7.1.10 FORMULAS . . . 34

7.1.11 DATA GENERATION . . . 38

7.1.12 MISSING DATA . . . 39

7.1.13 TRANSFORMATIONS . . . 42

7.1.14 DIRECT AND INDIRECT STANDARDIZATION . . . 45

7.1.15 SAMPLING SIMULATION . . . 47

7.1.16 SIMILARITY MATRIX . . . 50

7.2 WORKING WITH RESULTS SHEETS (REPORTS) . . . 60

7.3 MENU AND LANGUAGE SETTINGS . . . 61

7.3.1 LANGUAGE . . . 61

7.3.2 MENU . . . 61

8 ORGANIZATION OF WORK WITH THE PROGRAM 63

8.1 FORM OF DATA ORGANIZATION . . . 63

8.2 DATASHEET WORKING AREA . . . 65
8.2.1 Organising variables into sets . . . 65
8.2.2 Case activation/deactivation . . . 66
8.2.3 Selecting an area . . . 67
8.2.4 Saved selection . . . 68
8.2.5 Data filter . . . 69
8.3 RUNNING ANALYSES MULTIPLE TIMES . . . 72
8.4 ORGANISING REPORTS INTO SETS . . . 72
8.5 INFORMATION RETURNED TO THE REPORT . . . 73
8.6 MARKING STATISTICALLY SIGNIFICANT RESULTS . . . 73

9 PLOTS 74

9.1 PLOT GALLERY . . . 74
9.1.1 Column plots . . . 75
9.1.2 Estimator error plots . . . 80
9.1.3 Box-and-whisker plots . . . 81
9.1.4 Violin plots . . . 83
9.1.5 Point plots . . . 84
9.1.6 Line plots . . . 86
9.2 LOCAL LINEAR SMOOTHING TECHNIQUES . . . 87
9.2.1 LOWESS . . . 87
9.2.2 Kernel smoothing . . . 87
9.3 KERNEL ESTIMATION . . . 91
9.3.1 One-dimensional kernel estimator . . . 91
9.3.2 Bland-Altman plot . . . 95
9.4 Correlation matrices . . . 99

10 TEST POWER AND SAMPLE SIZE 108

10.1 Sample size determination . . . 108
10.2 Power and sample size for test . . . 111
10.2.1 Single-sample t-test . . . 113
10.2.2 T-test for dependent groups . . . 116
10.2.3 T-test for independent groups . . . 118
10.2.4 Chi-square test for single sample variance . . . 119
10.2.5 Chi-square test of two variances, Fisher-Snedecor . . . 121
10.2.6 Chi-square test of goodness of fit . . . 122
10.2.7 Chi-square test (RxC) . . . 123

Copyright ©2010-2026 U-smile Stats – All rights reserved

10.2.8 Test for two independent proportions, chi-square (2x2) . . . 125
10.2.9 One-way ANOVA for independent groups . . . 126
10.2.10 Test for one proportion . . . 128

11 DESCRIPTIVE ANALYSES **129**
11.1 Measurement scales . . . 129
11.2 Tables . . . 131
11.2.1 Frequency tables and empirical distributions of data . . . 131
11.2.2 Table report . . . 136
11.2.3 Analyses for contingency tables . . . 141
11.3 Descriptive statistics . . . 145
11.3.1 Measures of central tendency . . . 146
11.3.2 Other measures of location . . . 147
11.3.3 Measures of dispersion . . . 148
11.3.4 Other distribution attributes . . . 149
11.4 Descriptive summaries . . . 151

12 PROBABILITY DISTRIBUTIONS **157**
12.1 CONTINUOUS PROBABILITY DISTRIBUTIONS . . . 159
12.2 PROBABILITY DISTRIBUTION CALCULATOR . . . 162

13 Hypothesis testing **166**

14 NORMALITY DISTRIBUTION TESTS **169**
14.0.1 One-dimensional normality tests . . . 169
14.0.2 Multivariate normality tests . . . 178

15 COMPARISON - 1 GROUP **183**
15.1 PARAMETRIC TESTS . . . 184
15.1.1 The single-sample t-test . . . 184
15.1.2 The single-sample chi-square test for a population variance . . . 187
15.2 NON-PARAMETRIC TESTS . . . 191
15.2.1 The Wilcoxon test (signed-ranks) . . . 191
15.2.2 The chi-square goodness-of-fit test . . . 194
15.2.3 Tests for one proportion . . . 197

16 COMPARISON - TWO GROUPS **202**
16.1 PARAMETRIC TESTS . . . 203
16.1.1 The Fisher-Snedecor test . . . 203

Copyright ©2010-2026 U-smile Stats – All rights reserved

16.1.2 The t-test for independent groups . . . 205
16.1.3 The t-test with the Cochran-Cox adjustment . . . 205
16.1.4 The t-test for dependent groups . . . 210
16.2 NON-PARAMETRIC TESTS . . . 214
16.2.1 The Mann-Whitney U test . . . 214
16.2.2 The Wilcoxon test (matched-pairs) . . . 218
16.2.3 The Chi-square tests . . . 222
16.2.4 The Chi-square test for large tables . . . 223
16.2.5 The Fisher test for large tables . . . 226
16.2.6 The Chi-square test for small tables . . . 230
16.2.7 The Chi-square test corrections for small tables . . . 233
16.2.8 The Chi-square test for trend . . . 234
16.2.9 The Relative Risk and the Odds Ratio . . . 239
16.2.10 The Z test for two independent proportions . . . 241
16.2.11 The McNemar test, the Bowker test of internal symmetry . . . 246
16.2.12 The Z test for two dependent proportions . . . 252

17 COMPARISON - MORE THAN TWO GROUPS 256
17.1 PARAMETRIC TESTS . . . 257
17.1.1 The ANOVA for independent groups . . . 257
17.1.2 Contrasts and POST-HOC tests . . . 261
17.1.3 The ANOVA for independent groups with F^* and F'' corrections . . . 267
17.1.4 The Brown-Forsythe test and the Levene test . . . 274
17.1.5 The ANOVA for dependent groups . . . 275
17.1.6 Mauchly's sphericity . . . 280
17.1.7 The ANOVA for dependent groups with Epsilon correction and MANOVA . . . 281
17.2 NON-PARAMETRIC TESTS . . . 288
17.2.1 The Kruskal-Wallis ANOVA . . . 288
17.2.2 The Jonckheere-Terpstra test for trend . . . 295
17.2.3 The Conover ranks test of variance . . . 298
17.2.4 The Friedman ANOVA . . . 302
17.2.5 The Page test for trend . . . 309
17.2.6 The Durbin's ANOVA (missing data) . . . 311
17.2.7 The Skillings-Mack ANOVA (missing data) . . . 317
17.2.8 The Chi-square test for multidimensional contingency tables . . . 321
17.2.9 The Q-Cochran ANOVA . . . 323

Copyright ©2010-2026 U-smile Stats – All rights reserved

18 MULTIPLE COMPARISONS **328**

19 HOTELLING T-SQUARE **332**

19.0.1 Hotelling's T-square test for a single sample . . . 332
19.0.2 Hotelling's T-square test for dependent groups . . . 336
19.0.3 Hotelling's T-square test for independent groups . . . 340
19.0.4 Hotelling's T-square test for independent groups with correction for unequal variances . . . 345
19.0.5 Box's test for equality of covariance matrices . . . 347

20 UNIVARIATE MANOVA **349**

20.0.1 One-way MANOVA for independent groups . . . 349

21 ANALYSIS FOR STRATAS **356**

21.1 The Mantel-Haenszel method for several $2x2$ tables . . . 356
21.1.1 The Mantel-Haenszel odds ratio . . . 357
21.1.2 The Mantel-Haenszel relative risk . . . 361

22 CORRELATION **364**

22.1 PARAMETRIC TESTS . . . 365
22.1.1 The linear correlation coefficients . . . 365
22.1.2 The Pearson correlation coefficient significance . . . 366
22.1.3 The slope coefficient significance . . . 370
22.1.4 Prediction . . . 371
22.1.5 Comparison of correlation coefficients . . . 371
22.1.6 Comparison of the slope of regression lines . . . 373
22.2 NON-PARAMETRIC TESTS . . . 375
22.2.1 The monotonic correlation coefficients . . . 375
22.2.2 The t-test for the significance of the Spearman rank-order correlation coefficient . . . 376
22.2.3 The test for the significance of the Kendall's tau correlation coefficient . . . 379
22.2.4 Contingency table coefficients and their statistical significance . . . 381

23 AGREEMENT ANALYSIS **386**

23.1 PARAMETRIC TESTS . . . 387
23.1.1 The intraclass correlation coefficient and a test to examine its significance . . . 387
23.2 NON-PARAMETRIC TESTS . . . 394
23.2.1 The Kendall's concordance coefficient and a test to examine its significance . . . 394
23.2.2 The Cohen's Kappa coefficient and a test to examine its significance . . . 397
23.2.3 The Fleiss' Kappa coefficient and a test to examine its significance . . . 403

Copyright ©2010-2026 U-smile Stats – All rights reserved

24 DIAGNOSTIC TESTS **409**

24.0.1 Evaluation of a diagnostic test . . . 409
24.0.2 The ROC curve . . . 414
24.0.3 Selection of the optimal cut-off . . . 416
24.0.4 ROC curves comparison . . . 422

25 MULTIDIMENSIONAL MODELS **429**

25.0.1 Matching groups . . . 429
25.0.2 Preparation of variables for analysis in multivariate models . . . 441
25.1 MULTIPLE LINEAR REGRESSION . . . 445
25.1.1 Multiple linear regression . . . 445
25.1.2 Model-based prediction and test set validation . . . 461
25.1.3 Validation and prediction for multiple regression . . . 461
25.2 COMPARISON OF MULTIPLE LINEAR REGRESSION MODELS . . . 465
25.2.1 Comparing multiple regression models . . . 465
25.3 LOGISTIC REGRESSION . . . 474
25.3.1 Logistic regression . . . 474
25.3.2 Model-based prediction and test set validation . . . 498
25.3.3 Validation and prediction for logistic regression . . . 498
25.4 COMPARISON OF LOGISTIC REGRESSION MODELS . . . 506
25.4.1 Comparing logistic regression models . . . 506
25.4.2 Factorial ANOVA – GLM . . . 523
25.4.3 ANCOVA . . . 536
25.4.4 Mediation effect . . . 544

26 MIXED MODELS **549**

26.0.1 Introduction: dependent data, fixed effects and random effects . . . 549
26.1 MIXED MULTIPLE REGRESSION . . . 553
26.1.1 Mixed multiple regression (linear mixed model) . . . 553
26.1.2 Validation and prediction for mixed multiple regression . . . 568
26.1.3 Comparing mixed multiple regression models . . . 570
26.2 MIXED LOGISTIC REGRESSION . . . 575
26.2.1 Mixed logistic regression (generalized linear mixed model) . . . 575
26.2.2 Validation and prediction for mixed logistic regression . . . 584
26.2.3 Comparing mixed logistic regression models . . . 586
26.3 COX PH REGRESSION WITH RANDOM EFFECT . . . 591
26.3.1 Cox PH regression with random effect (frailty) . . . 591

Copyright ©2010-2026 U-smile Stats – All rights reserved

26.3.2 Comparing Cox PH regression models with random effect . . . 599

27 DIMENSION REDUCTION AND GROUPING 603

27.0.1 Principal component analysis . . . 603

27.1 CLUSTER ANALYSIS . . . 615

27.1.1 Hierarchical methods . . . 615

27.1.2 The k-means method . . . 621

28 SURVIVAL ANALYSIS 627

28.0.1 Life tables . . . 628

28.0.2 Kaplan-Meier curves . . . 632

28.0.3 Comparison of survival curves . . . 635

28.0.4 Cox proportional hazards regression . . . 644

28.0.5 Comparison of Cox PH regression models . . . 648

28.0.6 Cox regression with time-dependent variables . . . 658

28.0.7 Competing risks analysis – the Fine-Gray model . . . 664

29 META-ANALYSIS 671

29.0.1 Introduction to meta-analysis and data preparation . . . 671

29.0.2 Summary effect – fixed effect and random effect . . . 676

29.0.3 Heterogeneity testing . . . 680

29.0.4 Sensitivity testing . . . 682

29.0.5 Asymmetry testing (publication bias) . . . 684

29.0.6 Cumulative meta-analysis . . . 688

29.0.7 Group comparison . . . 690

29.0.8 Meta-regression . . . 700

30 RELIABILITY ANALYSIS 706

30.0.1 Scale reliability . . . 706

31 TEST SUMMARIES 713

32 THE WIZARD 722

33 OTHER NOTES 723

33.1 FILE FORMATS . . . 723

Copyright ©2010-2026 U-smile Stats – All rights reserved

1 SYSTEM REQUIREMENTS

To use PQStat, your computer must meet the following minimum requirements:

- Processor: Intel Pentium II 500MHz
- RAM: 256MB
- Graphics Card: SVGA 800x600
- Storage: 0.2GB
- If using CD version: CD-ROM
- Other : mouse and keyboard
- Microsoft Windows 10 or 11 (64-bit)

2 INSTALLATION

To start the installaton process, run the applicaton installer - PQStat-setup_x86-FULL (for the 64-bit version run PQStat-setup_x64-FULL.exe).

After doing so, a setup dialog box will appear. Press “Next” to continue. Further installation requires you to accept the Terms of Service. If you accept, select: ”I accept the terms of service” and press “Next”. Otherwise, select ”I do not accept the terms of service” and press : “Cancel” to abort the installation.

The following window will give you an option to change the default installation directory. Information about available disc space will also be displayed. Choosing the default installation directory is recommended.

Pressing “Next” will give you a choice of either a full installation or a version not including exemplary data sets. The data sets are used in the user guide.

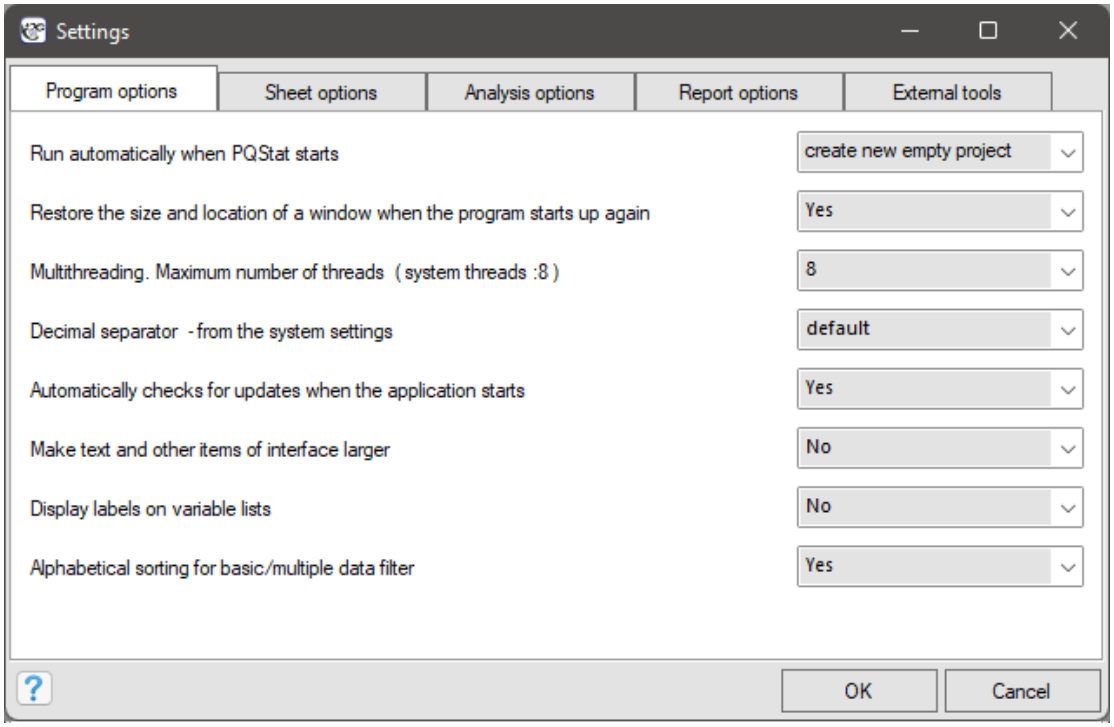
In the next section, you will be given a chance to change the shortcut name, which will be created in the Windows Start Menu.

After pressing “Next”, you will have an option to create a Desktop Shoutcut or add a shoutcut to the Quick Lunch toolbar. Press “Next” to continue.

The following window it the last one before the installation process begins. It displays a summary of installation options chosen so far. Pressing ''Instal'' will start the installation process.

Copyright ©2010-2026 U-smile Stats – All rights reserved

3 SETTINGS



General settings regarding program options, data sheets, carrying out analysis, returned reports and use of external tools can be changed by selecting menu Edit→Settings.

Program options

- Run automatically when PQStat starts - allows for one of three things to happen when starting the program : open a new project, open a recently used project, or do nothing.
- Restore the size end location of a window when the program starts up again - allows you to start the program so that its window is in the same place and is of the same size of the last window displayed immediately before closing the program.
- Multithreading. Maximum number of threads (system threads: 8) - provides the ability to execute tasks in multiple threads simultaneously (up to a maximum of eight threads).
- Decimal separator - from the system settings - can be set as comma or period, or by default chosen according to the settings of the operating system the program is running on.
- Automatiocally checks for updates when the aplication starts - gives you the option to check for and consequently install updates or to opt out of this service.
- Make tekst and other items of interface larger - allows you to increase the font size and the size of other interface elements, which will reduce the amount of information displayed but may make it easier to read.
- Display Labels on variable lists - Gives the option to display the labels of the variables (long names of the variables) in the variables selection lists or as a hint to the names displayed in those lists. This option will have the desired effect if the names of the variables have assigned labels.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Sheet options

- Maximum number of undo steps in sheet - the number of possible actions you can undo in the worksheet - it affects the speed of the program, so the recommended number is 10.
- The maximum number of cells to remember in one step - the number of worksheet cells, the simultaneous change of which can be considered as one action to be remembered by the programme. It affects the speed of the program therefore the recommended number of cells is 5000.
- Sheet font - allows you to set the font type that will be used to display the data contained in the datasheet.
- Action for duble-click column header - allows you to set what happens after double-clicking the column header. This can be: adjusting the columns width to fit the data inside, opening the Code/Label/Format window, opening the Inspection window.

Analysis options

- Default significance level for testing - gives the option to change the 0.05 significance level proposed as a standard in statistical analysis windows.
- MachineEpsilon - information about setting the Epsilon machine size or calculation precision (1E-16).
- Measure kurtosis - information about the setting of the kurtosis calculation method (g2, or Pearson's b2), where a normal distribution is characterized by value zero of kurtosis g2 or b2 close to the value of three.
- Sorting order for the contingency tables - gives you the option to sort rows and columns of a contingency table in ascending or descending order. This option not only affects the displayed results in tables and graphs, but also affects those analyses based on contingency tables for which the order of categories is important.
- Method to sort string values in alphabetical order - Allows you to sort data and results stored in text form naturally or alphabetically. Natural sort order is an improved alphabetical order, where multi-digit numbers are treated indivisibly, i.e. as if they were a single character. For example, in alphabetical sorting, "a11" will be sorted before "a2" because "1" is less than "2", whereas in natural sorting, "a2" is sorted before "a11" because "2" as less than "11"
- Labels for values. Show in analysis: - gives you the choice of displaying in values or labels corresponding to current values when using the filter or when the analysis requires setting individual values of the selected variable.

Report options

- Displays a value rounded to the specified number of decimal places - allows you to set the maximum number of decimal places reported for real numbers.
- Displays a percentage rounded to the specified number of decimal places - allows you to set the maximum number of decimal places reported percentages.
- Exponential notation for p-value - allows you to specify how the p-values of statistical tests are displayed in numeric form (with a defined number of decimal places) or in scientific notation.
- Color for p-value below the significance level - allows you to set the color that will be used to indicate statistically significant results at the set level.

Copyright ©2010-2026 U-smile Stats – All rights reserved

- Report name in the navigation tree - gives a choice of the type of information about executed analyses added to the name of analysis in navigation tree. It is possible to give only the name of the test or the name together with: the time of its execution or the description of the test, or the chosen filter, or the used grouping variable, or the names of the variables involved in the analysis.
- Show variable values as item labels - Gives the option to display the labels of the variables (long names of the variables) in analysis reports. This option will have the desired effect when the names of the variables have assigned labels.
- Report font - allows you to set the font that will be used to display the result descriptions included in the reports.
- Default plot size(width/height) - allows you to set the default number of pixels for the height and for the width of the graph.
- Default plot temlates - gives you the ability to pre-set chart options that are important to user.

External tools

- Path to gnuplot binary files - To be able to generate a 3D plot for the analysis of a nuclear 2D density estimator, the gnuplot program must be installed on the system.
- Path to IO module for SPSS Statistics - To be able to read sav files (IBM SPSS Statistics program data storage format), the IBM IO SPSS vendor module is required, so you have to download the $IO\_Module\_for\_SPSS\_Statistics\_20001.zip$. For more information, see

AI model source - PQStat can use a locally or remotely running language model (AI). Three independently stored model sources are available, and you can freely switch between them without losing the settings entered for the others:

– Local (this computer) - uses a model running on the same computer as PQStat. In the Local engine field we choose the program that serves the model:

· Ollama - requires filling in the API URL field (by default `http://localhost:11434`) and the Model field - the name of a model previously downloaded with the `ollama pull` command. The Refresh list button retrieves from the server the list of currently available (downloaded) models.

· LM Studio - uses the local LM Studio server, which serves the model through an OpenAI-compatible API (no key). LM Studio can be downloaded from `https://lmstudio.ai` (when PQStat does not detect the server, it offers to open this page). In LM Studio download a model, then enable the server: Settings → Local Model API → the Local API server switch (it should show Running); the Base URL field in the same window shows the address to enter in PQStat. With Just-in-time model loading enabled the model is loaded automatically at the first request. In PQStat enter the API URL (by default `http://localhost:1234/v1`) and select the Model - the Refresh list and Test connection buttons retrieve the list of models available in LM Studio. If the selected model supports images, check Model supports images (force).

The settings of both engines are remembered independently - switching the engine does not clear the address or model of the other one.

– Remote Ollama server - uses an Ollama server running on another computer, available on the local network or via VPN. In addition to the API URL and Model fields, you can provide an optional API key if the remote server is exposed behind a reverse proxy requiring authentication, and select Require valid certificate for encrypted (https) connections. The

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test connection button checks the availability of the server regardless of whether a model has already been selected, while Refresh list retrieves the list of models available on that server.

– Cloud (OpenAI-compatible API) - uses any external service providing an API compatible with the OpenAI standard (e.g. OpenAI, Google Gemini, LiteLLM API Gateway and other gateways of this type, including company proxy servers). Requires filling in the API URL (the drop-down list suggests the addresses of popular providers) and the API key (stored in encrypted form - the Show checkbox lets you temporarily view the entered value). After clicking Test connection, the program attempts to retrieve the list of available models from the server - if this succeeds, the Model field turns into a drop-down list to choose from. If a given service does not provide such a list in a standard way (e.g. Google's native API requires a different type of authentication for this than a plain key), the model name must be entered manually, and the connection test will then be performed as an actual request to the specified model. The Model supports images (force) checkbox lets you manually force the selected model to be treated as supporting images, when the program cannot detect this automatically.

AI settings for all users of the computer. By default the AI settings are stored separately for each Windows account. Checking Save as default for all users of this computer before clicking OK saves the selected model source (API URL, model, encrypted API key) to the file `C:\ProgramData\PQStat\AI\pqst`. These settings are used by every user of the computer who has not changed the AI settings in their own account. The computer settings are only defaults: a user can change them, and the changes apply to that user's account only. The label below the checkbox shows whether the computer settings or the user's own settings are in use; the Restore computer settings button returns to the defaults. An API key saved for the computer can be used by every user of that computer.

IT departments can set these values during installation by passing parameters to the installer: `/AIMODE=local|remote|cloud`, `/AIENGINE=ollama|lmstudio` (for `local`), `/AIURL=`*address*, `/AIMODEL=`*model*, `/AIKEY=`*key*, optionally `/AIVERIFYCERT=0` and `/AIVISION=1`, e.g.

```
PQStat-setup_x64-FULL.exe /VERYSILENT /AIMODE=remote /AIURL=https://llm.university.edu
/AIMODEL=gemma2:9b
```

Copyright ©2010-2026 U-smile Stats – All rights reserved

4 AI TOOLS

5 AI: RESULTS INTERPRETER

The AI Interpreter turns the results table of a chosen analysis into a written description, drafted in the style of the Results section of a scientific publication – ready to paste into a paper or report, or for further editing by the user. The description is generated by a language model (AI) based on the report contents, the type of test performed, and (optionally) the researcher's own guidance.

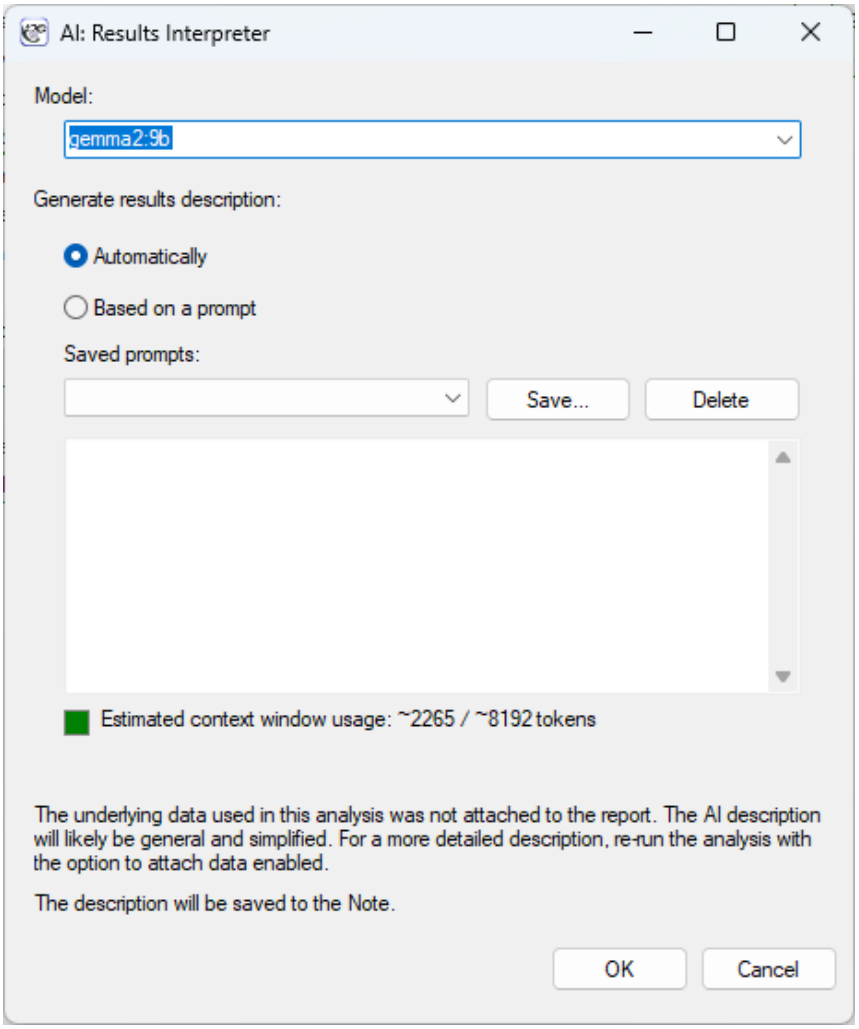
Note! This function requires a configured AI model source in Settings (the External tools tab) – a local Ollama server, a remote Ollama server, or a service compatible with the OpenAI API. See the SETTINGS chapter for details on configuring each source.

HOW TO RUN THE INTERPRETATION

The interpreter is invoked for the active report in one of three equivalent ways:

- the Help→AI: Results Interpreter menu,
- the report's context menu (right mouse button on the report) → AI: Results Interpreter,
- the button AI on the report toolbar.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Step 1. Make the report you want a description for the active one.

Step 2. Invoke the interpreter using one of the ways above.

Step 3. In the options window, choose an AI model (the one selected in Settings by default, but it can be changed for this one interpretation).

Step 4. Decide whether the description should be generated Automatically or Based on a prompt – i.e. with additional, custom guidance for the model (e.g. *"describe the result in one sentence"* or *"write it in Polish"*). A frequently used custom prompt can be saved under a chosen name with the Save... button and reused for the next analysis.

Step 5. If a chart is attached to the report and the selected model supports images, you can check Attach chart to the request so the model takes it into account in the description.

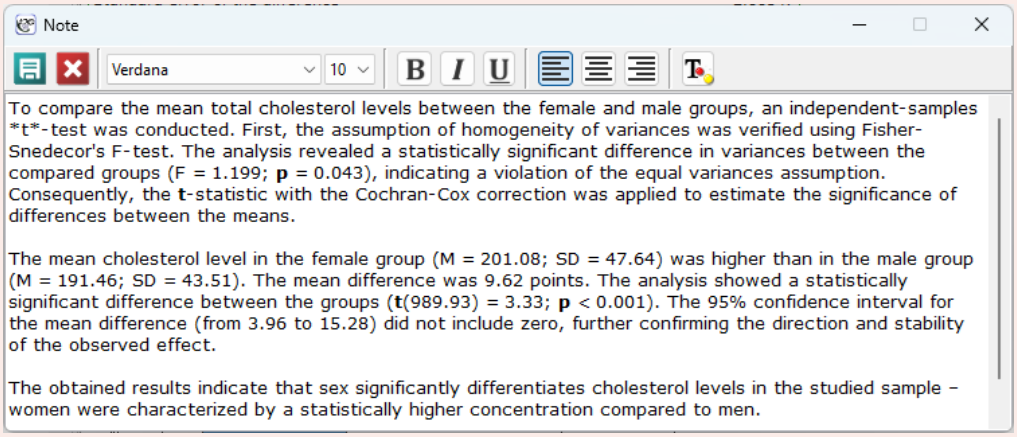
Step 6. Start the generation with the OK button.

Note! The bar below the model selection shows the estimated context-window usage of the selected model (green – comfortably within limits, orange – close to the limit, red – exceeded, which risks an incomplete or truncated response). Very large reports may require a model with a bigger context window.

PERFORMING AND INTERPRETING THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

EXAMPLE 5.1. After performing any statistical test (e.g. Student's t-test), we invoke the interpreter from the report's context menu and start the generation in Automatically mode.



The generated description is placed in the report's Note window (visible in the report or in the project tree, marked with a note icon), where it can be freely edited, supplemented, or corrected before it is finally saved.

BATCH INTERPRETATION OF MULTIPLE REPORTS

By selecting several items at once on the report list (the Batch view) and choosing AI: Results Interpreter from that list's context menu, descriptions for many reports can be generated at once, without an options window for each one individually – the interpretation then runs automatically, using the active model source settings from Settings.

CONVERSATION WITH THE MODEL FROM THE NOTE

The generated description need not be the end of the work with the model. In the report's Note window, below the note text field, there is a conversation panel: a field for the next question, the Ask button (or the Ctrl+Enter shortcut) and the New conversation button. The question is sent to the model together with the report and the whole conversation so far (the previous questions and answers), so the model "remembers" its previous answer – one can ask for an explanation of a fragment of the description, for shortening it, for a change of style, for a reference to a specific value in the table, etc. The answer is appended at the end of the note (preceded by the question text), where it can be edited just like the original description. The conversation history is saved in the project together with the report, so the conversation can be resumed after the file is reopened; the panel appears only in the notes of reports for which an AI description has been generated.

Because the whole conversation is sent to the model with every question, its length grows with each turn. Below the question field an indicator of the model's context window usage is shown (green, orange or red – as in the interpreter options window), taking into account the report, all previous questions and answers, and the question currently being typed. When the conversation no longer fits into the context window, the program warns before sending, and the New conversation button allows starting over: the model then receives only the report and the new question again, while the note text remains unchanged. Subsequent questions are directed to the active model source from Settings; a chart

Copyright ©2010-2026 U-smile Stats – All rights reserved

attached to the first request is not sent again.

FOR THOSE INTERESTED

Procedure: the description is generated from three elements combined into a single prompt passed to the model: (1) shared rules for drafting the result (Results-section style of a scientific publication, a ban on guessing values not present in the table, avoiding phrases such as ”we reject the null hypothesis”), (2) instructions specific to the particular type of test performed, and (3) the content of the results table itself, together with variable metadata (value labels), if available. The researcher’s own prompt – if given – takes precedence over the rules above, including with regard to the response language.

If the option to echo the raw data was attached to the report, only a sample (up to the first 10 rows) is passed in the prompt, so as not to overload the model’s context window.

The conversation from the note uses the same interfaces (Ollama, LM Studio, OpenAI-compatible API), which are stateless: the server does not remember previous requests, so with every question the program sends the first prompt (with the report) and all the previous question–answer pairs anew. The model ”sees” the earlier conversation only because it is attached to every subsequent request; hence the growing context window usage and the purpose of the New conversation button.

Copyright ©2010-2026 U-smile Stats – All rights reserved

6 DATA QUALITY CHECK

Before starting a statistical analysis, it is worth checking whether the loaded data contains errors that could distort the results – e.g. a numeric column that is in fact an encoded category, values clearly deviating from the rest, non-standard missing-data codes (e.g. -99) mixed in with real values, or single entries that do not fit the rest of the column (e.g. a negative value in an age column).

Data Quality Check scans all columns of the active spreadsheet, computes a concise statistical summary for each of them, and passes that summary to a language model (AI), which points out what looks suspicious in the data and why. The function does not modify the data – it only returns a descriptive list of observations for the user to review.

Note! This function requires a configured AI model source in Settings (the External tools tab) – a local Ollama server, a remote Ollama server, or a service compatible with the OpenAI API. See the SETTINGS chapter for details on configuring each source.

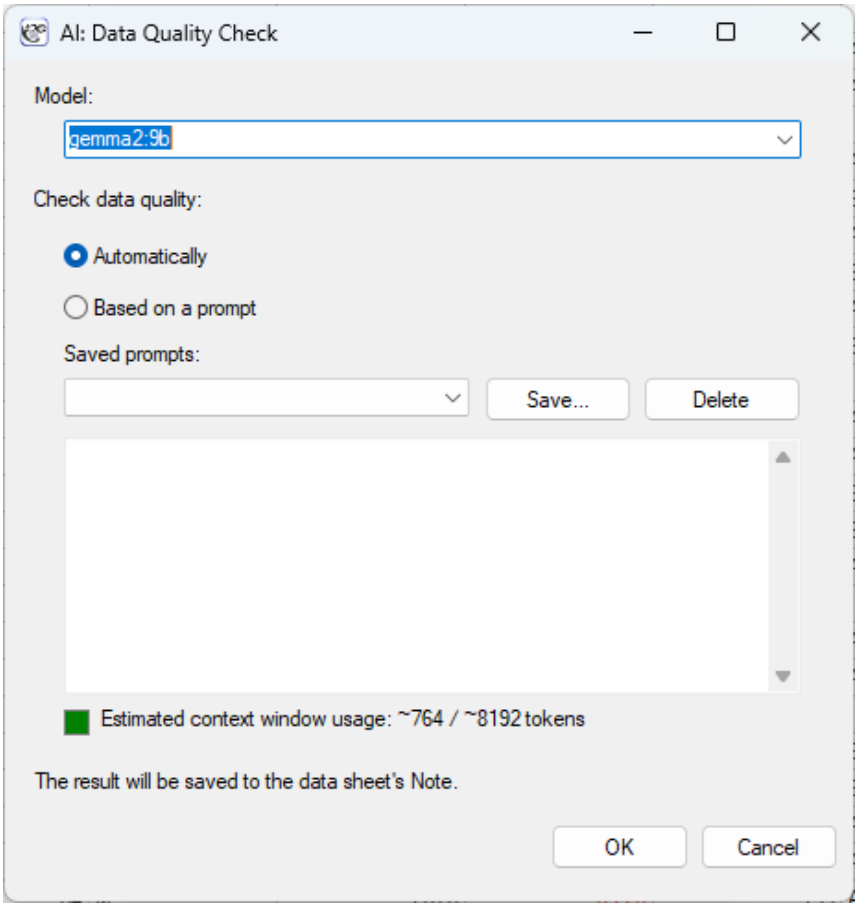
WHAT IS CHECKED

- **Numeric columns that look like an encoded category** – when a column contains only integers and the number of distinct values is small relative to the number of observations.
- **Outliers** – detected using the interquartile range (IQR) method: a value is considered an outlier if it lies below $Q_1 - 1.5 \cdot IQR$ or above $Q_3 + 1.5 \cdot IQR$.
- **Suspicious, repeated extreme values** – a value lying far outside the rest of the distribution that appears in the column more than once, which distinguishes it from a single, random deviation and suggests a non-standard missing-data code (e.g. -99, 999).
- **Plausibility of values given the column name** – e.g. whether an age column contains no negative values, or whether a column that looks like a yes/no question contains no other values. This kind of assessment requires understanding the meaning of the column name, so it is performed exclusively by the AI model based on the summary provided – it is not a rule checked directly by the program.

HOW TO RUN THE CHECK

The data quality check window is opened via the Data→Data Quality Check... menu.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Step 1. Make sure the data set you want to check is active in the spreadsheet.

Step 2. Select Data→Data Quality Check...

Step 3. In the options window, choose an AI model (or leave the default) and decide whether the description should be generated automatically or based on your own prompt with additional guidance.

Step 4. Start the analysis with the OK button.

Note! The function checks **all** non-empty columns of the active spreadsheet at once – there is no way to limit the check to selected columns.

PERFORMING AND INTERPRETING THE ANALYSIS

Example 6.1. (Data_Quality_Check.csv file)
The file contains 1000 rows (subjects) and 10 columns describing typical medical data: age, sex, height, weight, systolic blood pressure, blood type, presence of a chronic disease, number of visits, psychometric test score, and resting heart rate. Each column deliberately contains a few unusual values – the table below shows what kind of problem each of them represents. After loading the file, we select Data→Data Quality Check... and run the analysis with the default model.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Column	Type of problem	Example values
age	outlier / impossible values	-2, 0, 150, 199
sex	value outside the category set (F/M)	"3", "unknown"
height_cm	outliers (data-entry error)	17.0 (instead of 170), 1210.0
weight_kg	repeated suspicious code	-99 ($5\times$ in different rows)
systolic_bp	outliers	255, 38
blood_type	value outside the category set (0/A/B/AB)	"5", "Rh+"
chronic_disease	value outside the category set (yes/no)	"female", "2"
visit_count	value impossible for a count	-2, -1
psych_test_score	repeated suspicious code	999 ($4\times$ in different rows)
resting_heart_rate	outliers	2, 410

Expected check result

For each of the columns above, the check should point out exactly the kind of problem deliberately placed in it – e.g. for age: "the values -2, 0, 150 and 199 are inconsistent with the meaning of age"; for weight_kg: "the value -99 repeats several times and may be a non-standard missing-data code"; for chronic_disease: "the value *female* does not match the other yes/no values". The exact wording depends on the AI model used – the above is only the expected direction of the findings, not a literal response.

The result is saved to the data sheet's Note (visible in the project tree), where it can be freely edited or supplemented with your own comments before saving.

FOR THOSE INTERESTED

Procedure: For each column, the program independently computes: the number of observations, the number of missing values, the number of distinct values, and then – depending on the nature of the column – either numeric statistics (minimum, quartiles, maximum, outliers per the IQR rule) or a list of categories together with their counts. This concise summary (not the raw data) is what is sent to the AI model, together with an instruction to point out only those columns where something genuinely raises concern, and not to guess anything beyond what follows from the summary provided.

Note on performance: Because it is the summary that is analysed, not the full content of the spreadsheet, the size of the request sent to the model stays small even for spreadsheets with many rows – it grows with the number of **columns**, not the number of observations.

Copyright ©2010-2026 U-smile Stats – All rights reserved

7 PROGRAM OPERATION

Work in the PQStat program is based on projects - each project is a single file. **A project** works similarly to a workbook (notebook) and contains 3 basic elements:

1. Datasheets (including map sheets and matrices) - a single project can hold up to 1000 of them,
2. Result sheets (reports) - a single datasheet can produce up to 2000 of them,
3. Project manager - a window in which you rename datasheets and result sheets, add your own descriptions and notes, and export the contents of the project.

Up to 255 projects can be worked on simultaneously. The first project with an empty data sheet is opened automatically when you start the program, if this option is set in the Program Settings window.

Further projects can be created via the menu

- File→New project (Ctrl+N) 

- File→New datasheet (Ctrl+D) 

Created projects (files with pqs, pqx extensions) we open via:

- File→Open project (Ctrl+O) 

- File→Open examples 

- applies to examples that come together with the program,
- dragging the project file into the application window,
- double-clicking the project file.

The project can be saved via:

- menu File→Save project(Ctrl+S) 

- File→Save project as... 

- Save project button in the Project Manager.

Saving a project saves all project components to a file with the extension pqs or pqx.

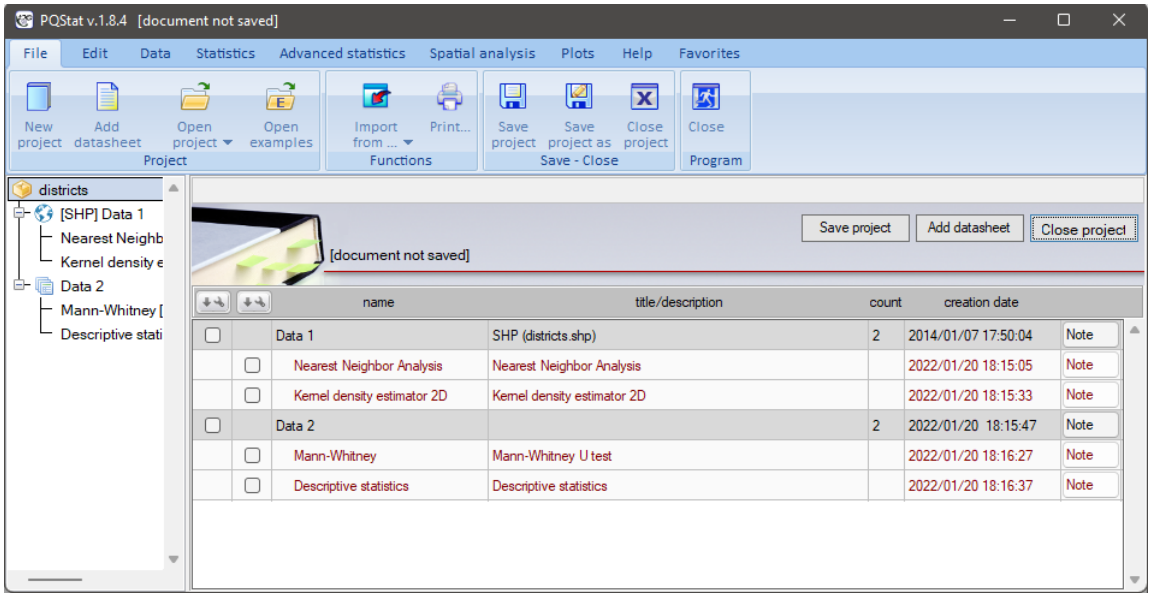
The project can be closed via:

- menu File→Close project 

- Close project button in the Project Manager.

For easy navigation, the Project Manager is displayed after selecting the appropriate project. In this window you can save as well as delete the selected project, add or delete a datasheet, delete a result reports as well as add notes. The project name is also the name of the project file (pqs/pqx).

Copyright ©2010-2026 U-smile Stats – All rights reserved



Copyright ©2010-2026 U-smile Stats – All rights reserved

7.1 WORKING WITH DATASHEETS

The most important part of any project is the datasheet. Every open project must have at least one.

ADDING, REMOVING AND EXPORTING DATASHEETS

The first blank datasheet is opened automatically with a new project.
Subsequent datasheets can be added to the project via:

- menu File→New datasheet (Ctrl+D) 

- New datasheet button in the Project Manager.

Datasheets can be deleted via:
- context menu Delete Sheet (Shift+Del) on the name of a datasheet in the Navigation tree,
-  button →Delete in the Project Manager, if a datasheet is selected. **Note!** If reports or a map are attached to a datasheet, deleting the datasheet also deletes all reports and the map assigned to it.

Datasheets can be described in the Project Manager by adding a name, title or a note.

Data sheets created in PQStat can be exported to csv (txt) , dbf and xls formats. Exporting data is done in the Project Manager via the  button→Seve Sheet to..., if a datasheet is selected.

INPUTTING DATA INTO A DATASHEET

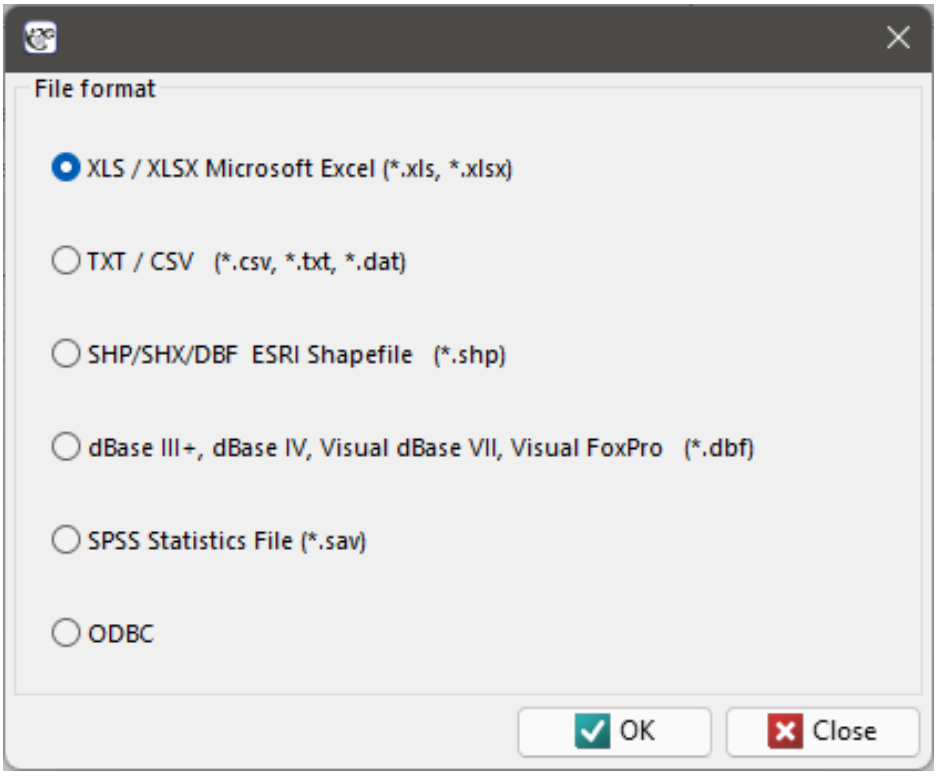
The datasheets are empty when created. The user enters data themselves, copies previously prepared data from another data sheet, or imports it. The amount of data a worksheet can hold is limited to 4 million rows and 1000 columns. Each cell can contain a maximum of 40 characters.

IMPORTING DATA

Data can be easily imported from files saved in formats such as:
- *.xls/*xlsx,
- *.txt/*.csv with internal character encoding UTF8, Windows-1250,
- *.shp (SHP/SHX/DBF ESRI Shapefile),
- *.dbf (dBase III, dBase IV, dBase VII),
- *.sav (SPSS),
- *.dbf (FoxPro).

To import click menu File→Import from ... 

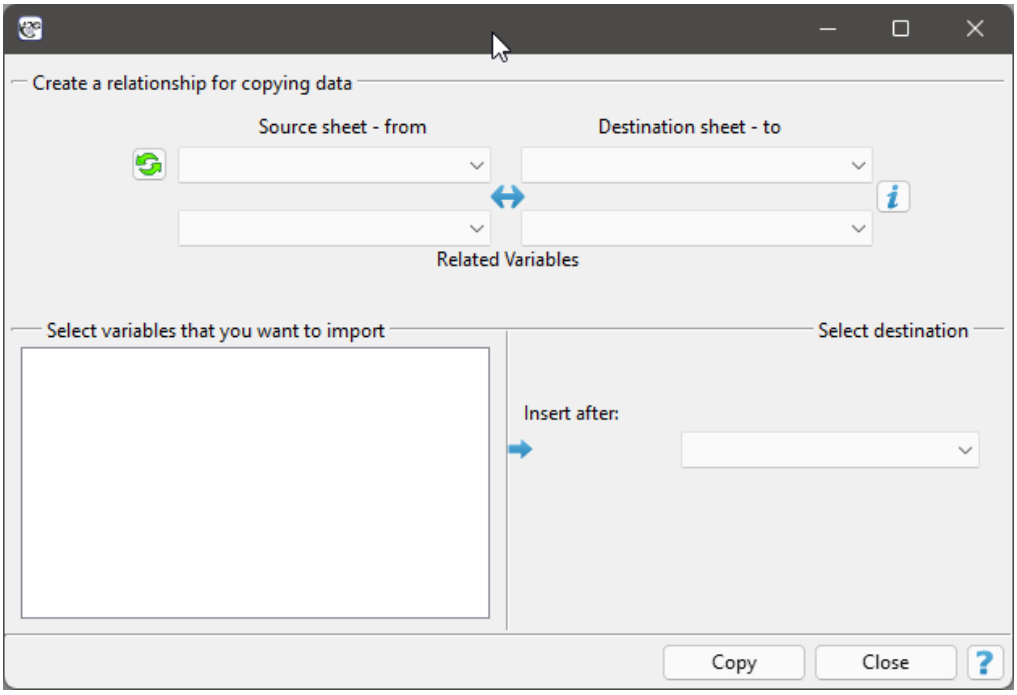
Copyright ©2010-2026 U-smile Stats – All rights reserved



In the import window you can preview the imported data and check the result of the import in advance, depending on the set options for data interpretation. It is worth checking that special characters are displayed correctly in the preview - this helps you avoid misinterpretation of the encoding. For large files, the preview window contains only the initial portion of the file data. **Note!** In Microsoft Office Excel 2000-2007, the default character encoding is Windows-1250. Importing data from Microsoft Excel spreadsheets covers only cell values - it is not possible to import formatting or formulas. **Copying data with relation**

Data from another worksheet can be copied into the selected data sheet based on relations. This type of data copying is performed by selecting menu: Data→Copying with relation... 

Copyright ©2010-2026 U-smile Stats – All rights reserved

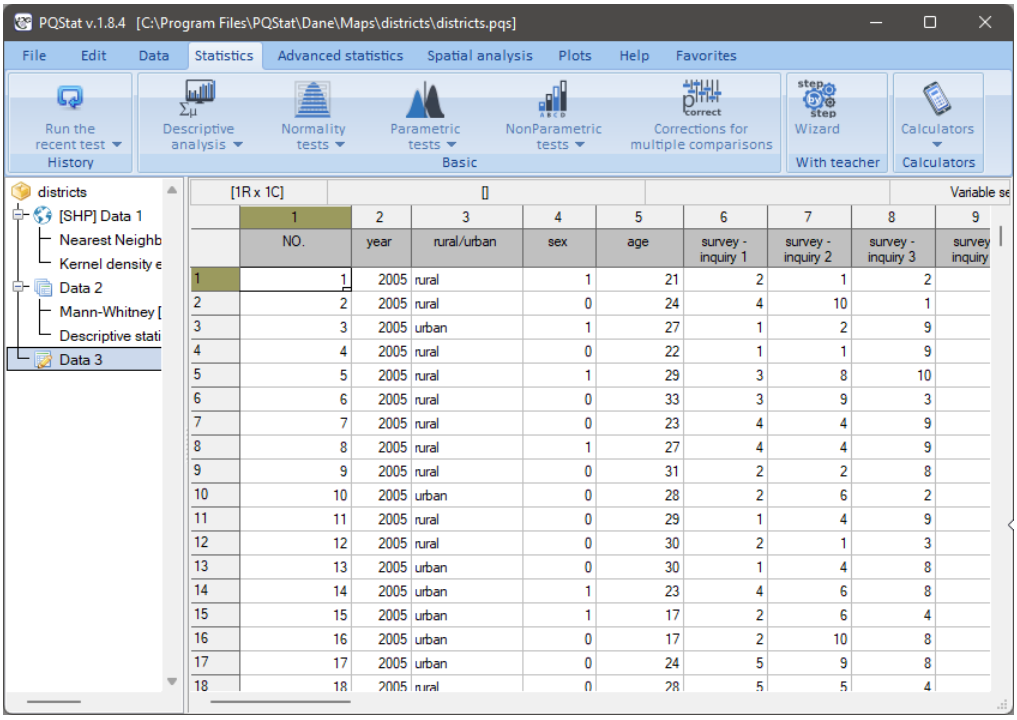


To build a relation, you must specify the data sheet from which you are copying and the data sheet in which you will place the copied data. Both of these datasheets must have the same key, i.e. a variable which values identify each row in the datasheet. It is required that the key for the source sheet is unique. Linking is done on a one-to-many basis, which means that one row in the source datasheet can be linked to multiple rows in the target datasheet. The keys of both datasheets should be selected as Related Variables. For such a relation, indicate the variables to be copied and the column after which to place the copied variables.

DATASHEET WINDOW

The rows and columns of the datasheet are indicated by consecutive natural numbers. Each column in the space marked in gray can be given its own heading. At the top of the datasheet is the Message bar. This is where the current user information is displayed. The left part of the bar informs about the size of the selected area [number of rows, number of columns], the middle part displays the value located in the selected cell, and the right part is for the user's information concerning, for example, the statistical analysis being performed.

Copyright ©2010-2026 U-smile Stats – All rights reserved

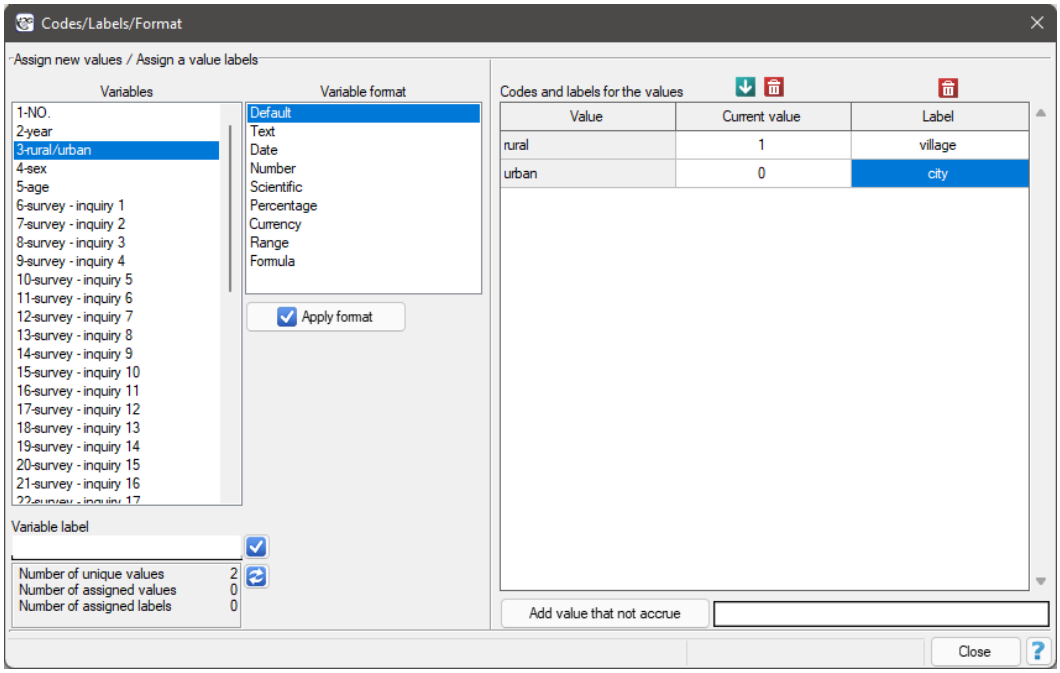


	1 NO.	2 year	3 rural/urban	4 sex	5 age	6 survey - inquiry 1	7 survey - inquiry 2	8 survey - inquiry 3	9 survey inquiry
1	1	2005	rural	1	21	2	1	2	
2	2	2005	rural	0	24	4	10	1	
3	3	2005	urban	1	27	1	2	9	
4	4	2005	rural	0	22	1	1	9	
5	5	2005	rural	1	29	3	8	10	
6	6	2005	rural	0	33	3	9	3	
7	7	2005	rural	0	23	4	4	9	
8	8	2005	rural	1	27	4	4	9	
9	9	2005	rural	0	31	2	2	8	
10	10	2005	urban	0	28	2	6	2	
11	11	2005	rural	0	29	1	4	9	
12	12	2005	rural	0	30	2	1	3	
13	13	2005	urban	0	30	1	4	8	
14	14	2005	urban	1	23	4	6	8	
15	15	2005	urban	1	17	2	6	4	
16	16	2005	urban	0	17	2	10	8	
17	17	2005	urban	0	24	5	9	8	
18	18	2005	rural	0	28	5	5	4	

VARIABLE PROPERTIES

For each column of the worksheet, we can set its properties such as codes, labels and format. Setting the properties of a variable is possible by

- selecting Variable Properties → Codes/Labels/Format from the context menu on the number above the column header,
- by double-clicking on the number above the column header - if it was specified in the program settings (corresponding double-click action).



Copyright ©2010-2026 U-smile Stats – All rights reserved

Codes and labeles for the values – are assignable to each value in the column.

> **Current value** - By filling in the codes we set the values to be valid in the given column. As a result the values taken for calculations will be changed (in the background) to assigned codes (valid values).
> **Label** - The values you enter under Lable are used in reports and charts for the user-defined result report.

Variable label - is assigned to the header of a given column. This is usually a brief description of the variable's contents. The variable label is used instead of the column header (variable name) in reports and charts for clearer description of the results. The use of the variable label is optional and depends on the program settings.

Variable format
Each worksheet cell (including a column heading) may contain a maximum of 40 characters. Texts containing national characters are also allowed. Entered values can be formatted as:

- **default** - default format is a format in which the program automatically recognizes the contents of the cell in the range - numeric data, text data;
- **text** - in text format, data are interpreted as text (alignment to the left edge of the cell);
- **date** - In date format, numeric data is interpreted as consecutive date values, so value 1 means 1899.12.31, value 2 means 1900.01.01 etc. Depending on the selected date format, you can also enter data in text format, these are:

  2010.12.31
  31.12.2010
  12.31.2010
  2010/12/31
  31/12/2010
  12/31/2010
  2010-12-31
  31-12-2010
  12-31-2010
  Monday...
  January...

  For the format Monday..., the value 1 indicates Monday, ..., 7 indicates Sunday, for the format January..., the value 1 indicates January, ..., 12 indicates December.

- **time** - In the time format, numeric data is interpreted as consecutive time values, the fractional part of the number means the number of milliseconds since midnight divided by the total number of milliseconds of the day (86400000), so the value 0.000694444 means 00:01:00, the value 0.041666667 means 01:00:00, the value 0.999988426 means 23:59:59. Depending on the selected time format, it is also possible to enter data in text format, these are:

  18:31:58
  18:31
  12/31/2010 18:31
  12/31/2010 18:31:58

Copyright ©2010-2026 U-smile Stats – All rights reserved

- **numerical** - real numbers in this format are in decimal form with either a comma or a period separating the whole from the fraction (depending on the settings you have chosen in Settings in Decimal Separator), you can set the number of decimal places and the thousandths separator;
- **scientific** - that is, $M \cdot 10^E$, where the base is the mantissa of M, and the exponent E is an integer; that is, using M times 10 to the power of E, where the base is the mantissa of M and the exponent E is an integer; just as in numerical format, it is possible to set the number of decimal places;
- **percentage** - changing a number to a percentage by multiplying it by 100 and displaying it with the % symbol; as in numeric format, it is possible to set the number of decimal places;
- **currency** - are used for monetary values - this allows you to add a currency symbol; as in the numeric format, it is possible to set the number of decimal places;
- **range** - written with upper and lower limits; as in the numeric format, it is possible to set the number of decimal places;
- **formula** - values calculated according to the formula assigned to the column; the value is automatically recalculated when any of the input data is changed.

When you open a new sheet, a **default format** is set for each cell.

The entire header row is permanently set to text format. User-defined formats can be set for the rest of the sheet. Formatting is not done for a single cell, but for the entire column (except for its header).

In the worksheet, you can specify the **column width** using the mouse. To do this, use the mouse pointer to drag a line dividing the columns, narrowing or widening the column to the left of the selected line.

Additionally, you can set a different background color in each cell of the worksheet (after selecting the area to be changed). You can do it via:
- menu File→Fill color ... 

- command Fill color in the cell's context menu.

EDITING DATA

Selection of a consistent area in a worksheet can be done with mouse or keyboard (Arrow keys + Shift). During selection, the size of the selection (number of rows and columns) is continuously displayed in the message bar. You can select the entire worksheet by simply clicking in the upper left corner of the worksheet with the mouse, or by selecting menu Edit→Select all (Ctrl+A). Whole rows or whole columns are selected by selecting their headers.

Copying or moving cells is done through the copy, cut and paste commands.

The copy, cut and paste commands are available in several places:
- in menu Edit,
- in the context menu of cells,
- on the toolbar  ,
- in the context menu for rows and columns,

Copyright ©2010-2026 U-smile Stats – All rights reserved

- via hotkeys: copy (Ctrl+C), cut (Ctrl+X), paste (Ctrl+V).

Deleting data from cells can be performed via menu Edit→Delete (Del)

Undoing the most recently performed operation can be done via menu Edit→Undo (Ctrl+Z). By default, the Program remembers the last 10 operations involving 5000 cells in each cell. You can change these settings in the Settings window. It should be noted, however, that increasing these values results in greater use of computer memory by the program.

Inserting and deleting rows and columns

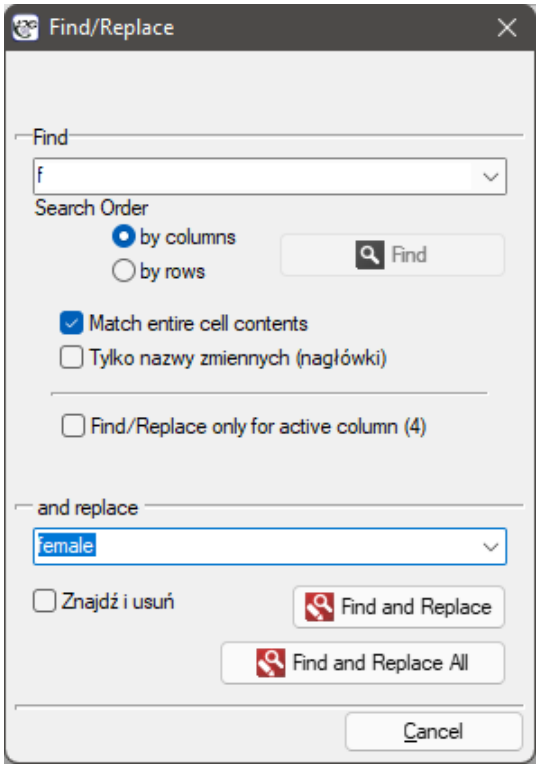
You can insert blank rows or columns above or to the left of an existing row or column. This will move the cells down or to the right. To insert a row(s), select the row(s) above which you want to insert a new row(s), and then choose Insert row from the context menu on the selected row number. Inserting columns is done in the same way.

Rows and columns can also be deleted by selecting them and choosing Delete row/Delete column from the context menu in the row or column number..

Finding/replacing a cell value

To find or replace the entire contents of a cell with another value, use the Find/Replace window, which is accessible via menu Edit→Find/Replace (Ctrl+F)  .

The upper part of the Find/Replace window is used for searching and the lower part is used for replacing cell values.



Copyright ©2010-2026 U-smile Stats – All rights reserved

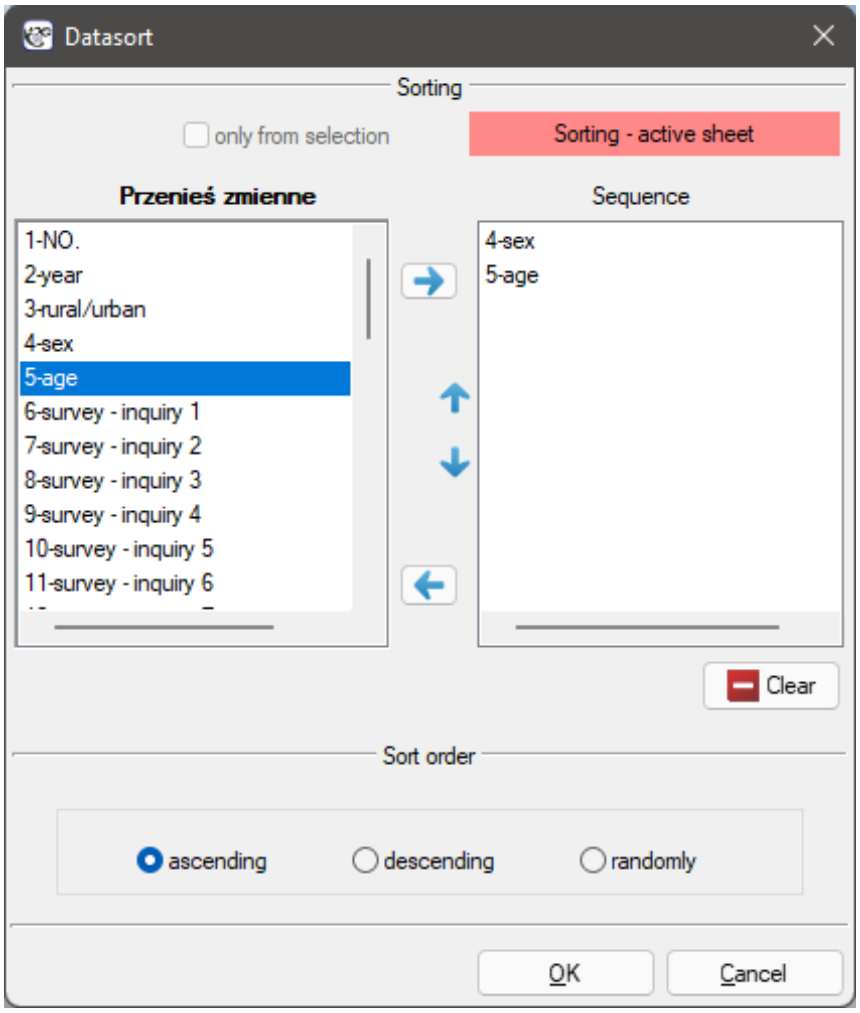
To search for data, enter a search string in the upper part of the window, select the search order and choose the Find button.

To find and replace the entire contents of a cell with another value, fill in the upper and lower parts of the window. Fill in the upper part of the window as you would for a data search. In the lower part of the window type the data that you want to replace and choose Find and Replace or Find and Replace All when you want to replace all the occurrences of the searched data. Searching as well as replacing is done in direct view mode of the operations performed on the sheet.

SORTING DATA



Sorting options are available by selecting menu Data→Sort... or Sort... functionality from the context menu on the number above the column header. Normally you sort the whole datasheet (this is the default setting for sorting), but if you start sorting by selecting a piece of data, then in the sorting window you will have the option to limit the sorting area only to the selection.



In the sorting window, use the arrows to move from the Move Variables box to the Sequence box those

Copyright ©2010-2026 U-smile Stats – All rights reserved

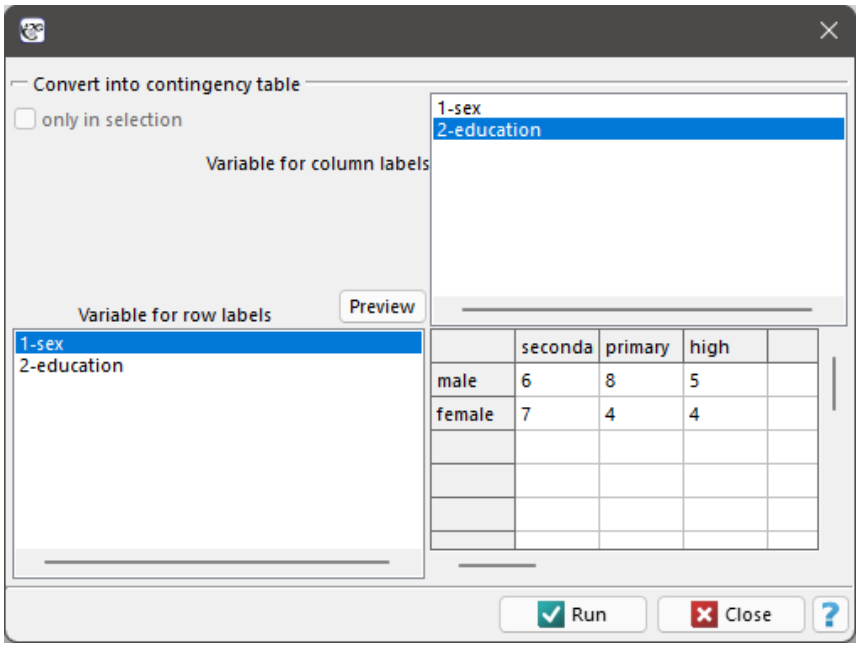
variables by which you want to sort the data, then select Sort Order and confirm your selection by pressing the Ok button.
You can sort by at most 3 columns (variables). If you sort data by more than one variable, sorting is done in the order in which the variables are placed in the Order box.

CONVERTING RAW DATA INTO A CONTINGENCY TABLE



The operation of changing raw data into a contingency table is available after selecting menu Data→Create table... Normally, the entire data sheet is available for this operation (this is the default setting), but if you start the transformation by selecting a piece of data, then in the data transformation window you will have the option to limit the available area to the selection only.

You design the contingency table by selecting the variables that form the row and column labels. If the preview of the table is consistent with the expected result, confirm your choice with Run. The returned result will be placed in a new sheet.

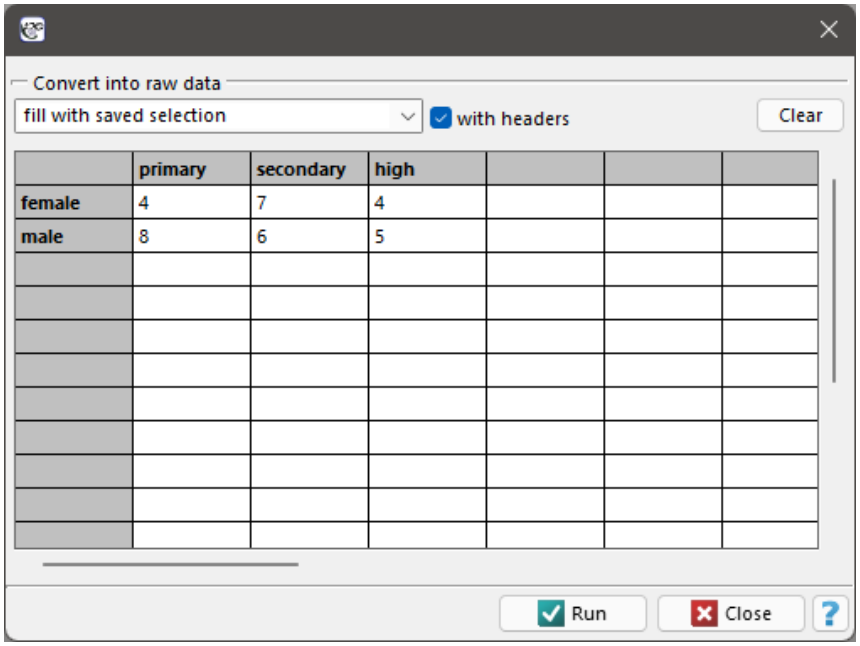


CONVERTING A CONTINGENCY TABLE INTO RAW DATA



The operation of changing a contingency table into raw data is available after selecting menu Data→Create raw data... In the data transformation window, enter the appropriate numbers and row and column headings and confirm your choice with Run. The returned result will be placed in a new sheet.

Copyright ©2010-2026 U-smile Stats – All rights reserved



If we are converting a table that is in the data sheet, then we select that table (with or without headers) before converting it to raw data. This table will then be automatically placed in the data transformation window. It is also possible to use other tables marked as fill with saved selection.

RESHAPING DATA BETWEEN THE WIDE AND THE LONG FORMAT

The same measurements can be arranged in a sheet in two ways:

- **wide format** – one row is one object (e.g. a patient) and the consecutive measurements of that object (e.g. before treatment, after one month, after one year) are stored in consecutive columns; this is the layout used e.g. by the ANOVA for dependent groups and the tests for dependent groups,
- **long format** – one row is one measurement; the object is described by an identifier column (repeated as many times as it has measurements), next to it there is a column with the name or number of the measurement and one common value column; this layout is required by mixed models and by most analyses in which a measurement is a separate observation.

An example of the same data in both formats:

wide format			
ID	sex	before	after
1	F	120	112
2	M	135	130

long format			
ID	sex	variable	value
1	F	before	120
1	F	after	112
2	M	before	135
2	M	after	130

Both conversions are available in the Data menu. The result is always placed in a **new sheet** of the project and the source sheet stays unchanged. Deactivated rows and hidden columns are skipped, as in

Copyright ©2010-2026 U-smile Stats – All rights reserved

the statistical analyses.
An example of data from the cisnienie.pqs file, where we examine two types of drugs (A and B) and blood pressure at consecutive visits: Wide format:

CS before	CS in the middle	CS after	treatment
166	160	160	Drug A
144	140	120	Drug A
188	170	150	Drug A
170	161	160	Drug A
156	150	130	Drug A
160	150	120	Drug A
148	130	126	Drug A
150	150	150	Drug A
178	160	156	Drug A
180	160	140	Drug A
190	140	138	Drug A

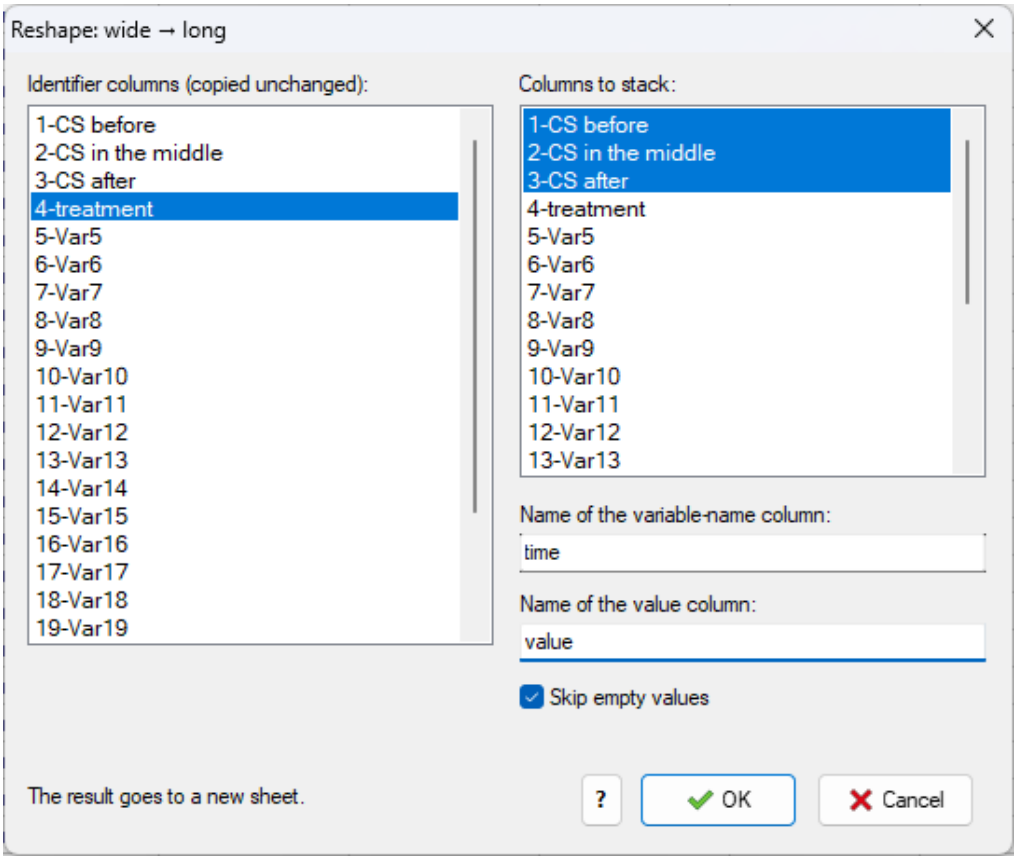
Long format:

treatment	time	value
Drug A	CS before	166
Drug A	CS in the middle	160
Drug A	CS after	160
Drug A	CS before	144
Drug A	CS in the middle	140
Drug A	CS after	120
Drug A	CS before	188
Drug A	CS in the middle	170
Drug A	CS after	150
Drug A	CS before	170

Wide → long format

The command Data→Reshape: wide → long... stacks the selected columns one below another (example file: pressure.pqs).

Copyright ©2010-2026 U-smile Stats – All rights reserved



In the reshape window we indicate:

- **Identifier columns** – columns describing the object (e.g. identifier, sex, group) that are to be copied unchanged and repeated in every row produced from the same object; their selection is optional,
- **Columns to stack** – columns with the measurements; at least one must be selected and the same column cannot be both an identifier and a value column,
- **Name of the variable-name column** (by default `variable`) – this column will hold the name of the source column the value comes from,
- **Name of the value column** (by default `value`),
- **Skip empty values** – when checked (the default) measurements without a value produce no rows; when unchecked every object gets exactly as many rows as there are selected columns and missing values remain empty cells.

Confirm with OK. Each source row yields as many rows as there are stacked columns and the rows are ordered object by object, in the order of the selected columns. Column names in the new sheet are automatically made distinct if they would repeat.

Note! If the variable-name column is to serve as the time or measurement number in a mixed model, it is convenient to give the source columns numeric names (e.g. `0`, `1`, `12`) or to recode them afterwards with formulas.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Long → wide format

The command Data→Reshape: long → wide... spreads one value column into several columns, one for each variable name. In the reshape window we indicate:

- **Identifier columns** – columns whose common combination of values defines one object (e.g. the patient identifier, or the identifier and sex); all rows with the same combination are merged into one wide-format row; when no column is selected every source row stays a separate row,
- **Column with variable names** – each unique value of this column becomes a separate column of the new sheet (in the order of first appearance in the data); rows with an empty name are skipped,
- **Column with values** – its content fills the new columns.

The variable-name column and the value column must differ and cannot be identifier columns at the same time. Confirm with OK. The new sheet contains the identifier columns (values from the first occurrence of the object) and the columns created from the variable names; combinations for which there was no measurement in the data remain empty cells.

Note! If several values occur in the data for the same combination of identifiers and the same variable name (e.g. several measurements of the same variable in one object, or too few identifier columns), none of them is lost: the first goes to the column named after the variable and the following ones to additional columns `name_2`, `name_3`, …placed next to it. The program reports the number of such repetitions. If the further occurrences should be distinguished differently (e.g. by the visit number), add the corresponding column to the identifiers.

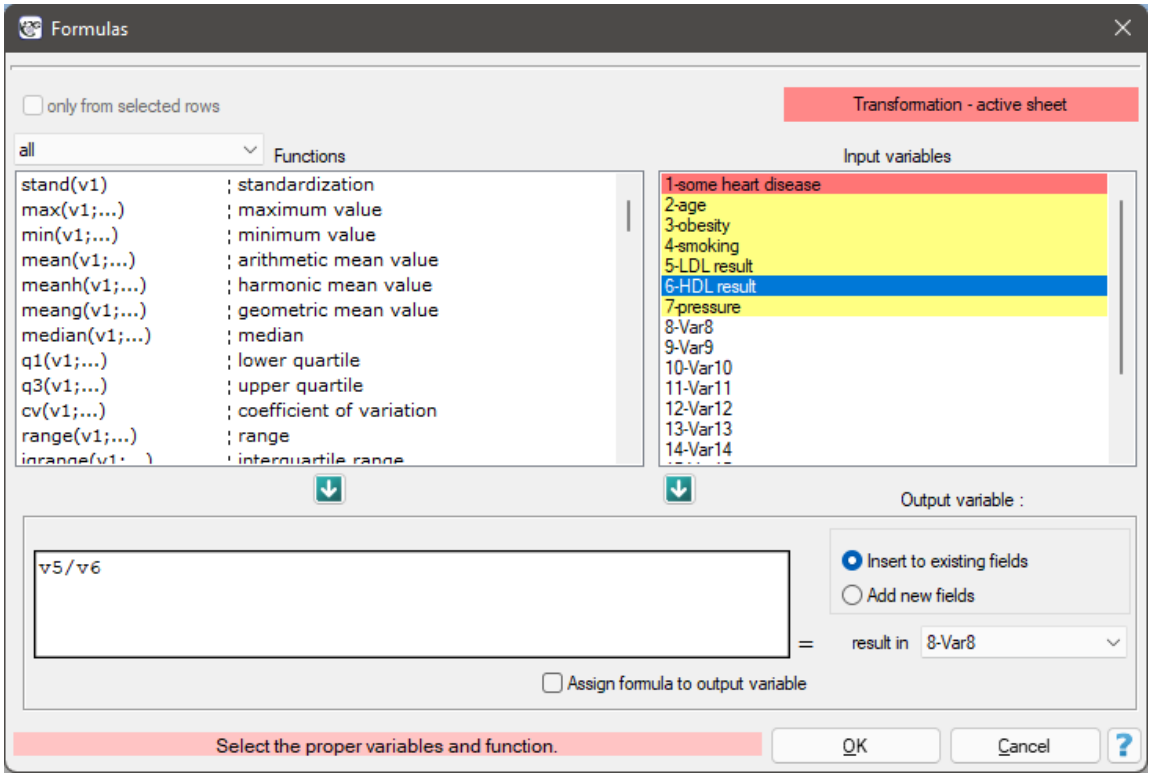
FORMULAS



Defining a formula is a way to recalculate data, resulting in new values in variables.

The window for defining formulas is opened via Data→Formulas...

Copyright ©2010-2026 U-smile Stats – All rights reserved



Formulas assigned to a datasheet variable as the format of that variable are stored with the worksheet data. Their result is automatically recalculated when any of the input data is changed. Formula can be assigned in th Formula... window or by setting the Column format (Ctrl+F10).

Creating formulas

Formulas are entered in the edition box.

- The variables referred to in the formulas are entered with their numbers, e.g `v1+v2`
- Text values are entered using an apostrophe, e.g. 'house'.
- You can enter functions by double-clicking on the name of the selected function - then the name will appear in the formula edition box, or you can enter the name yourself in the edition box, though the function name is case-insensitive. The function arguments are given in brackets using the syntax given in the function description.

Results of formulas

The results of formulas will be displayed in the selected column.

If the function's arguments include values that it cannot interpret, the program displays a message asking whether to ignore the uninterpreted variables. If you choose Yes, the formula will be recalculated by omitting the unexpanded data. If you choose No, the formula returns error (NA). For example, for the values in columns v1, v2 and v3 respectively: 1, 2, 'ada', the sum function `sum(v1;v2;v3)` will return a result of 3 - when omitting the uninterpreted value 'ada', or will return NA - when not omitting this value from the calculation.

An empty value (no data) will be returned only if all arguments used in the formula are empty.

You can limit the number of rows involved in a formula by selecting the appropriate number of rows in the datasheet and choosing only from selected rows in the formula window.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Operators

$+$ addition,
$-$ subtraction,
$*$ multiplication,
$/$ division,
% modulo division (resulting in the remainder of the division),
$>$ greater,
$<$ less,
$=$ equal.

Mathematical functions

Mathematical functions require numeric arguments.

ln(v1) - outputs the natural logarithm of the given number,
log10(v1) - outputs the logarithm of the base 10 for a given number,
logn(v1) - outputs the logarithm of n for the given number,
sqr(v1) - outputs the square of the given number,
sqrt(v1) - outputs the square root of the given number,
fact(v1) - outputs the power of a given number,
degrad(v1) - outputs the angle in radians (the argument of the function is in degrees),
raddeg(v1) - outputs the angle in degrees (the argument of the function is in radians).,
sin(v1) - outputs the sine of the given angle, (the argument of the function is in radians),
cos(v1) - outputs the cosine of the given angle, (the argument of the function is in radians),
tan(v1) - outputs the tangent of the given angle, (the argument of the function is in radians),
ctng(v1) - outputs the cotangent of the given angle, (the argument of the function is in radians),
arcsin(v1) - outputs the arcus sine of the given angle, (the argument of the function is in radians),
arctan(v1) - outputs the arcus tangent of the given angle, (the argument of the function is in radians),
exp(v1) - outputs the value of the number e raised to the power specified by the given value,
frac(v1) - outputs the fractional part of a given number,
int(v1) - outputs the integer part of a given number,
abs(v1) - outputs the absolute value of the specified number,
odd(v1) - if given number is even, outputs 1, 0 otherwise,
sum(v1;...) - outputs the result of adding the specified numbers,
multip(v1;...) - outputs the result of multiplication of specified numbers,
power(v1;n) - outputs the result of raising a number to the power of n,
norme(v1;...) - outputs the Euclidean norm of the vector,
round(v1;n) - outputs a number rounded to n decimal places.

Statistical functions

Statistical functions require numeric arguments.

stand(v1) - outputs the standardized value of the specified variable,
max(v1,...) - outputs the largest value,
min(v1,...) - outputs the smallest value,
mean(v1,...) - outputs the value of the arithmetic mean,
meanh(v1,...) - outputs the value of the harmonic mean,
meang(v1,...) - outputs the value of the geometric mean,
median(v1,...) - outputs the median value,
q1(v1,...) - outputs the value of the bottom quartile,
q3(v1,...) - outputs the value of the top quartile,

Copyright ©2010-2026 U-smile Stats – All rights reserved

cv(v1,...) - outputs the value of the coefficient of variation,
range(v1,...) - outputs the value of the interval,
iqrange(v1,...) - outputs the value of the quartile interval,
variance(v1,...) - outputs the variance value,
sd(v1,...) - outputs the value of the standard deviation.

Text functions

Text functions work on any string.

upperc(v1) - converts characters from a string to uppercase,
lowerc(v1) - converts characters from a string to lowercase,
clean(v1) - removes characters that cannot be printed,
trim(v1) - removes leading and trailing spaces,
length(v1) - outputs the length of the string,
search('abc';v1) - outputs the position of the beginning of the searched text,
concat(v1;...) - combines texts,
compare(v1;...) - compares texts,
copy(v1;i;n) - returns a portion of text starting from the i-th character, where n is the number of returned characters,
count(v1;...) - outputs the number of cells that are not empty,
counte(v1;...) - outputs the number of cells that are empty,
countn(v1;...) - outputs the number of cells that contain numbers.

Date and time functions

Date and time functions should be used on data formatted as date or time (see chapter 7.1.4). If this is not the case, the program tries to automatically recognize the format, and if not possible gives the value NA.

year(v1;) - outputs the year corresponding to the date,
month(v1;) - outputs the month corresponding to the date,
day(v1;) - outputs the day corresponding to the date,
hour(v1;) - outputs the time corresponding to the specified time,
minute(v1;) - outputs the minute corresponding to the specified time,
second(v1;) - outputs the second corresponding to the specified time,
yeardiff(v1;v2) - outputs the number of years separating two dates,
monthdiff(v1;v2) - outputs the number of months separating two dates,
weekdiff(v1;v2) - outputs the number of weeks separating two dates,
daydiff(v1;v2) - outputs the number of days between two dates,
hourdiff(v1;v2) - outputs the number of hours between the two times,
minutediff(v1;v2) - outputs the number of minutes separating the two times,
seconddiff(v1;v2) - outputs the number of seconds separating the two times,
compdate(v1;v2) - compares dates and outputs the number 1 when v1>v2; 0 when v1=v2, -1 when v1<v2.

Logic functions

if(question;'yes - answer';'no - answer') - a question is formulated as an expression that can be true or false; the function outputs one value if the expression is true and the other if it is false,
and - conjunction operator - returns true (1) when all conditions it combines are true, false (0) otherwise,
or - the alternative operator returns true (1) when at least one of its conditions is true, false (0) otherwise,
xor - disjunctive alternative operator - returns true (1) when one of the conditions it combines is

Copyright ©2010-2026 U-smile Stats – All rights reserved

true, false (0) otherwise,
not - negation operator used in a conditional statement if,
empty(v1) - outputs true (1) when empty cells are present, false (0) otherwise,
text(v1) - outputs true (1) when text is present, false (0) otherwise,
number(v1) - outputs true (1) when a number is present, false (0) otherwise.

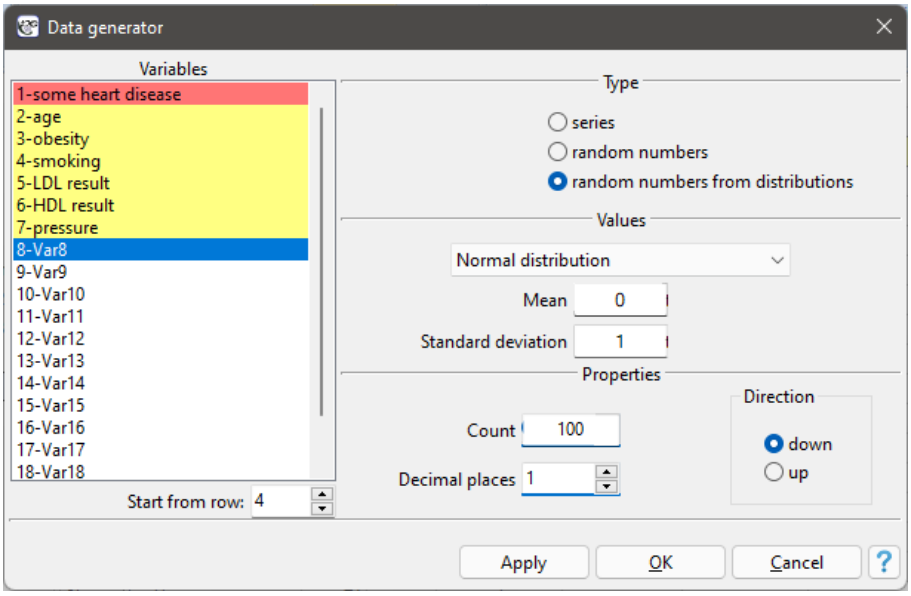
DATA GENERATION



There are two methods of data generation:

1. The first method uses simple dragging of the contents from the selected cells to the neighboring cells using the mouse pointer. This method lets you generate the same values (text or numbers) in neighboring columns or rows.
   To generate, start by selecting the cell with the appropriate data , then use the mouse pointer depicted by the + sign to grab the bottom right corner of that cell and drag through the cells you want to fill. Dragging a single cell can be done in any direction (up, down, left and right). It is also possible to drag different values placed in one column (left or right) or in one row (up or down).

2. The second method generates numerical data in columns as data series, random values, and random values from an appropriate data distribution.
   To generate numeric data, select the cell from which you want to start filling in the datasheet and open the numeric data generation window from the menu Data→Generate...



You start by selecting the variable in which the generated data will be placed.
In the middle part of the window, depending on the settings of the method of data generation selected above, we set:

- For generating data series:
  - Start value - the first value to be generated,
  - Increment - the value by which the subsequent generated data is to vary.

Copyright ©2010-2026 U-smile Stats – All rights reserved

- For generating random values:
  - Lower limit - the beginning of the interval from which the values will be randomly selected,
  - Upper limit - The end of the interval from which the values will be randomly selected.
- For generating random values from a distribution, select the type of distribution (Normal distribution, Chi-square distribution) and enter its parameters.

The amount of data generated depends on the value the user enters in the Count box, and the precision depends on the setting of the Decimal places box. The data will be filled in starting from the active cell in either down or up - depending on the selected option. Finally, confirm your selection with Apply.

MISSING DATA

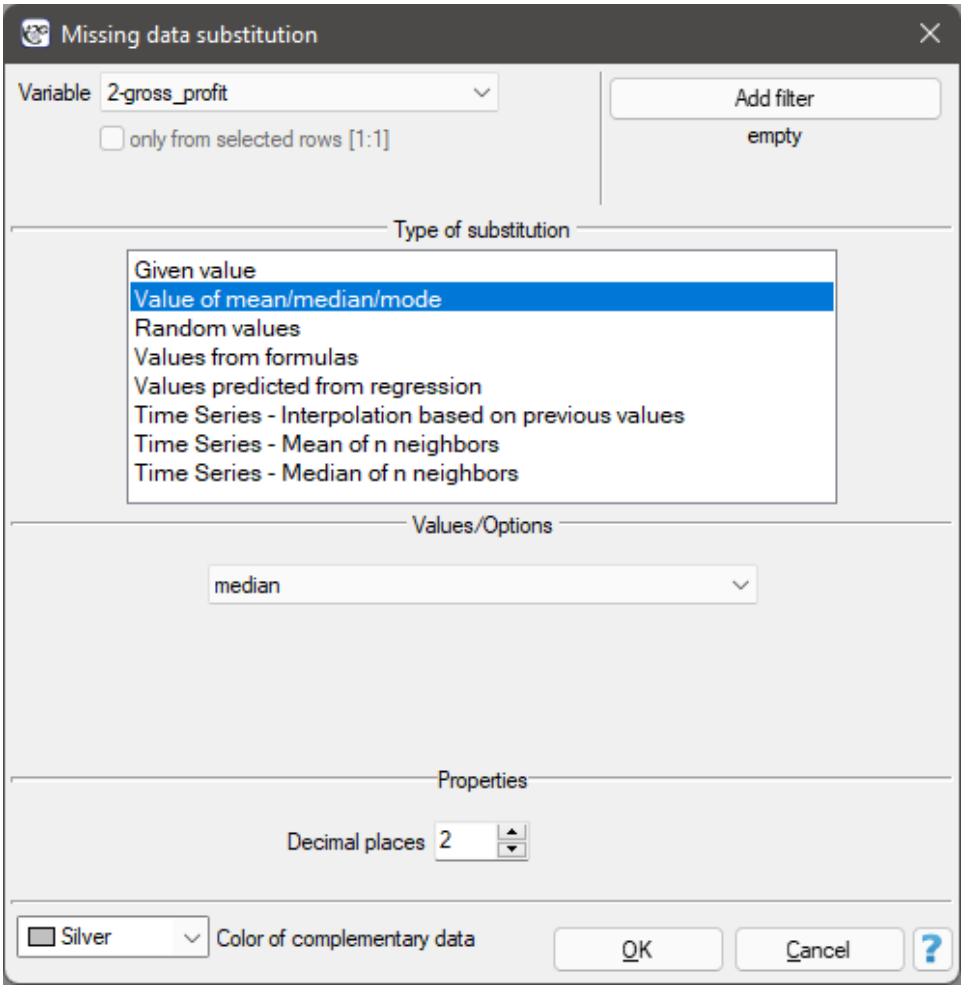


In research we very often encounter missing data, this is natural in particular for survey data. There are situations in which missing data provides valuable information. For example: the number of missing data items in response to a question about support for political parties gives an idea about the number of undecided people who do not like (or do not admit to liking) certain political groups. Small numbers of missing data are not a problem in statistical analyses. A large number of them, however, may cast doubt on the reliability of the research. It is worth at the very beginning of the work to make sure that there is as little missing data as possible. Of course, it is best to find information about the actual value that should be put in the place of missing data, but this is not always possible.

How missing data are estimated depends primarily on the nature of the data. The program proposes several ways to impute missing data for individual variables.

The Missing data substitution settings window accessed via menu Data→Missing data...

Copyright ©2010-2026 U-smile Stats – All rights reserved



1. Filling with one value

   Selecting one of the following options will replace all missing data in the selected column with the same value:

   - specified by the user,
   - the arithmetic mean calculated from the data,
   - the geometric mean calculated from the data,
   - the harmonic mean calculated from the data,
   - the median,
   - the mode (unless it is multiple).

2. Filling with multiple values

   Selecting one of the following options will replace the missing data in the selected column with multiple (usually different) values. These values can be predicted from the column for which the missing data is filled, but they can also be predicted from the values of other columns (variables). You can replace missing data with values:

   - random from the data;
   - random from normal distribution - normal distribution is defined by the mean and standard deviation of the existing data;

Copyright ©2010-2026 U-smile Stats – All rights reserved

- with random values from an interval specified by the user;
- calculated from user functions - this option allows you to use data from other variables to predict the missing value in the selected column;
- predicted from the regression model - this option allows to predict the value of missing data based on the multiple regression model (the functioning of multiple regression is described in the section 25.1 Multiple linear regression);
- interpolation based on neighboring values - applies to time series - so the user must indicate the time variable indicating the order of data; interpolation involves the determination of values for the missing data in such a way that they are graphically located on the straight line connecting values for data adjacent to the missing data;
- the average of n neighbors - applies to time series - Thus, the user must indicate the time variable that tells the order of the data; interpolation works by determining the average of the values for n neighbors preceding and n neighbors immediately following the missing data;
- median from n neighbors - applies to time series - thus the user must indicate the time variable telling about the order of the data; interpolation works by determining the median from the values for n neighbors preceding and from n neighbors immediately following the missing data;

Note! In order to be able to distinguish imputed data from real data, the replaced cells are marked with a chosen color.

EXAMPLE 7.1. (missingData-publisher.pqs file)
Analysis of the publisher.pqs file with no missing data is discussed in Multiple linear regression. This time we are going to deal with a datasheet in which there are missing data in the column containing gross profit from book sales. For these missing data, the actual values are known (datasheet: "REAL VALUES"), so you can compare the values generated by the program for the missing data with the actual values to compare the results obtained by different techniques. In the example, we will use 2 ways of replacing missing data: replacing with the median value and the value determined by the regression model. The other options are left to you to work on your own.

Replacing missing data with the median value is done on datasheet 1 called "Insert the median". Set the variable prepared to be inputed as gross profit in the Missing data window and select the method of replacement as the median value. This will result in a value of 46 850 dollars being inputed in place of the missing data.

It is suspected that profits are higher when dealing with books from known authors (coded as 1) and lower when dealing with those from unknown authors (coded as 0). So we calculate the median gross profit separately for books by known and unknown authors. We perform the imputation on the datasheet named "Insert two medians". We set the filter twice for the variable defining the authors' popularity (variable 7) - once giving the value 1 and once giving the value 0. The resulting median gross profit in the group of books by popular authors is about 51 000 dollars, and among those by lesser-known authors it is about 34 000 dollars.

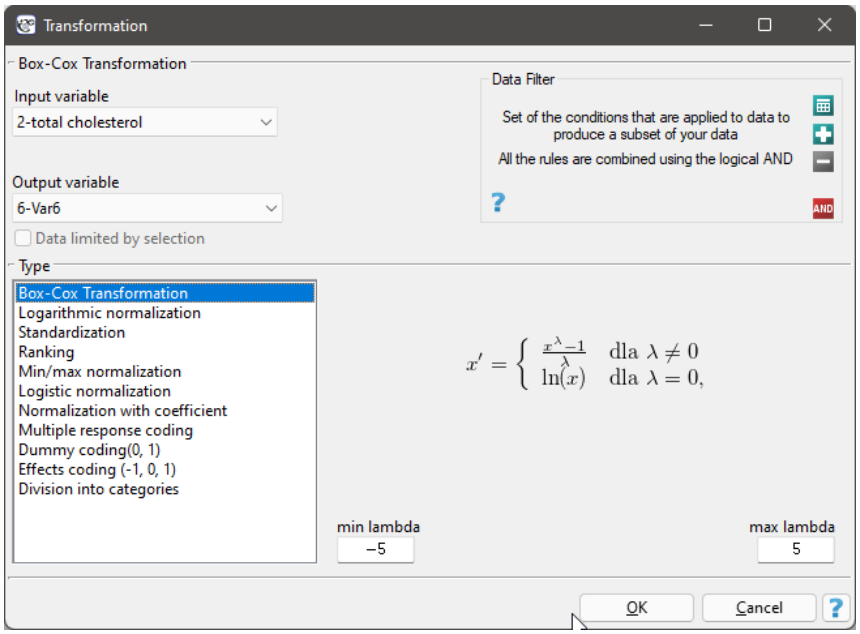
Another way to replace missing data, is to use a regression model. Select the "Insert from regression" datasheet and once again select the gross profit variable as the variable to be inserted, and select "Values predicted from regression" as the method of substitution. There will be more variables used to predict the value of gross profit this time: production costs (variable 3), advertising costs (variable 4) and authors' popularity (variable 7). This time the results seem to be less different from the real values, unfortunately the result for the item number 35 is missing, because for this book we had no information about the cost of production, on which, among other things, we wanted to base the prediction.

Copyright ©2010-2026 U-smile Stats – All rights reserved

TRANSFORMATIONS



The transformation window is accessed via Data→Transform...



Data transformation is the alteration of data so that it meets certain criteria, such as meeting the criteria for normality of distribution or extending within a specified range.

Box-Cox transformation

The Box-Cox transformation introduced by Box and Cox in 1964 [25] brings the data to a normal distribution through a transformation based on the coefficient λ. Positive data are required to perform the transformation. If the data are not positive, it is recommended to first transform them to positive numbers using the min-max normalization method.

The Box-Cox transformation is expressed by the formula:

$$x' = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(x) & \text{for } \lambda = 0, \end{cases} \tag{1}$$

where the value of λ is determined as the maximum value of the log-likelihood function (LL) in the interval specified by the researcher. The default range for searching for λ values is the range [-5, 5], and the LL function is described by the formula:

$$LL = -\frac{n}{2} \ln(sd_{pop}^2) + (\lambda - 1) \sum \ln x \tag{2}$$

where:
n - sample size,
sd_{pop} - population standard deviation.

Note! If min-max normalization was used before the Box-Cox transformation, you can return to the previous range after the Box-Cox transformation by applying that normalization again.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Logarithmic normalization

The logarithmic transformation can be used to reduce the skewness of the distribution i.e. when we are dealing with a lognormal distribution.

$$x' = \ln x \tag{3}$$

Standardization

Standardization, is a transformation of data that results in a variable having a mean of 0 and a standard deviation of 1.

$$x' = \frac{x - \bar{x}}{sd} \tag{4}$$

Ranking

Ranks - are consecutive numbers (usually natural) assigned to the values of ordered measurements of the variable under study. They are often used in those nonparametric tests that rely solely on the order of items in the sample. Assigning ranks calculated according to a variable is called **ranking**. Ranking can be done for variables sorted ascendingly (this is the default setting) or descendingly.

Recurring values of a variable are assigned a **tied rank**. The tied rank can be a/an:
- arithmetic mean calculated from the proposed consecutive natural numbers for repeated values - this is the default setting;
- lower rank, i.e., the smallest limit proposed for consecutive repeated values of natural numbers;
- the upper rank, meaning the largest proposed for consecutive repeated values of natural numbers.

For example, for a variable with the following values: 8.6, 5.3, 8.6, 7.1, 9.3, 7.2, 7.3, 7.4, 7.3, 5.2, 7, 9.9, 8.6, 5.7 the following ranks are assigned:

sorted values of the variable	ranks
5.2	1
5.3	2
5.7	3
7	4
7.1	5
7.2	6
7.3	7.5
7.3	7.5
7.4	9
8.6	11
8.6	11
8.6	11
9.3	13
9.9	14

While for the variable with a value of 7.3 a tied rank calculated as the arithmetic mean of the numbers:7 and 8 is assigned, and for a variable with value 8.6 a tied rank calculated from the numbers: 10, 11, 12 is assigned.

Min/max normalization

The min/max normalization through a linear function puts the data into a user-specified range (new_{min}, new_{max}). You should know the range that the data can cover. If you do not know the

Copyright ©2010-2026 U-smile Stats – All rights reserved

range, you can use the largest and smallest value in the analyzed set (in the Transformation window, then select the Calculate from sample option)..

$$x' = \frac{x - \min}{\max - \min} \cdot (new_{\max} - new_{\min}) + new_{\min} \tag{5}$$

Logistic normalization

Normalization using a logarithmic (S-shaped) function puts the standardized data into the indicated range.

$$x' = \frac{e^x}{1 - e^x} \tag{6}$$

If you want to stretch the transformed data over a range other than the specified one, then enter the span of the new range in the Transformation window.

Normalizing function with coefficient

This normalization brings the standardized data into the indicated range using an S-shaped function with a changing normalization factor α.

$$x' = \frac{x}{\sqrt{x^2 + \alpha}} \tag{7}$$

Increasing the α value creates a graph with a smoother slope.
If you want to stretch the transformed data over a range other than the specified one, then enter the span of the new range in the Transformation window.

Multiple response coding

This type of coding allows the answers given to multiple-choice questions to be prepared in such a way as to facilitate their further statistical processing. As a result of applying this transformation, a selected variable with k-possible answers is broken down into k new variables. It is necessary to specify which character (or set of characters) is a separator of particular categories. For example, respondents were asked what kind of alcohol they drink? The data is stored in Alcohol column, separating multiple answers with semicolon sign. This way of storing data does not even allow for a simple summary. Among other things, it is not possible to quickly count how many people drink wine. After recoding the multiple responses, three new columns were obtained - one for each possible answer. Each of these columns can now be statistically analyzed.

Alcohol	Alcohol(beer)	Alcohol(wine)	Alcohol(vodka)
beer;wine	1	1	0
wine	0	1	0
wine	0	1	0
beer	1	0	0
vodka;wine	0	1	1
wine;vodka	0	1	1
beer;vodka	1	0	1
beer;wine;vodka	1	1	1

Dummy coding

Transforming a variable with k categories by dummy coding allows you to obtain $k - 1$ dummy variables. This form of transformation is primarily used in regression models. A detailed description of this type of transformation can be found in 25.0.2 PREPARING VARIABLES FOR ANALYSIS IN MULTI-DIMENSIONAL MODELS.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Effect coding

Transforming a variable with k categories by effect coding yields $k-1$ dummy variables. This form of transformation is used primarily in regression and ANOVA models. A detailed description of this type of transformation can be found in 25.0.2 PREPARING VARIABLES FOR ANALYSIS IN MULTI-DIMENSIONAL MODELS.

Division into categories

This way of preparing data allows for any division of variables, e.g. total cholesterol can be divided according to the current standards (choose Manual division, set the number of categories and enter their limits ourselves and give appropriate labels to each category). However, if we do not have an idea for dividing our data, we can use the automatic division options presented in the window. Possible ways of dividing a variable:

- **Natural breaks (Jenks)** - a method of dividing a variable into classes such that the variance within classes is minimized and the variance between classes is maximized.
- **Division by Quantiles** - a method of dividing a variable into classes of equal frequency.
- **Standard Deviation** - a method of dividing a variable into classes based on its distance from the mean by 1, 2, or more standard deviations.
- **Standard error of the mean** - a method of dividing a variable into classes based on the distance from the mean by 1, 2, or more standard errors of the mean.
- **Manual** - a method of dividing a variable into classes according to any division entered manually by the user.

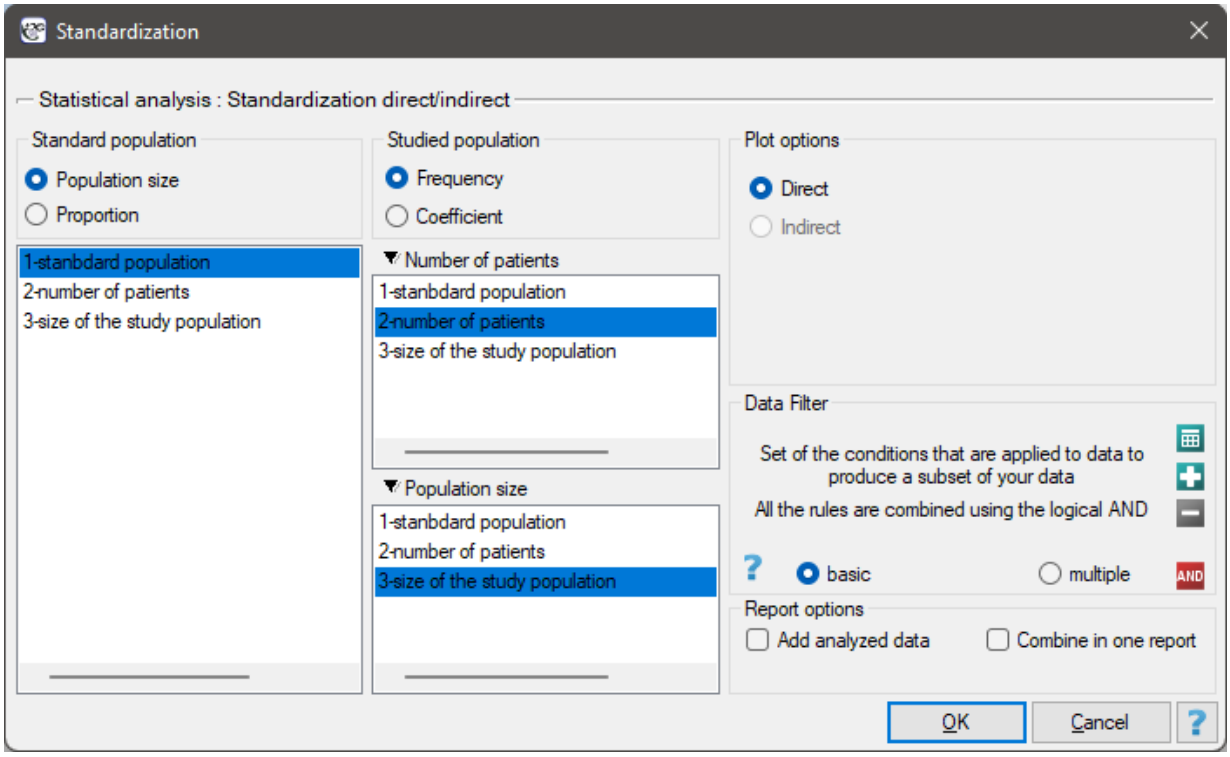
In the division window, it is also possible to select Add color scheme then the column that will store the new data will be color coded according to the indicated scheme.

EXAMPLE 7.2. (normalization.pqs file)
Perform a transformation on the variables contained in the file:
a) Transform the value of triglycerides using the Box-Cox transformation and then check with the appropriate test whether the data have a normal distribution.
b) Transform the value of triglycerides using the logarithmic transformation and then check with the appropriate test whether the data have a normal distribution.
c) Using min-max normalization, transform the selected variables to the range [0,10].
d) Using logistic normalization, transform the selected variables to the specified range.
e) Using normalization with a coefficient, transform the selected variables to the specified range. Do it several times, changing the value of the coefficient α.
f) Standardize all data that are normally distributed.
g) Transform the variable showing how body weight changed during the diet so that it represents a normal distribution.
h) The question about past infectious diseases was a multiple choice question. Prepare the obtained answers to this question so that they can be further statistically processed i.e. record each of the multiple answers in a different column.
i) Prepare the education variable so that it is stored using dummy variables with dummy coding.
j) Prepare the total cholesterol variable by dividing it into 3 classes according to the percentiles (quartiles). Give the created classes labels : ”low”, ”average”, ”high” and choose the color scheme.

DIRECT AND INDIRECT STANDARDIZATION

The direct and indirect standardization window for epidemiological coefficients is opened via Data→Standardization...

Copyright ©2010-2026 U-smile Stats – All rights reserved



Indirect and direct standardization applies to frequency coefficients, e.g. prevalence rates. Direct comparison of raw rates between different populations, usually living in different geographical areas and differing in factors that may influence the rates (e.g. age), may hide the differences. In order to remove the influence of population structure on these confounding characteristics, comparisons can be made using standardized coefficients. Standardization offers a mechanism to ”filter out” the influence of a known confounding factor (e.g. age) and makes standardized coefficients obtained from different populations comparable. A necessary step in the process of standardizing coefficients is the selection of a standard population. The standard population for a population occupying a certain geographical area at a certain time may be a population covering a larger geographical area, including the study area, e.g. if the study population is the population of Wielkopolska voivodeship, the population of Poland may be used as a standard population. It is also possible to select a completely different population, geographically distant from the study population. However, it is advisable that the population chosen should be the reference population not only for the study in hand, but also for many other researchers. This is because it gives the opportunity to compare the results of studies using the same standard population. When choosing a population, it is recommended to pay attention to several aspects of the selection, such as:

- if several populations are being compared, a shared standard population minimizes the variance (variability) of the resulting standard coefficients;
- in trend analysis, the recommended standard population is the one that represents the average structure for the analyzed time period;
- the standard population should be as similar as possible to the studied population;
- the same standard population should be consistently chosen to ensure comparability of studies (choosing a different standard population than the one commonly used means that all historical data would have to be recalculated).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Age and gender are the most commonly used characteristics for standardization, however, the standardization can also be based on other, arbitrary characteristics, which due to their obvious impact on the phenomenon should be ”filtered out” from the study. Such features are called disturbing or confounding features. When choosing a characteristic for which we want to standardize, it should be remembered that standardization will be possible if we have sufficient information about the distribution of that characteristic in the population under study and in the standard population (Table 1). In addition, standardization by selected trait, e.g., by age, compensates to some extent for the influence of other age-related confounding factors, such as lifestyle, and standardization by gender compensates for those factors that are gender-related, such as occupation. The compensation of other factors is thus an important aspect in the selection of the trait against which standardization is performed.

Types of standardization:

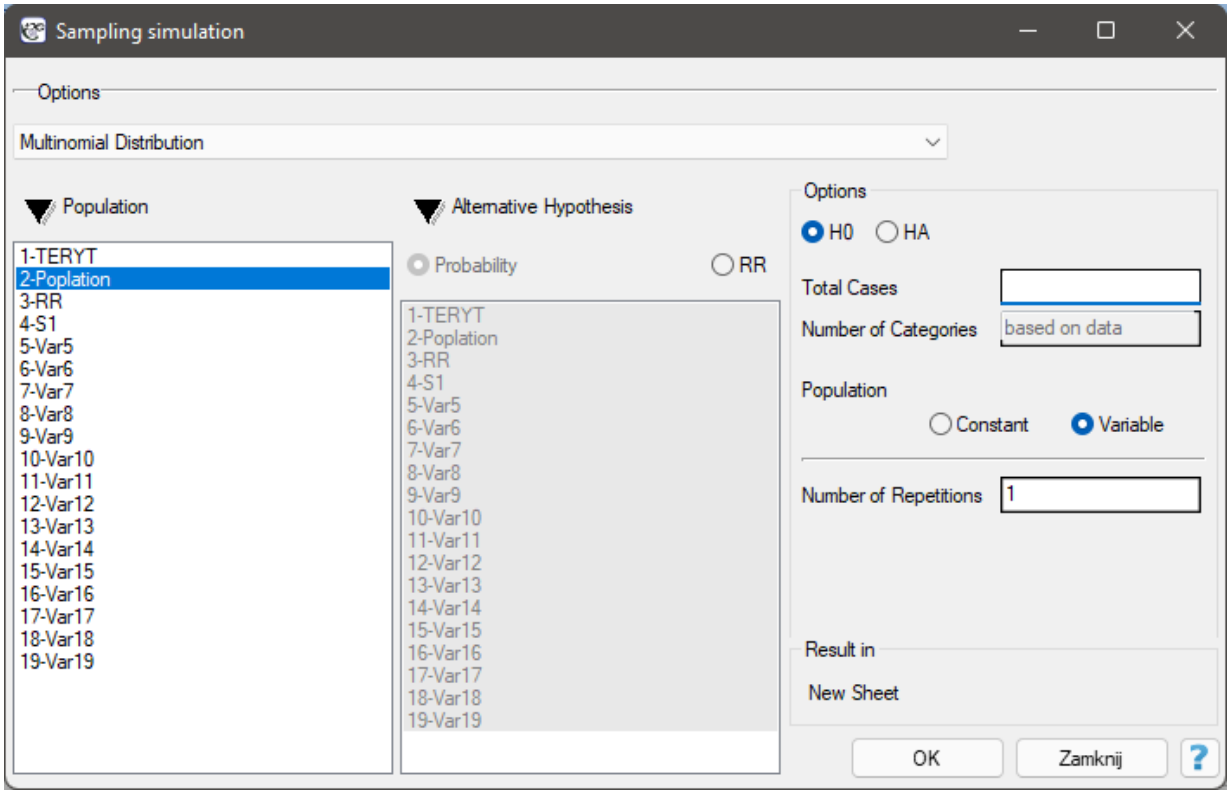
- **direct standardization** – the standardized prevalence rate obtained with this method gives what the prevalence of the disease would look like in the study population if it had the structure (e.g., age structure) of the reference population;
- **indirect standardization** – the standardised prevalence rate obtained with this method gives an indication of what the prevalence of a disease in the study population would look like if the prevalence of the disease in the study population were the same in particular categories (e.g. age categories) as in the reference population.

SAMPLING SIMULATION



The sampling window is opened via Data→Sampling simulation ...

Copyright ©2010-2026 U-smile Stats – All rights reserved



Sampling simulation is a way of generating multinomial distribution data. It involves assigning a given number of cases to categories, in a user-specified manner. The generated data is returned in a new datasheet. The generation can be repeated, so that the datasheet will have many generated columns depending on the number of repetitions of this operation set in the sampling window.

Options:

H0 - the null hypothesis assumes an even distribution of all cases across categories.

HA - the alternative hypothesis assumes an uneven distribution of cases. Selecting this option requires indicating categories with higher probability or relative risk. Information about the defined probability or relative risk for each category should be entered into the selected datasheet column prior to conducting the analysis.

Probability should be defined as a value between 0 and 1, with the sum of the probabilities given for all categories being 1.

Relative Risk defines risk relative to other categories and is a value greater than 1 for the increased risk category and a fraction less than 1 for the reduced risk category.

Setting the probability or relative risk values to the same level for all categories, is the same as the distribution for H0.

Constant population - assumes that the user is interested in distributing the cases according to the proposed distribution.

Variable population - assumes that the user is interested in distributing the cases such that the proportion of cases to the population is distributed according to the proposed distribution.

EXAMPLE 7.3. (simulations.pqs file)
As a basis for the simulation, the population of Wielkopolska in 2013 was used, which according to

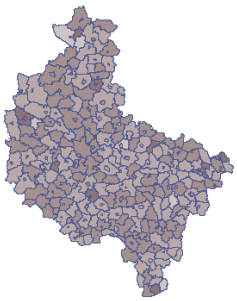
Copyright ©2010-2026 U-smile Stats – All rights reserved

the CSO was N = 3467016 people. The voivodeship is divided into 315 municipalities. The municipalities differ significantly in the number of inhabitants. The most populous municipality (the capital of the province) 548028 inhabitants, the least populous 1454 inhabitants, the median and quartiles are respectively: 6298 (4462; 9621) inhabitants. Assuming that in 2013 there were 6934 residents of the voivodeship with disease X, it is necessary to simulate the distribution of the sick people in such a way as to obtain:

1. Random distribution (based on data from the ”Random” datasheet)
2. Four times higher disease frequency in the indicated municipalities than in the rest of the voivodeship (based on data from the ”Clusters” datasheet)

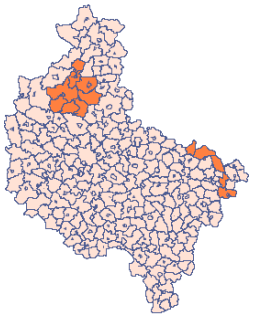
Ref 1. It should be noted that an uniform random distribution of 6934 patients does not imply a similar number of patients in each municipality. It is known that municipalities with a larger number of those at risk should have a corresponding larger number of patients than those with a smaller population. It is therefore of interest to distribute the patients in such a way that the ratio of patients to population is relatively constant. This implies accepting the null hypothesis H0 and the population variable. The number of individual municipalities was recorded in a column named: population.

The data drawn based on these assumptions are presented in the first column of the new datasheet. To be able to observe the random distribution of illness rates across municipalities, copy the resulting data into the ”Random” datasheet of column ”S1”. The formula in column 7 will then be recalculated (you can view and change the formula by setting Codes/Labels/Format in the column properties). On the map, the result is shown using the Map manager from the Spatial Analysis menu. The proportion of patients to population in each municipality is then plotted. An example of the result is shown on the map below.



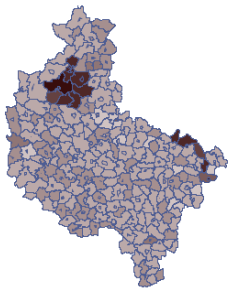
Ref 2. In the ”Clusters” datasheet, as in the previous task, the frequency for the study population is given. This time the higher frequency is expected in some municipalities (indicated on the map), so in addition, in the next column of the datasheet, the value of the relative risk for individual municipalities is presented, setting it to 4, for municipalities with increased risk and 1 for the remaining municipalities.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Appropriate sampling requires that you select the alternative hypothesis HA (by selecting the relative risk column) and the population variable (by indicating the population size column of the municipalities). The data drawn under these assumptions are presented in the first column of the new datasheet.

To be able to observe the distribution of the coefficient, assuming greater risk in the indicated municipalities, copy the result obtained to the datasheet ”Clusters” column ”S1”. The formula in column 7 will then be recalculated. On the map, the obtained result is presented using the Map manager from the Spatial Analysis menu. The proportion of ill people to population in each municipality is then plotted. An example of the result is shown on the map below.

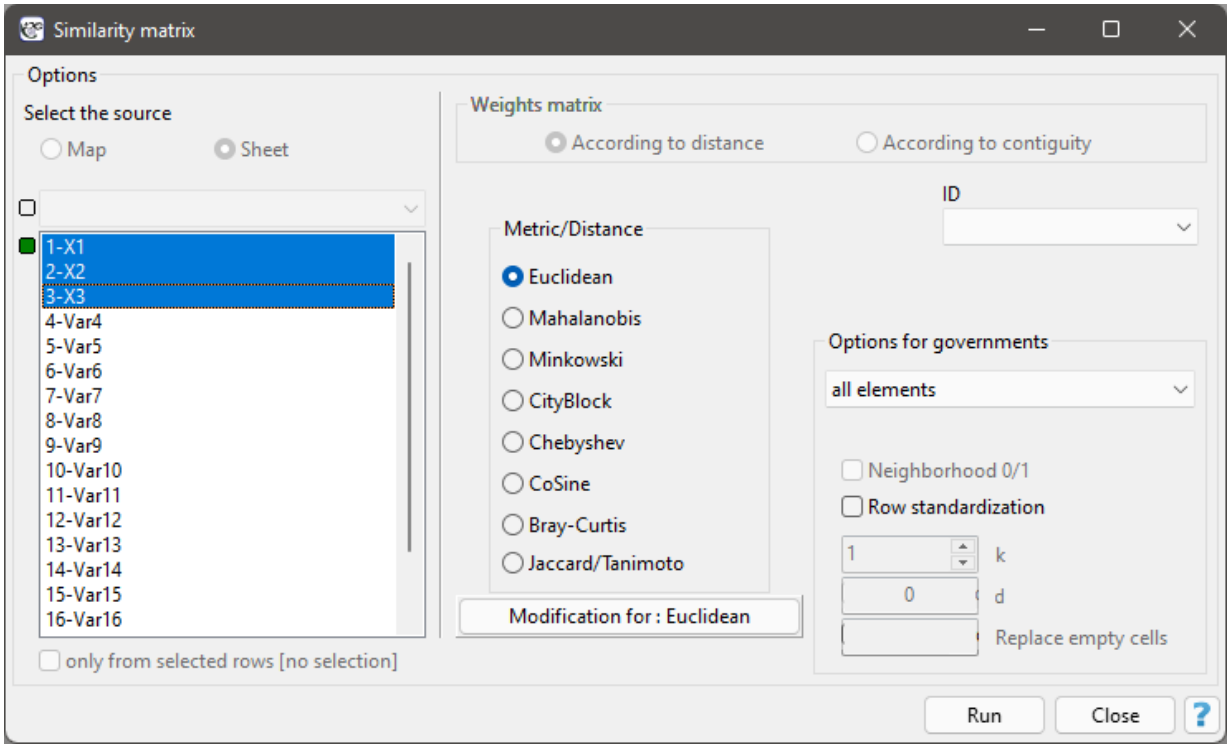


SIMILARITY MATRIX

The relationship between objects can be expressed by their distances or more generally by their dissimilarity. The further apart the objects are, the more dissimilar they are, while the closer together they are, the greater their similarity. It is possible to examine the distance of objects in terms of many features, e.g. when compared objects are cities, their similarity can be defined, among others, based on: length of the road that connects them, population density, GDP per capita, pollution emissions, average real estate prices, etc. With so many different features you have to choose the measure of distance in such a way, that it best reflects the actual similarity of objects.

The window with the settings for the macierzy podobieństwa options is opened via Data→Matrices...

Copyright ©2010-2026 U-smile Stats – All rights reserved



We express the dissimilarity/similarity of objects by means of distances which are most often **metrics**. However, not every measure of distance is a metric. For a distance to be called a metric it must meet 4 conditions:

1. the distance between objects cannot be negative: $d(x_1, x_2) \geq 0$,
2. the distance between two objects is 0 if and only if they are identical: $d(x_1, x_2) = 0 \iff x_1 = x_2$,
3. the distance must be symmetric, i.e., the distance from object x_1 to x_2 must be the same as from x_2 to x_1: $d(x, y) = d(y, x)$,
4. the distance must meet the triangle requirement: $d(x, z) \leq (x, y) + d(y, z)$.

Note! Metrics should be calculated for features with similar value ranges. Otherwise, features with higher ranges will have a greater impact on the similarity score than those with lower ranges. For example, when calculating the similarity of people based on body weight and age, body weight in kilograms (range 40-150) will have a greater impact on the result than age in years (range 18-90). To keep the influence of each feature balanced, it is worth normalizing/standardizing each of them before the analysis. If you want to decide the strength of this influence yourself, you can enter your own weights for the chosen metric after standardization. **Distance/Metrics:**

Euclidean

When talking about distance without defining its type, one assumes that it is Euclidean distance - the most common type of distance that is a natural part of real world models. Euclidean distance is a metric and is given by the formula:

$$d(x_1, x_2) = \sqrt{\sum_{k=1}^{n} (x_{1k} - x_{2k})^2}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

Minkowski

The Minkowski distance is defined for parameters p and r equal to each other - it is then a metric. This type of metric allows one to control the similarity calculation process by specifying values of p and r included in the formula:

$$d(x_1, x_2) = \sqrt[p]{\sum_{k=1}^{n} |x_{1k} - x_{2k}|^r}$$

Increasing the parameter r increases the weight assigned to the difference between objects for each feature, changing p gives more/less importance to closer/farther objects. If r and p are equal to 2, Minkowski distance reduces to Euclidean distance, if they are equal to 1, to CityBlock distance, and with these parameters approaching infinity, to Chebyshev metric.

CityBlock (otherwise: Manhattan distance or cab distance)

This is a distance that allows you to move in only two directions perpendicular to each other. This type of distance is similar to moving on perpendicularly intersecting streets (a square street grid that resembles the layout of Manhattan). This metric is given by the formula:

$$d(x_1, x_2) = \sum_{k=1}^{n} |x_{1k} - x_{2k}|$$

Chebyshev

The distance between the objects being compared is the greatest of the distances obtained for each characteristic of those objects:

$$d(x_1, x_2) = \max_k |x_{1k} - x_{2k}|$$

Mahalanobis

Mahalanobis distance is also called statistical distance. It is a distance weighted by the covariance matrix, by which objects described by mutually correlated characteristics can be compared. The use of Mahalanobis distance has two main advantagesi:

1) Variables for which larger variances or larger ranges of values are observed do not have an increased effect on the Mahalanobis distance score (since when using a covariance matrix you standardize the variables using the variance located on the diagonal). As a result, there is no requirement to standardize/normalize the variables before proceeding with the analysis.

2) It takes into account the mutual correlation of the characteristics describing the compared objects (using the covariance matrix it uses the information about the relationship between the characteristics located outside the diagonal of the matrix).

$$d(x_1, x_2) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}$$

The measure calculated in this way meets the conditions of the metric.

CoSine

The Cosine distance should be calculated using positive data because it is not a metric (it does not meet the first condition: $d(x_1, x_2) \geq 0$). So if you have features that also take negative values you should transform them beforehand using, for example, normalization to an interval spanned

Copyright ©2010-2026 U-smile Stats – All rights reserved

by positive numbers. The advantage of this distance is that (for positive arguments) it is limited to the range [0, 1]. The similarity of two objects is represented by the angle between two vectors representing the features of those objects.

$$d(x_1, x_2) = 1 - K,$$

where K is the similarity coefficient (cosine of the angle between two normalized vectors):

$$K = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2} \sqrt{\sum_{k=1}^{n} x_{2k}^2}}$$

Objects are similar when the vectors overlap - then the cosine of the angle (similarity) is 1 and the distance (dissimilarity) is 0. Objects are different when the vectors are perpendicular - then the cosine of the angle (similarity) is 0 and the distance (dissimilarity) is 1.

Example - comparing texts
Text 1: several people got on at this stop and one person got off at the next stop
Text 2: at the bus stop, one lady got off and several got on
One wants to know how similar the texts are in terms of the number of the same words, but is not interested in the order in which the words occur.

You create a list of words from both texts and count how often each word occurred:

SŁOWA	na	tym	przystanku	wsiadło	kilka	osób	a	następnym	wysiadła	jedna	Pani
tekst 1	2	1	1	1	1	1	1	1	1	1	0
tekst 2	1	0	1	1	1	0	1	0	1	1	1

The cosine of the angle between the vectors is 0.784465, so the distance between them is not large $d(tekst_1, tekst_2) = 1 - 0,784465 = 0.215535$.
In a similar way, you can compare documents by the occurrence of keywords to find those most relevant to the search term.

Bray-Curtis

The Bray-Curtis distance (measure of dissimilarity) should be calculated using positive data because it is not a metric (it does not meet the first condition: $d(x_1, x_2) \geq 0$). If you have features that also take negative values, you should transform them beforehand using, for example, normalization to an interval spanned by positive numbers. The advantage of this distance is that (for positive arguments) it is limited to the interval [0, 1], where 0 - means that the compared objects are similar, 1 - dissimilar.

$$d(x_1, x_2) = \frac{\sum_{k=1}^{n} |x_{1k} - x_{2k}|}{\sum_{k=1}^{n} (x_{1k} + x_{2k})} \tag{8}$$

In calculating the similarity measure BC, we subtract the Bray-Curtis distance from the value 1:

$$BC = 1 - d(x_1, x_2) \tag{9}$$

Jaccard

Jaccard's distance (measure of dissimilarity) is calculated for binary variables (Jaccard, 1901), where 1 indicates the presence of a feature 0- its absence.

		object 1	
		1	0
object 2	1	a	b
	0	c	d

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Jacckard distance is expressed by the formula:

$$d(x_1, x_2) = 1 - J. \tag{10}$$

where:

$J = \frac{a}{a+b+c}$ - Jaccard similarity coefficient.

Jaccard similarity coefficient is in the range [0,1], where 1 means the highest similarity, 0 - the lowest. Distance (dissimilarity) is interpreted in the opposite way: 1 - means that the compared objects are dissimilar, 0 - that they are very similar. The meaning of Jaccard's similarity coefficient is well described by the situation involving the choice of goods by customers. By 1 we denote the fact that the customer bought the given product, 0 - the customer did not buy this product. Calculating the Jaccard coefficient you compare 2 products to find out what part of the customers buy them in together. Of course we are not interested in information about customers who did not buy either of the compared items. Instead, we are interested in how many people who choose one of the compared products choose the other one at the same time. The sum $a+b+c$ - is the number of customers who chose either of the compared items, a - is the number of customers choosing both items at the same time. The higher Jaccard's similarity coefficient, the more inseparable the products are (the purchase of one is accompanied by the purchase of the other). The opposite will happen when we get a high Jaccard dissimilarity coefficient. It will indicate that the products are highly competitive, i.e. the purchase of one will result in the lack of purchase of the other.

The formula for Jaccard's similarity coefficient can also be written in general form:

$$J = \frac{\sum_{k=1}^{n} x_{1k}x_{2k}}{\sum_{k=1}^{n} x_{1k}^2 \sum_{k=1}^{n} x_{2k}^2 - \sum_{k=1}^{n} x_{1k}x_{2k}}$$

proposed by Tanimoto (1957). An important feature of Tanimoto's formula is that it can also be computed for continuous features.
For binary data, Jaccard's and Tanimoto's dissimilarity/similarity formulas are the same and meet the conditions of the metric. However, for continuous variables, Tanimoto's formula is not a metric (does not meet the triangle condition).

Example - comparison of species
We study the genetic similarity of members of three different species - in terms of the number of genes they share. If a gene is present in an organism, we give it the value 1, 0 - in the opposite case. For the sake of simplicity only 10 genes are analysed.

GENES	gene1	gene2	gene3	gene4	gene5	gene6	gene7	gene8	gene9	gene10
specimen1	0	1	1	1	1	1	1	0	1	0
specimen2	0	0	1	1	1	1	1	0	1	0
specimen3	1	0	1	1	0	0	1	0	0	0

The calculated similarity matrix is as follows:

OBJECTS	specimen1	specimen2	specimen3
specimen1	0	0.857143	0.375
specimen2	0.857143	0	0.428571
specimen3	0.375	0.428571	0

Specimens 1 and 2 are most similar and 1 and 3 are least similar: - The Jaccard similarity of specimen 1 and specimen 2 is 0.857143, i.e., just over 85% of the genes found in the two compared

Copyright ©2010-2026 U-smile Stats – All rights reserved

species are shared by them.
- The Jaccard similarity of specimen1 and specimen 3 is 0.375, meaning that more than 37% of the genes found in the two compared species are shared by them.
- The Jaccard similarity of specimen 2 and specimen 3 is 0.428571, i.e. almost 43% of the genes found in the two compared species are common to them.

Similarity matrix options are used to indicate how the elements in the matrix are returned. By default, all elements of the matrix are returned as they were computed according to the adopted metric. You can change this by setting:

Elements of a matrix:

- minimum - means that in each row of the matrix only the minimum value and the value on the main diagonal will be displayed;
- maximum - means that in each row of the matrix only the maximum value and the value on the main diagonal will be displayed;
- k minimum - means that each row of the matrix will display as many smallest values as indicated by the user by entering the value of k and the value on the main diagonal;
- k maximum - means that each row of the matrix will display as many largest values as indicated by the user by entering k and the value on the main diagonal;
- elements below d - means that in each row of the matrix those elements will be displayed whose value is smaller than the user specified value d and the value on the main diagonal;
- elements above d - means that each row of the matrix will display those elements whose value is greater than the user-specified value of d and the value on the main diagonal;

Neighborhood 0/1

Choosing the option Neighborhood $0/1$, we replace the values inside the matrix with 1, and empty spaces with 0. This way, we mark, for example, whether the objects are neighbors (1) or not (0), that is, we determine the neighborhood matrix.

Row standardization

Row standardization means that each item of the matrix is divided by the sum of the matrix row. The resulting values are between 0 and 1.

Replace empty cells

Opcja Replacacing empty cells allows you to enter the value to be placed in the matrix in place of any empty items.

The selected object ID allows you to name the rows and columns of the similarity matrix according to the naming stored in the indicated variable.

EXAMPLE 7.4. (file: flatsSimilarities.pqs)
In real estate valuation procedures, for both substantive and legal reasons, the issue of similarity plays an important role. For example, it is an essential prerequisite for grouping objects and assigning them to an appropriate segment.

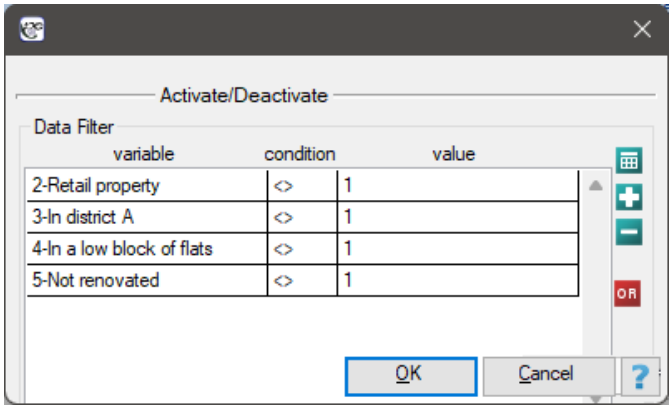
Let's assume that a real estate agent is approached by a person looking for an apartment, who defines those features that the apartment must have and those that have a big influence on the purchase decision but are not decisive. The features that the property must have are:

Copyright ©2010-2026 U-smile Stats – All rights reserved

- a residential property,
- located in district A,
- located in a low-rise multi-family housing area (up to 5 storeys),
- not renovated (average or deteriorated condition).

Data for these locations are summarized in a table, where 1 indicates that the location meets the search conditions, 0 that it does not.

Those locations that do not meet the search conditions will be excluded from the analysis by deactivating the corresponding rows. Deactivate the rows that do not meet any of the search conditions via menu Edit→Activate/Deactivate (filter)....



Remember to combine the deactivation conditions with the alternative (replace AND with OR).
As a result, 11 locations were identified (locations 10, 12, 17, 35, 88, 101, 105, 122, 130, 132, 135) that fit this segment (meeting all 4 conditions).

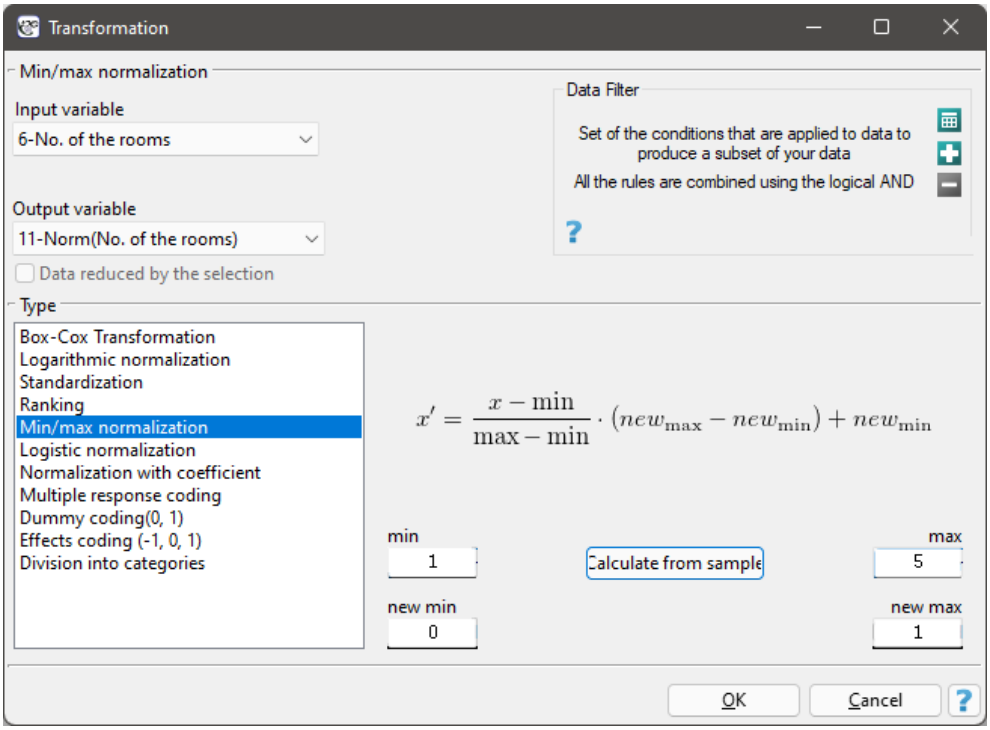
Now we will consider those features that have a strong influence on the customer's decision, but are not decisive:

- Number of rooms = 3;
- The floor on which the apartment is located = 0;
- Age of the building in which the apartment is located = approx. 3 years;
- Proximity of the A district to the city center (time it takes to get to the city center) = approx. 30 min;
- Proximity to public transport station = approx. 80 m.

Copyright ©2010-2026 U-smile Stats – All rights reserved

	Number of rooms	Apartment floor	Age of the building	Proximity to the city center	Proximity to the public transport station
Wanted	3	0	3	30	80
Lokal 10	2	1	1	0	150
Lokal 12	1	2	1	0	200
Lokal 17	3	1	7	20	500
Lokal 35	2	0	6	5	100
Lokal 88	3	4	6	5	200
Lokal 101	4	2	10	0	10
Lokal 105	2	2	6	0	50
Lokal 122	1	0	6	5	100
Lokal 130	2	0	10	0	20
Lokal 132	3	5	6	30	400
Lokal 135	3	1	6	5	100

Note that the last feature, the distance of the public transportation station, is expressed by much larger numbers than the other features of the compared locations. As a result, this feature will have a much greater impact on the distance matrix than the other features. To prevent this from happening before the analysis, normalize all the features by choosing a common range from 0 to 1 for them - to do this, use the Data→Normalization/Standardization... menu. In the normalization window, set "Number of rooms" as the input variable and an empty variable called "Norm(Number of rooms)" as the output variable; the normalization type is normalization min/max; the values of min and max are calculated from the sample by selecting the Calculate from sample button - the normalization result will be returned to the datasheet when the Ok button is clicked. The normalization is repeated for the following variables ie: "Floor", "Building Age", "Distance to Center" and "Station Distance".



The normalized data is shown in the table below.

Copyright ©2010-2026 U-smile Stats – All rights reserved

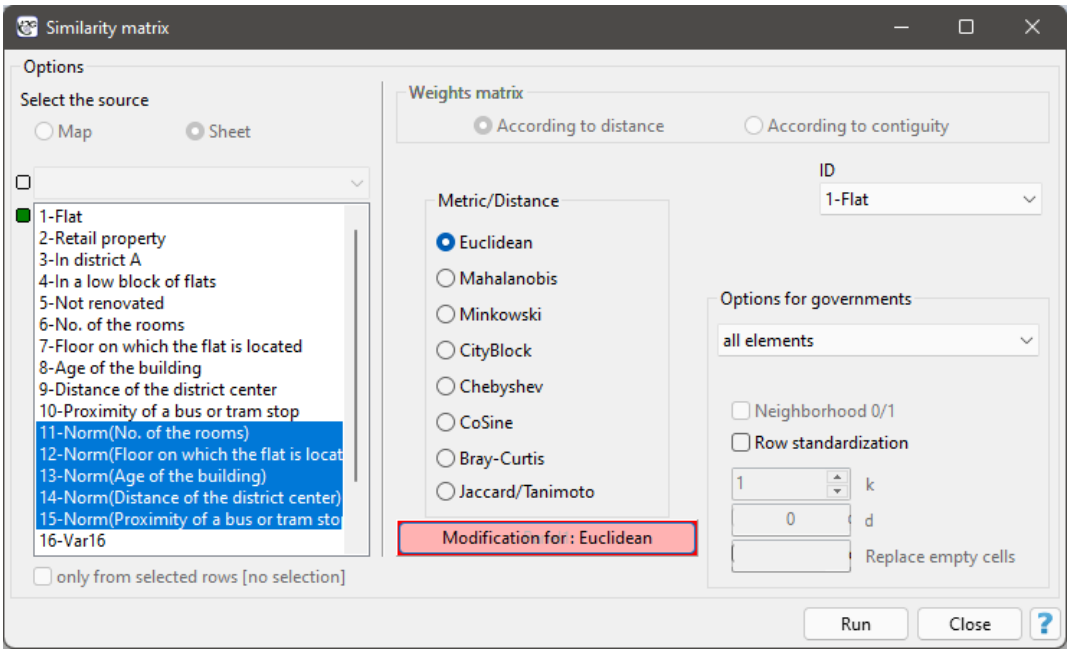
	Norm(Number of rooms)	Norm(Apartment floor)	Norm(Age of the building)	Norm(Proximity to the city center)	Norm(Proximity to the public transport station)
Poszukiwany	0,666666667	0	0,222222222	1	0,142857143
Lokal 10	0.333333333	0.2	0	0	0.285714286
Lokal 12	0	0.4	0	0	0.387755102
Lokal 17	0.666666667	0.2	0.666666667	0.666666667	1
Lokal 35	0.333333333	0	0.555555556	0.166666667	0.183673469
Lokal 88	0.666666667	0.8	0.555555556	0.166666667	0.387755102
Lokal 101	1	0.4	1	0	0
Lokal 105	0.333333333	0.4	0.555555556	0	0.081632653
Lokal 122	0	0	0.555555556	0.166666667	0.183673469
Lokal 130	0.333333333	0	1	0	0.020408163
Lokal 132	0.666666667	1	0.555555556	1	0.795918367
Lokal 135	0.666666667	0.2	0.555555556	0.166666667	0.183673469

Based on the normalized data, we will determine the locations that most closely matches the customer's request. To calculate the similarity we will use the Euclidean distance metric. The smaller the value, the more similar the units are. The analysis can be performed assuming that each of the five features mentioned by the client are equally important, but it is also possible to indicate those features that should have greater influence on the result of the analysis. We will construct two Euclidean distance matrices:

(1) The first matrix will contain the Euclidean distances calculated from the equivalently treated five features;

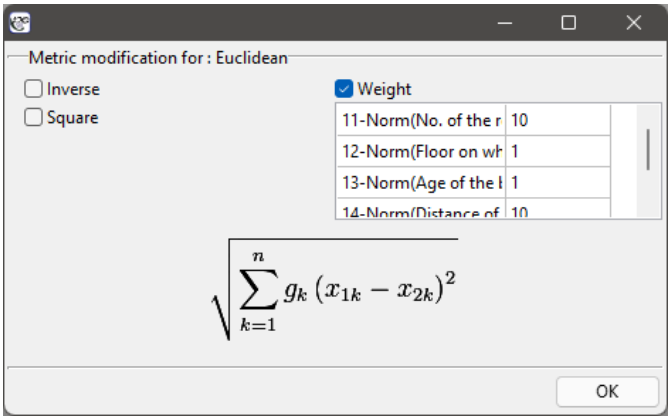
(2) The second matrix will contain the Euclidean distances, in the construction of which the number of rooms and the distance to the city center will be the most important.

To build the first matrix, in the similarity matrix window, we select 5 normalized variables labeled as Norm, the Euclidean metric, and as Object Identifier the variable " Location".



Copyright ©2010-2026 U-smile Stats – All rights reserved

To build the second matrix, we use the same settings in the similarity matrix window as we did to build the first matrix, but in addition, we select the Modification for : Euclidean button and in the modification window, we enter larger values for ”Number of rooms” and ”Distance to city center”, e.g., equal to 10, and smaller values for the other features, e.g., equal to 1.



This will result in two matrices. In each of them, the first column refers to how similar it is to the location the customer is looking for:

Euclidean	**Wanted**	...
Wanted	0	...
Location 10	1.10	...
Location 12	1.31	...
Location 17	1.04	...
Location 35	0.96	...
Location 88	1.23	...
Location 101	1.38	...
Location 105	1.18	...
Location 122	1.12	...
Location 130	1.32	...
Location 132	1.24	...
Location 135	0.92	...

Euclidean with scales	**Wanted**	...
Wanted	0	...
Location 10	3.35	...
Location 12	3.84	...
Location 17	1.44	...
Location 35	2.86	...
Location 88	2.78	...
Location 101	3.45	...
Location 105	3.37	...
Location 122	3.39	...
Location 130	3.43	...
Location 132	1.24	...
Location 135	2.66	...

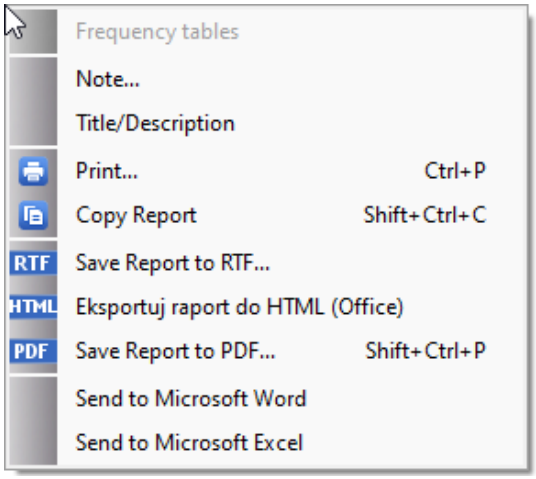
According to the unmodified Euclidean distance, location 35 and location 135 should most closely match the client’s requirements. When the weights are taken into account, locations 17 and 132 are the closest to the client’s requirements - these are the locations that are primarily similar in terms of the number of rooms required by the client (3) and the indicated distance to the center, with the other 3 features having a smaller impact on the similarity score.

Copyright ©2010-2026 U-smile Stats – All rights reserved

7.2 WORKING WITH RESULTS SHEETS (REPORTS)

A report is a project element used to store the results of performed statistical analyses. It is included in the project automatically and assigned to the active datasheet when the selected statistical procedure is completed. A report is not editable, except for its chart and title. You can edit the graph by double-clicking the mouse or through the right-click context menu. Editing of the title is done in Project Manager by adding or changing the description.

The most important report related operations can be done through the right-click context menu in the report window:



- **Printing**
  Printing options are available through:
  – context menu,
  – menu File→Print...

- **Exporting and sending reports**
  Reports created in ℙℚStat can be exported to a file in *.rtf, *html and *.pdf formats. They can also be exported to Word or Excel documents.
  If you are exporting in the Project Manager, you can put the reports in separate files or in one common file. To do this, select the chosen reports and then select the button and export to a file or files of the chosen format. Exporting of individual reports can be done separately through the context menu in the report window.

- **Describing reports**
  Reports can be described in Project Manager or the report window by adding a title or a note.

- **Editing charts**
  Editing the chart regarding its general and specific options is available through the context menu in the report window.

- **Copying reports**
  Using the system clipboard you can also transfer analysis results to other programs such as Word, Excel.

- **Deleting reports**
  Deleting a report is possible via:

Copyright ©2010-2026 U-smile Stats – All rights reserved

– context menu Delete report (Shift+Del) on the report name in the Navigation tree,
– Project Manager.

Note! If map layers are attached to a report, deleting the report also deletes all layers assigned to it. **Changing the report order** is possible via the context menu of the right mouse button Move up (Ctrl+Up) or Move down (Ctrl+Down) on the report name in the Navigation tree.

Adding information to the report name in the Navigation tree such as:

- time of generation,
- description,
- filter,
- grouping variable name,
- variable name.

is possible after selecting the corresponding option in the program settings window.

7.3 MENU AND LANGUAGE SETTINGS

LANGUAGE

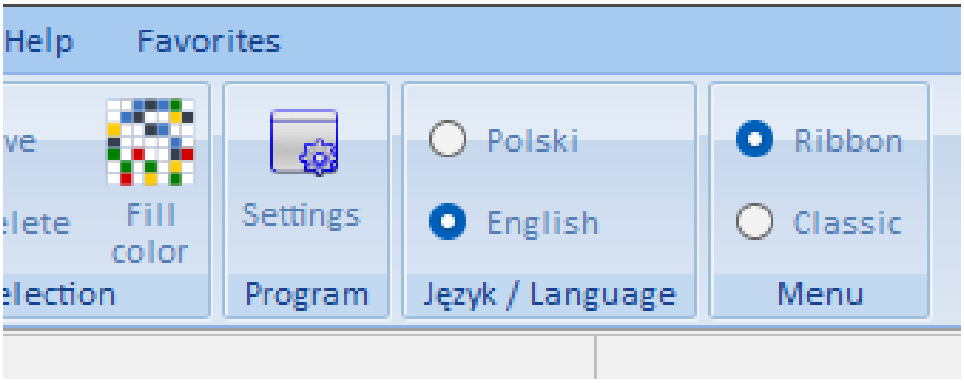
Both the program interface and the reports can be displayed in Polish and English. Changing the selected language does not require a restart of the program and can be done by selecting Język/Language from the Edit menu. Reports opened after a language change will be automatically translated (except for the procedure name, which is a description and is user editable).

MENU

Program menus can be displayed as Classic or as Ribbon. You can switch the way the menu is displayed by selecting the Edit menu and then the desired menu type.

Favorites
The Favorites menu provides quick access to frequently performed statistics and other operations. To configure the menu for your own needs, use the Favorites search box - you can search for any menu item and then add it to the menu or remove it from it.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Copyright ©2010-2026 U-smile Stats – All rights reserved

8 ORGANIZATION OF WORK WITH THE PROGRAM

All statistical analysis procedures are available in the Statistics menu.

8.1 FORM OF DATA ORGANIZATION

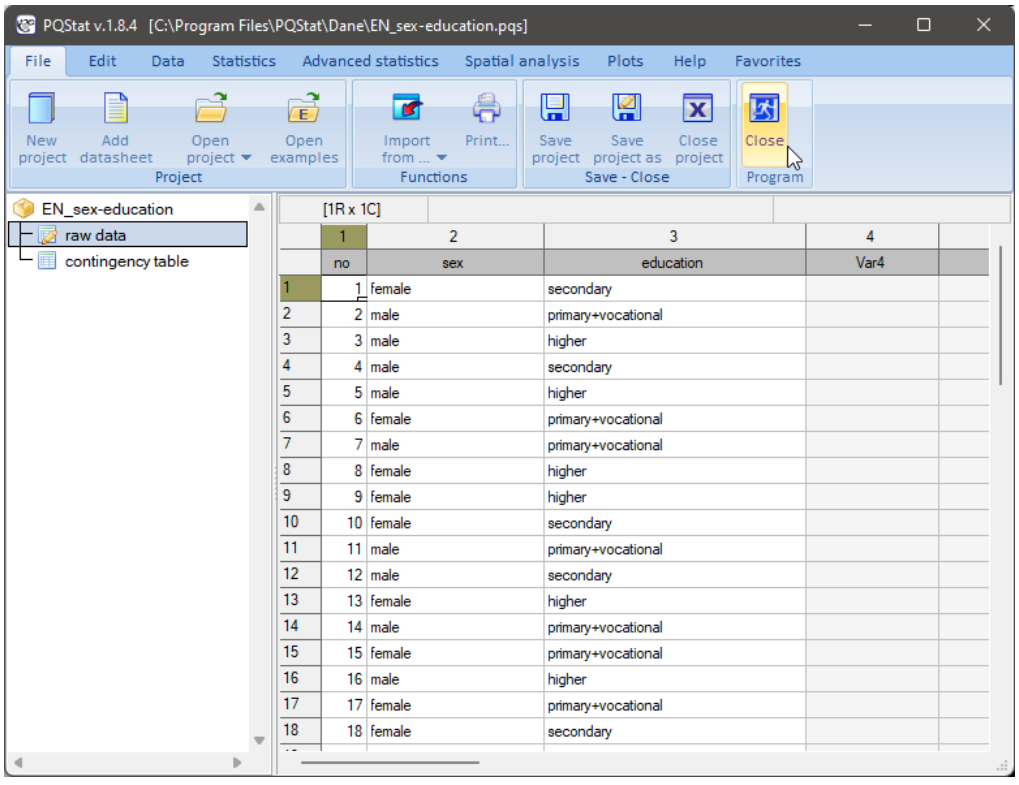
The form of data organization depends on the statistical procedures that the user plans to perform.

A statistical analysis may concern data collected in the form of a contingency table or in raw form. The data can be converted from one form to the other:

- from a contingency table to raw form – via the Data menu → Convert to raw data...,
- from raw form to a contingency table – via the Data menu → Convert to table...

1. Data in the form of raw records is data organised in such a way that each row contains information about a different studied object (patient, company, etc.)

   EXAMPLE 8.1. *Raw data* (sex-education.pqs file)

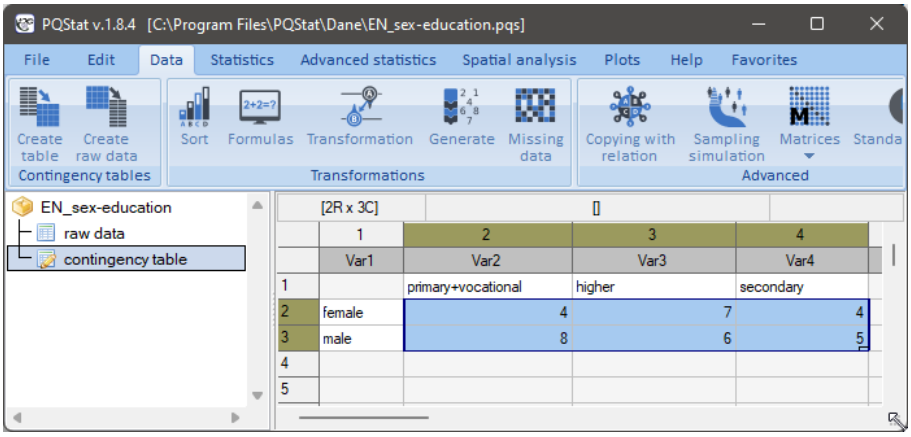


	1	2	3	4
	no	sex	education	Var4
1	1	female	secondary	
2	2	male	primary+vocational	
3	3	male	higher	
4	4	male	secondary	
5	5	male	higher	
6	6	female	primary+vocational	
7	7	male	primary+vocational	
8	8	female	higher	
9	9	female	higher	
10	10	female	secondary	
11	11	male	primary+vocational	
12	12	male	secondary	
13	13	female	higher	
14	14	male	primary+vocational	
15	15	female	primary+vocational	
16	16	male	higher	
17	17	female	primary+vocational	
18	18	female	secondary	

2. A contingency table is a table presenting the joint distribution of two variables. Its interior consists of the observed frequencies (natural numbers).

   EXAMPLE 8.2. *Contingency table* (sex-education.pqs file)

Copyright ©2010-2026 U-smile Stats – All rights reserved



Copyright ©2010-2026 U-smile Stats – All rights reserved

8.2 DATASHEET WORKING AREA

Usually, when performing a statistical analysis, we have access to the whole working area of the datasheet. However, the user can freely narrow this area by indicating a specific part of the sheet. Such an indication can be made in several ways:

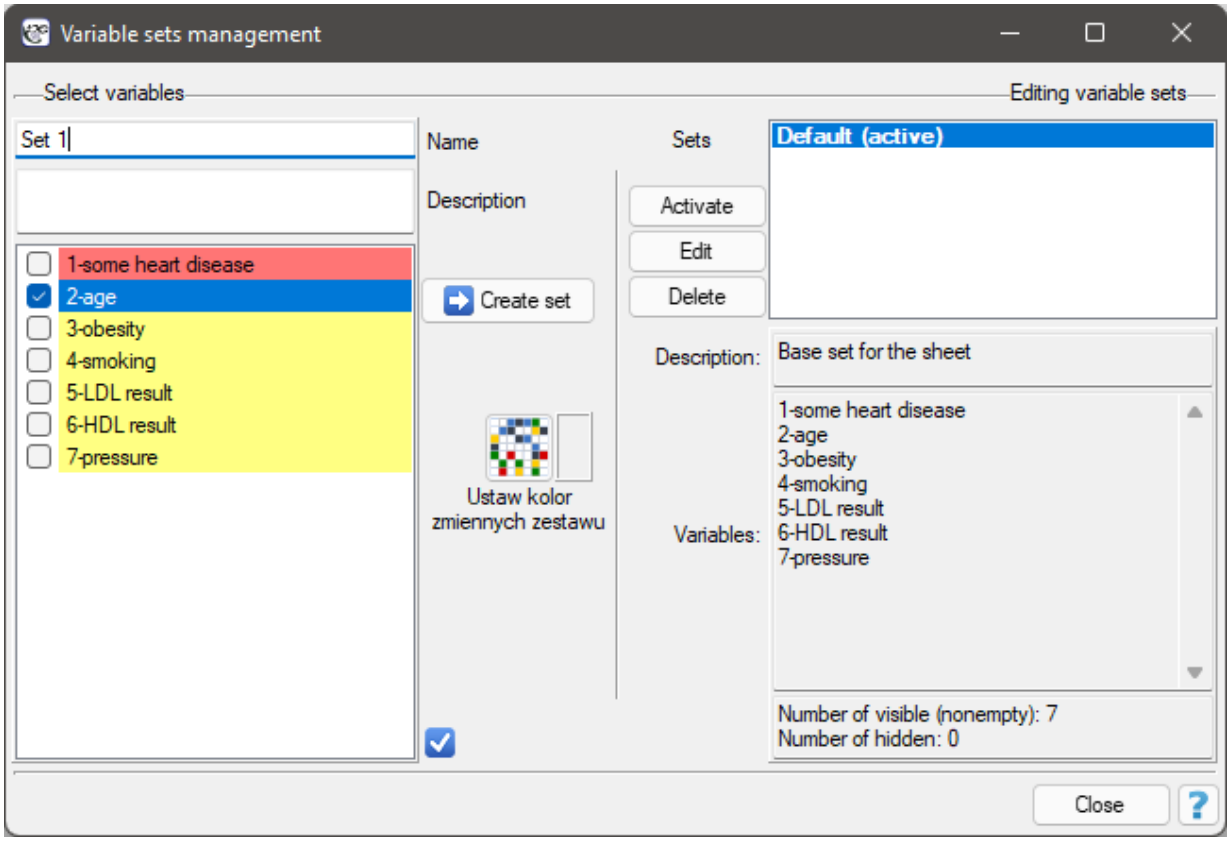
Organising variables into sets

Variable sets are subsets of variables. Defining various sets is intended to make it easier to select variables in analysis windows by limiting the displayed list of variables. The selected variable set is displayed in the variable lists of every analysis dialog box, as well as in the datasheet.

By default, the default set of variables is used. The default set is a set made up of all the variables.

DEFINING VARIABLE SETS

To define your own variable sets, double-click the name of the active set at the top of the datasheet window.



In the Variable Set Management window, we give the name of the new set (and, optionally, its description), and select on the list the variables that are to make up the set. We confirm the choice with the Create Set button. A variable may belong to multiple sets.

Copyright ©2010-2026 U-smile Stats – All rights reserved

EDITING VARIABLE SETS

To edit a set, in the Variable Set Management window we choose a set from the list of sets and confirm with the Edit button. In the variable selection window, the variables already selected for the set will appear checked, along with the unchecked variables that do not belong to the set. After checking/unchecking the variables as required, we confirm the choice with the Save Changes button.

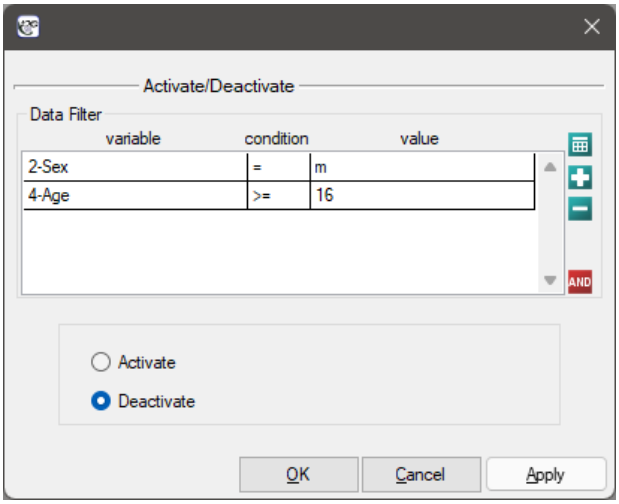
ACTIVATING THE SELECTED VARIABLE SET

Activating a set is done in the Variable Set Management window, or through the context menu at the top of the datasheet, on the name of the selected set.

Case activation/deactivation

Case activation/deactivation is a global option that overrides other area restrictions available in the program. Cases (rows) indicated as deactivated are shaded in the datasheet and do not take part in statistical analyses.
To activate or deactivate selected cases, choose one of the following options:

- indicate the rows in the datasheet and choose the Activate/Deactivate option from the context menu on their name;
- select the Edit →Activate/Deactivate (filter)... menu



EXAMPLE 8.3. (filter.pqs file)
We intend to carry out a number of statistical analyses on the data from the filter.pqs file. These analyses are not to concern boys who are 16 years old or older. For this purpose, we define the rows that will be excluded from the analysis: we click the button and set a rule for the variable gender; we click the button again and set a rule for the variable age. Remember: for this task to be carried out correctly, all filter conditions should be joined by a conjunction (indicated by the AND sign). We leave the Deactivate option selected, and confirm the analysis conditions set this way with the OK button.

When narrowing the working area of the datasheet, remember that filter conditions can be joined by a conjunction or by an alternative. We switch between the alternative and the conjunction

Copyright ©2010-2026 U-smile Stats – All rights reserved

using the AND OR buttons.
More examples of how to use the filter can be found here.

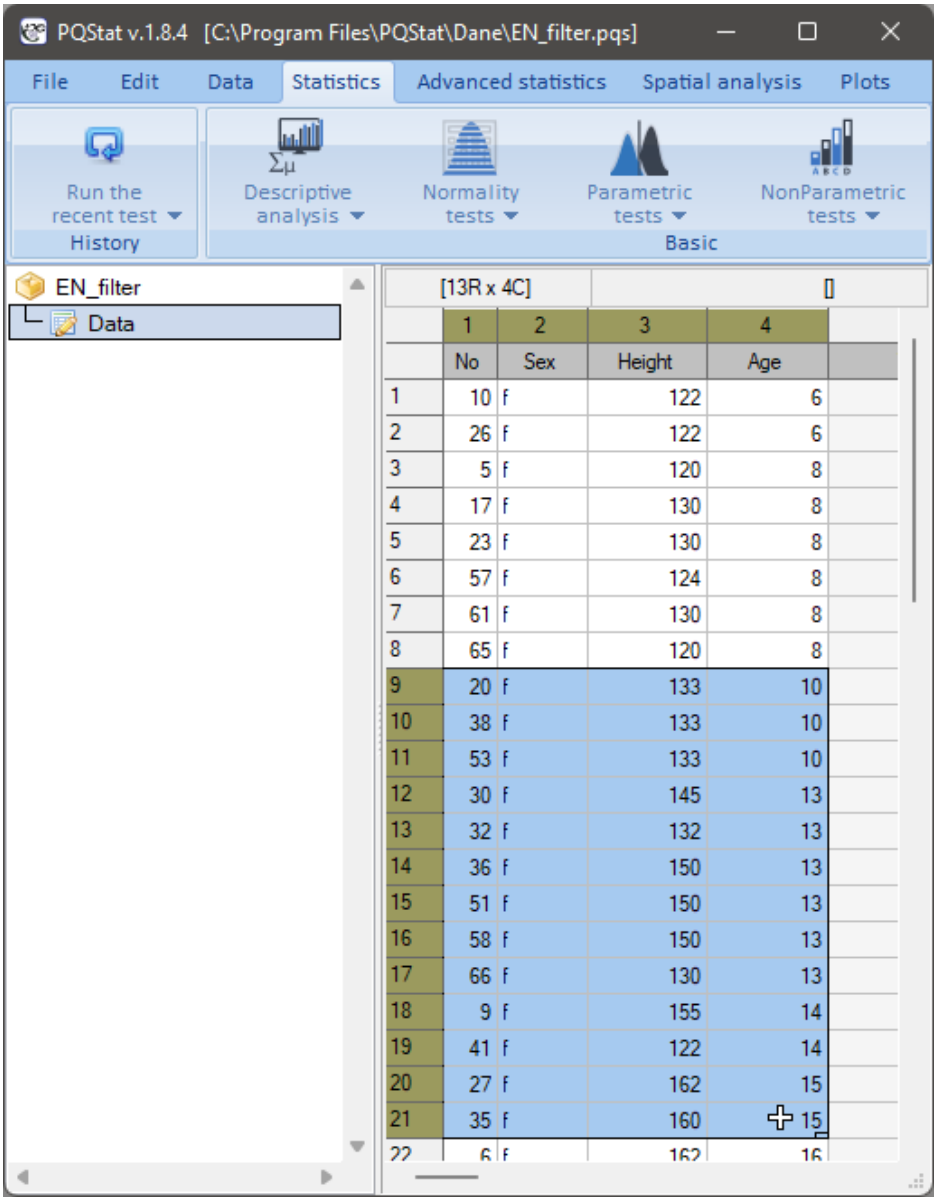
To activate all cases, select the Edit →Activate all menu.

Selecting an area

Selecting a coherent area causes the analysis we subsequently choose to be performed only on the rows contained in the selection, and to have available only the columns that lie inside the selection and contain data.

EXAMPLE 8.4. (filter.pqs file)
We want to determine the descriptive statistics for the height of girls between 10 and 15 years of age. We sort the data by the gender column and by the age column, and then select a coherent area containing the column with the height of girls between 10 and 15 years of age, and choose Statistics→Descriptive analyses→Descriptive statistics from the menu.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In the descriptive statistics test options window, we select the procedures we want to perform (e.g. we tick the mean, standard deviation, minimum and maximum) and the variable for the analysis (the height column), and confirm the choice with the OK button.

Narrowing the working area of the datasheet by selecting a coherent fragment of the data causes the following message to be shown in the analysis window:
Data restricted by selection

Saved selection

If saved selection ranges are assigned to the datasheet, they are marked with a border. They can be used in those analyses that allow the data to be entered directly in the analysis window. In that case, using the fill from saved selection button, the data lying within the chosen range can be pasted in.

Copyright ©2010-2026 U-smile Stats – All rights reserved

EXAMPLE 8.5. (layers.pqs file)
We want to determine the statistics related to the Odds Ratio (OR) analysis for several strata. We will use data stored in the form of 10 tables – they are highlighted (bordered). From the Statistics menu, we choose Analysis for strata → Mantel-Haenszel OR/RR. In the test options window, we tick contingency table and then set the number of strata to 10. Each of the resulting strata can now be filled from the saved selection. Once we have filled in all the tables, we perform the analysis by clicking the OK button.

Note! To assign another selection to the datasheet, choose Save Selection (Ctrl+T) from the Edit menu, and to remove the assigned selections, choose Delete Selections.

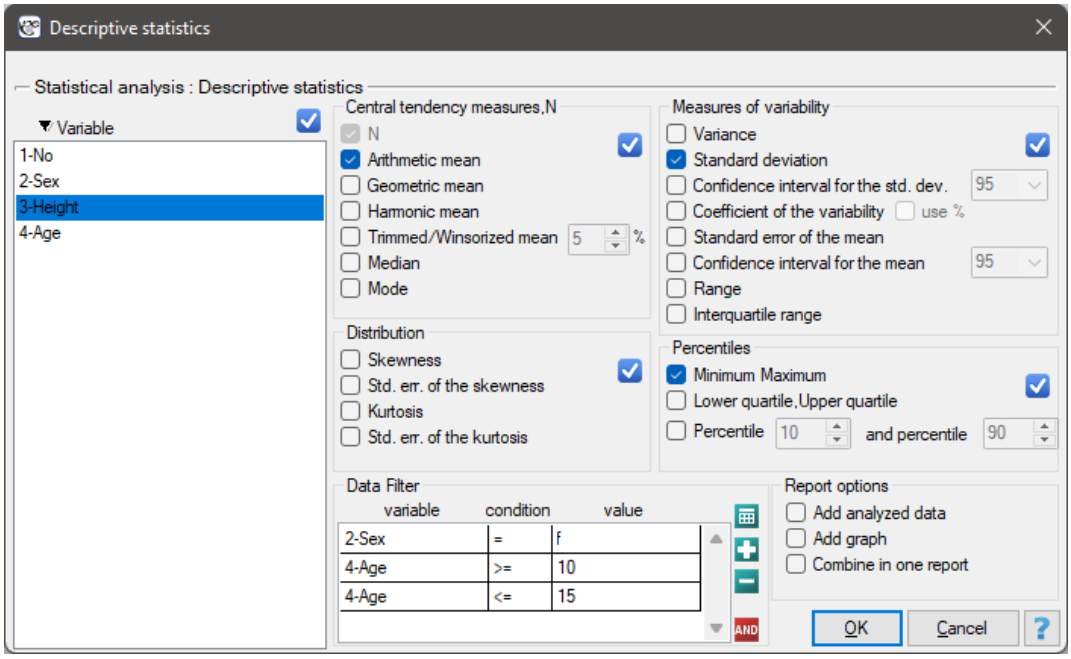
Data filter

Data Filter is an option available once any statistical analysis has been selected. Setting it reduces the number of rows taking part in that analysis. Either a basic filter or a multiple filter can be set.

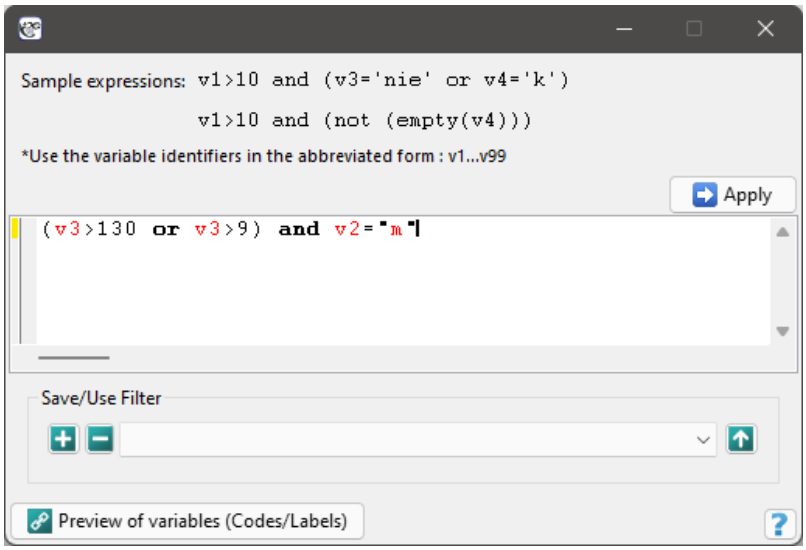
- **Basic filter** lets you select one particular subgroup. The choice can be made in two ways:
  - **Automatic method** – this way we can impose one condition on the collected data using the ➕ button, or several conditions by pressing ➕ again. These conditions will be joined by an alternative or a conjunction depending on whether we choose the AND or OR button. Conditions are removed using the ➖ button.

    EXAMPLE 8.6. *Automatic basic filter* (filter.pqs file)
    We want to determine the descriptive statistics for the height of girls between 10 and 15 years of age. From the menu, we choose Statistics→Descriptive analyses→Descriptive statistics. In the descriptive statistics test options window, we select the procedures we want to perform (e.g. we tick the mean, standard deviation, minimum and maximum) and the variable for the analysis (the height column). We set the filter by adding rules with the ➕ button. First we set a rule for the variable gender, choosing the equals sign as the condition and the letter g (girls) as the value. Next we add another condition and set a filter for the variable age, choosing >= as the condition and entering 10 as the value. In the same way we add the condition age <=15. Remember: for this task to be carried out correctly, all filter conditions should be joined by a conjunction (indicated by the AND sign). We confirm the analysis conditions set this way with the OK button.

Copyright ©2010-2026 U-smile Stats – All rights reserved



– **Advanced method** – this way we can write any rule, whether a fully simple one or a more complex one. We switch to the advanced method by toggling the button to and then selecting . This opens a window in which we can build the filter formula.



When setting up an advanced filter, we use variable numbers preceded by the letter *v*. The variable numbers and their contents can be checked using the Preview of variables (Codes/Labels) button.

The following logical functions are available:
and – the conjunction operator – checks whether all the conditions it joins are true,
or – the alternative operator – checks whether at least one of the given conditions is true,
xor – the exclusive-alternative operator – checks whether exactly one of the given conditions is true,
not – the negation operator,
empty(v1) – a function checking whether empty cells occur,
text(v1) – a function checking whether the cells contain text,
numer(v1) – a function checking whether the cells contain a number.

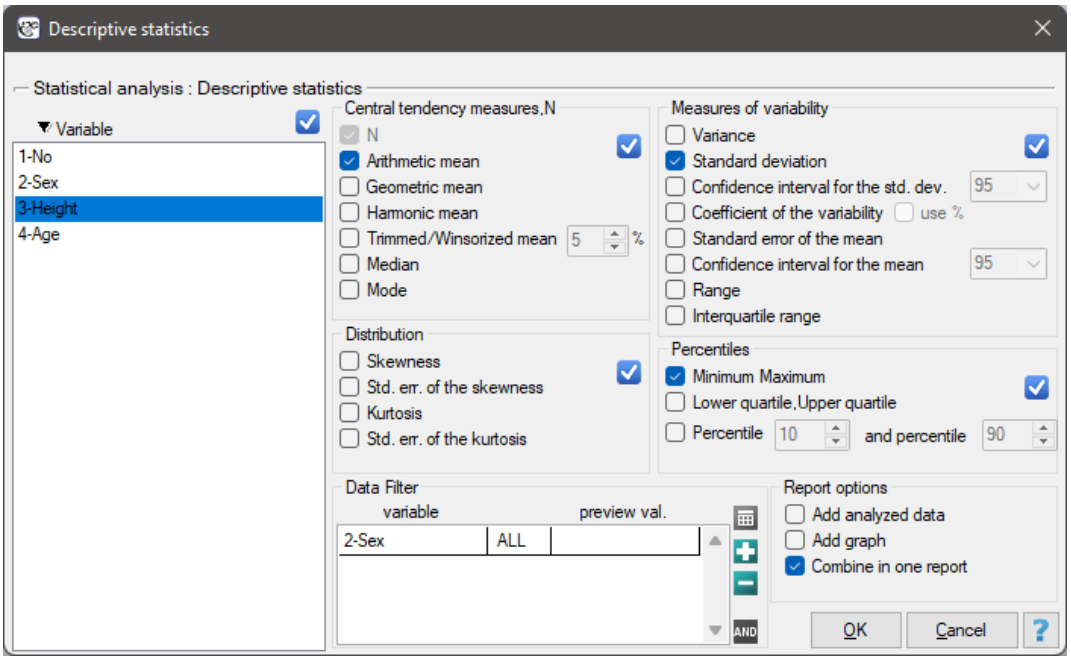
Copyright ©2010-2026 U-smile Stats – All rights reserved

A finished filter can be used by selecting the Apply button, or saved for later by selecting the ➕ button. Saved filters can be used by sending them to the formula window with the ⬆ button.

EXAMPLE 8.7. *Basic advanced filter* (filter.pqs file)
We want to determine the descriptive statistics for the height of boys who are taller than 130 cm or older than 9 years. From the menu, we choose Statistics→Descriptive analyses→Descriptive statistics. In the descriptive statistics test options window, we select the procedures we want to perform (e.g. we tick the mean, standard deviation, minimum and maximum) and the variable for the analysis (the height column). We set the filter by entering the formula (v3>130 or v4>9) and v2='m'. We then choose Apply and perform the analysis.

- **Multiple filter** – imposes one condition on the collected data, dividing it into several subgroups. The chosen analysis is performed repeatedly, separately for each subgroup.

EXAMPLE 8.8. *Multiple filter* (filter.pqs file)
We want to determine the descriptive statistics for the height of girls, and separately for the height of boys. From the menu, we choose Statistics→Descriptive analyses→Descriptive statistics. In the descriptive statistics test options window, we select the procedures we want to perform (e.g. we tick the mean, standard deviation, minimum and maximum) and the variable for the analysis (the height column). We set the multiple filter by adding a rule with the ➕ button and choosing the variable Gender. We confirm the chosen analysis options with the OK button. As a result we obtain 2 reports: a separate one for boys and a separate one for girls.



Copyright ©2010-2026 U-smile Stats – All rights reserved

8.3 RUNNING ANALYSES MULTIPLE TIMES

To make it easier to run the same analyses multiple times, you can:

1. Use the option of remembering the current and previous analyses. PQStat automatically remembers the most recently performed analysis and the options set in its window. A quick way to return to it is provided by the button (ribbon menu ->) placed on the toolbar.
2. Use the option of remembering the analysis based on the report. Together with the analysis report, PQStat automatically remembers the options set in the analysis window. A quick way to return to it is provided by the button, found in the right-click context menu on a report in the project's Navigation Tree.
3. Select multiple variables in the analysis window, so that the analysis is carried out multiple times. The results for each of the analyses performed are returned in successive reports.
4. Use the multiple filter so that the analysis is carried out separately for individual subsets of the data. The results for each of the analyses performed are returned in successive reports.

8.4 ORGANISING REPORTS INTO SETS

When there is a larger number of reports, it is convenient to group them into sets. Prepared sets make it easier both to navigate through the results and to speed up transferring the results to other programs such as Excel or Word – in that case, entire sets of reports are transferred, not just individual reports.

CREATING SETS

Through the Create empty context menu option in the Navigation Tree, on the datasheet or on an existing set.

PLACING REPORTS INTO SETS

Reports can be placed into a set:

- at the moment of their creation – from within the analysis – for this purpose, in the analysis window we click the button and choose the set the created analysis report is to go into, or before running the analysis, point the cursor at that set in the Navigation Tree;
- after the analysis has been performed – using drag and drop, or the context menu on the name of the report(s) in the Navigation Tree, or through the Project Manager.

EDITING SETS

Through the context menu in the Navigation Tree, on the datasheet or on an existing set – Edit sets.

DELETING SETS

Through the context menu in the Navigation Tree, on the name of the set – Delete set (Shift+Del).

Copyright ©2010-2026 U-smile Stats – All rights reserved

8.5 INFORMATION RETURNED TO THE REPORT

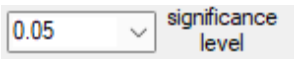
In addition to the basic settings concerning the statistical analyses performed, in the test window we have the option of:

- Including the analysed data in the report.
  Depending on the test being performed, the analysed data can be returned to the report:
  - in the form of raw records,
  - in the form of a contingency table.

- Including a chart in the report.
  For the appropriate chart to appear in the report, in the window of the chosen statistical analysis we tick the Add chart option.
- Combining the results of multiple analyses into a single report.
  By selecting multiple variables, or by applying a multiple data filter, several analyses of the same type can be carried out at the same time. To make browsing the results easier, all the analyses carried out this way can be returned to a single report. For this purpose, in the analysis window we tick the Combine into one report option.

8.6 MARKING STATISTICALLY SIGNIFICANT RESULTS

In the report, the p value of the statistical test performed is marked in red when it is lower than the significance level set by the user. By default, a significance level of $\alpha = 0.05$ is selected for all tests. This setting can be changed permanently in the Settings window, or temporarily (until the application is closed) in the window of the chosen test.



Copyright ©2010-2026 U-smile Stats – All rights reserved

9 PLOTS

The PQStat program lets you create column plots, error plots, box-and-whisker plots, point plots, and line-point plots.

The plots options window is opened via the Plots menu.

The basic plot parameters can be changed directly in the plot window. The individual tabs of this window serve different purposes:

- the Plot General Options tab lets you change general parameters such as titles, background, axes, gridlines, or the legend;
- the Plot Detailed Options tab lets you change the appearance of the plotted object itself, e.g. its shape, style, or colors;
- the Other tab lets you add extra elements to the plot, e.g. an additional line.

Plots showing the results of statistical analyses are available in the window of the given statistical analysis, under the Include plot option.

WHAT YOU CAN DO WITH A FINISHED PLOT

The plot is placed in the report, where it can be, from the context menu (right-click on the plot):

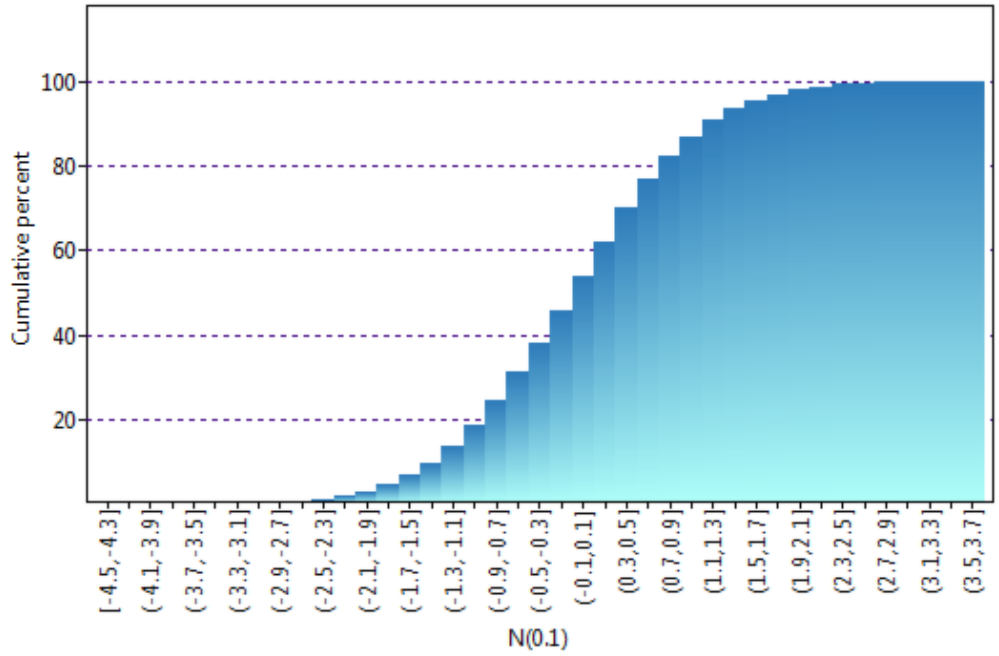
- saved - the Save plot as... option;
- printed - the Print plot option;
- copied - the Copy plot option;
- edited - this applies to the Plot General Options and Plot Detailed Options. To edit a plot, simply double-click it with the mouse, or select the Edit plot option from the context menu. The plot editing window also lets you save the plot in high resolution.

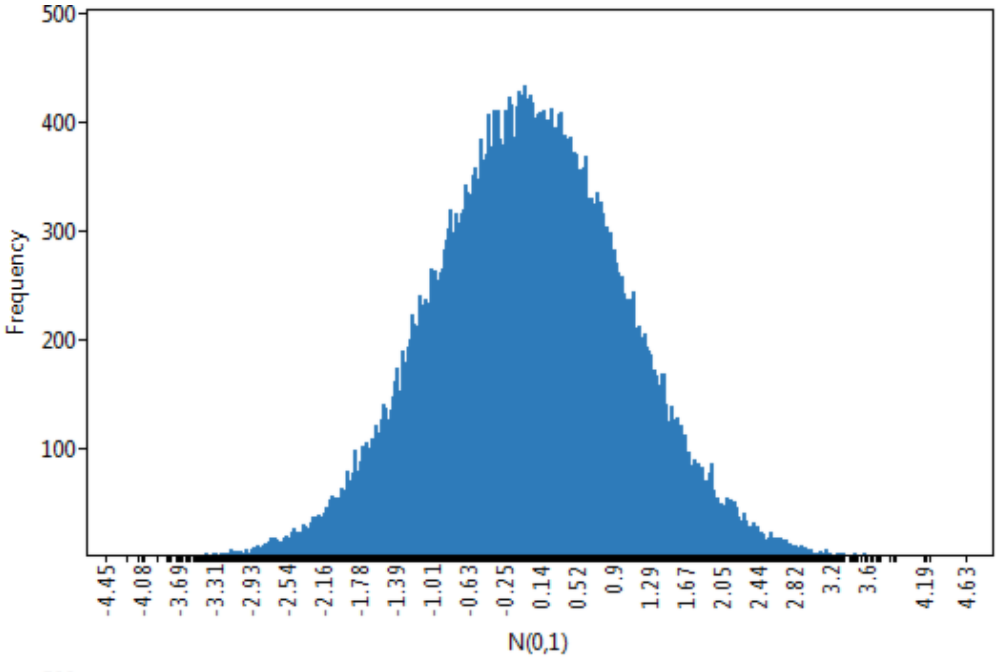
9.1 PLOT GALLERY

The types of plots available depend on the analysis being performed. Examples of the individual plot types are shown below.

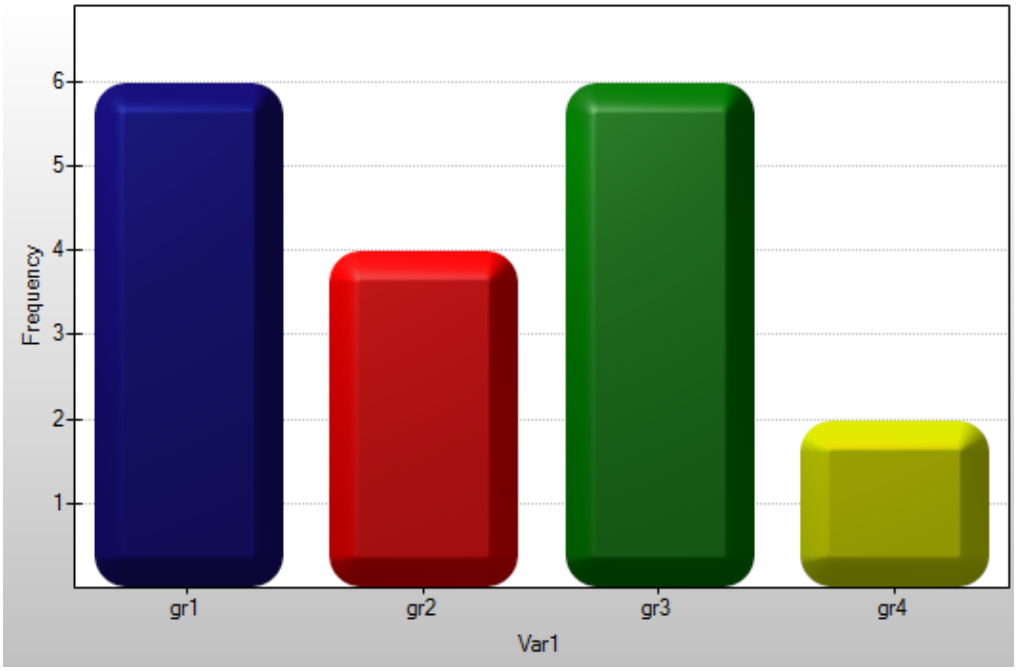
Copyright ©2010-2026 U-smile Stats – All rights reserved

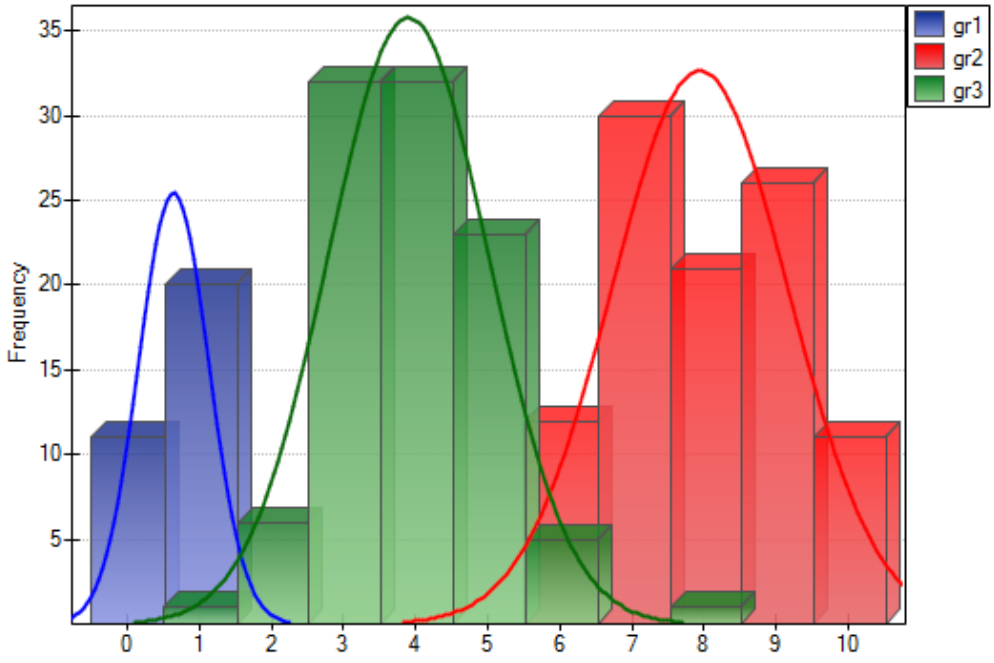
Column plots



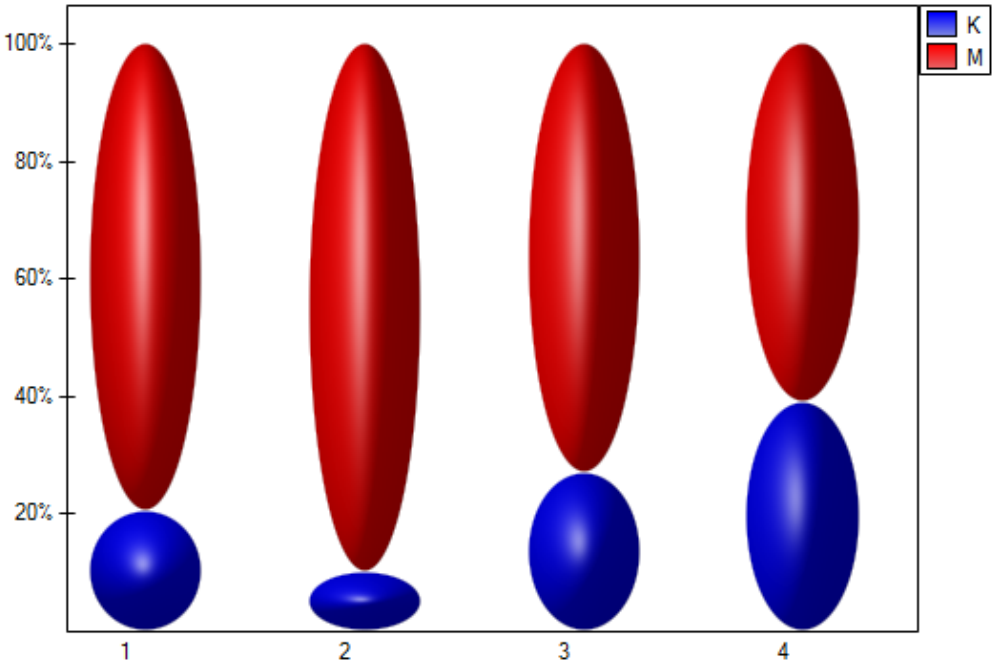


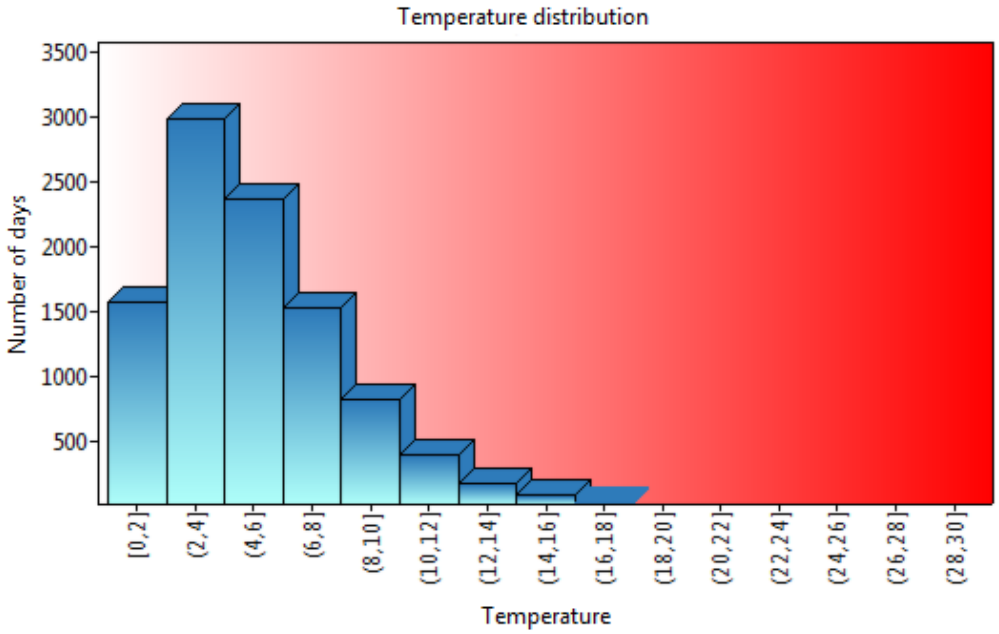
Copyright ©2010-2026 U-smile Stats – All rights reserved



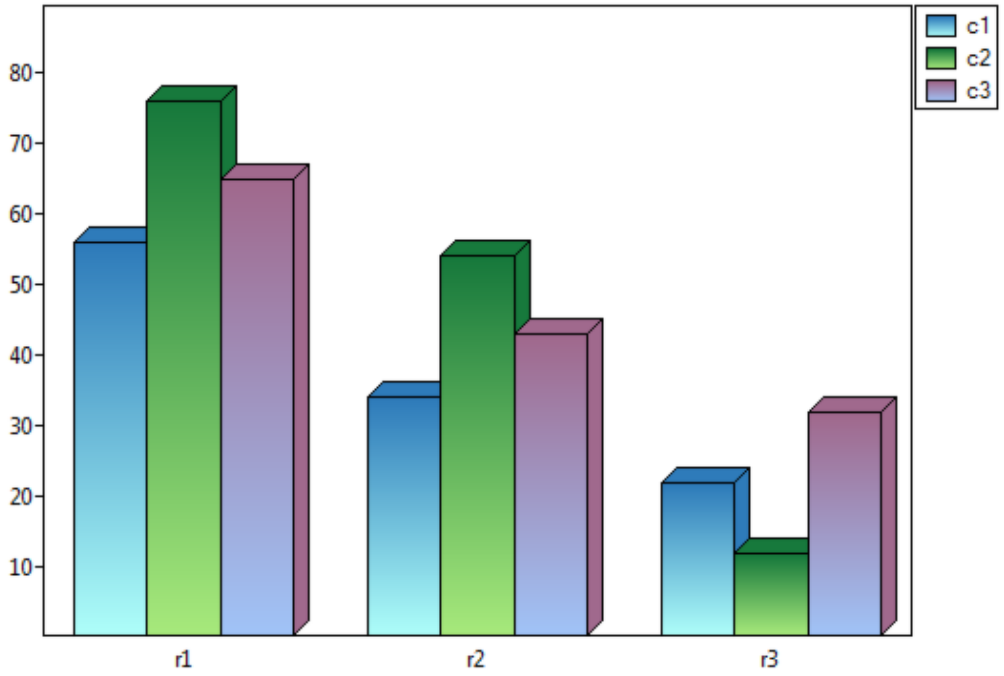


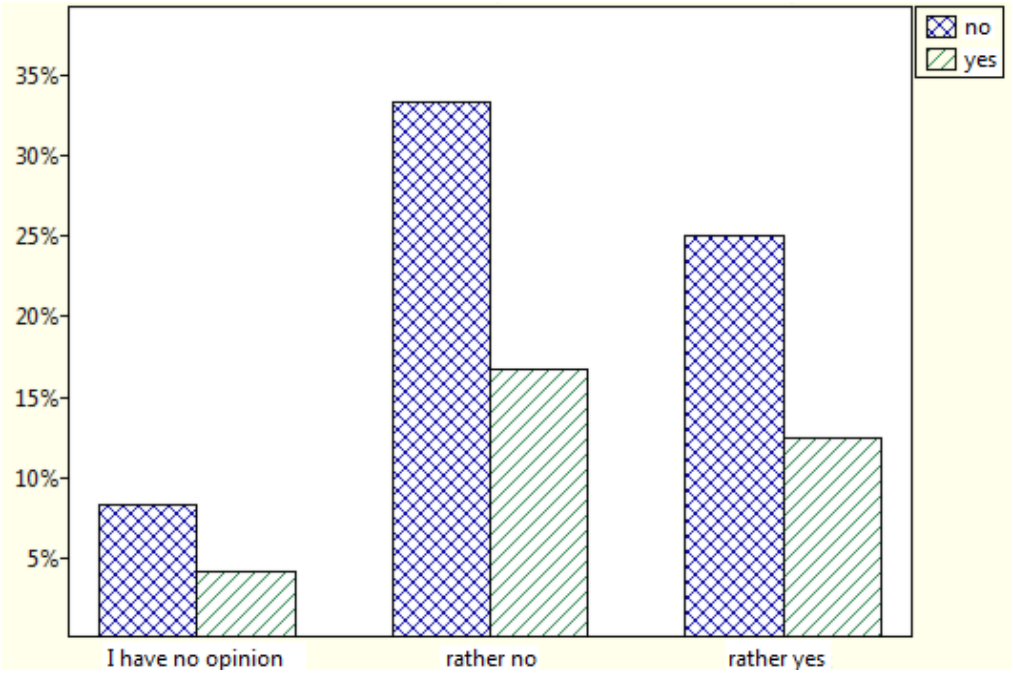
Copyright ©2010-2026 U-smile Stats – All rights reserved



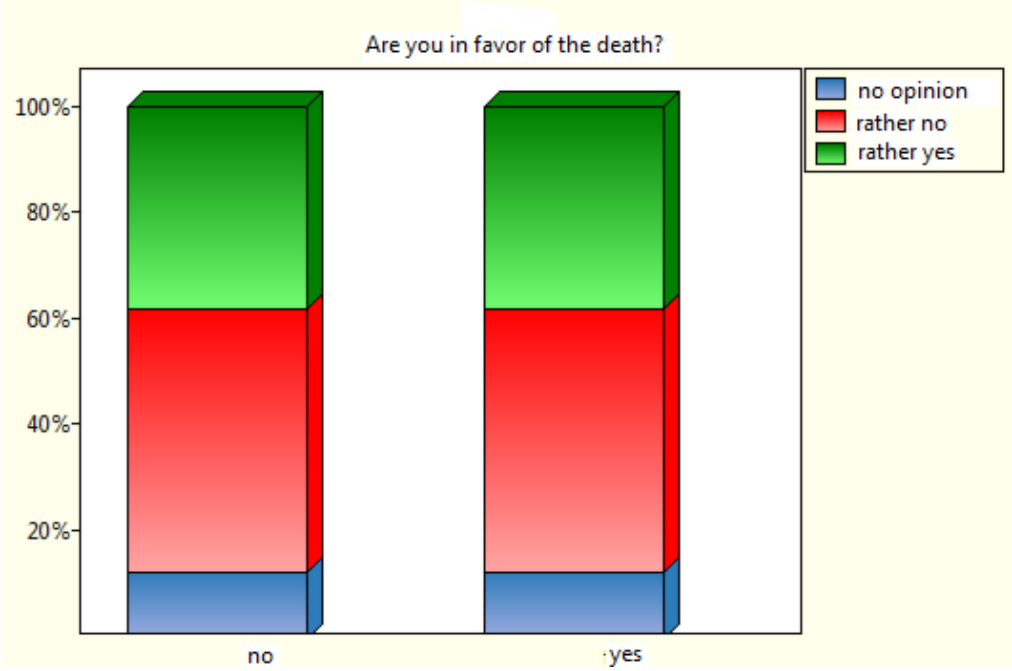


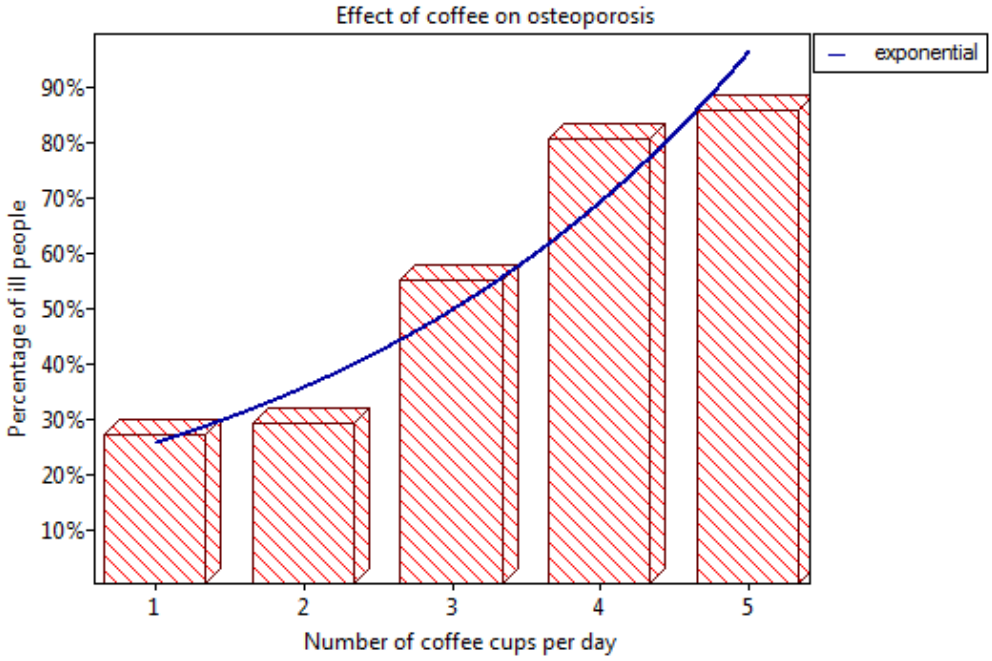
Copyright ©2010-2026 U-smile Stats – All rights reserved





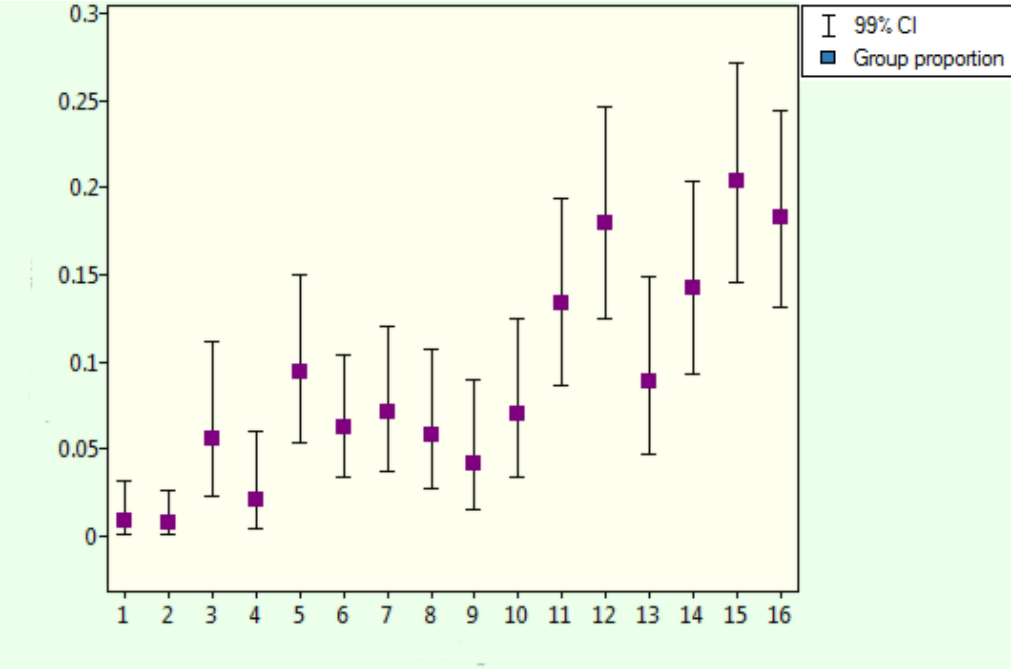
Copyright ©2010-2026 U-smile Stats – All rights reserved

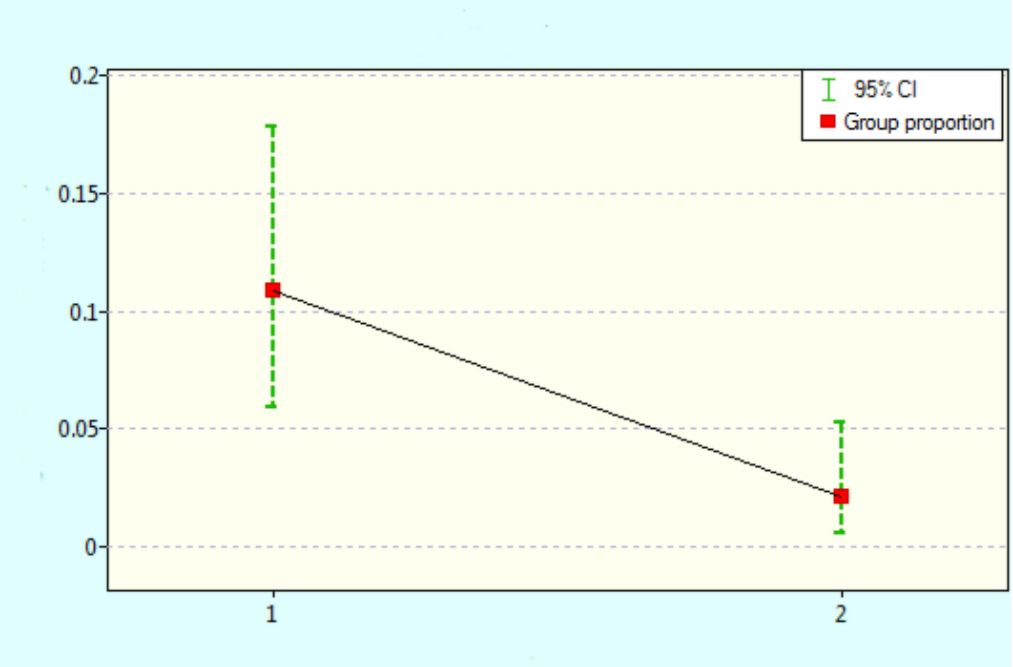




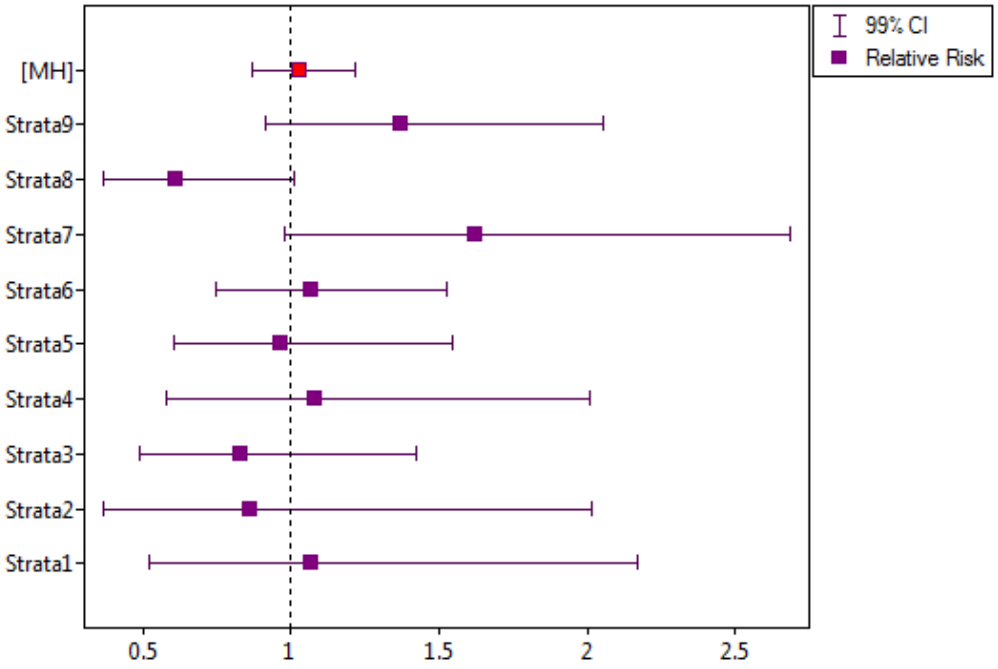
Copyright ©2010-2026 U-smile Stats – All rights reserved

Estimator error plots

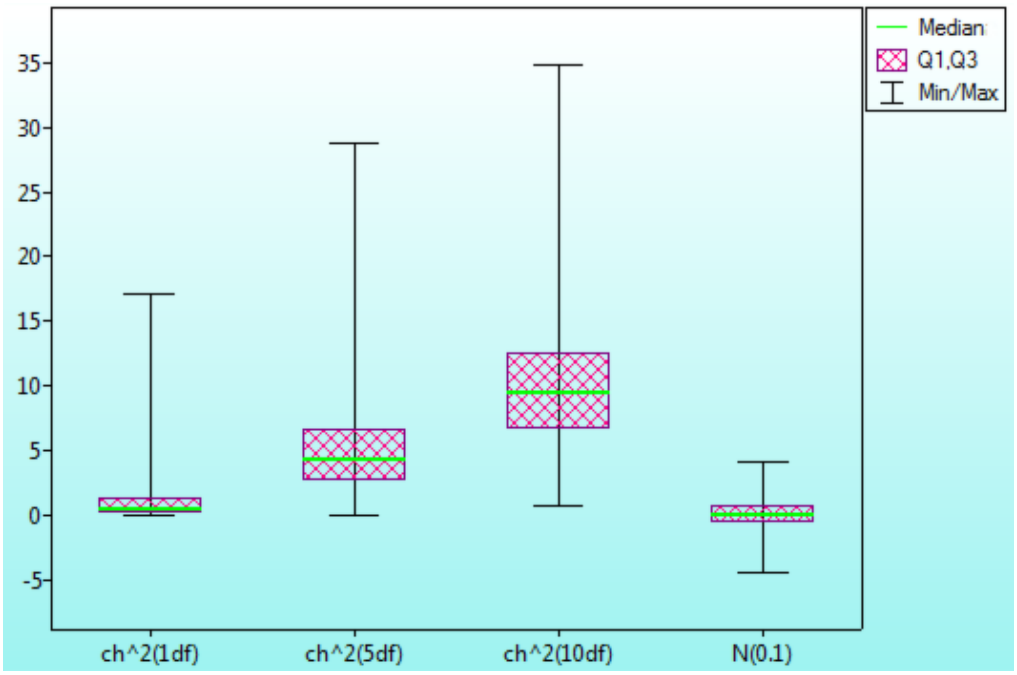




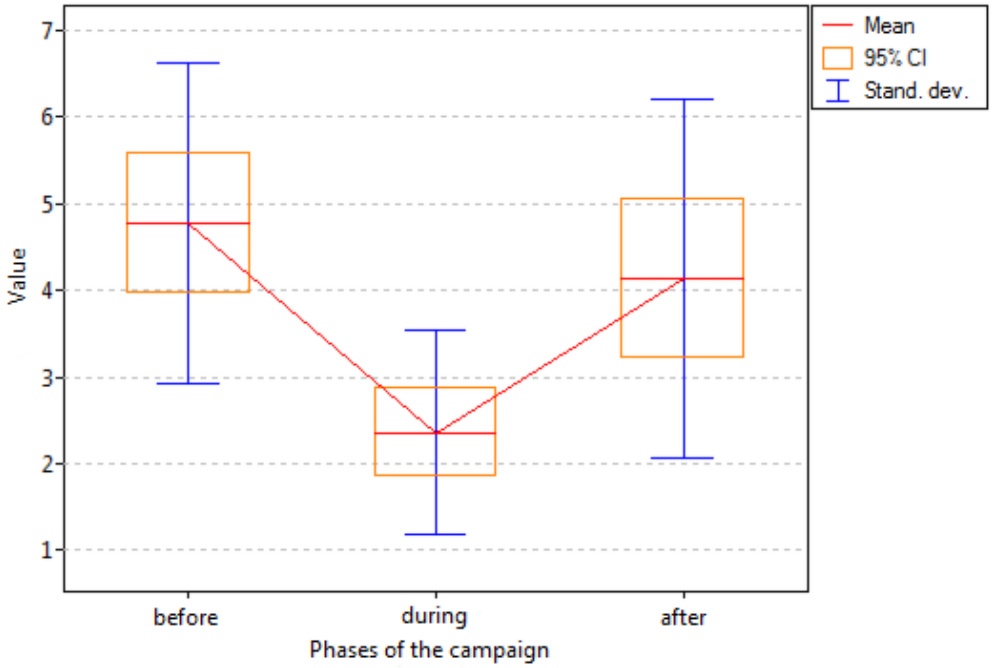
Copyright ©2010-2026 U-smile Stats – All rights reserved

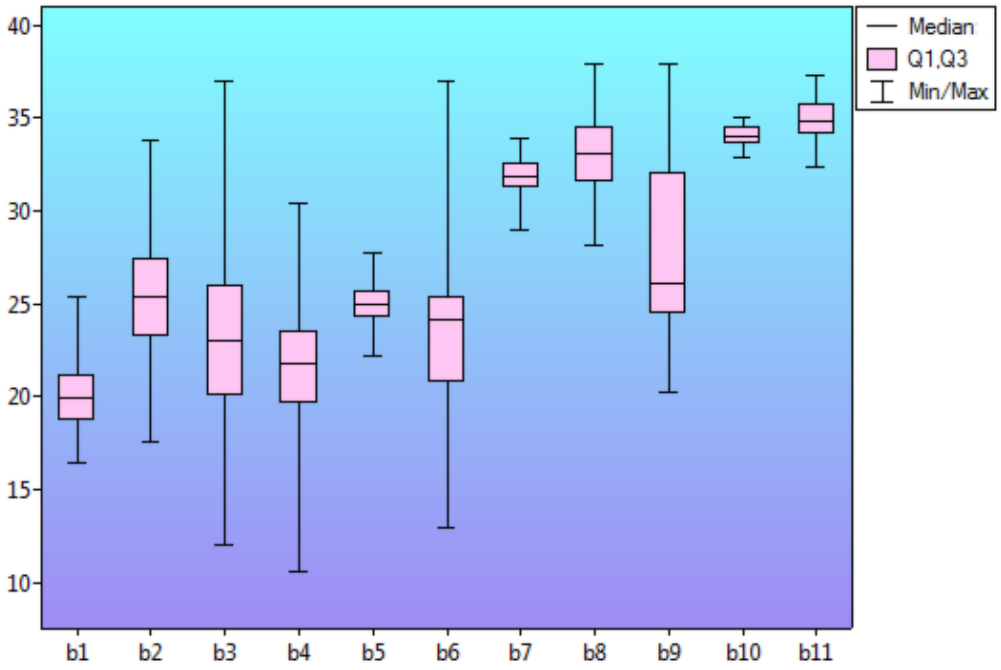


Box-and-whisker plots



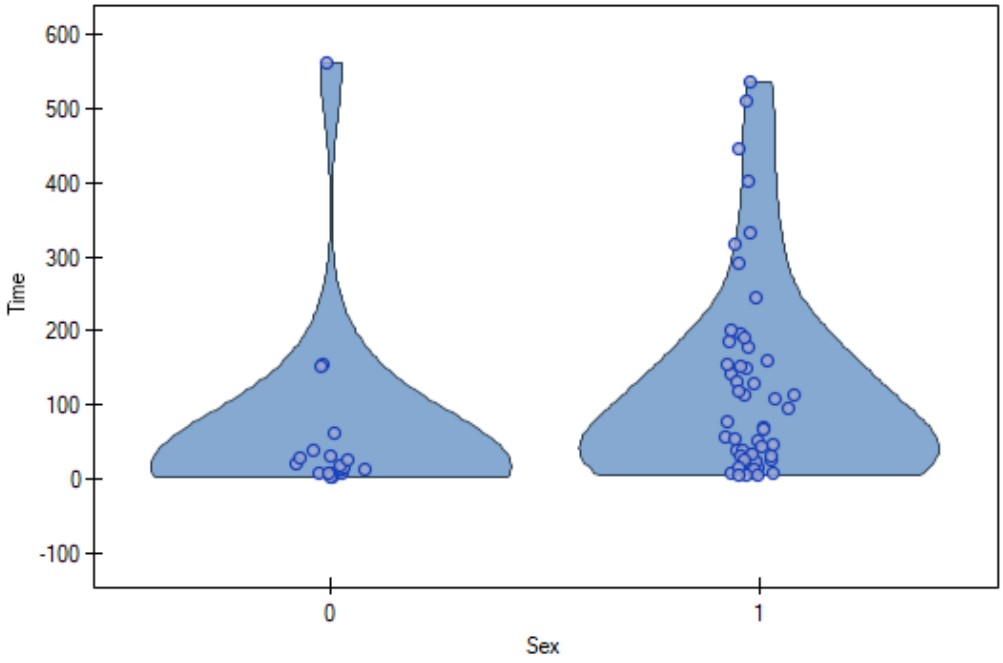
Copyright ©2010-2026 U-smile Stats – All rights reserved

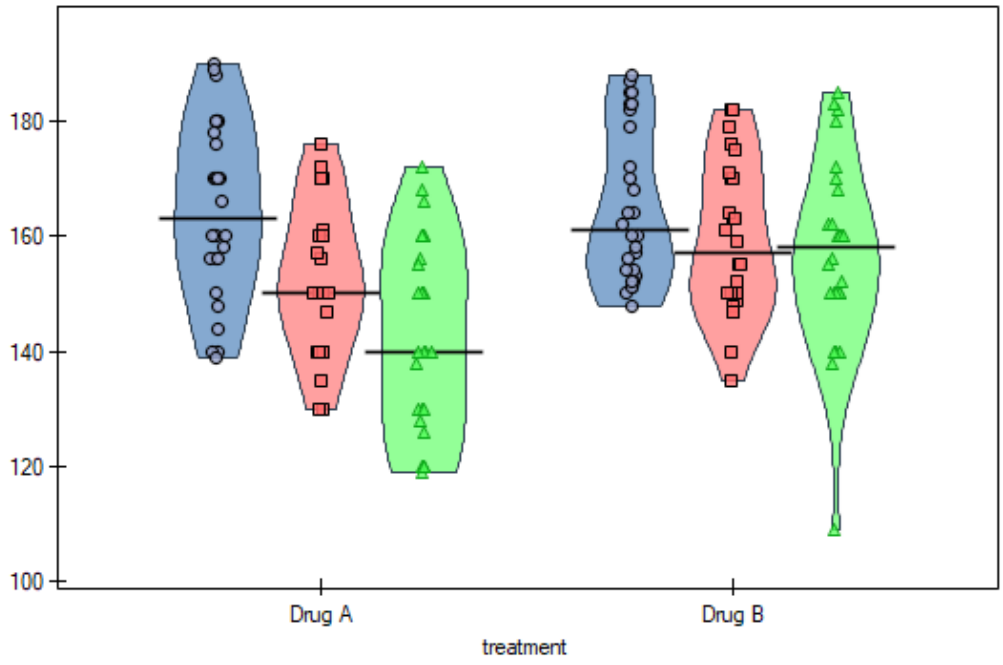




Copyright ©2010-2026 U-smile Stats – All rights reserved

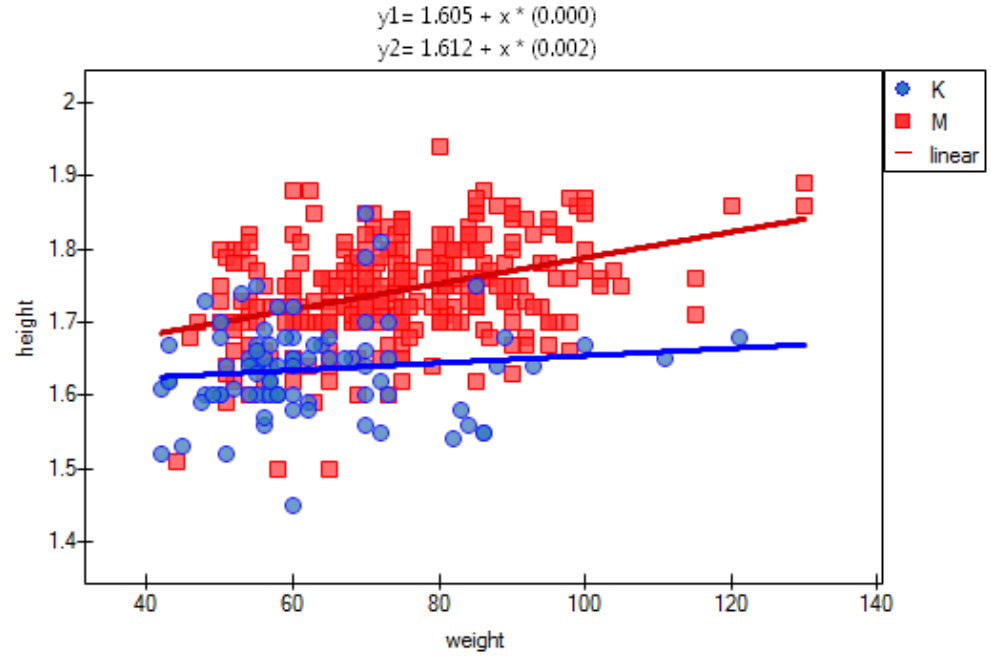
Violin plots

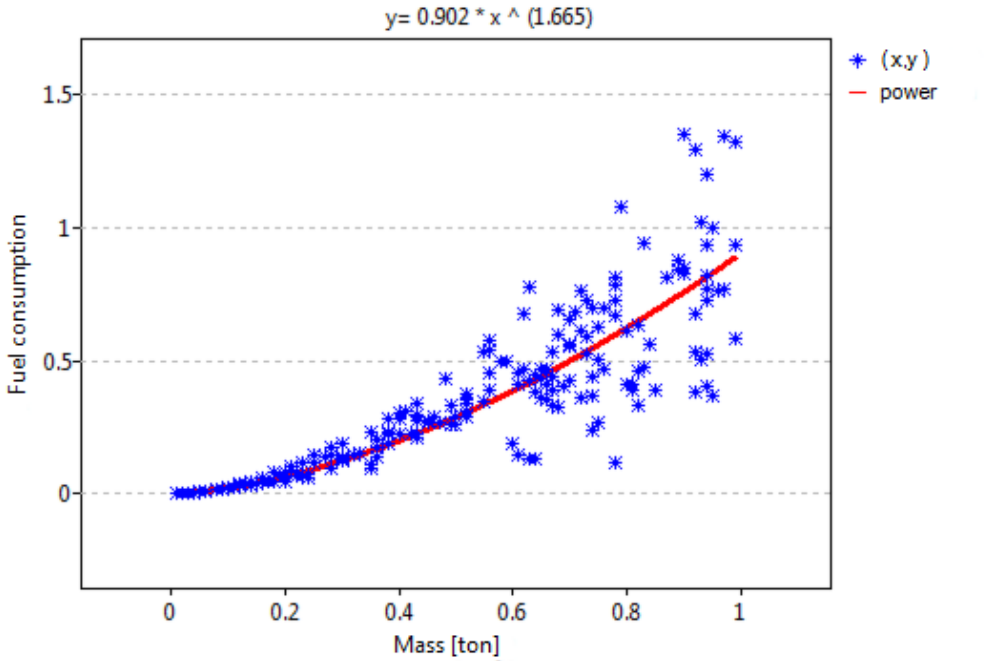




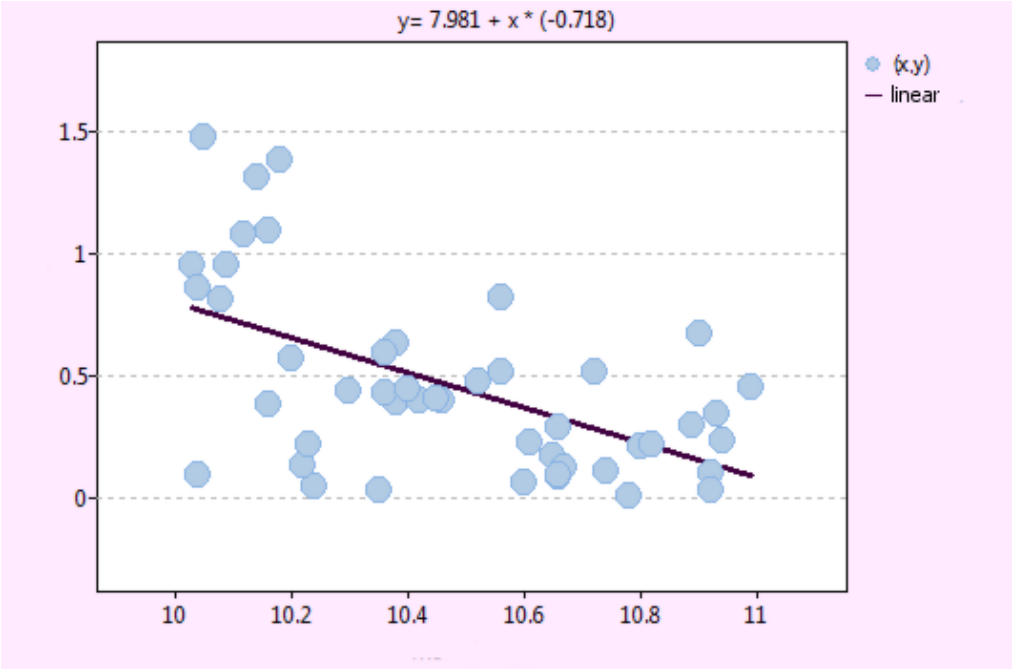
Copyright ©2010-2026 U-smile Stats – All rights reserved

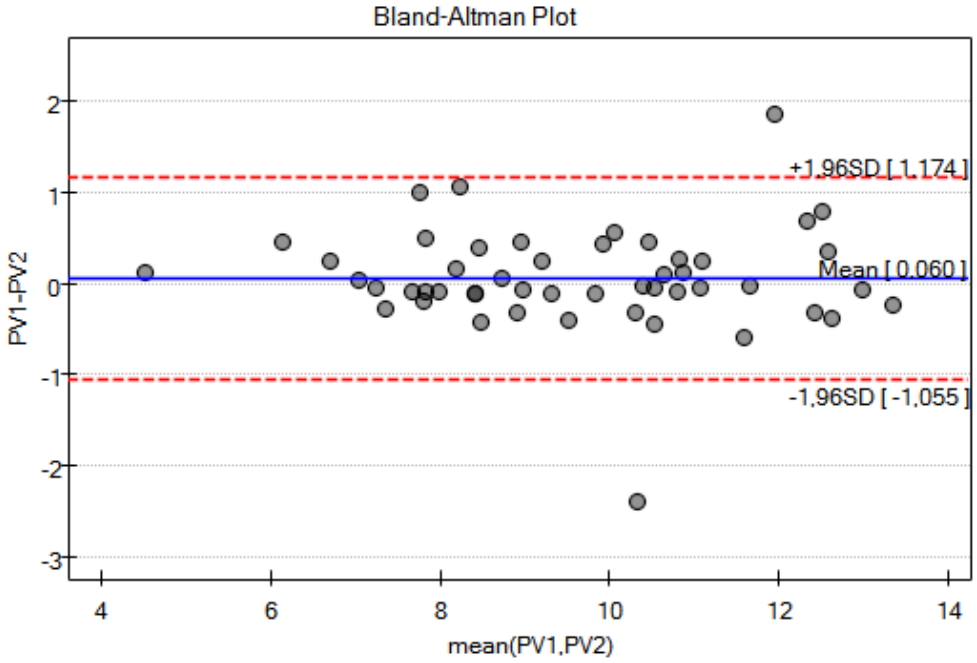
Point plots





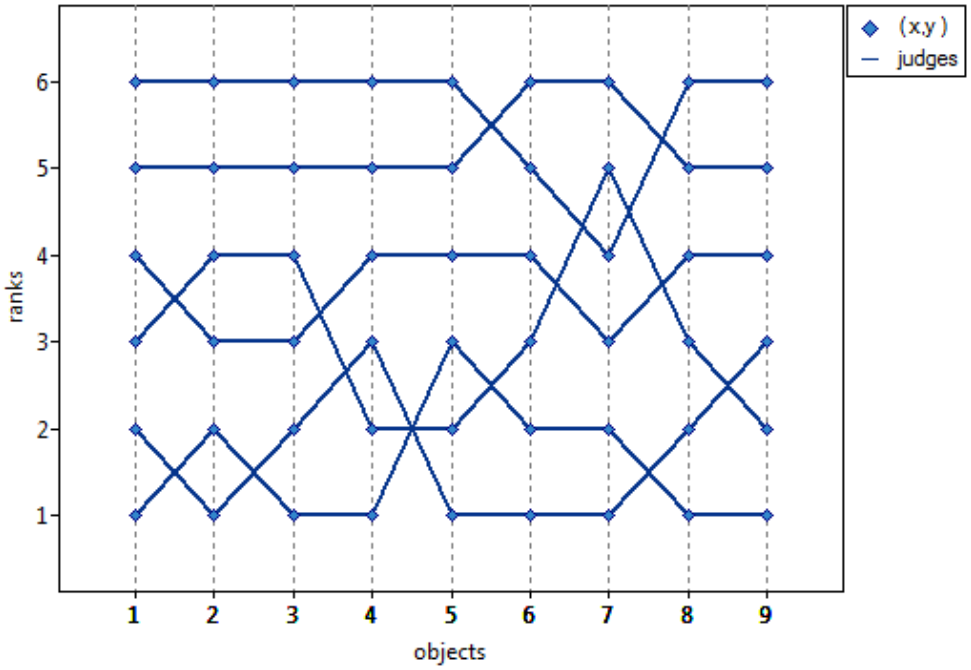
Copyright ©2010-2026 U-smile Stats – All rights reserved

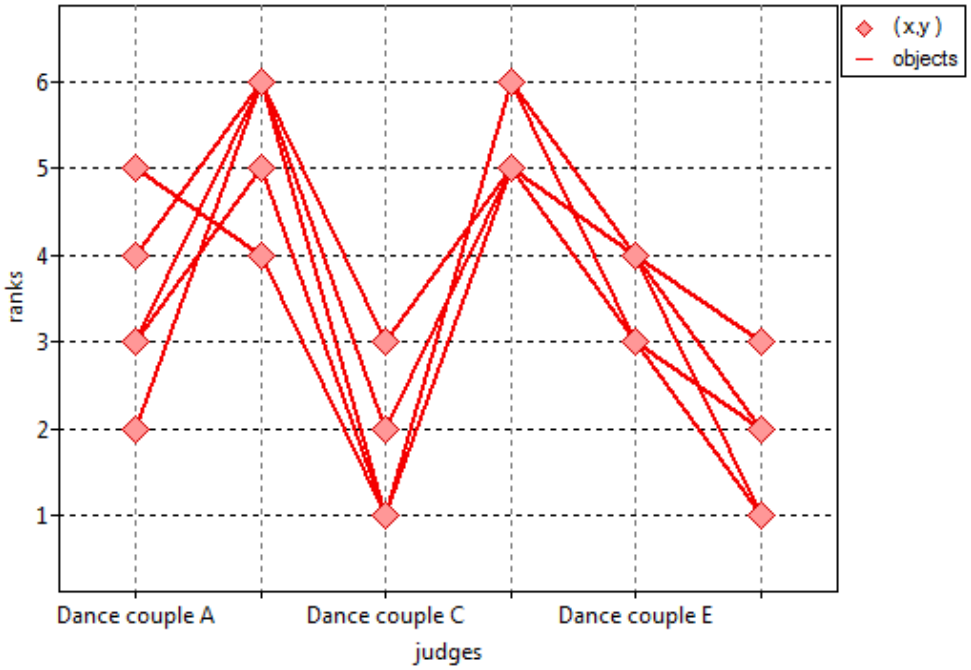




Copyright ©2010-2026 U-smile Stats – All rights reserved

Line plots





Copyright ©2010-2026 U-smile Stats – All rights reserved

9.2 LOCAL LINEAR SMOOTHING TECHNIQUES

LOWESS

WHEN TO USE:

LOWESS *(locally weighted scatterplot smoothing)*, also known as LOESS *(locally estimated scatterplot smoothing)*, lets you draw a smoothed trend line through the cloud of points on a scatter plot, built in a non-parametric way (without assuming in advance any particular form of the relationship, e.g. linear or polynomial). It is one of many ”modern” modelling methods based on the least squares method, combining the simplicity of linear regression with the flexibility of nonlinear regression. Locally weighted regression estimation was independently introduced in several different fields at the end of the 19th and the beginning of the 20th century (Henderson, 1916[81]; Schiaparelli, 1866[151]). In the statistical literature the method was introduced independently, from various points of view, at the end of the 1970s, among others by Cleveland, 1979[39] – and it is this work on which the method used in PQstat is based.

The basic principle is that a smooth function can be well approximated by a low-degree polynomial (the program uses a linear function, i.e. a first-degree polynomial) in the neighbourhood of any point x. In other words: instead of fitting a single function to all the data at once, a separate local linear function is fitted at every point, built from the neighbouring observations – the closer an observation is to the given point, the greater its influence on the result at that point. The detailed algorithm, together with the formulas, is described in the ”For those interested” section.

Kernel smoothing

WHEN TO USE:

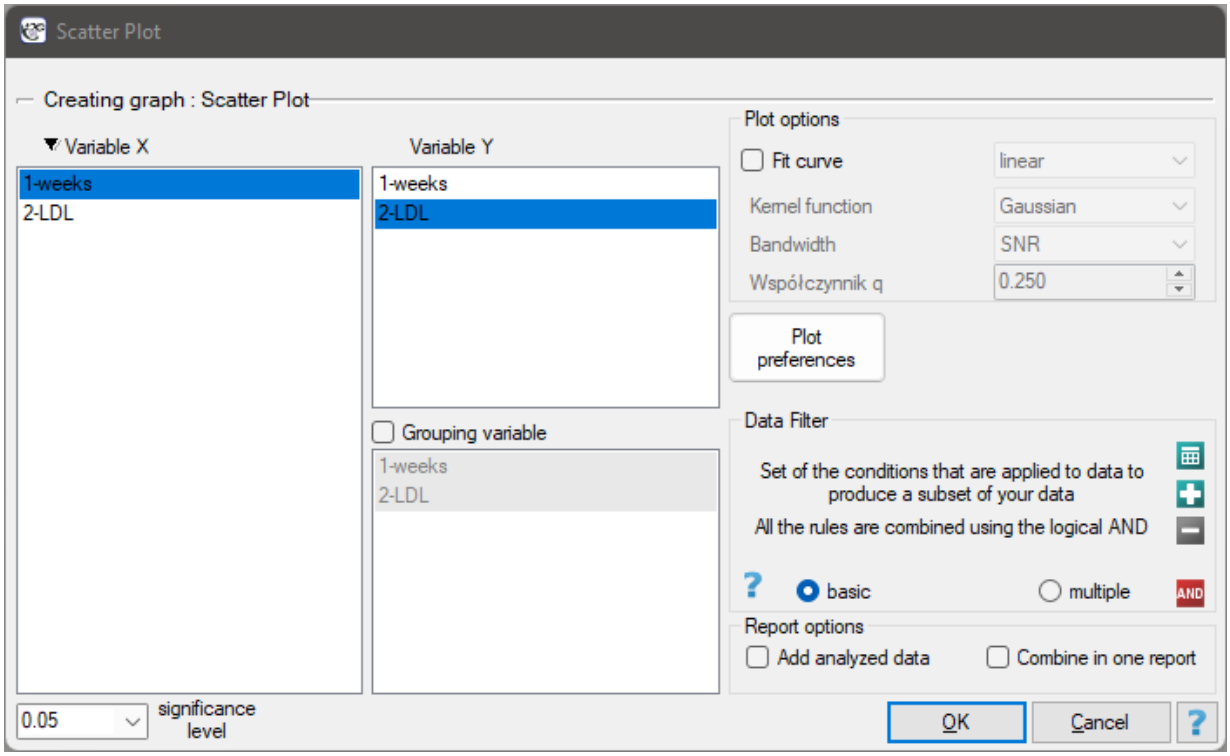
An alternative way of smoothing a scatter plot is kernel smoothing, sometimes called the Nadaraya-Watson estimate (Nadaraya, 1964[127]; Watson, 1964[172]). Like LOWESS, it does not assume any particular form of the relationship in advance – at every point a weighted average of the neighbouring observations is calculated, with the weights assigned according to the kernel function described in the Kernel estimation chapter. The exact formula is given in the ”For those interested” section.

The shape of the resulting curve is decisively influenced by the **smoothing parameter** h *(bandwidth)* – the higher its value, the greater the degree of smoothing. This parameter can be set freely (a user value) or selected automatically using one of the methods: SNR, SROT or OS. The choice of **kernel function** – Gaussian, uniform (rectangular), triangular, Epanechnikov or quartic/biweight (fourth degree) – has a much smaller influence on the result. A description of the individual smoothing parameters and kernel functions can be found in the aforementioned Kernel estimation chapter.

HOW TO SET UP THE ANALYSIS

The window with the settings of a scatter plot option with a function fitted using the LOWESS method or by kernel smoothing can be found in various analyses. This plot can also be produced via the menu Plots→Scatter plot.

Copyright ©2010-2026 U-smile Stats – All rights reserved

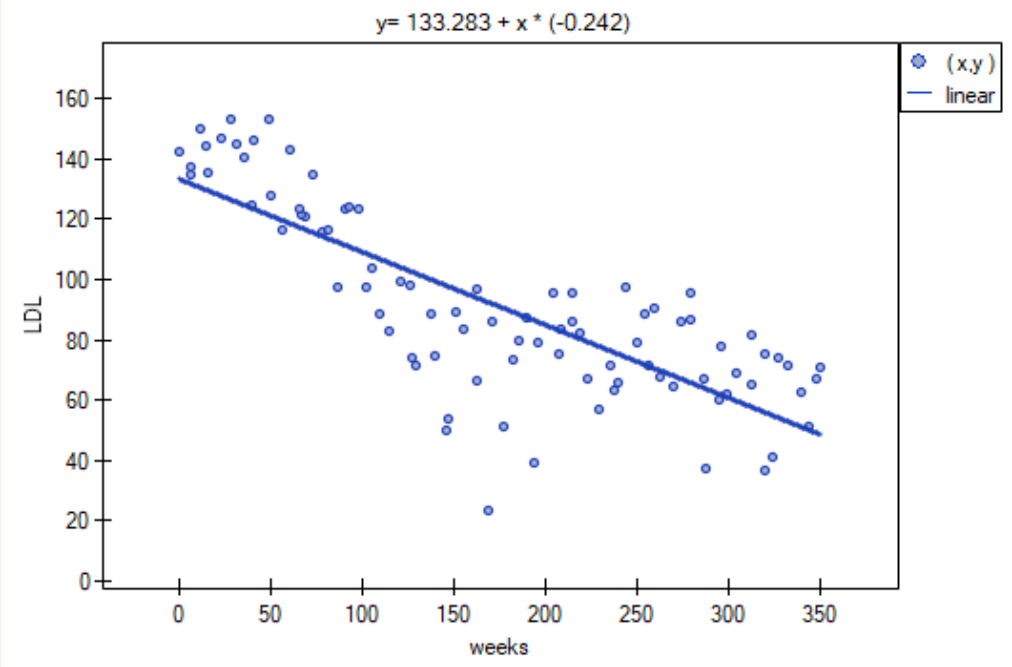


PERFORMING AND INTERPRETING THE ANALYSIS

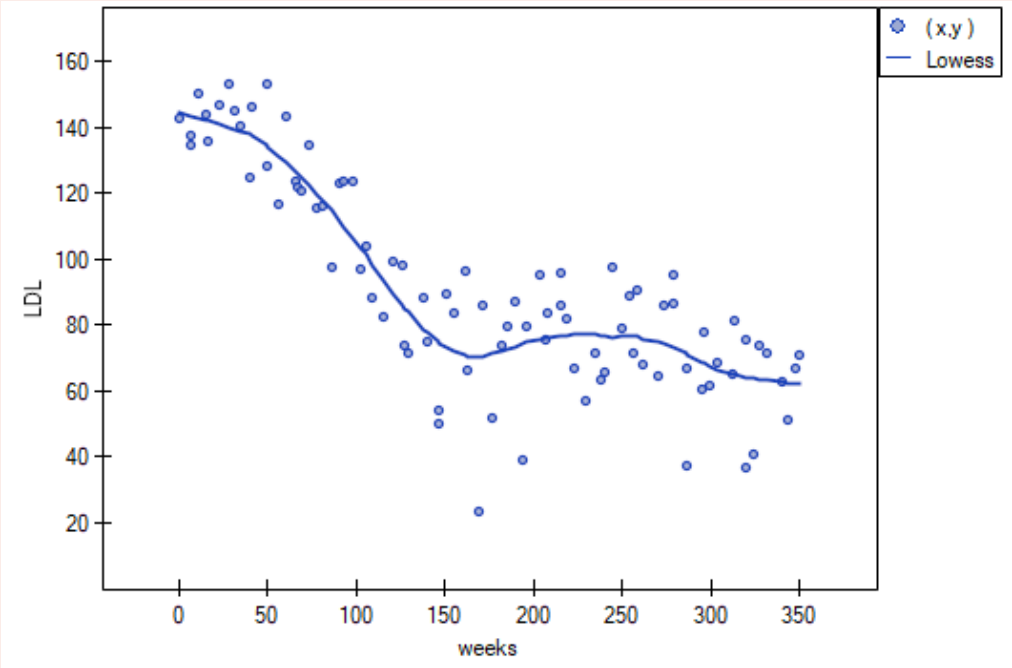
EXAMPLE (22.2) cont. *(LDL weeks.pqs file)*
The effectiveness of a new therapy aimed at lowering LDL cholesterol levels was studied. 88 people at various stages of the treatment were examined. We will check whether, as the duration of the treatment increases (time in weeks), the LDL cholesterol level decreases and stabilises.

The results were first presented by fitting a straight line indicating the direction of the relationship under study.

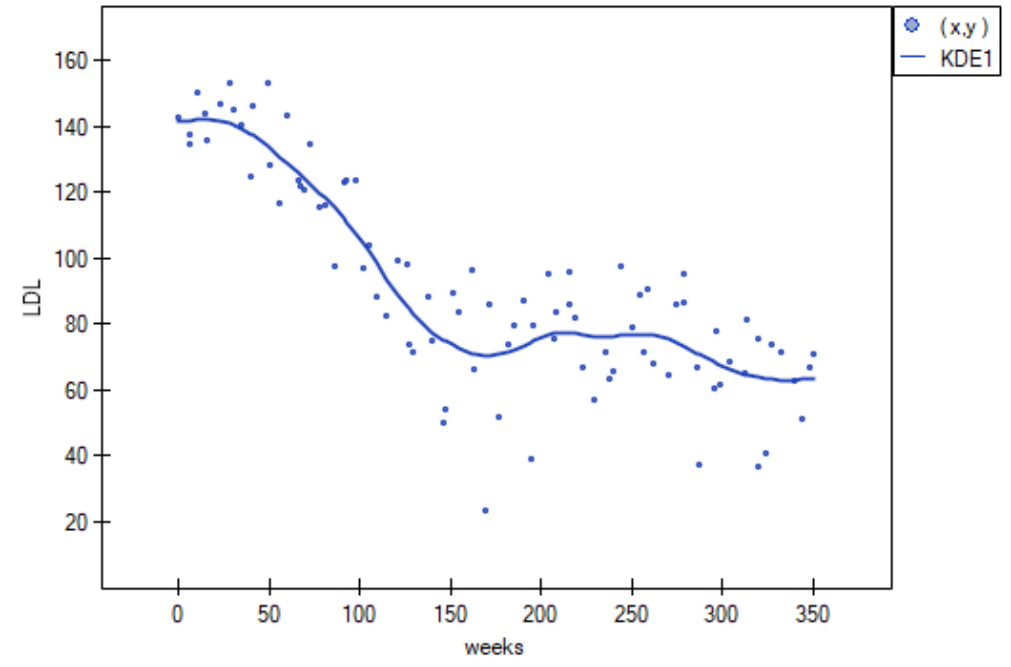
Copyright ©2010-2026 U-smile Stats – All rights reserved



However, this way of presenting the data does not fully reflect the relationship taking place. From the arrangement of the points it can be seen that the relationship is initially decreasing, and after 150 weeks it begins to stabilise. The relationship is presented again, using the LOWESS method and Gaussian kernel smoothing.



Copyright ©2010-2026 U-smile Stats – All rights reserved



Both methods, i.e. both LOWESS and kernel smoothing, gave similar results and described the data much better, showing an initial decline in LDL followed by stabilisation at a level close to 70 mg/dl.

FOR THOSE INTERESTED

The LOWESS algorithm
Procedure:

1. For every point of the dataset we build a window containing the neighbouring elements. How many elements fall into the window is determined by the smoothing parameter q. The higher its value, the smoother the resulting function. If, for example, this parameter is 0.2, then the window will contain about 20% of the data, and these will be used to build the polynomial (here a monomial, i.e. a linear function).
   Because the window has to remain symmetric, i.e. the data in the window should lie both above and below the point for which the model is being built, the number of elements in the window should be odd. Therefore, the number of elements obtained from the parameter q is rounded up to the nearest odd number. This produces a window containing the element x_0 together with the corresponding number $k(x_0)$ of elements located before it and after it in the ordered dataset.

$$k(x_0) = \frac{roundUpNpar(q * n) - 1}{2}$$

   For example, if the window is to contain 7 elements, it will contain the element x_0 together with the three preceding and three following elements of the sample. The window determined for the first and last elements of the sample is of the same size, but the element in question is not placed symmetrically within it, i.e. not in the middle.

2. At every point of the dataset a low-degree polynomial (here a linear function) is fitted to the subset of data located in the window. The fit is made using the weighted least squares method, which

Copyright ©2010-2026 U-smile Stats – All rights reserved

gives greater weight to points near the point whose response is being estimated, and smaller weight to points further away. In this way a different polynomial formula (here a linear function) is assigned to each point.
The weights used in the weighted least squares method can be set quite flexibly, but points distant from the point in question must have a smaller weight than points nearby. Here the weights proposed by Cleveland are used, the so-called tricube weights

$$w = (1 - d^3)^3,$$

where the distance d from the point x_0 was 0 for points outside the designated window, and was given as the actual distance between points but normalised to the interval [0,1] for points located within the window, so that the maximum distances in every window were the same.

3. At every point, based on the polynomial formula assigned to it (here a linear function), the value of the function $\hat{y}_i$ is calculated. Based on the points x_i and the points $\hat{y}_i$ estimated by this method, a smoothed function fitted to the data is drawn.

Kernel smoothing
Formulas:
The kernel estimate is a weighted average of the observations within the smoothing window:

$$\hat{y}_i = \frac{\sum_{i=1}^{2} K_h(t_i) y_i}{\sum_{i=1}^{2} K_h(t_i)},$$

where $K_h(t_i)$ is the kernel function described in the Kernel estimation chapter.

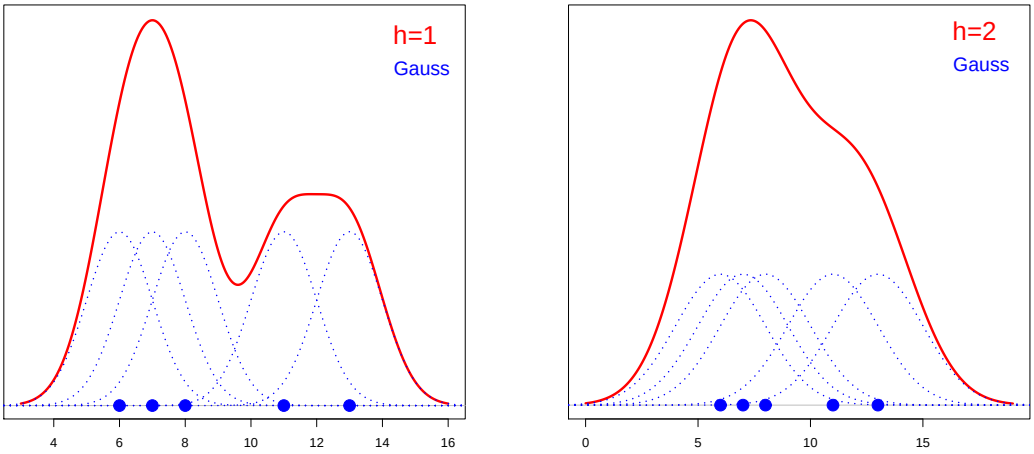
9.3 KERNEL ESTIMATION

One-dimensional kernel estimator

WHEN TO USE:

The one-dimensional kernel density estimator lets you approximate the shape of a data distribution with a smoothed curve, built in a non-parametric way (without assuming any particular form of the distribution). It is an alternative to the traditional histogram – where a histogram presents the distribution as a set of step-shaped bars (depending on an arbitrarily chosen bin width), the kernel estimator gives a smooth, continuous curve, which lets it reflect the actual shape of the distribution more faithfully.

Copyright ©2010-2026 U-smile Stats – All rights reserved



For every point x in the range of the data, the value of the estimator is determined, i.e. the local density at that point. It is obtained by summing the values of the kernel functions $K_h(t_i)$ built around the individual observations:

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(t_i)$$

If we assign weights w_i to the individual cases, a weighted kernel density estimator can be built, defined by the formula:

$$\hat{f}_K(x) = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i K_h(t_i)$$

DEFINITIONS OF MEASURES

The shape of the resulting curve is influenced by two elements: the choice of the **smoothing factor** and the choice of the **kernel function**.

The smoothing factor (h) *(bandwidth)* determines the degree of smoothing of the curve – the higher its value, the more strongly smoothed (more "flattened") the resulting estimator is. This factor can be set in several ways:

User – any positive value of the smoothing factor, specified by the user.

User scaled – a value set so that, when the kernel function is changed, a similar level of smoothing to the one previously chosen for the Gaussian kernel is preserved. Thanks to this rescaling of the factor (Scott, D. W. 1992[152]), it is convenient to switch between different kernels without having to re-select the smoothing. The conversion is based on the standard deviation:

$$h_2 = \frac{\sigma(K_{h_1})}{\sigma(K_{h_2})} h_1$$

SNR – a factor determined automatically by Silverman's method (Silverman B. W. 1986[157] pp. 45 and 47) with Jones' correction (Jones M. C. et al. 1996[92]), using the standard deviation from the sample (rather than the population), as proposed by Silverman:

$$h_{SNR} = 1.06 sd \cdot n^{1/5}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

For a kernel other than Gaussian, the smoothing factor is additionally rescaled (Scott D. W., 1992[152]).

SROT – a factor determined automatically by Silverman's method (Silverman B. W. 1986[157] p. 48) with Jones' correction (Jones M. C. et al. 1996[92]):

$$h_{SROT} = 0.9 \min\left(sd, \frac{IQR}{1.34}\right) n^{1/5}$$

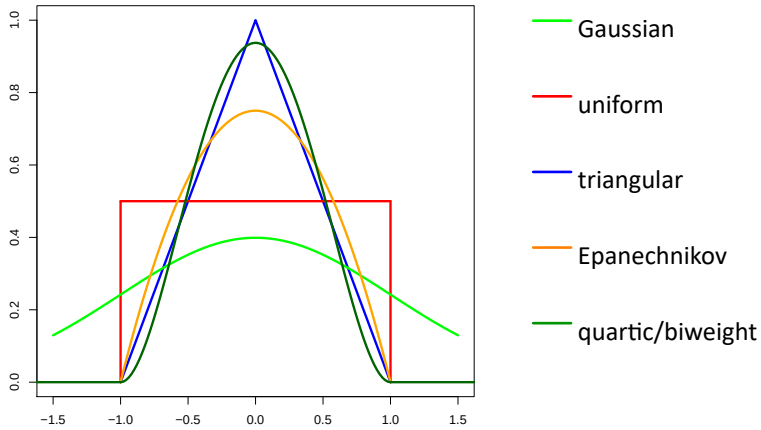
For a kernel other than Gaussian, the smoothing factor is additionally rescaled (Scott D. W., 1992[152]).

OS – a factor determined automatically by the method of Terrell and Scott (Terrell G. R. and Scott D. W. 1985[166], Terrell G. R. 1990[165] p. 470):

$$h_{OS} = 1.144 sd \cdot n^{1/5}$$

For a kernel other than Gaussian, the smoothing factor is additionally rescaled (Scott D. W., 1992[152]).

The kernel function has less influence on the result than the smoothing factor, but it also shapes the resulting estimator. The kernel is a probability density function built around each data point x_i – usually a symmetric function, reaching its maximum at the point x_i and decreasing as the distance d_i from that point grows. The distance from the point being analysed is modified by the smoothing factor h according to the formula $t_i = \frac{d_i}{h}$. Depending on your needs, the kernel function can take the form of the Gaussian, uniform (rectangular), triangular, Epanechnikov or quartic/biweight (fourth-degree) function – the exact formulas can be found in the "For those interested" section.



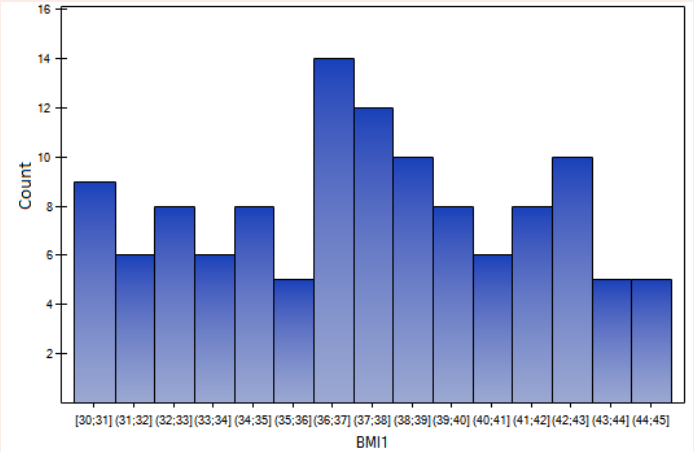
HOW TO SET UP THE ANALYSIS

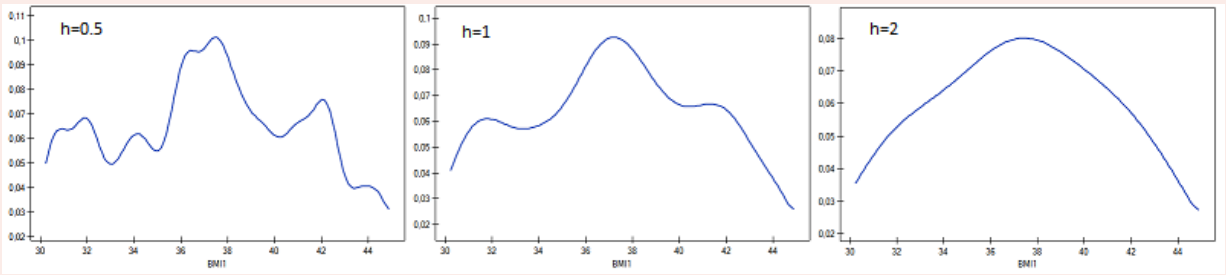
The kernel density estimator is available as a graphical option used in many different analyses describing the distribution of a variable (among others, when creating a histogram). In the settings window of such a plot you can choose the kernel function and the way the smoothing factor is selected (a user value, a scaled user value, or one of the automatic methods: SNR, SROT, OS).

Copyright ©2010-2026 U-smile Stats – All rights reserved

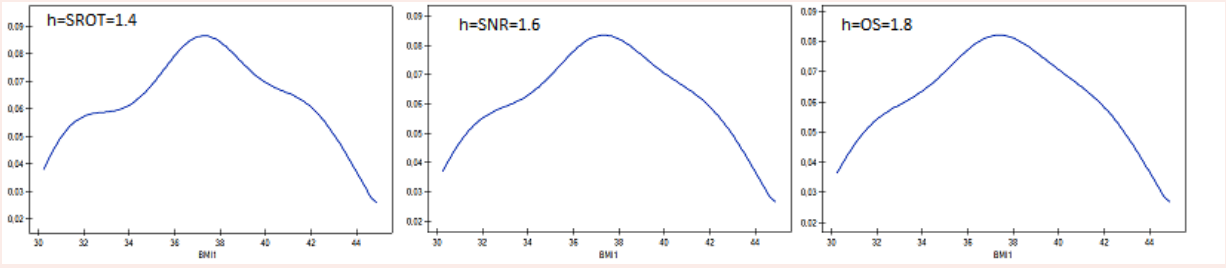
PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 9.1. (BMI.pqs file)
Values of the weight-to-height index BMI1 were calculated for a certain group of obese people. Their distribution was first presented using a histogram with values divided into bins 1 BMI unit wide. The same data were also visualised using a kernel density estimator, choosing the Gaussian kernel function and setting the smoothing factors to, respectively: 0.5, 1, 2.





It can be seen that the higher the smoothing factor, the smoother (less ”jagged”) the resulting density curve, although this can mask finer features of the distribution, e.g. its multimodality. For this data, the smoothing factors suggested by the automatic SROT, SNR and OS methods reach values between 1.4 and 2.



Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Gaussian kernel function

$$K_h(t_i) = \frac{1}{h\sqrt{2\pi}} \exp(-\frac{t_i^2}{2})$$

Uniform kernel function (rectangular)

$$K_h(t_i) = \begin{cases} \frac{0,5}{h} & \text{if } t_i < 1 \\ 0 & \text{if } t_i \geq 1 \end{cases}$$

Triangular kernel function

$$K_h(t_i) = \begin{cases} \frac{1-t_i}{h} & \text{if } t_i < 1 \\ 0 & \text{if } t_i \geq 1 \end{cases}$$

Epanechnikov kernel function

$$K_h(t_i) = \begin{cases} \frac{3}{4}\frac{1-t_i^2}{h} & \text{if } t_i < 1 \\ 0 & \text{if } t_i \geq 1 \end{cases}$$

Quartic or biweight kernel function (fourth degree)

$$K_h(t_i) = \begin{cases} \frac{15}{16}\frac{(1-t_i^2)^2}{h} & \text{if } t_i < 1 \\ 0 & \text{if } t_i \geq 1 \end{cases}$$

Bland-Altman plot

WHEN TO USE:

The Bland-Altman plot (Bland and Altman, 1986[22], 1999[8]) is used for the graphical assessment of agreement between two methods of measuring the same quantity, e.g. two different blood pressure monitors, two laboratory analysers, or a new method compared with the one currently in use (the so-called gold standard). In clinical medicine, measurements taken on a living organism are constantly changing and their true value is unknown (e.g. arterial blood pressure), which calls for the continual refinement and development of new, better measurement tools. Usually, when a new method is created, its results are compared with an already established method. The new method cannot be expected to give exactly the same result as the method used so far, but it is worth checking how much the results obtained actually differ. For a new method to replace an old one, the difference between the results of the two methods should be small enough not to create a problem for clinical interpretation – for example, in blood pressure measurements a difference of 20mmHg is already large enough that it could affect the decision to start treatment. Statistical methods cannot directly answer the question of how large a difference between methods is acceptable for them to be considered concordant, but an appropriate graphical illustration of the differences obtained and their possible limits of variability makes such a decision easier.

DEFINITIONS OF MEASURES

The Bland-Altman plot is a scatter plot in which:

X axis – the mean of the measurements for the two compared methods;

Copyright ©2010-2026 U-smile Stats – All rights reserved

Y axis – the difference between the measurements for the two compared methods;

Mean difference – if the results obtained by the new method are consistently greater/smaller than those obtained by the old method, a shift called bias occurs, i.e. the line representing the mean difference is not at the 0 level, but is shifted significantly up or down from it;

95% limits of agreement – if the differences are normally distributed, 95% of the differences will lie within the interval (Mean difference $\pm$ 1.96SD), where SD is the standard deviation of the differences.

Note! The interval of agreement defined above is not the same thing as a confidence interval. Nor is there a requirement that the data themselves be normally distributed – it is enough that the distribution of the differences does not deviate significantly from a normal distribution. Conformity of the differences with a normal distribution can be checked using tests for normality, or by graphical interpretation of that distribution.

Precision of the limits of agreement – the confidence interval for the limits of agreement, i.e. the interval of accuracy with which we determine these limits on the basis of a representative sample. The larger the sample and the smaller the variance of the differences, the greater the precision obtained.

HOW TO SET UP THE ANALYSIS

The Bland-Altman plot is available as an additional graphical option in the analysis windows used for comparing two dependent series of measurements of the same quantity (e.g. when comparing the results of two instruments or two assessment methods). Simply select the appropriate plot option in the window of the chosen analysis to obtain an agreement plot for the compared methods.

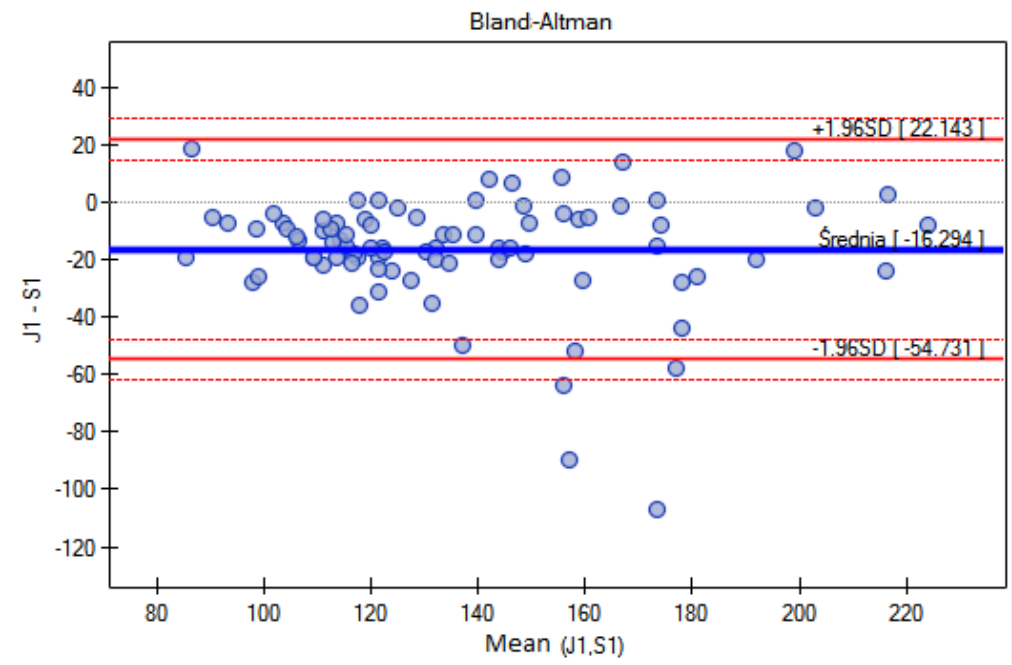
PERFORMING AND INTERPRETING THE ANALYSIS

Example 9.2. (pressure_gauge.pqs file)
This example is taken from the work of Bland and Altman (1999[8]). It compared a semi-automatic blood pressure monitor (S) with the classic blood pressure monitor (J) that had traditionally been used before. For this purpose, systolic blood pressure was measured for 85 patients using both monitors. An excerpt of the data is shown below.

J1	S1
100	122
108	121
76	95
108	127
124	140
122	139
116	122
114	130
100	119

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Bland-Altman plot for the collected data shows that the semi-automatic monitor (S) gives readings that are on average 16.3mmHG higher than the classic monitor (the line for the mean difference lies 16.3 below the line of absolute agreement at the 0 level). The span of the agreement interval is as much as 76.9mmHG.



For people with hypertension (systolic pressure $\geq$ 140), pressure changes can be quite large, so the measured differences may be distorted by genuine spikes in pressure. For this reason, a subgroup of people with normal blood pressure and a subgroup with hypertension were separated out on the basis of the mean pressure value. We can produce the plot separately for each subgroup by setting a multiple filter on the variable *group* in the test window. The agreement between the methods for people with normal blood pressure will then be markedly better (a narrower agreement interval).

The Bland-Altman plot for repeated measurements

The repeatability of measurements is an important, but often overlooked, aspect of studying the agreement between methods. A method with greater repeatability is more precise. If the measurements of one of the compared methods are not repeatable (i.e. repeated measurements taken on the same objects tend to give different results), its agreement with the other method will be low. If the repeatability of both methods is poor, their agreement will be even lower. As a consequence, when the repeatability of the old method is poor, the agreement of the new method may turn out to be low even if the new method itself has high repeatability. This is why, although in practice a single measurement is usually taken for each object (patient), in studies aimed at estimating agreement it is recommended that measurements be taken several times – this makes it possible to take the repeatability of the results obtained into account in the agreement study.

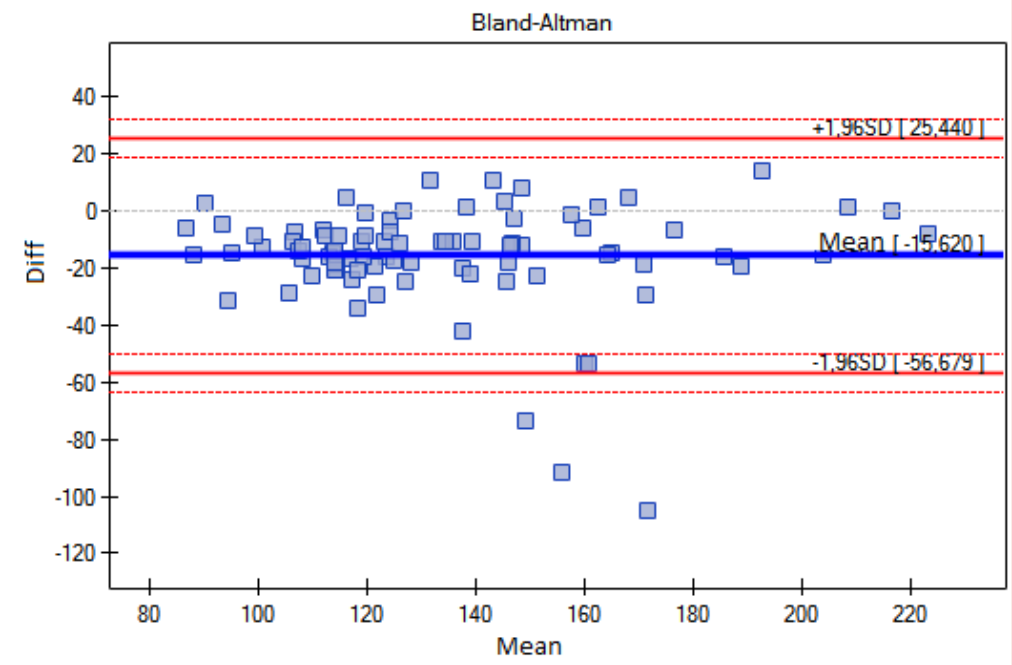
Note! By repeated measurements here we mean measurements taken independently on the same objects, under the same conditions.

Copyright ©2010-2026 U-smile Stats – All rights reserved

EXAMPLE (9.2) continued *(pressure_gauge.pqs file)*
In comparing the agreement between measurements taken with the two blood pressure monitors, the repeatability of both methods was also taken into account. For this reason, the study was repeated two more times, ultimately yielding, for each patient, 3 measurements taken with the semi-automatic monitor and 3 measurements taken with the classic monitor. An excerpt of the data is shown below.

J1	J2	J3	S1	S2	S3
100	106	107	122	128	124
108	110	108	121	127	128
76	84	82	95	94	98
108	104	104	127	127	135
124	112	112	140	131	124
122	140	124	139	142	136
116	108	102	122	112	112
114	110	112	130	129	135
100	108	112	119	122	122

This time the limits of agreement are somewhat wider than when a single measurement was used for each method – the span of the agreement interval is as much as 82.11mmHG. This happens because we are now taking the degree of repeatability of the measurements into account. Unfortunately, taking several repetitions into account increases the width of the limits, but the results obtained better represent reality – disregarding the degree of repeatability would amount to assuming 100% repeatability, which is almost impossible under real conditions.



As before, it is recommended to repeat the analysis separately for people with hypertension and those with normal blood pressure.

Copyright ©2010-2026 U-smile Stats – All rights reserved

9.4 Correlation matrices

WHEN TO USE:

When we are interested in the correlation between many variables at once, a convenient way to visualise it is to present the correlation coefficients for all pairs of variables in a single table (matrix). Instead of repeating the same analysis for every pair of variables in turn, we get an immediate, complete picture of the relationships within the dataset being studied. Depending on the scale on which the data were collected, PQStat lets you choose one of the following coefficients:

- r-Pearson (interval scale)
- r-Spearman (ordinal scale or stronger)
- tau-Kendall (ordinal scale or stronger)
- C-Pearson (nominal scale or stronger)
- V-Cramer (nominal scale or stronger)
- Phi (nominal scale or stronger)
- Q-Yule (nominal scale or stronger)

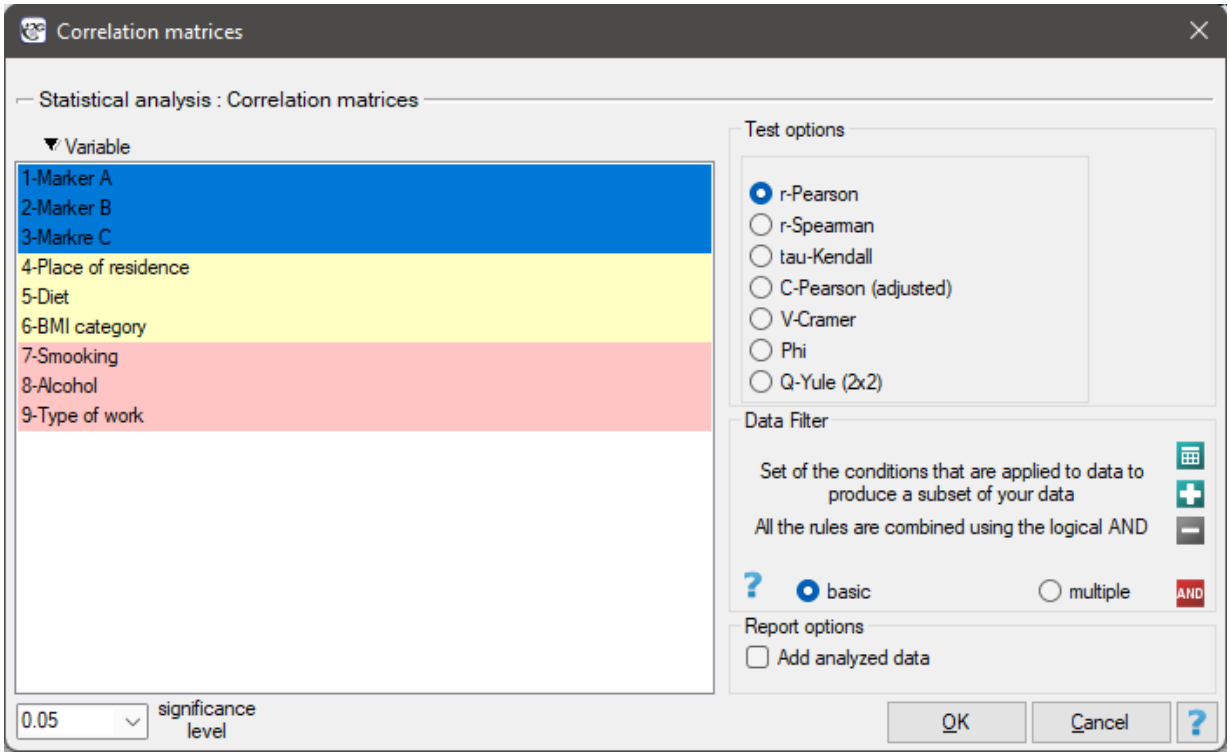
The analysis produces two matrices: a matrix of correlation coefficients and a matrix of p values for the test checking the statistical significance of each of these coefficients (for r-Pearson, r-Spearman and tau-Kendall these are their dedicated tests, and for the nominal scale it is the chi-square test). At the intersection of two variables in the coefficient matrix is the value of their mutual correlation, and the corresponding p value is found in the same place in the second matrix. The colour of a cell in the coefficient matrix is graded from blue (negative coefficients) to red (positive coefficients), which lets you immediately spot the strongest relationships at a glance.

Note! Correlation is determined separately for every pair of variables, so missing data are omitted pairwise. If you want cases with missing data in other variables (outside a given pair) to be removed from the analysis as well, you should use the advanced filter.

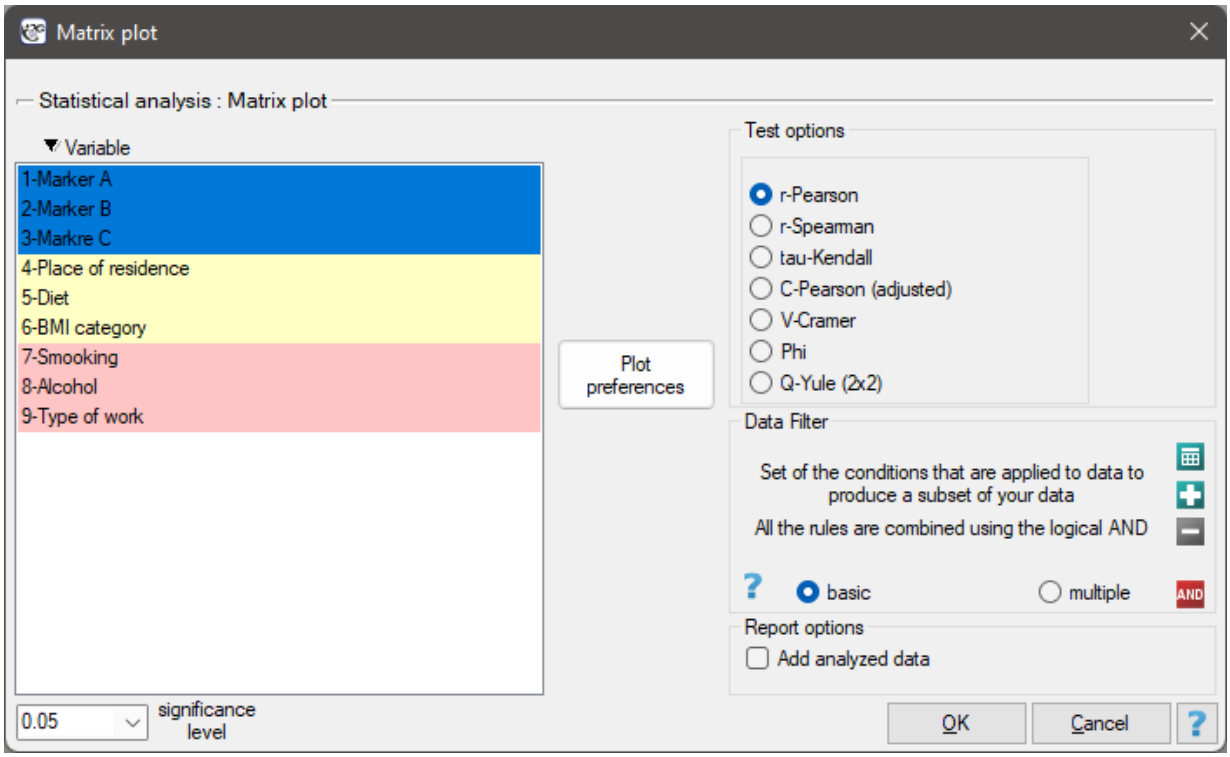
HOW TO SET UP THE ANALYSIS

The window with correlation matrix settings is opened via Statistics→Calculators→Correlation matrices.

Copyright ©2010-2026 U-smile Stats – All rights reserved



The window with the settings for the matrix plot of the correlation matrix is opened via Plots→Matrix plot.



PERFORMING AND INTERPRETING THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

EXAMPLE 9.3. (markers others.pqs file)
An excerpt from a larger study on cancer is presented. The data collected concern a group of 100 people. The study measured, among other things, the values of tumour markers (interval scale), determined BMI categories for the patients, and asked for opinions on the possible influence of place of residence and diet on health (ordinal scale); it also recorded the patients' answers about smoking, alcohol consumption and the type of work performed (nominal scale).
Conducting multivariate analyses often requires first checking the mutual correlations between the variables. For the purposes of the further analysis:
(1) we will check the correlation within each of these groups,
(2) we will check the correlation between all the variables.

(1) Correlation within the individual groups of variables

For the interval scale, assuming normality of the distribution, we will check the correlation using Pearson's linear correlation coefficient.

Marker A and marker C: $r = 0.8995$**,** $p < 0.0001$

The strongest and statistically significant correlation in this group of variables. The weakest, statistically non-significant correlation concerns marker B and marker C ($r = 0.0753$, $p = 0.4567$).

r-Pearsona

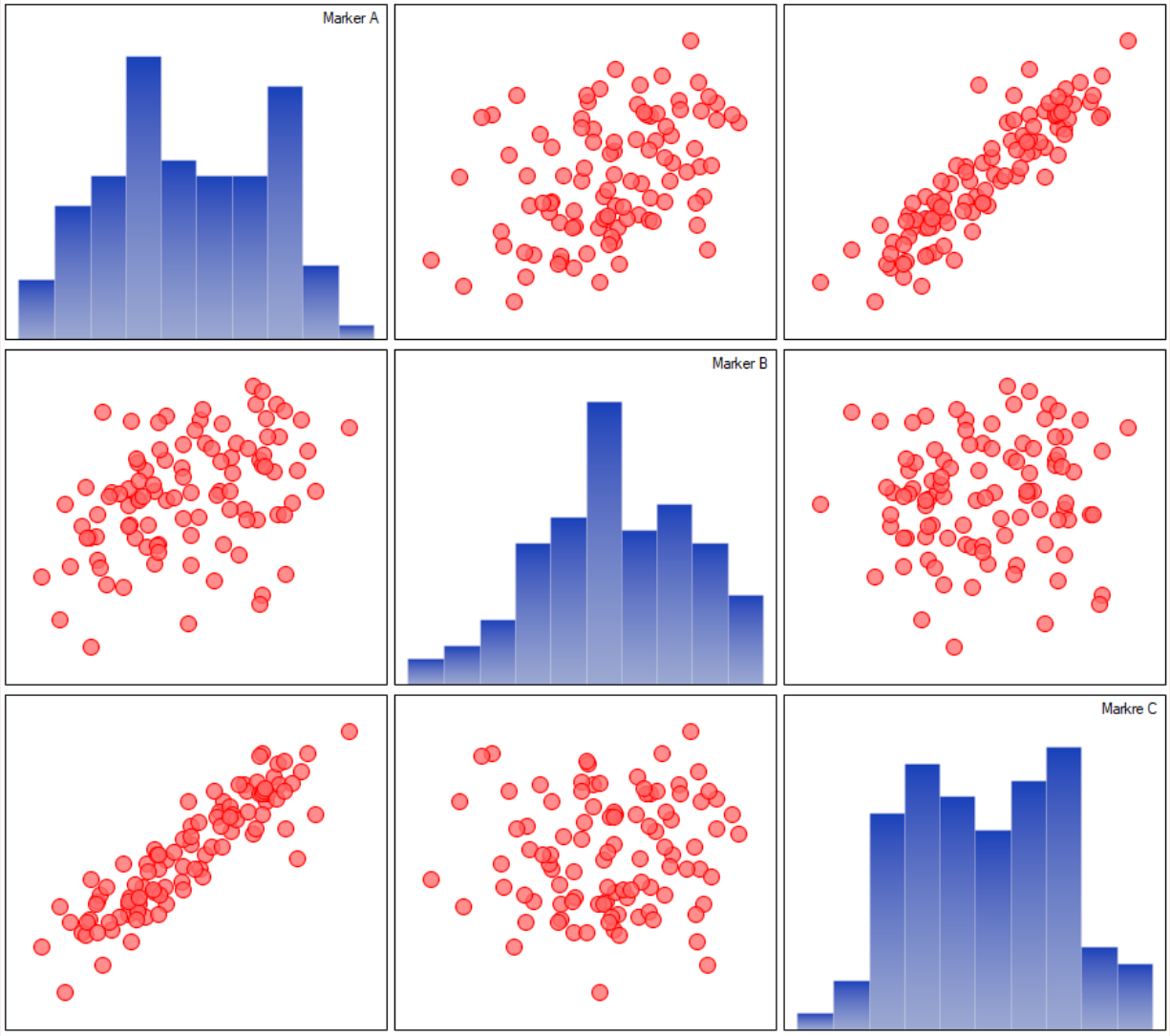
	Marker A	Marker B	Markre C
Marker A		**0.4027**	**0.8995**
Marker B	**0.4027**		0.0753
Markre C	**0.8995**	0.0753	

P-value

	Marker A	Marker B	Markre C
Marker A		<0.0001	<0.0001
Marker B	<0.0001		0.4567
Markre C	<0.0001	0.4567	

The correlations described can be observed on scatter plots (the X axis of these plots is the variable described in the columns, the Y axis – in the rows), and the distributions of the individual variables – on column charts.

Copyright ©2010-2026 U-smile Stats – All rights reserved



For the ordinal scale we will check the correlation using the Spearman correlation coefficient.

Diet and place of residence: $r = 0.2634$, $p = 0.0081$

The only statistically significant correlation in this group of variables.

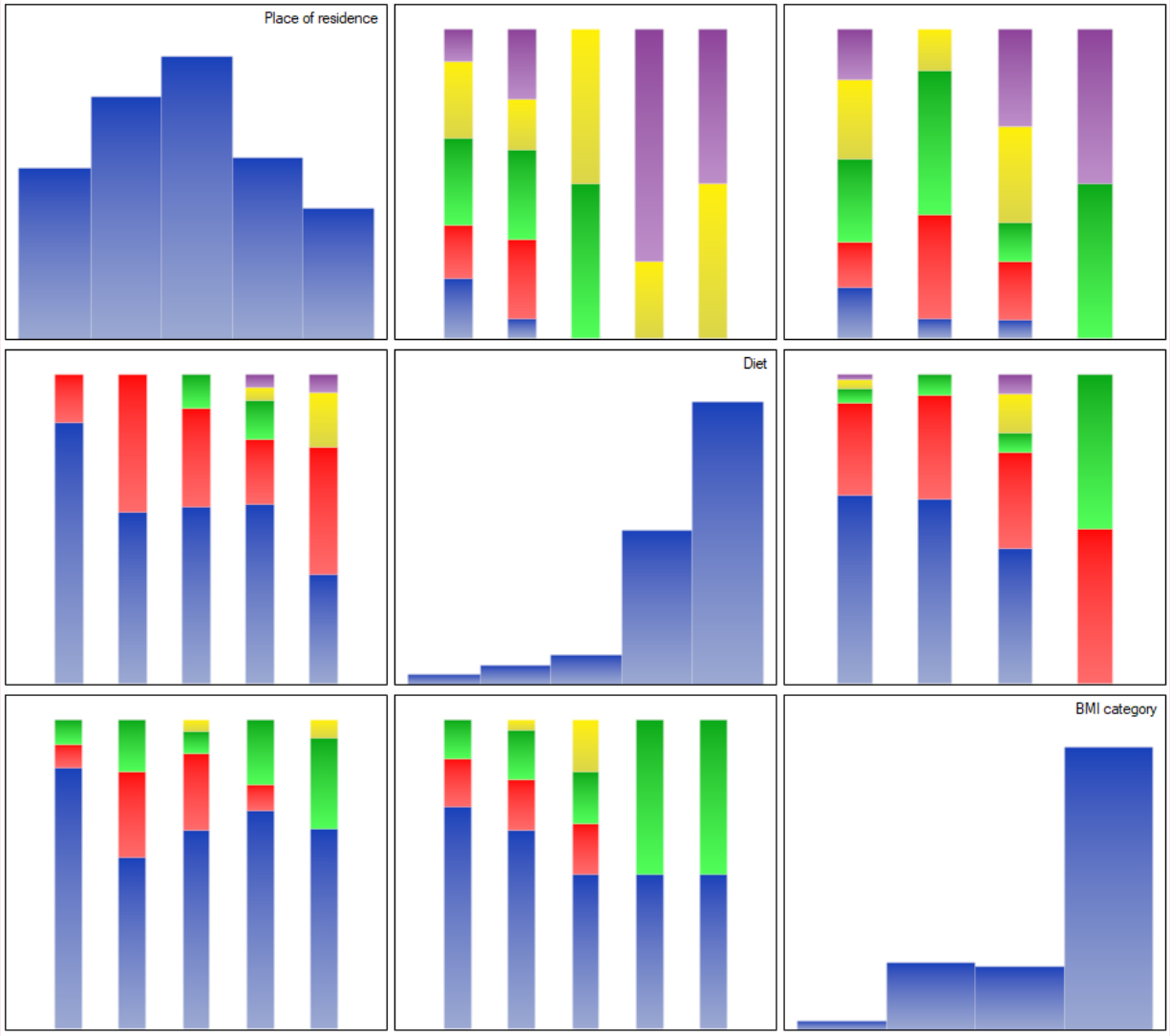
Copyright ©2010-2026 U-smile Stats – All rights reserved

r-Spearman	Place of resi	Diet	BMI categor
Place of residen		**0.2634**	0.0746
Diet	**0.2634**		0.1735
BMI category	0.0746	0.1735	

P-value	Place of resi	Diet	BMI categor
Place of residen		0.0081	0.4608
Diet	0.0081		0.0842
BMI category	0.4608	0.0842	

The correlations described can be observed on stacked column charts (the X axis of these plots is the variable described in the columns, the legend is the variable described in the rows), and the distributions of the individual variables – on the column charts located on the main diagonal.

Copyright ©2010-2026 U-smile Stats – All rights reserved



For the nominal scale we will check the correlation using the C-Pearson coefficient corrected for table size. We did not obtain any statistically significant correlations in this group of variables.

C-Pearson (adju

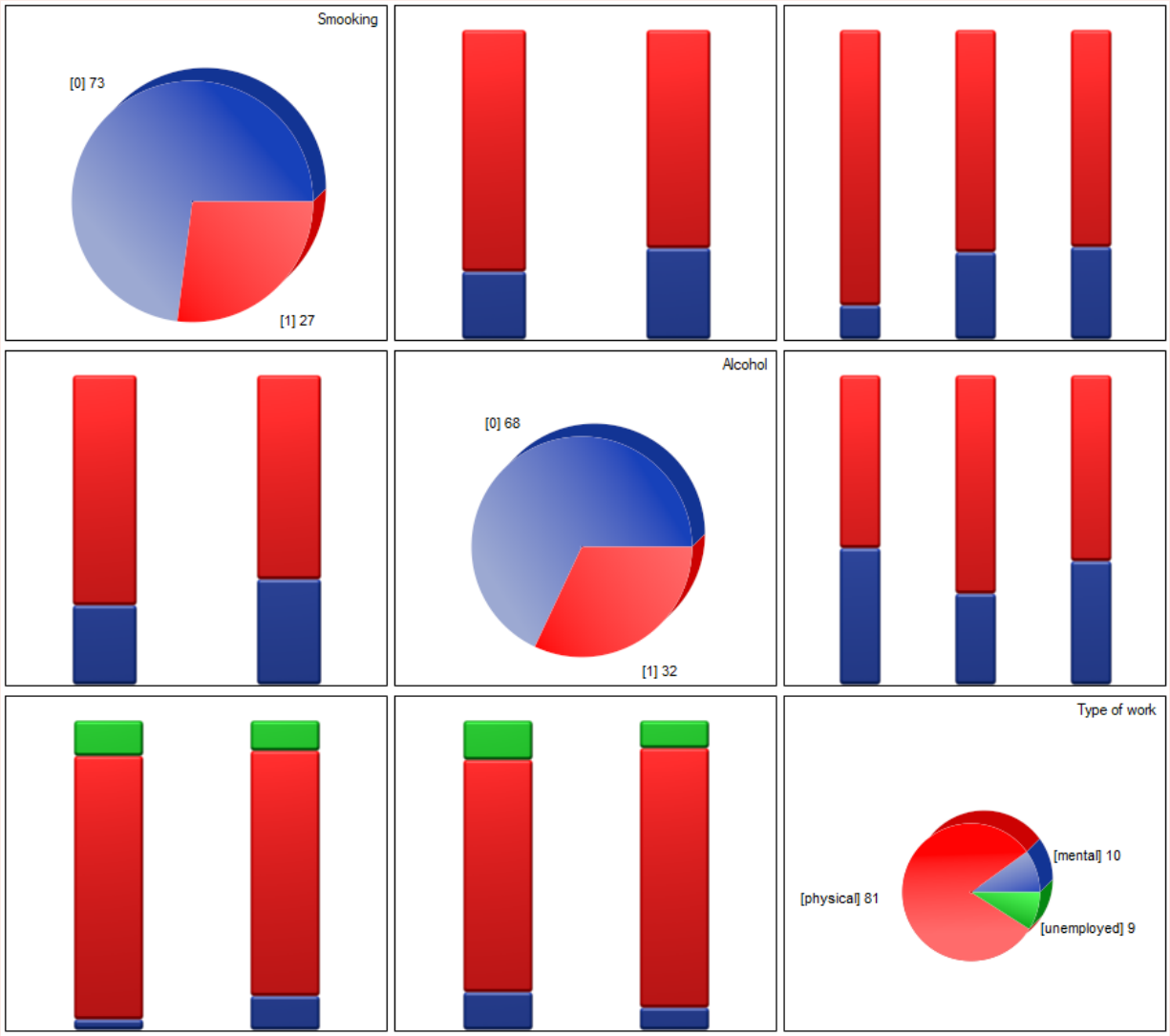
	Smooking	Alcohol	Type of work
Smooking		0.1116	0.1589
Alcohol	0.1116		0.1504
Type of work	0.1589	0.1504	

P-value

	Smooking	Alcohol	Type of work
Smooking		0.4284	0.5277
Alcohol	0.4284		0.5645
Type of work	0.5277	0.5645	

Copyright ©2010-2026 U-smile Stats – All rights reserved

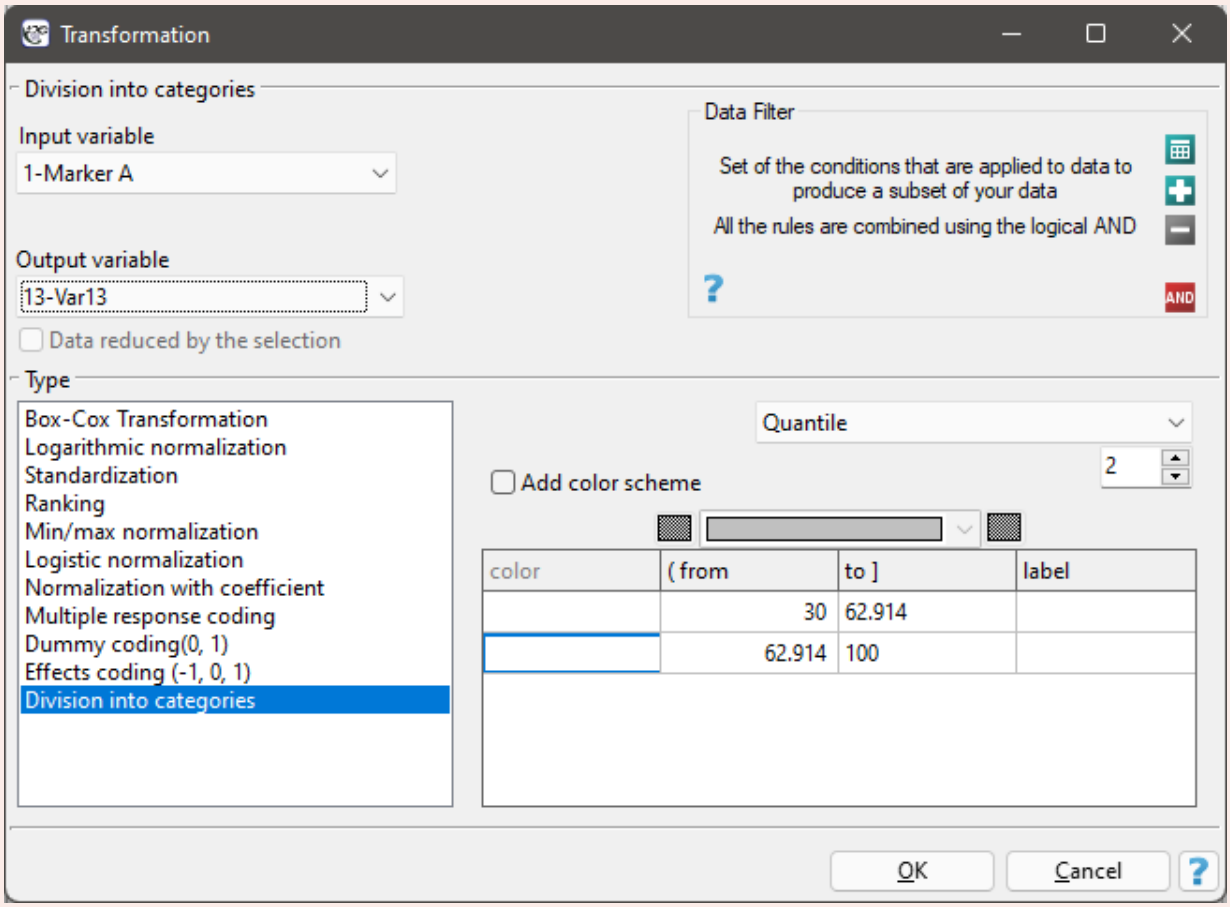
Any correlations could be observed on stacked column charts (the X axis of these plots is the variable described in the columns, the legend is the variable described in the rows), and the distributions of the individual variables – on the column charts located on the main diagonal.



(2) Correlation between all the variables

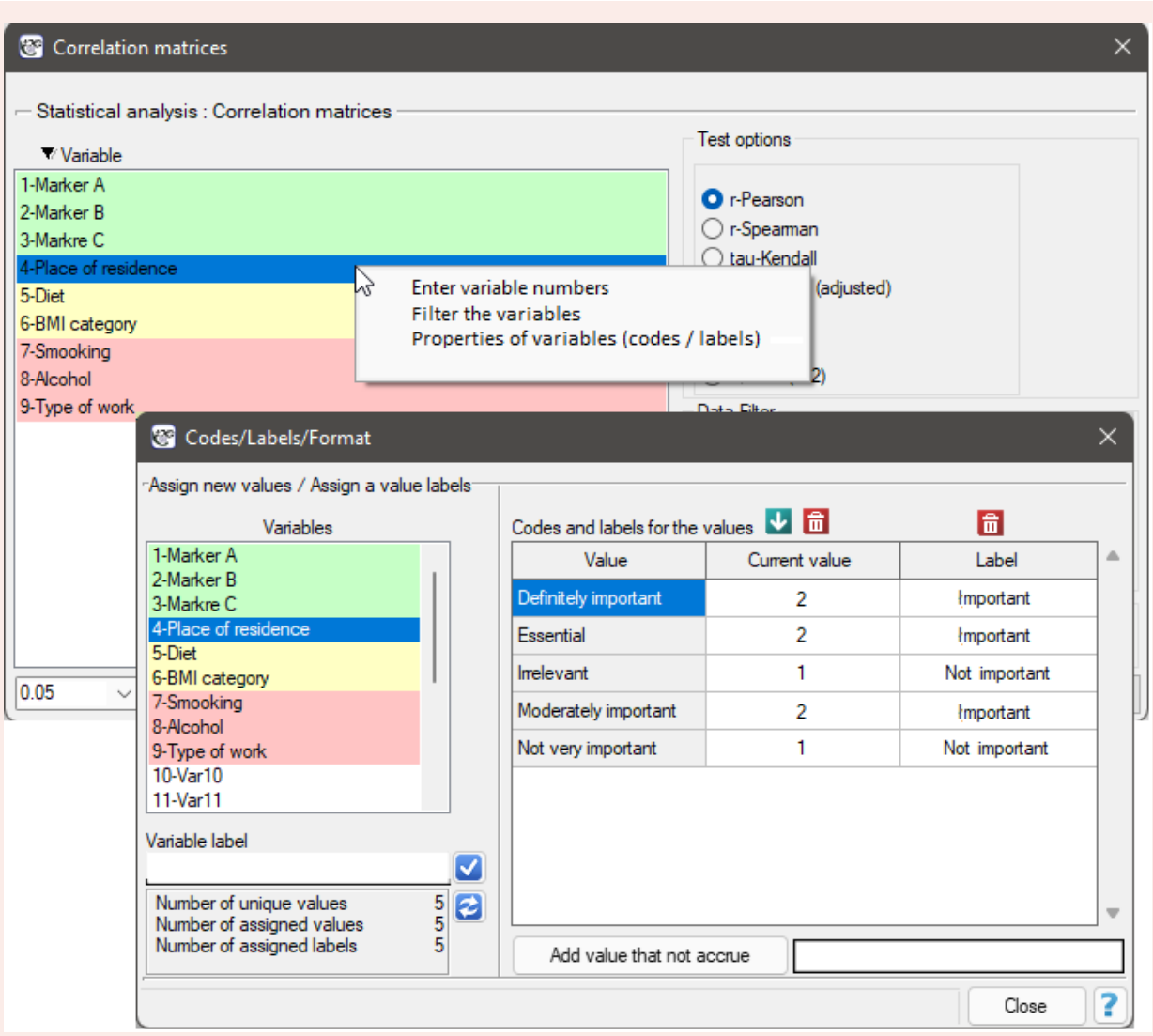
The easiest way to determine the correlation between variables measured on different scales is to bring them to the same scale. To do this, we will recode the interval data by dividing it into two categories – ”low” and ”high” – e.g. by quantiles. We can do this automatically in the transformation window via the menu Data→Transform.

Copyright ©2010-2026 U-smile Stats – All rights reserved



The ordinal data will also be divided into two categories, but this division will be made by selecting Variable Properties (Codes/Labels) in the analysis window via the context menu (right mouse button) and entering only the two valid values and two labels.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In the results we will show only the correlation matrix (without a plot), because a plot for so many variables would not be readable enough.

Copyright ©2010-2026 U-smile Stats – All rights reserved

C-Pearson (adju

	Place of resi	Diet	BMI categor	Smooking	Alcohol	Type of work	Range Quan	Range Quan	Range Quan
Place of residen		**0.4102**	**0.3824**	0.1335	0.1761	0.107	0.0287	0.143	0.0965
Diet	**0.4102**		0.3357	0.083	0.0102	**0.4159**	<0.0001	0.2349	0.1329
BMI category	**0.3824**	0.3357		0.273	0.2072	0.2357	0.1541	0.2271	0.329
Smooking	0.1335	0.083	0.273		0.1116	0.1589	0.0953	0.1583	0.192
Alcohol	0.1761	0.0102	0.2072	0.1116		0.1504	<0.0001	0.239	0.0218
Type of work	0.107	**0.4159**	0.2357	0.1589	0.1504		0.0497	0.1669	0.0353
Range Quantile[	0.0287	<0.0001	0.1541	0.0953	<0.0001	0.0497		**0.3813**	**0.91**
Range Quantile[	0.143	0.2349	0.2271	0.1583	0.239	0.1669	**0.3813**		0.1686
Range Quantile[	0.0965	0.1329	0.329	0.192	0.0218	0.0353	**0.91**	0.1686	

P-value

	Place of resi	Diet	BMI categor	Smooking	Alcohol	Type of work	Range Quan	Range Quan	Range Quan
Place of residen		0.0024	0.0484	0.3431	0.2094	0.7497	0.8389	0.3093	0.4942
Diet	0.0024		0.113	0.5565	0.9424	0.0088	1	0.0921	0.3452
BMI category	0.0484	0.113		0.2757	0.5333	0.6975	0.7528	0.4491	0.126
Smooking	0.3431	0.5565	0.2757		0.4284	0.5277	0.4992	0.2601	0.1705
Alcohol	0.2094	0.9424	0.5333	0.4284		0.5645	1	0.0863	0.8772
Type of work	0.7497	0.0088	0.6975	0.5277	0.5645		0.9401	0.4935	0.9693
Range Quantile[	0.8389	1	0.7528	0.4992	1	0.9401		0.0051	<0.0001
Range Quantile[	0.3093	0.0921	0.4491	0.2601	0.0863	0.4935	0.0051		0.2298
Range Quantile[	0.4942	0.3452	0.126	0.1705	0.8772	0.9693	<0.0001	0.2298	

10 TEST POWER AND SAMPLE SIZE



There are several ways to approach determining the sample size. One possibility is to estimate how large a sample should be to reflect the population well. Another possibility is to estimate the sample size for the situation of using specific statistical tests – in that case, besides the required sample size, we may also be interested in the power of those tests. The first, and seemingly easier, approach is presented in the subsection Sample size determination, the second in the subsection Power and sample size for test.

10.1 Sample size determination

WHEN TO USE:

Studying an entire population is usually neither practical nor possible, so a subset of it – a sample – is selected. The sample is of course smaller than the population, but it should reflect it well. One of the key aspects when planning a study, besides the randomness of the sample, is determining its size – so that inference about the population based on the sample is reliable.

If we are interested in ensuring that the proportion of a certain characteristic, or the mean value, calculated for a sample reflects the proportion or mean value in the population with as small an error as possible, we can, before starting the study, estimate how many people/observations are needed for this. In PQStat, the required sample size can be determined in this way both for estimating a **proportion** and for estimating a **mean value** in the population.

Copyright ©2010-2026 U-smile Stats – All rights reserved

HOW TO SET UP THE ANALYSIS

The window with the Sample size determination settings is opened via the menu Advanced statistics→Test power and sample size→Sample size determination. Regardless of whether we are estimating the required sample size for a proportion or for a mean, we need to provide the assumed margin of error E, i.e. the maximum deviation we are willing to accept between the result obtained from the sample and the true value in the population.

- For a **proportion** we additionally provide the expected proportion of the characteristic in the population – if it is unknown, the program uses the most conservative value $p = 0.5$, for which the required sample size is the largest.
- For a **mean** we additionally provide the standard deviation of the characteristic in the population, known e.g. from earlier studies.

If we know the size of the entire population (and, in particular, when it is relatively small compared to the calculated sample size), it is worthwhile to apply the so-called finite population correction (FPC) (Lenth (2001)[104], Armitage and Colton (2009)[12]) – the program allows this correction to be applied once the population size is provided.

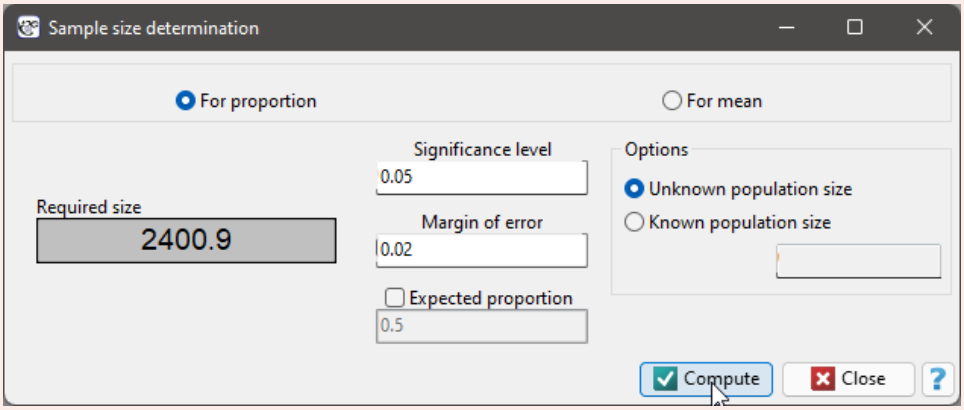
PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 10.1. Estimating a proportion

Population: People eligible to vote in the Polish presidential election.

We are interested in the support for individual candidates.

How many people should be selected so that the resulting percentage has an error of at most 2%?



Required sample size = 2401 people

With a sample of at least 2401 people, we will have 95% confidence that the error of the support for the selected presidential candidate does not exceed 2%. This means that in 95% of experiments consisting of drawing a random 2401-element sample from the population, the error of the support estimate for a given candidate will not exceed 2%, but in 5% of such experiments it may be greater than 2%.

Copyright ©2010-2026 U-smile Stats – All rights reserved

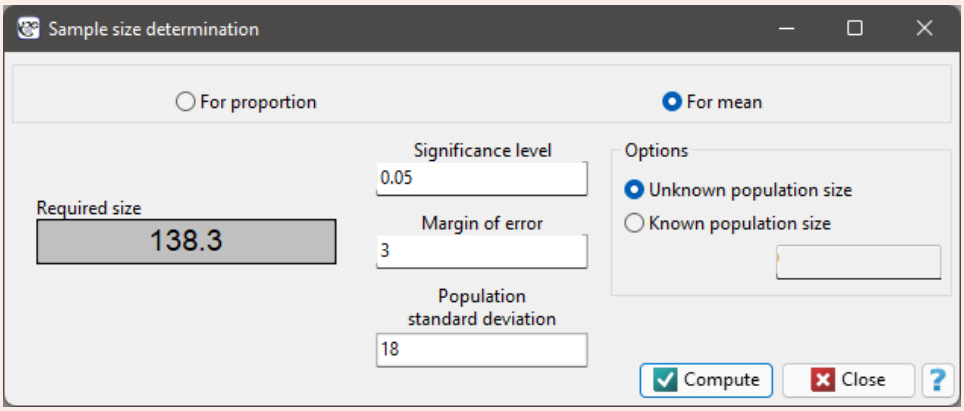
When choosing the size of the acceptable error, it is worth checking whether the election also includes candidates with low support (close to the assumed margin of error). If so, it is worth reducing the assumed error – the consequence of this will be an increase in the required sample size.

EXAMPLE 10.2. Estimating a mean value

Population: People suffering from hypertension in Poland in 2005-2010, aged 20-40.

We are interested in the mean body weight of these people.

How many people should be selected so that the mean weight has an error of at most $3kg$? We know that the population standard deviation of body weight for these people is $18kg$.



Required sample size = 139 people

To be 95% confident that the population mean lies within the margin of error ($\pm 3kg$) built around the mean of our sample, we need to collect at least 139 people.

FOR THOSE INTERESTED

Sample size for estimating a population proportion

Formulas: assuming the possibility of an estimation error of size E, we determine the required sample size n_0 – for an unknown population size, or n_{FPC} – for a known population size.

$$n_0 = \frac{Z^2 p(1-p)}{E^2}$$

where:

p – the expected proportion in the population, as specified by the researcher; if this quantity is unknown, the estimated required sample size will be increased so that it is sufficient for any possible proportion, i.e. the value $p = 0.5$ will be used.

When we know the population size (in particular – when it is relatively small compared to n_0, i.e. when $n_0/N > 5\%$), the so-called finite population correction (FPC) (Lenth (2001)[104], Armitage and Col-

Copyright ©2010-2026 U-smile Stats – All rights reserved

ton (2009)[12]) should be used, given by the formula:

$$n_{FPC} = \frac{n_0 N}{n_0 + (N-1)}$$

Sample size for estimating a population mean
Formulas: assuming the possibility of an estimation error of size E, we determine the required sample size n_0 – for an unknown population size, or n_{FPC} – for a known population size.

$$n_0 = \frac{Z^2 \sigma^2}{E^2}$$

where:

σ – the population standard deviation, known from earlier studies.

When we know the population size (in particular – when it is relatively small compared to n_0, i.e. when $n_0/N > 5\%$), the so-called finite population correction (FPC) should be used, given by the formula:

$$n_{FPC} = \frac{n_0 N}{n_0 + (N-1)}$$

10.2 Power and sample size for test

WHEN TO USE:

Before conducting a study (or after having conducted it), it is worth knowing whether the statistical test used even had a chance of detecting the difference, relationship or dependency we are interested in. This is exactly what **test power analysis** and determining the **required sample size** are for.

When testing hypotheses, we can make two kinds of mistakes: conclude that something differs, even though there is in fact no difference (**Type I error**, denoted α), or overlook a real difference, concluding that it does not exist (**Type II error**, denoted β). **Test power** is the ability of a test to detect a difference, relationship, agreement or other dependency described in the alternative hypothesis – in other words, it is the probability that the test correctly "notices" an effect that truly exists in the population. Not all tests are equally effective – some are "more sensitive" (more powerful), others "less sensitive" (weaker), so when several tests are available for the same problem, it is worth choosing the one with greater power.

Test power can be checked **a priori**, i.e. before collecting data – then we are planning how many people/objects we will need to have a real chance of detecting the effect we are interested in. It can also be done **a posteriori**, i.e. after collecting the sample – this is often done by paper reviewers or by the researchers themselves, to assess how credible the obtained (especially statistically non-significant) result is. Low power means that the results may be ambiguous – another researcher repeating the experiment could easily obtain a different result. High power means that the result is solid – it will be difficult to challenge it. For example: if, using a test with 80% power, we show that two groups of students differ statistically significantly in the number of correctly solved tasks, this means that when repeating the same experiment (the same sample size, the same test, the same significance level), other researchers would confirm this difference in 80% of cases.

Copyright ©2010-2026 U-smile Stats – All rights reserved

In PQStat, we can go two ways: calculate **test power**, by specifying the sample size, or calculate the **required sample size**, by specifying the test power we want to achieve.

ASSUMPTIONS

- knowledge of the rules for using the test for which we are calculating power or sample size (its hypotheses, assumptions, way of interpreting the result),
- the ability to determine the effect size that we want to be able to detect – ideally based on pilot data, earlier studies, or subject-matter knowledge about the phenomenon under study.

DEFINITIONS OF MEASURES

Test power and the required sample size are closely related to each other, and additionally depend on several other elements that need to be determined:

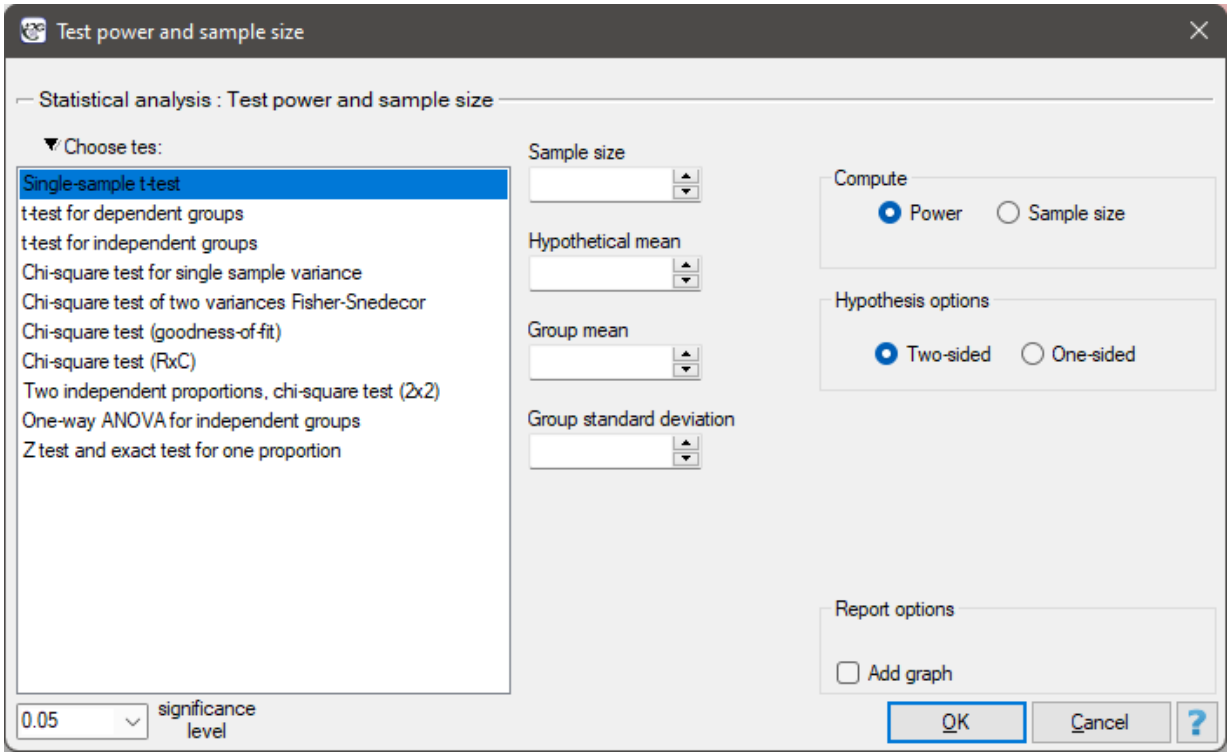
- **Effect size** ES – that is, how large a difference, relationship or agreement we consider worth detecting. The larger the effect, the easier it is to demonstrate – so the test power increases and the required sample size decreases. For example, it is easier to convince someone that two groups differ when the difference is five correctly solved tasks than when it is only one task. The effect size usually has to be standardized (e.g. using the standard deviation or a correlation coefficient), so additional information about the variability of the data is often needed for the calculation.
- **Statistical significance level** α (Type I error) – the higher the significance level, the greater the test power. However, this is an element over which we have only apparent control – there are rarely situations in which we are allowed to change it freely, and if we are, it is usually in the direction of decreasing it (see Multiple comparisons). By convention, $\alpha = 0.05$ is used.
- **Direction of the hypothesis**, i.e. a two-sided hypothesis (equality in the null hypothesis) or a one-sided hypothesis (a $<$ or $>$ sign in the null hypothesis). A one-sided hypothesis gives greater power, but is used less often – in practice we rarely assume in advance that one group has no chance at all of "beating" the other; more often we give both groups equal chances.
- **Test power** is expressed by the simple formula $Power = 1 - \beta$, where β (Type II error) is the probability of accepting the null hypothesis when it is in fact false.

Test power is also closely related to the **sample size** n – the larger the sample, the greater the power: the more people/objects we collect for the study, the easier it will be to demonstrate that the detected differences between groups are not the result of chance, but actually occur in the populations. Therefore, instead of (or alongside) test power, a researcher may want to determine the **required sample size** needed to achieve a given test power for a given effect size.

HOW TO SET UP THE ANALYSIS

The window with the settings for test power and the required sample size for the selected test is opened via the menu Advanced statistics→Test power and sample size→Power and sample size for test.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Power analysis is directly related to specific statistical tests, so in this window we first choose the test we are interested in, and then – depending on our needs – provide the sample size (to obtain test power) or the desired test power (to obtain the required sample size), as well as other information specific to the given test: the effect size (or the data from which it can be calculated), the significance level α and the direction of the hypothesis. Below, the data required for the power and sample size analysis for each test available in the program is described in detail, together with examples of its application.

Single-sample t-test

WHEN TO USE:

We perform the power and required sample size analysis for the single-sample t-test when we are planning (or have already performed) a comparison of a sample mean with a given hypothetical value, and want to know how large a sample to collect (or what power we have to detect a difference). Before starting the calculations, it is worth becoming familiar with the rules for using this test.

In this case, the given effect size is the size of the difference between the mean of the studied population and the hypothetical mean value. Test power and the required sample size are calculated based on the noncentral t-distribution.

HOW TO SET UP THE ANALYSIS

To determine the test power and the required sample size, we need to provide:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test power	Group size
Group size	Test power
Hypothetical mean	
Group mean	
Group standard deviation	
Significance level α	

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 10.3. We want to check (at a significance level of 0.05) whether the average waiting time for a delivery from a certain courier company is 3 days (i.e. 72 hours).

A priori study – we are planning to conduct a study, but have not yet collected data. We set the test power we want to achieve at 80%. The standard deviation we provide should reflect the differences in delivery time we expect in the planned study – we assume (based on the experience of the company's employees) that it will be 1 day (24 h).

a) What will be the required sample size if we assume that the effect size at which we would like to obtain statistical significance is 12 hours (0.5 days)?
b) What will be the required sample size if we assume that the effect size at which we would like to obtain statistical significance is 6 hours (0.25 days)?

Answer a) We know that the range from 2.5 to 3.5 days falls within the accepted margin of error. So we enter 3 days as the hypothetical mean and 2.5 days (or 3.5 days) as the mean of the studied group.

Test power and sample size	
Significance level	0.05
Single-sample t-test	
Hypothesis	two-sided
Moc - set	0.8
Hypothetical mean - set	3
Group mean - set	2.5
Standard deviation - set	1
Effect size (difference of the means)	0.5
Approximate power	0.8078
Sample size	34

Required sample size = 34 deliveries

This many deliveries need to be collected to prove that an effect exceeding 12 hours is statistically significant.

Answer b) We know that the range from 2.75 to 3.25 days falls within the accepted margin of error. So we enter 3 days as the hypothetical mean and 2.75 days (or 3.25 days) as the mean of the studied group.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test power and sample size	
Significance level	0.05
Single-sample t-test	
Hypothesis	two-sided
Moc - set	0.8
Hypothetical mean - set	3
Group mean - set	3.25
Standard deviation - set	1
Effect size (difference of the means)	0.25
Approximate power	0.8015
Sample size	128

Required sample size = 128 deliveries

This many deliveries need to be collected to prove that an effect exceeding 6 hours is statistically significant. **A**

posteriori study – we have collected data for the study – our sample contains 22 deliveries (data in the file kurier.pqs). Based on the collected data we determine the mean number of days waiting for delivery and the standard deviation of the group. In our case these are: mean=3.727273, deviation=1.906925.

a) What is the power of the analysis performed?
b) What would the power be if we increased the sample size to 100, leaving the remaining assumptions unchanged?

Answer a) The power of the analysis performed is only 0.400302.

Test power and sample size	
Significance level	0.05
Single-sample t-test	
Hypothesis	two-sided
Frequency - set	22
Hypothetical mean - set	3
Group mean - set	3.7273
Standard deviation - set	1.9069
Effect size (difference of the means)	0.7273
Power	0.4003

Power = 0.400302

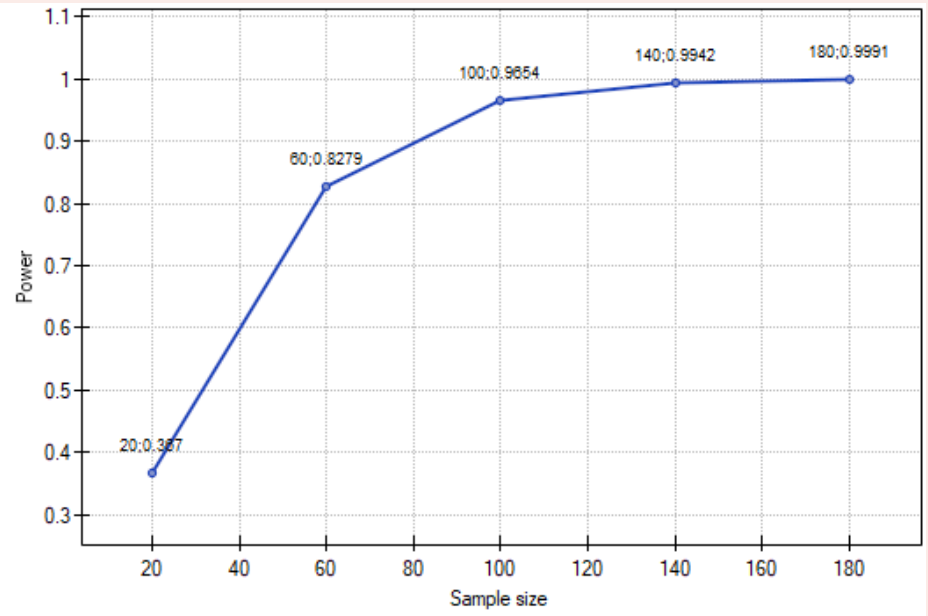
Many random samples of size 22 (about 60% of such samples) will not lead to confirmation of the alternative hypothesis. An-

swer b) The power of our analysis will increase to 0.965364 when its sample size increases to 100 elements and the assumptions of the analysis do not change.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test power and sample size	
Significance level	0.05
Single-sample t-test	
Hypothesis	two-sided
Frequency - set	100
Hypothetical mean - set	3
Group mean - set	3.7273
Standard deviation - set	1.9069
Effect size (difference of the means)	0.7273
Power	0.9654

How the power of the analysis changes with a changing sample size, with the other assumptions unchanged, can be seen on the chart.



T-test for dependent groups

WHEN TO USE:

We perform the power and required sample size analysis for the t-test for dependent groups when planning or evaluating a study in which the same people/objects are measured twice (e.g. before and after an intervention). Before starting the calculations, it is worth becoming familiar with the rules for using this test.

In this case, the given effect size is the size of the difference between the means that we expect to obtain in the population. Test power and the required sample size are calculated based on the noncentral t-distribution.

Copyright ©2010-2026 U-smile Stats – All rights reserved

HOW TO SET UP THE ANALYSIS

To determine the test power and the required sample size, we need to provide:

Test power	Group size
Group size	Test power
Group (study) mean 1	
Group (study) mean 2	
Standard deviation of the difference (or deviations of both groups and Pearson correlation coefficient)	
Significance level α	

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 10.4. We want to check (at a significance level of 0.05) whether treating an eating disorder at a certain clinic produces a significant reduction in body weight after just 30 days of following a new type of diet. We consider a change in BMI of half a unit to be a significant change in body weight. How large a sample should be collected for a difference of this magnitude to be statistically significant in a t-test for dependent groups?

Since we do not have data from a pilot study, we will base the basic data for the calculations on the experience and estimates of the clinic staff. We assume that the average BMI of a treated person is 35 – we enter this value in the box for the first mean. Since a change in BMI of less than half a unit is not clinically significant, only a decrease below 34.5 (or an increase above 35.5) will be considered significant. So we enter the value 34.5 (or 35.5) as the second mean. We presume that the standard deviation of the difference (BMI before and BMI after) may be quite large, since the group will usually include people who are disciplined about following the diet, and those who nevertheless allow themselves extra snacks between meals. We therefore set the deviation at 2.5. The power of the analysis we want to obtain is 80%.

Test power and sample size		
Significance level	0.05	0.05
t-test for dependent groups		
Hypothesis	two-sided	one-sided
Moc - set	0.8	0.8
Group mean 1 - set	35	35
Group mean 2 - set	34.5	34.5
SD of difference - set	2.5	2.5
Effect size (difference of the means)	0.5	0.5
Approximate power	0.8017	0.8002
Sample size	199	156

Required sample size = 199 people (two-sided hypothesis) or 156 people (one-sided hypothesis)

The two-sided hypothesis assumes that BMI may decrease or increase as a result of the diet; the one-sided hypothesis assumes only a decrease in BMI.

If

Copyright ©2010-2026 U-smile Stats – All rights reserved

we assumed that the group would be more disciplined and the standard deviation of the difference would be 1.5 BMI units, then the sample could be slightly smaller.

Test power and sample size		
Significance level	0.05	0.05
t-test for dependent groups		
Hypothesis	two-sided	one-sided
Moc - set	0.8	0.8
Group mean 1 - set	35	35
Group mean 2 - set	34.5	34.5
SD of difference - set	1.5	1.5
Effect size (difference of the means)	0.5	0.5
Approximate power	0.8023	0.806
Sample size	73	58

Required sample size = 73 people (two-sided hypothesis) or 58 people (one-sided hypothesis)

A smaller assumed standard deviation of the difference makes it possible to achieve the same power with a smaller sample.

T-test for independent groups

WHEN TO USE:

We perform the power and required sample size analysis for the t-test for independent groups when planning or evaluating a study comparing means in two independent groups (e.g. ill and healthy). Before starting the calculations, it is worth becoming familiar with the rules for using this test.

In this case, the given effect size is the size of the difference between the means that we expect to obtain between the populations. Test power and the required sample size are calculated based on the noncentral t-distribution.

HOW TO SET UP THE ANALYSIS

To determine the test power and the required sample size, we need to provide:

Test power	Group size
Group size 1 Group size 2	Test power
Group mean 1	
Group mean 2	
Common standard deviation (or deviations of both groups*)	
Significance level α	

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! *When the sizes of both groups are known, the common standard deviation is determined using an exact formula; when the sizes are not known, the simplified Cohen formula is used (Cohen, J. (1988)[47]) – both formulas are given in the "For those interested" section at the end of the chapter.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 10.5. We are studying men suffering from disease X and healthy men. We want to check (at a significance level of 0.05) whether the ill differ from the healthy in HDL cholesterol level. We consider a difference of 2 mg/dl to be clinically significant. How large a sample should be collected for a difference of this magnitude to be statistically significant in the t-test for independent groups?

We enter 40 mg/dl as the mean HDL for ill men and 42 mg/dl for healthy men. We set the ratio of the two group sizes to 1, since we assume equal group sizes. We have data from a pilot study, so we provide the standard deviations for the ill (13 mg/dl) and healthy (11 mg/dl), indicating the additional options . The power of the analysis we want to obtain is 80%.

Test power and sample size		
Significance level	0.05	0.05
t-test for independent groups		
Hypothesis	two-sided	two-sided
Moc - set	0.8	0.8
Sample size ratio - set	1	2
Group mean 1 - set	42	42
Group mean 2 - set	40	40
Pooled standard deviation - set	12.0416	12.0416
Effect size (difference of the means)	2	2
Approximate power	0.8007	0.8004
Sample size	571 / 571	855 / 428

Required sample size = 571 people (equal groups, n1/n2=1) or n1=855, n2=428 (n1/n2=2)

The less equal the group sizes are meant to be, the larger the total number of people needed to maintain the same test power.

Chi-square test for single sample variance

WHEN TO USE:

We perform the power and required sample size analysis for the chi-square test for single sample variance when planning or evaluating a study comparing the variability (standard deviation) in a sample with a given hypothetical value. Before starting the calculations, it is worth becoming familiar with the rules for using this test.

In this case, the given effect size is the ratio of the standard deviation of the studied population to the

Copyright ©2010-2026 U-smile Stats – All rights reserved

hypothetical standard deviation. Test power and the required sample size are calculated based on the chi-square distribution.

HOW TO SET UP THE ANALYSIS

To determine the test power and the required sample size, we need to provide:

Test power	Group size
Group size	Test power
Hypothetical standard deviation	
Group standard deviation	
Significance level α	

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 10.6. Before starting production of the next batch of a certain cough syrup, control measurements should be taken of the volume of syrup poured into bottles. The bottles should contain 200 ml of syrup. According to the technical documentation of the dispensing device, the permissible spread of syrup volume, measured by the standard deviation, is 1 ml. It should be checked (at a significance level of 0.05) whether the tested device is working correctly. Will a sample of 20 bottles be sufficient to demonstrate an excessive device error, if one exists? A standard deviation exceeding 1.2 ml is considered an excessive device error.

Since we expect the standard deviation of the dispensing device to match the documentation, we enter the value 1 ml as the hypothetical value. An excessive error occurs when the deviation exceeds 1.2 ml, and we enter this value in the field for the standard deviation of the group of bottles we are going to test.

If the sample consists of 20 bottles, the resulting power for the two-sided hypothesis is only 0.25, and for the one-sided hypothesis 0.34.

Test power and sample size		
Significance level	0.05	0.05
Chi-square test for single sample variance		
Hypothesis	two-sided	one-sided
Frequency - set	20	20
Hypothetical standard deviation - set	1	1
Standard deviation - set	1.2	1.2
Effect size (ratio of deviations)	0.8333	0.8333
Power	0.2482	0.3405

Power = 0.25 (two-sided) / 0.34 (one-sided) for n=20

Low values – fewer than 35% of random samples of this size will detect a spread exceeding 0.2 ml. 20 bottles is too small a group to prove an excessive device standard deviation, if one truly exists. We would like to obtain a standard power. To obtain a power of 80%, we change the program settings and calculate the required sample size.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test power and sample size		
Significance level	0.05	0.05
Chi-square test for single sample variance		
Hypothesis	two-sided	one-sided
Moc - set	0.8	0.8
Hypothetical standard deviation - set	1	1
Standard deviation - set	1.2	1.2
Effect size (ratio of deviations)	0.8333	0.8333
Approximate power	0.8027	0.8032
Sample size	115	92

Required sample size = 115 (two-sided) / 92 (one-sided) at 80% power

This many bottles need to be checked in order to demonstrate that the permissible spread has been exceeded while maintaining 80% test power.

Chi-square test of two variances, Fisher-Snedecor

WHEN TO USE:

We perform the power and required sample size analysis for the Fisher-Snedecor test when planning or evaluating a study comparing the variability (standard deviation) between two independent samples. Before starting the calculations, it is worth becoming familiar with the rules for using this test.

In this case, the given effect size is the ratio of the standard deviation of the first population to that of the second. Test power and the required sample size are calculated based on the F-Snedecor distribution.

HOW TO SET UP THE ANALYSIS

To determine the test power and the required sample size, we need to provide:

Test power	Group size
Group size 1 Group size 2	Test power
Group 1 standard deviation	
Group 2 standard deviation	
Significance level α	

PERFORMING AND INTERPRETING THE ANALYSIS

Example 10.7. Before starting production of the next batch of a certain cough syrup, control measurements should be taken of the volume of syrup poured into bottles. The bottles should contain 300

Copyright ©2010-2026 U-smile Stats – All rights reserved

ml of syrup. Two dispensing devices are used at the bottling plant. We want to check (at a significance level of 0.05) whether the spread of syrup volume, measured by the standard deviation, is the same for both devices. A small pilot study was conducted, and it was established that the standard deviation for the first device is 1.32, and for the second 1.1. If the difference is small, i.e. the ratio of the two deviations is below 1.2 (as in the pilot study), both devices will be used interchangeably; if not, the one with the smaller deviation from the mean will be chosen. How many randomly selected bottles should be measured to be able to show that a ratio of 1.2 is statistically significant?

We enter the values of the standard deviations obtained in the pilot study and assume 80% test power.

Test power and sample size		
Significance level	0.05	0.05
Chi-square test of two variances Fisher-Snedecor		
Hypothesis	two-sided	two-sided
Moc - set	0.8	0.8
Sample size ratio - set	1	2
Standard deviation 1 - set	1.32	1.32
Standard deviation 2 - set	1.1	1.1
Effect size (ratio of deviations)	0.8333	0.8333
Approximate power	0.8015	0.8006
Sample size	239 / 239	363 / 182

Required sample size = n1=n2=239 (equal groups, n1/n2=1) or n1=363, n2=182 (n1/n2=2)

This many bottles must be measured for each device in order to show a statistically significant ratio of standard deviations equal to 1.2 while maintaining 80% test power.

Chi-square test of goodness of fit

WHEN TO USE:

We perform the power and required sample size analysis for the chi-square goodness-of-fit test when planning or evaluating a study checking whether the distribution of data in a sample agrees with an expected (theoretical or hypothetical) distribution. Before starting the calculations, it is worth becoming familiar with the rules for using this test.

In this case, the given effect size, ϕ, is the square root of the ratio of the chi-square test statistic to the size of the studied group – the formula is given in the "For those interested" section at the end of the chapter. Test power and the required sample size are calculated based on the noncentral chi-square distribution.

HOW TO SET UP THE ANALYSIS

To determine the test power and the required sample size, we need to provide[47]:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test power	Group size
Group size Number of categories	Test power
Effect size ϕ	
Significance level α	

Chi-square test (RxC)

WHEN TO USE:

We perform the power and required sample size analysis for the chi-square test for RxC tables when planning or evaluating a study checking the relationship between two categorical variables shown in a table with several rows and columns. Before starting the calculations, it is worth becoming familiar with the rules for using this test.

In this case, the given effect size, ϕ, is the square root of the ratio of the chi-square test statistic to the size of the studied group – the formula is given in the "For those interested" section at the end of the chapter. Test power and the required sample size are calculated based on the noncentral chi-square distribution.

HOW TO SET UP THE ANALYSIS

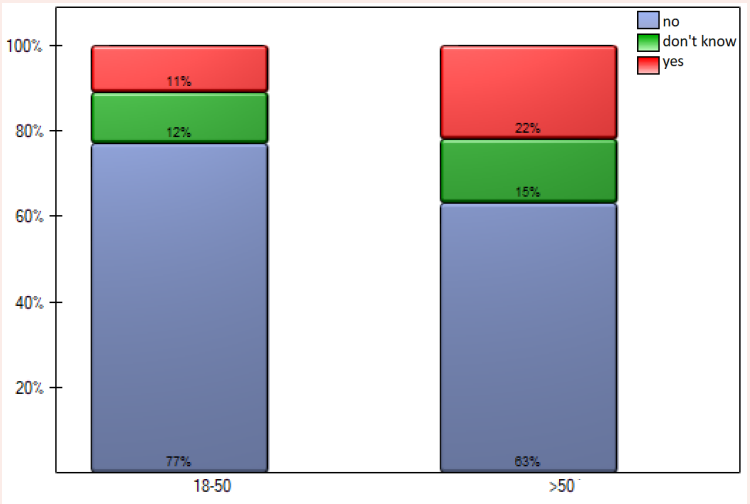
To determine the test power and the required sample size, we need to provide[47][133]:

Test power	Group size
Group size Number of categories (rows) Number of categories (columns)	Test power
Effect size ϕ	
Significance level α	

PERFORMING AND INTERPRETING THE ANALYSIS

Example 10.8. There are plans to conduct a broad study showing the state of knowledge of the Polish population about ways of fighting common viruses. As part of the project, it should be established whether educational activities informing about the ineffectiveness of antibiotic therapy in viral infections were equally effective in the older age group (i.e. over 50 years old) as in younger adults (18-50 years old). A pilot study was conducted, and a randomly selected sample of 200 people was asked the question: "Do antibiotics fight viruses?". Respondents could choose one of three answers: "no", "I don't know", "yes". The results of the pilot study were prepared for publication. Below is an excerpt from the description included in the paper:

Copyright ©2010-2026 U-smile Stats – All rights reserved



	no	dont konow	yes
18-50	77	12	11
>50	63	15	22

The p-value obtained in the chi-square test was statistically non-significant: p=0.0672. The paper's reviewer was rightly surprised by the fact that twice as many people over 50 incorrectly indicated that antibiotics fight viruses (22% versus 11%), yet this difference was not statistically significant.

As suggested by the reviewer, it should be checked whether the lack of statistical significance of this difference is due to too low a test power, and how large a sample should be to obtain, with the same percentages, a chi-square test power of 80%.

Preparing an answer for the reviewer

We will determine the ϕ coefficient for the obtained test (menu: Chi-square, Fisher, OR/RR→Correlation coefficients...→Phi). We obtain ϕ=0.1643. Based on a 200-element sample, with data placed in a table with two rows and three columns, and the given value of the ϕ coefficient, we determine the power of the chi-square test.

Test power and sample size	
Significance level	0.05
Chi-square test (RxC)	
Hypothesis	two-sided
Frequency - set	200
Number of Categories (rows) - set	2
Number of Categories (columns) - set	3
Effect size (Phi) - set	0.1643
Chi-square test statistic value	5.3989
Power	0.5368

Power = 0.5368

A low value, which confirms concerns about an insufficiently large sample.

If, for a sample of a different size, we obtain the same distribution of data, this means that we also obtain

Copyright ©2010-2026 U-smile Stats – All rights reserved

the same coefficient ϕ. To determine the sample size that would give us 80% power of the chi-square test, we again enter the coefficient ϕ=0.1643.

Test power and sample size	
Significance level	0.05
Chi-square test (RxC)	
Hypothesis	two-sided
Moc - set	0.8
Number of Categories (rows) - set	2
Number of Categories (columns) - set	3
Effect size (Phi) - set	0.1643
Chi-square test statistic value	9.637
Approximate power	0.8001
Sample size	357

Required sample size = 357 people (at 80% power)

This many respondents are needed to demonstrate statistical significance with 80% test power while keeping the same distribution of answers. Since this is only a pilot study, we plan to increase the study group to 357 people in the main study.

Already at this point we notice that when the undecided respondents (i.e. those who chose the answer "I don't know") are omitted and the analysis is repeated, significant differences can be shown (chi-square, p=0.0251). In the group of decided respondents, the percentages choosing the incorrect answer are more than twice as high to the disadvantage of people over 50 years old (12.5% vs 25.9%).

Test for two independent proportions, chi-square (2x2)

WHEN TO USE:

We perform the power and required sample size analysis for the chi-square test for 2x2 tables and the Z test for two independent proportions when planning or evaluating a study comparing two percentages (proportions) in two independent groups. Before starting the calculations, it is worth becoming familiar with the rules for using both these tests.

In this case, the given effect size is the difference between the specified proportions. Test power and the required sample size are calculated based on the normal distribution.

HOW TO SET UP THE ANALYSIS

To determine the test power and the required sample size, we need to provide[37]:

Test power	Group size
Group size	Test power
Specified proportion in the first group	
Specified proportion in the second group	
Significance level α	

Copyright ©2010-2026 U-smile Stats – All rights reserved

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 10.9. Consider a study evaluating the effectiveness of aspirin in reducing mortality due to myocardial infarction. Earlier studies indicate that the percentage of deaths due to myocardial infarction is 0.015 for non-users and 0.001 for users of aspirin. The researchers want to determine the minimum sample size required to detect an absolute difference of $|0.001 - 0.015| = 0.014$ with 80% power using a two-sided test at a significance level of 5%.

Test power and sample size	
Significance level	0.05
Two independent proportions, chi-square test (2x2)	
Hypothesis	two-sided
Moc - set	0.8
Sample size ratio - set	1
Group proportion 1 - set	0.001
Group proportion 2 - set	0.015
Effect size (proporcja 1 - proporcja 2)	-0.014
Approximate power	0.8004
Sample size	635 / 635

Required sample size = 635 people in each group (equal groups)

This many people need to be collected in each of the groups to show statistical significance of a proportion difference of 0.014 at 80% test power.

One-way ANOVA for independent groups

WHEN TO USE:

We perform the power and required sample size analysis for one-way ANOVA for independent groups when planning or evaluating a study comparing means in three or more independent groups. Before starting the calculations, it is worth becoming familiar with the rules for using this method.

In this case, the given effect size is RMSSE, i.e. a standardized measure used in ANOVA to describe the overall level of effect in the population. Test power and the required sample size are calculated based on the noncentral F-Snedecor distribution.

HOW TO SET UP THE ANALYSIS

To determine the test power and the required sample size, we need to provide:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test power	Group size
Group size	Test power
Group 1 mean	
Group 2 mean	
Group 3 mean	
Group 4 mean	
Group 1 standard deviation	
Group 2 standard deviation	
Group 3 standard deviation	
Group 4 standard deviation	
Significance level α	

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 10.10. The FVC parameter was studied for patients with heart defects (heart defect A, heart defect B and heart defect C). We want to find out (at a significance level of 0.05) whether the patients differ in the values of this parameter. To check this, a pilot study was first conducted. Based on the results of this study, the expected effect sizes were established, i.e.:
– for heart defect A: mean = 3.8, standard deviation = 1.1,
– for heart defect B: mean = 4.5, standard deviation = 0.6,
– for heart defect C: mean = 4.2, standard deviation = 0.9.
How many people should be collected, if these quantities remain at the same level, to prove that there are statistically significant differences?

We enter the values of the means and standard deviations.

Test power and sample size	
Significance level	0.05
One-way ANOVA for independent groups	
Hypothesis	two-sided
Power - set	0.8
Number of groups - set	3
Mean 1 - set	3.8
Mean 2 - set	4.5
Mean 3 - set	4.2
Standard deviation 1 - set	1.1
Standard deviation 2 - set	0.6
Standard deviation 3 - set	0.9
Effect size (RMSSE)	0.3943
Approximate power	0.8131
Sample size	33

Required sample size = 33 people in each group (at 80% power)

This many people need to be collected in each of the studied groups to achieve a test power of 80% while maintaining the expected effect sizes.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test for one proportion

WHEN TO USE:

We perform the power and required sample size analysis for the test for one proportion when planning or evaluating a study comparing a proportion (percentage) in a sample with a given expected value. Before starting the calculations, it is worth becoming familiar with the rules for using this test.

In this case, the given effect size is the size of the difference between the specified proportion in the studied population and the hypothetical, expected proportion. Test power and the required sample size are calculated based on the normal distribution, when using the asymptotic test, or the binomial distribution, when using the exact test.

HOW TO SET UP THE ANALYSIS

To determine the test power and the required sample size, we need to provide:

Test power	Group size
Group size	Test power
Expected proportion	
Group proportion	
Significance level α	

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 10.11. 10 randomly selected families with children under 10 years old, living in Poznań, were asked about the planned educational future of their children. Of these, 6 families planned to educate their children at university. How large a sample would we need to collect to be able to state, at a significance level of 0.05 and a test power of 0.8, that more than 50% of families in Poznań with children under 10 years old already plan to have their children educated at university in the future?

We enter 0.5 as the expected proportion, and 0.6 as the group proportion.

Test power and sample size		
Significance level	0.05	0.05
Z test and exact test for one proportion		
Hypothesis	two-sided	one-sided
Power - set	0.8	0.8
Expected proportion - set	0.5	0.5
Group proportion - set	0.6	0.6
Effect size (difference of the proportions)	0.1	0.1
Approximate power	0.8003	0.8013
Sample size **(asymptotic)**	194	153

Required sample size = 194 families (two-sided hypothesis) or 151 families (one-sided hypothesis)

This many families need to be collected to show, at 80% test power, that the percentage of those planning higher education for their children exceeds 50%.

Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

T-test for independent groups – common standard deviation
Formulas: when the sizes of both groups are known, the common standard deviation is given by the formula:

$$SD_p = \sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}},$$

when the sizes are not known, the simplified Cohen formula is used (Cohen, J. (1988)[47]):

$$SD_p = \sqrt{\frac{sd_1^2 + sd_2^2}{2}}.$$

where:
n_1, n_2 – the sizes of the first and second sample,
sd_1, sd_2 – the standard deviations in the first and second sample.

Chi-square goodness-of-fit test and chi-square test (RxC) – effect size ϕ
Formulas: the given effect size ϕ is the square root of the ratio of the chi-square test statistic to the size of the studied group:

$$\phi = \sqrt{\frac{\chi^2}{n}}.$$

11 DESCRIPTIVE ANALYSES

The data collected by a researcher should first of all be described. Which measures are used for this description depends on how the measurements were made, i.e. on the so-called measurement scale of the variable.

11.1 Measurement scales

Correctly determining the type of analysis to be performed depends on the scale on which the collected data are expressed. There are 3 main measurement scales:

1. **Interval scale** (*ang. interval scale*)

   A variable is expressed on an interval scale when:

   - it can be ordered,
   - it is possible to calculate by how much one element is greater than another, and this difference has a real-world interpretation. A unit of measurement is usually defined.

   Example: the mass of an object [kg], the area of an object [m], time [years], speed [km/h], etc.

2. **Ordinal scale** (*ang. ordinal scale*)

   A variable is expressed on an ordinal scale when:

Copyright ©2010-2026 U-smile Stats – All rights reserved

- it can be ordered, i.e. the order in which the elements occur matters,
- it is not meaningful to determine the difference or the ratio between two values.

Example: education level, the order of competitors on a podium, etc.

Note! If a variable is expressed on an ordinal scale, then, for calculations to be performed on it correctly, it should be recorded using numbers. These numbers are conventional identifiers denoting the order of the elements.

3. **Nominal scale** (*ang. nominal scale*)

   A variable is expressed on a nominal scale when:

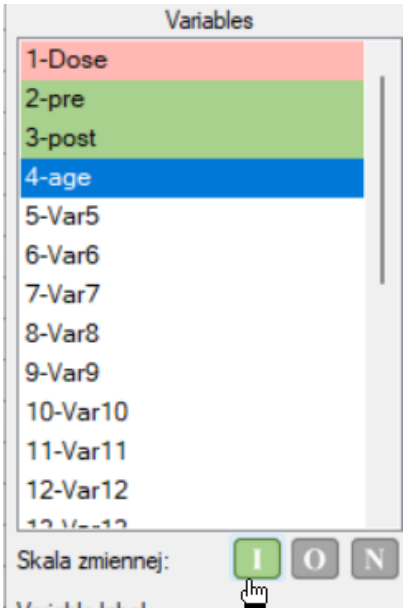
   - it cannot be ordered, i.e. there is no ordering that follows from the nature of the phenomenon,
   - it is not meaningful to determine the difference or the ratio between two values.

   Example: sex, country of residence, etc.

   Note! If a variable is expressed on a nominal scale, it may be recorded using text labels. Even if the values of a nominal variable are expressed numerically, these numbers are only conventional identifiers – so no arithmetic operations may be performed on them, nor may they be compared.

Before starting an analysis, it is recommended to assign a measurement scale to each variable. Doing so will give the variable headers a colour corresponding to that scale, i.e. green = **interval scale**, yellow = **ordinal scale**, red = **nominal scale**. The colour of the variables (and thus their scale) will be visible in the data sheet and in the list of variables in the analysis windows.

A scale can be assigned to a selected variable in the variable options window Codes/Labels/Format or from the context menu on the header of the selected variable.



Quantitative data comprise any information that can be quantified, counted or measured and given a numerical value. Quantitative data include the interval scale and, sometimes, also the ordinal scale.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Qualitative data are descriptive in nature, expressed in linguistic categories rather than numerical values. Qualitative data include the nominal scale and, sometimes, also the ordinal scale.

An ordinal scale that has many objectified categories is quantitative data, e.g. the SF-36 quality-of-life scale (0 to 100 points). If, however, there are so few categories that they can be described in words, e.g. education (primary, secondary, higher), it is qualitative data.

11.2 Tables

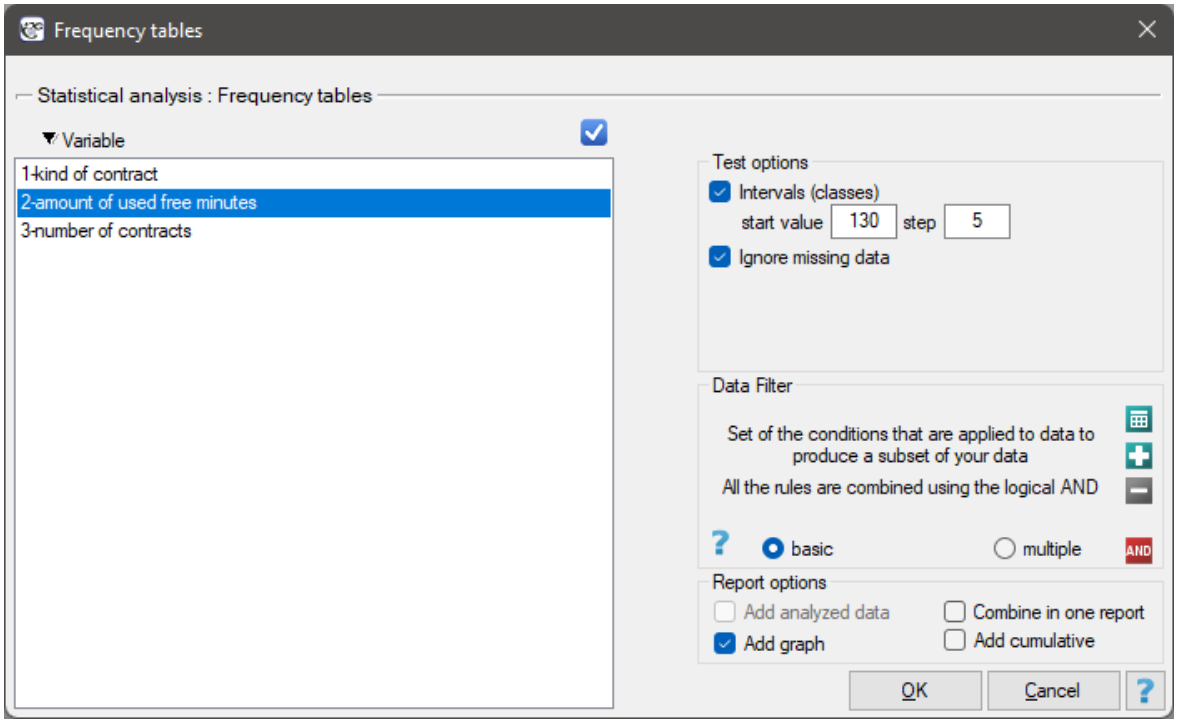
Frequency tables and empirical distributions of data

WHEN TO USE:

The foundation of statistical research is determining the **empirical distribution**, i.e. the distribution of a feature observed in the sample – in other words, checking how often the individual values of the feature recur. Such a distribution is presented in the form of a **frequency table** or a chart (histogram). For small data sets, frequency tables can present all the data – the so-called point frequency distribution; for larger data sets, so-called interval (class) frequency distributions are created instead.

HOW TO SET UP THE ANALYSIS

Statistics → Descriptive analyses → Frequency tables.



In this window we select the variable to analyse and the analysis options. Depending on the options chosen, the returned result can be sorted treating the variable's values as text or as numbers. If there are empty cells in the analysed column, they may be included in the analysis or omitted. The analysis

Copyright ©2010-2026 U-smile Stats – All rights reserved

result appears in the report attached to the data sheet for which it was performed.

If, in addition, we want the data to be illustrated with a column chart or a histogram, we select the Add graph option in the Frequency tables window.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 11.1. (distribution.pqs file)
A mobile network operator carries out a series of surveys on how customers use the number of "free minutes" included in their subscription. Customers can use up to 190 such minutes each month. The survey was based on a random sample of 200 customers. Among other things, the following information was analysed:
– the type of subscription purchased,
– the number of free minutes used,
– the number of subscriptions registered to a given customer (does not apply to companies).

We want to present the distribution of:

1. type of subscription,
2. number of free minutes used,
3. number of subscriptions registered to private individuals.

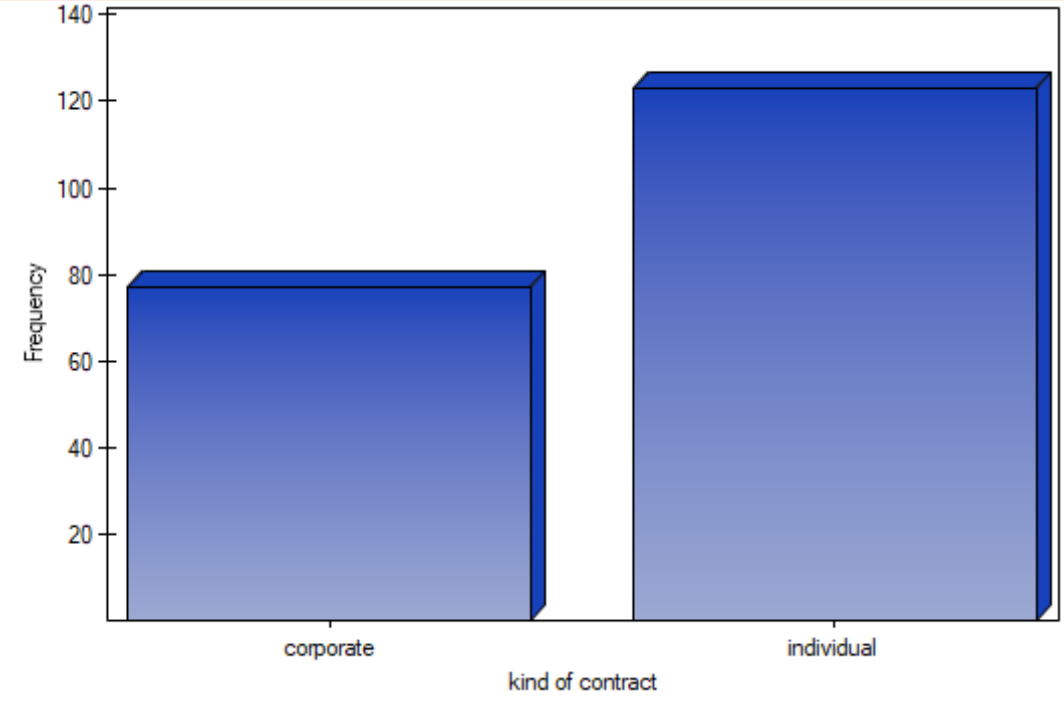
We open the Frequency tables window.

1. We select the Variable to analyse: "type of subscription" and check the Add graph option. We then confirm the selected settings with the OK button and obtain the result as a report:

Frequency tables

Analysed variables	kind of contract	
Number of missing data	0	
Number of deactivated cases	0	
	Frequency	Percent
corporate	77	38.5%
individual	123	61.5%
Total	200	100%

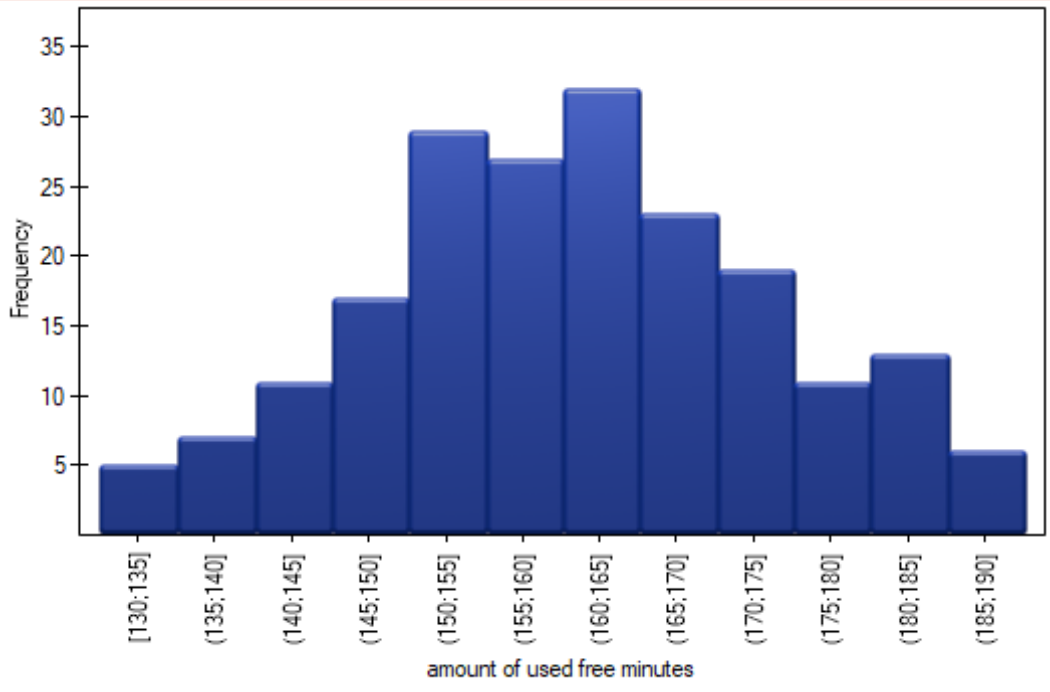
Copyright ©2010-2026 U-smile Stats – All rights reserved



2. We resume the analysis with the Run the recent test button. We select the Variable to analyse: "number of minutes used" and check the Intervals (classes) option, setting the start value to, say, 130 and the step to 5. We may also check the Add graph option. We then confirm the selected options with the OK button and obtain the result as a report:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Frequency tables		
Analysed variables	amount of used	
Number of missing data	0	
Number of deactivated cases	0	
	Frequency	Percent
[130;135]	5	2.5%
(135;140]	7	3.5%
(140;145]	11	5.5%
(145;150]	17	8.5%
(150;155]	29	14.5%
(155;160]	27	13.5%
(160;165]	32	16%
(165;170]	23	11.5%
(170;175]	19	9.5%
(175;180]	11	5.5%
(180;185]	13	6.5%
(185;190]	6	3%
Total	200	100%



3. We resume the analysis with the button. We set a filter so that the analysis is performed only for private individuals. We select the Variable to analyse: "Number of subscriptions". Since this variable also contains missing values, the result obtained may or may not include these missing cases in the analysis, depending on the chosen option for ignoring empty cases:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Frequency tables		
Analysed variables	number of conti	
Data Filter:	kind of contra	
Number of missing data	5	
Number of deactivated cases	0	
	Frequency	Percent
1	94	79.661%
2	12	10.169%
3	11	9.322%
4	1	0.847%
Total	118	100%

Frequency tables		
Analysed variables	number of conti	
Data Filter:	kind of contra	
Number of missing data	5	
Number of deactivated cases	0	
	Frequency	Percent
1	94	76.423%
2	12	9.756%
3	11	8.943%
4	1	0.813%
empty	5	4.065%
Total	123	100%

EXAMPLE 11.2. (fertilisers.pqs file)
An experiment was conducted to study the microbiological condition of soil under perennial ryegrass cultivation fed with biologically active fertilisers. The soils were fertilised with various microbiological preparations and fertilisers, after which the number of microorganisms present per 1 gram of soil dry matter was calculated. We want to know the frequency of occurrence of actinomycetes per 1 gram of dry matter of nitrogen-fertilised soil – we are interested in how often the studied sample contained 0 to 20 actinomycetes, more than 20 to 40, more than 40 to 60, and so on. In the data sheet we select only the first 54 rows, which meet the assumptions of the analysis (these are the actinomycetes fertilised with nitrogen), and open the Frequency tables window.

In the options window we select the variable to analyse: Number of microorganisms, and then set the class intervals so that the start value is 0 and the step is 20. A message should appear at the top of the window: Data restricted by selection. We confirm the choice with the OK button and obtain the result as a report:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Frequency tables		
Analysed variables	Number of micr	
Number of missing data	0	
Number of deactivated cases	0	
	Frequency	Percent
[0;20]	1	1.852%
(20;40]	3	5.556%
(40;60]	6	11.111%
(60;80]	21	38.889%
(80;100]	16	29.63%
(100;120]	4	7.407%
(120;140]	3	5.556%
Total	54	100%

Table report

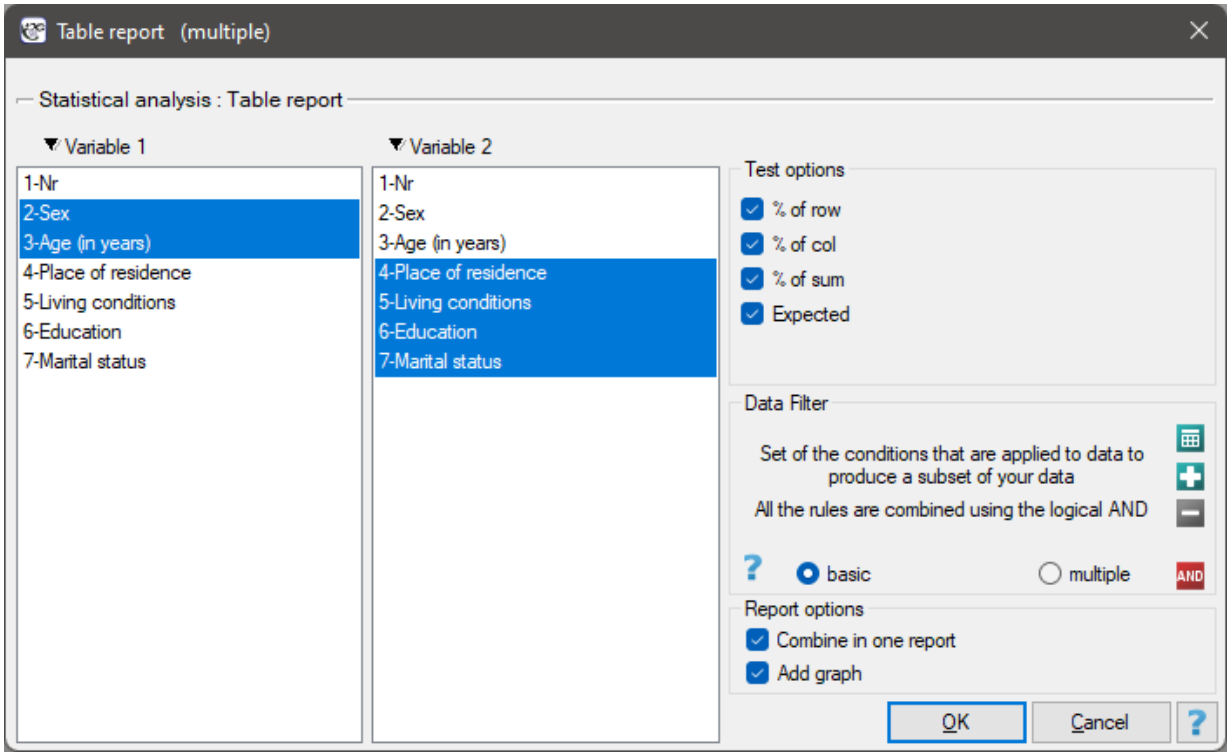
WHEN TO USE:

The table report makes it possible to summarise a large amount of data simultaneously in the form of two-way tables (tables of two features). For example, in this way we can present the distribution of age groups by place of residence, education, and so on. Each table is presented as counts within particular categories and can additionally be summarised by calculating row percentages, column percentages, or percentages of the grand total, and it is also possible to determine the table of expected frequencies. For such tables an automatic summary in the form of a column chart is also available.

HOW TO SET UP THE ANALYSIS

Statistics → Descriptive analyses → Table report.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 11.3. (Tables.pqs file)
We need to summarise, in table form, the distribution of sex by place of residence, social and living conditions, education, and marital status, as well as the distribution of age groups against the same characteristics. As a result we will obtain 4 tables for each pair of features, i.e. 8 tables in total for all the pairs, together with the corresponding charts. Only the breakdown by sex is shown below:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Data			↙Place of resid	
↓Sex	village	small town	big city	Total
man	47	135	24	206
woman	67	180	51	298
Total	114	315	75	504
% of row	village	small town	big city	
man	22.816%	65.534%	11.65%	
woman	22.483%	60.403%	17.114%	
% of col	village	small town	big city	
man	41.228%	42.857%	32%	
woman	58.772%	57.143%	68%	
% of sum	village	small town	big city	
man	9.325%	26.786%	4.762%	
woman	13.294%	35.714%	10.119%	
Expected:	village	small town	big city	
man	46.595238	128.75	30.654762	
woman	67.404762	186.25	44.345238	

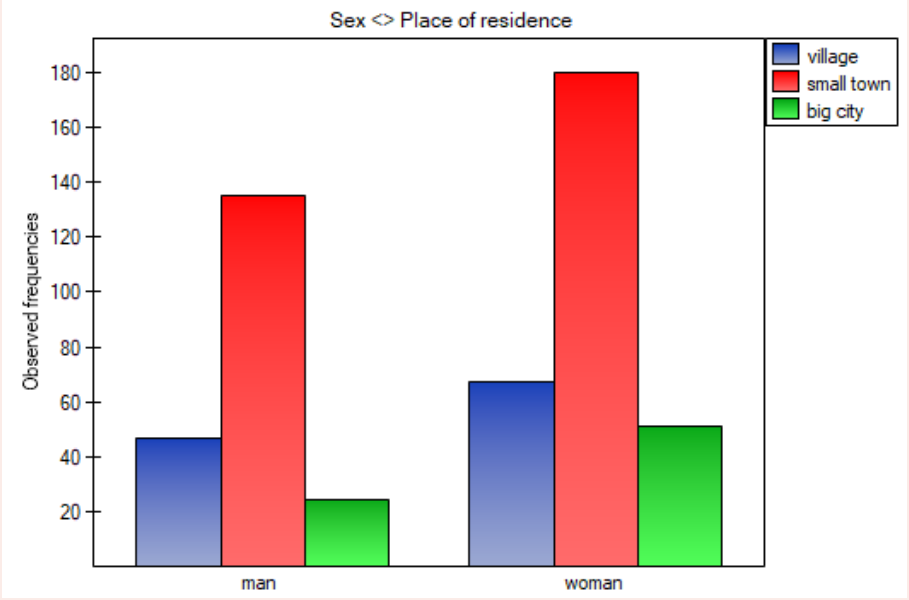
ID2: Data				↙Living cond		
↓Sex	very bad	bad	average	good	very good	Total
man	2	6	59	106	33	206
woman	11	11	84	148	44	298
Total	13	17	143	254	77	504
% of row	very bad	bad	average	good	very good	
man	0.971%	2.913%	28.641%	51.456%	16.019%	
woman	3.691%	3.691%	28.188%	49.664%	14.765%	
% of col	very bad	bad	average	good	very good	
man	15.385%	35.294%	41.259%	41.732%	42.857%	
woman	84.615%	64.706%	58.741%	58.268%	57.143%	
% of sum	very bad	bad	average	good	very good	
man	0.397%	1.19%	11.706%	21.032%	6.548%	
woman	2.183%	2.183%	16.667%	29.365%	8.73%	
Expected:	very bad	bad	average	good	very good	
man	5.3135	6.9484	58.4484	103.8175	31.4722	
woman	7.6865	10.0516	84.5516	150.1825	45.5278	

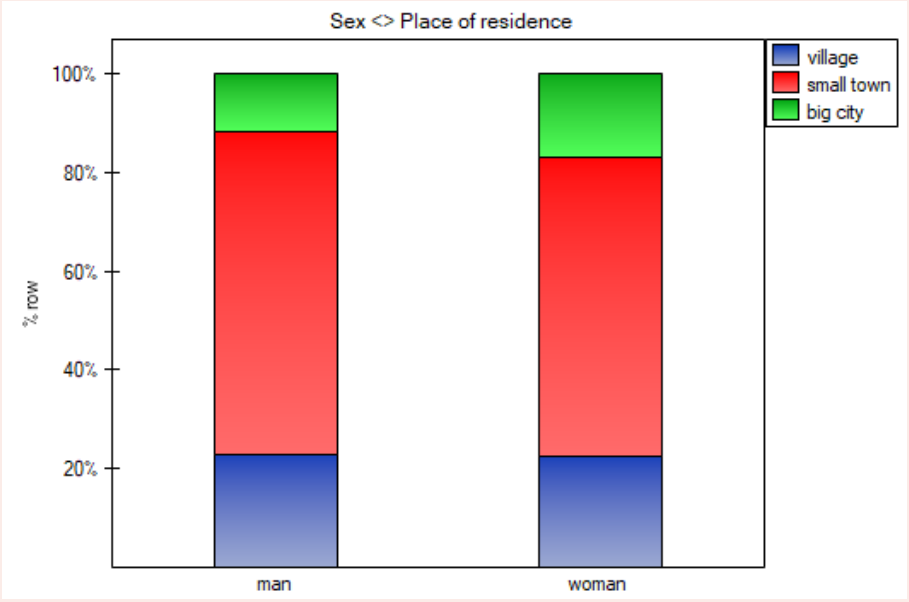
Copyright ©2010-2026 U-smile Stats – All rights reserved

ID3: Data				↙Education	
↓Sex	primary	vocational	secondary	high	Total
man	33	76	58	39	206
woman	90	59	106	43	298
Total	123	135	164	82	504
% of row	primary	vocational	secondary	high	
man	16.019%	36.893%	28.155%	18.932%	
woman	30.201%	19.799%	35.57%	14.43%	
% of col	primary	vocational	secondary	high	
man	26.829%	56.296%	35.366%	47.561%	
woman	73.171%	43.704%	64.634%	52.439%	
% of sum	primary	vocational	secondary	high	
man	6.548%	15.079%	11.508%	7.738%	
woman	17.857%	11.706%	21.032%	8.532%	
Expected:	primary	vocational	secondary	high	
man	50.2738	55.1786	67.0317	33.5159	
woman	72.7262	79.8214	96.9683	48.4841	

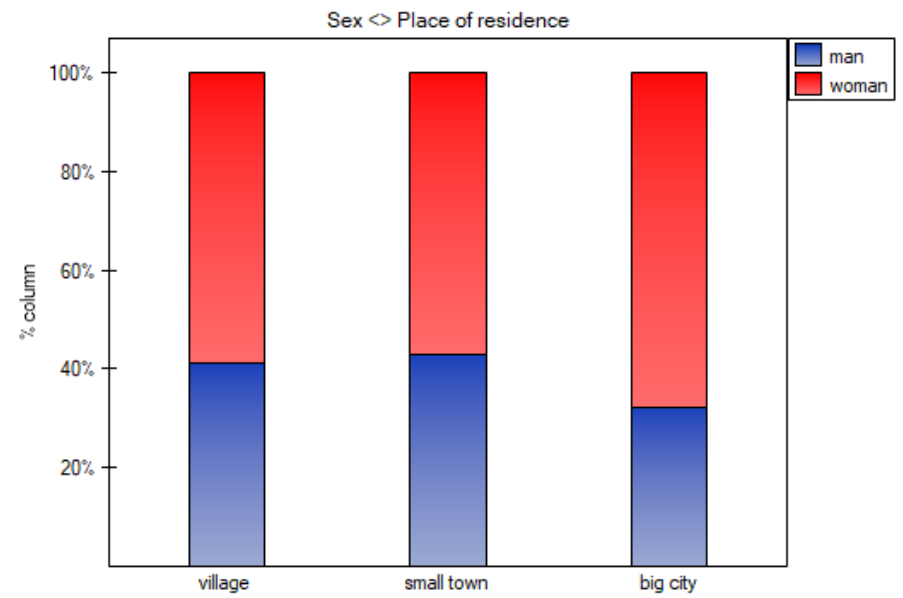
ID4: Data				↙Marital stat	
↓Sex	divorcee	married	single	widow / widc	Total
man	18	131	11	46	206
woman	16	112	24	146	298
Total	34	243	35	192	504
% of row	divorcee	married	single	widow / widc	
man	8.738%	63.592%	5.34%	22.33%	
woman	5.369%	37.584%	8.054%	48.993%	
% of col	divorcee	married	single	widow / widc	
man	52.941%	53.909%	31.429%	23.958%	
woman	47.059%	46.091%	68.571%	76.042%	
% of sum	divorcee	married	single	widow / widc	
man	3.571%	25.992%	2.183%	9.127%	
woman	3.175%	22.222%	4.762%	28.968%	
Expected:	divorcee	married	single	widow / widc	
man	13.8968	99.3214	14.3056	78.4762	
woman	20.1032	143.6786	20.6944	113.5238	

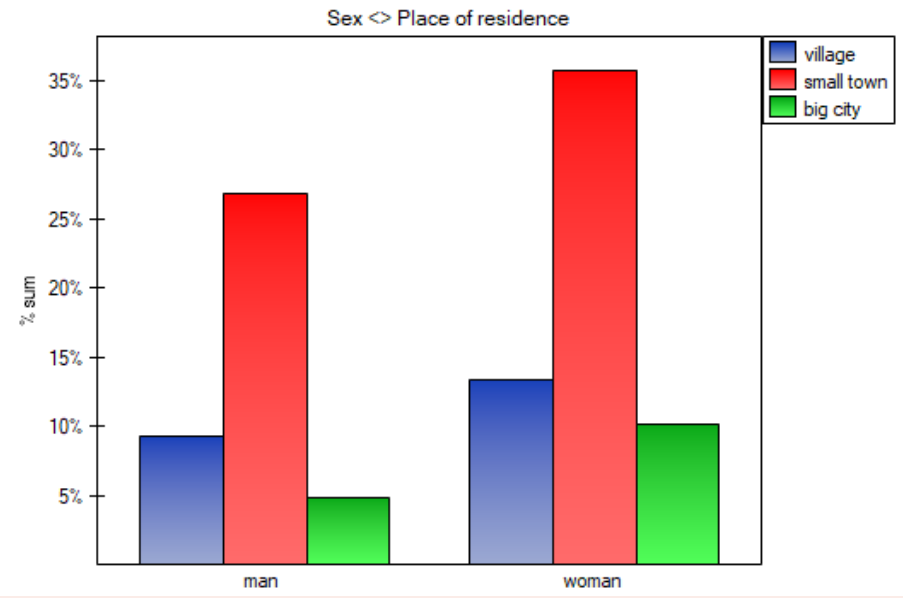
Copyright ©2010-2026 U-smile Stats – All rights reserved





Copyright ©2010-2026 U-smile Stats – All rights reserved





For the breakdown by age group, age categories were first created using codes/labels/format.

Analyses for contingency tables

WHEN TO USE:

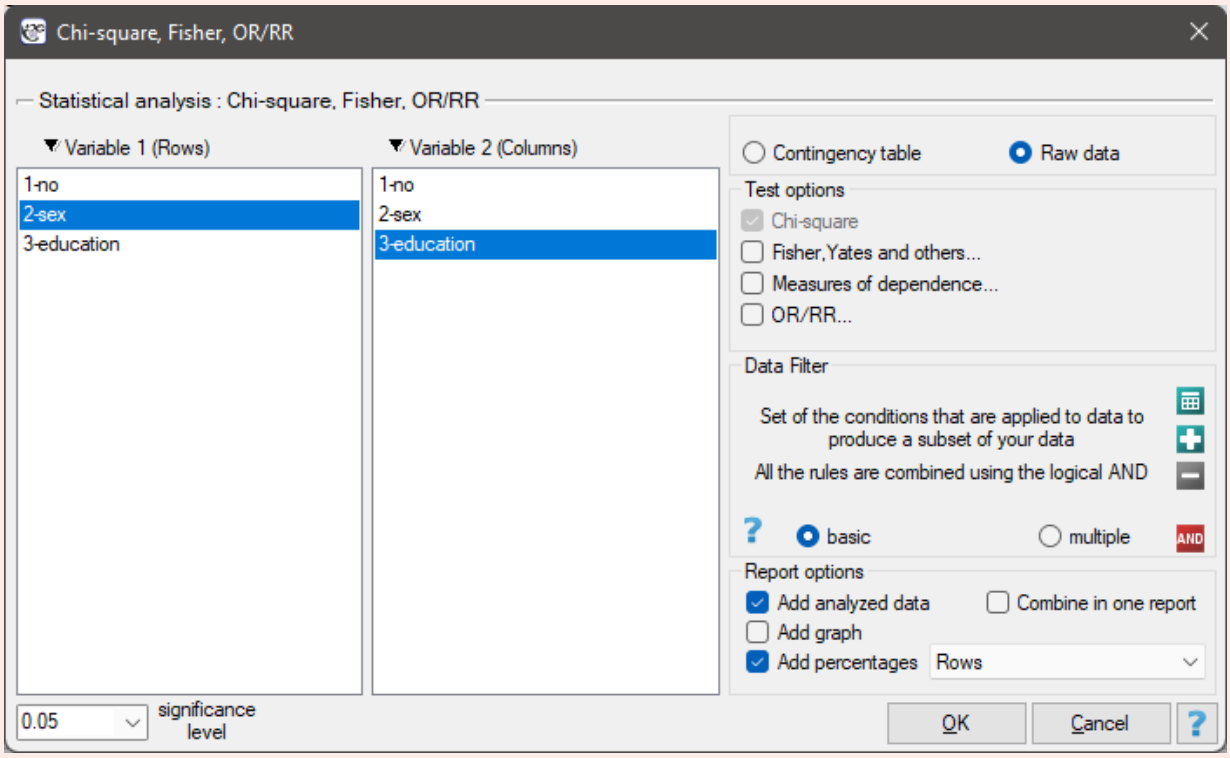
Analyses for contingency tables are used when we want to examine the relationship between two (or more) qualitative features – e.g. whether the type of medication taken is associated with the occurrence of complications. They are calculated on the basis of data collected in a contingency table or directly

Copyright ©2010-2026 U-smile Stats – All rights reserved

from data in raw form. It is also possible to transform data from a contingency table into raw form, or vice versa.

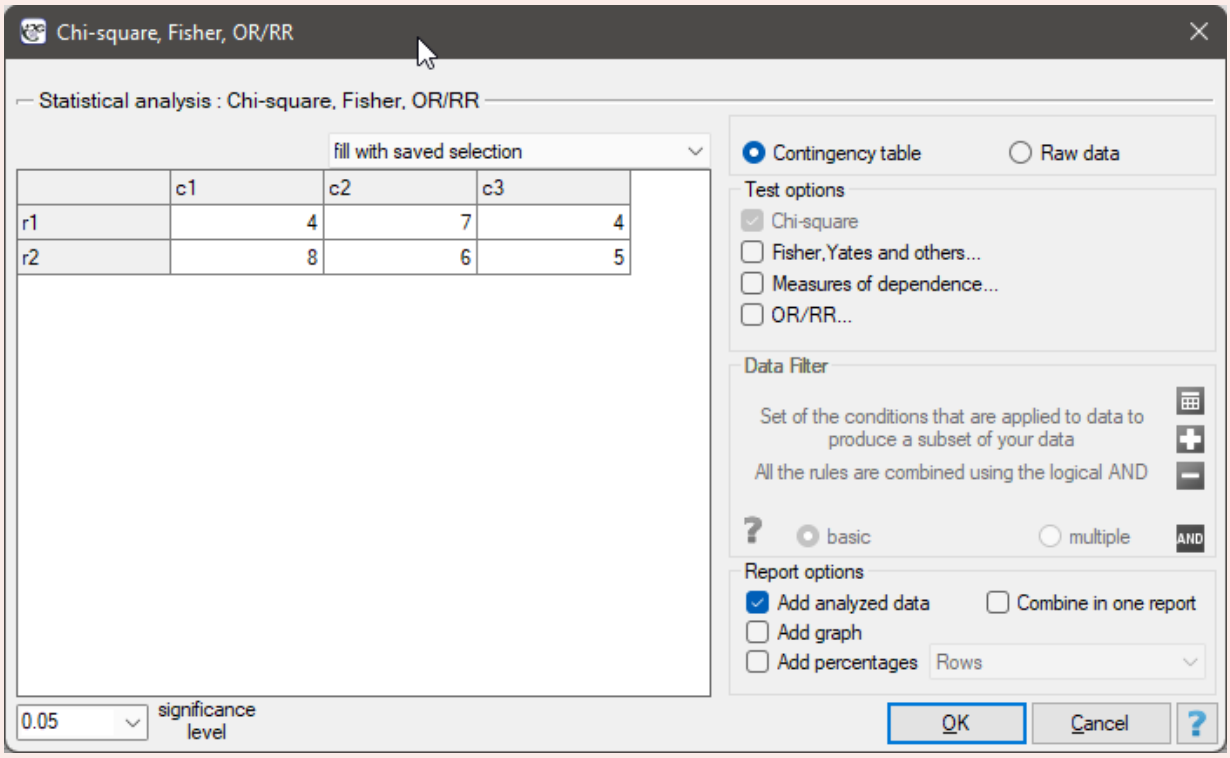
EXAMPLE 11.4. (sex-education.pqs file)
Consider a sample of 34 people ($n = 34$). We study 2 features of these people (X=sex, Y=education). Sex occurs in 2 categories (X_1=female, X_2=male), education in 3 categories (Y_1=primary + vocational, Y_2=secondary, Y_3=higher).

For raw data, when we open the options window of a test, e.g. the χ^2 test for $C \times R$ tables, the raw data option will be selected automatically.



For data already collected in a contingency table, it is a good idea to select that data (the numerical values, without headers) before opening the test window. In that case, once the test window is opened, the contingency table option will be selected automatically, and the selected data will be displayed.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In the test window we can always change the automatically detected setting for how the data are organised, and we can also enter data into the contingency table directly from the window.

Note! Cochran's condition is the basic condition for using many statistical tests based on contingency tables, e.g. the chi-square test. This condition assumes large expected frequencies. According to Cochran's 1952 interpretation[42], none of the expected frequencies may be < 1, and no more than 20% of the expected frequencies may be < 5. Information on whether this condition is satisfied (or not) by the data collected in the table can be returned to the report.

Basic tests for contingency tables:

- Chi-square goodness-of-fit test
- Chi-square test for RxC (2x2) tables and its corrections
- Chi-square trend test for Rx2 tables
- McNemar's test, Bowker's test of internal symmetry
- Chi-square test for multidimensional contingency tables
- Cochran's Q ANOVA
- The Mantel-Haenszel method for multiple 2x2 tables

Coefficients for contingency tables:

- Relative Risk and Odds Ratio

Copyright ©2010-2026 U-smile Stats – All rights reserved

- Contingency coefficients: Yule's Q, Phi, Cramer's V, Pearson's C
- Cohen's Kappa coefficient of agreement
- Fleiss' Kappa coefficient
- Sensitivity and specificity, PPV, NPV, LR(+), LR(-), prevalence, accuracy

DEFINITIONS OF MEASURES

The results report can also include a basic summary of the tables:

- **Contingency table of observed frequencies** – i.e. data in the form of a contingency table. Such a table shows the distribution of observations for several features (several variables). A table for 2 features (X, Y), of which the first has r possible categories and the second c, is shown below (table 11.1).

  Tabela 11.1. The $r \times c$ contingency table of observed frequencies

Frequencies observed O_{ij}		Feature Y				
		Y_1	Y_2	...	Y_c	Total
Feature X	X_1	O_{11}	O_{12}	...	O_{1c}	$\sum_{j=1}^{c} O_{1j}$
	X_2	O_{21}	O_{22}	...	O_{2c}	$\sum_{j=1}^{c} O_{2j}$
	...	...	...	...	...	...
	X_r	O_{r1}	O_{r2}	...	O_{rc}	$\sum_{j=1}^{c} O_{rj}$
	Total	$\sum_{i=1}^{r} O_{i1}$	$\sum_{i=1}^{r} O_{i2}$	...	$\sum_{i=1}^{r} O_{ic}$	$n = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}$

  Observed frequencies O_{ij} ($i = 1, 2, \ldots, r; \quad j = 1, 2, \ldots, c$) show the frequency of occurrence of the individual categories for both features.
  For such a table to be returned by the program, the option Include analysed data must be selected in the test window. For the data from example (11.4), the contingency table of observed frequencies is as follows:

Data			↙education	
↓sex	primary+vocat	secondary	higher	Total
male	8	5	6	19
female	4	4	7	15
Total	12	9	13	34

- **Contingency table of expected frequencies** – for each contingency table of observed frequencies, a corresponding table of **expected frequencies:** E_{ij} can be created (table 11.2) – i.e. the frequencies we would expect in the individual cells of the table if both features were independent of each other (formula in *For those interested*).

  Tabela 11.2. The $r \times c$ contingency table of expected frequencies

Frequencies expected E_{ij}		Feature Y			
		Y_1	Y_2	...	Y_c
Feature X	X_1	E_{11}	E_{12}	...	E_{1c}
	X_2	E_{21}	E_{22}	...	E_{2c}
	...	...	...	...	...
	X_r	E_{r1}	E_{r2}	...	E_{rc}

Copyright ©2010-2026 U-smile Stats – All rights reserved

For the data from example (11.4), the contingency table of expected frequencies is as follows:

Expected:	primary+vocat	secondary	higher
male	6.7059	5.0294	7.2647
female	5.2941	3.9706	5.7353

- **Contingency table of percentages calculated from the column totals**. For the data from example (11.4), this table is as follows:

% of col	primary+vocat	secondary	higher
male	66.667%	55.556%	46.154%
female	33.333%	44.444%	53.846%

- **Contingency table of percentages calculated from the row totals**. For the data from example (11.4), this table is as follows:

% of row	primary+vocat	secondary	higher
male	42.105%	26.316%	31.579%
female	26.667%	26.667%	46.667%

- **Contingency table of percentages calculated from the grand total of rows and columns**. For the data from example (11.4), this table is as follows:

% of sum	primary+vocat	secondary	higher
male	23.529%	14.706%	17.647%
female	11.765%	11.765%	20.588%

11.3 Descriptive statistics

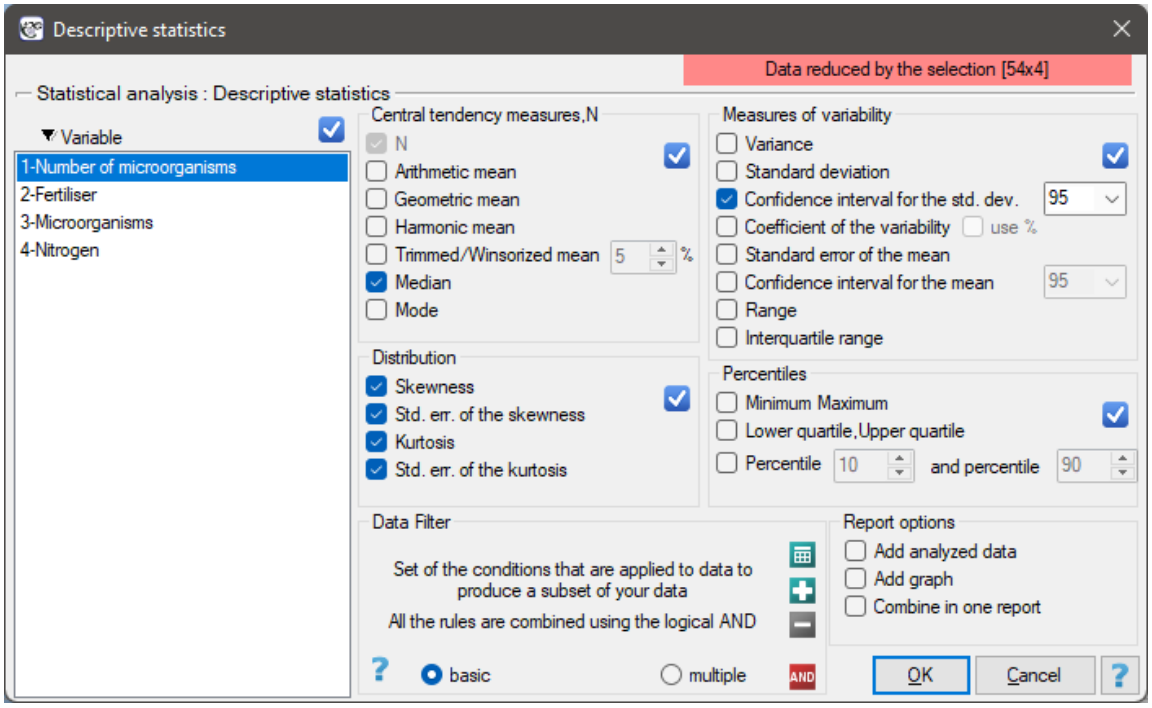
WHEN TO USE:

The purpose of descriptive statistics methods is to summarise a data set using a few numerical characteristics, e.g. the mean value, the median, or the standard deviation, and to draw basic conclusions and generalisations about the data set on that basis – without this step it is difficult even to begin to understand the data we are dealing with.

HOW TO SET UP THE ANALYSIS

Statistics → Descriptive analyses → Descriptive statistics.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In this window we select the variable to analyse and the analysis options, and check the descriptive statistics measures we are interested in. Individual statistics or whole groups of them can be selected using the ☑ button. We confirm the selection with the OK button. The analysis result appears in the report attached to the data sheet for which it was performed.

If, in addition, we want the data to be illustrated with a box-and-whisker plot, we check the Add graph option in the Descriptive statistics window.

Measures of central tendency

Measures of central tendency are so-called average measures, characterising the mean or typical level of the values of a feature.

DEFINITIONS OF MEASURES

The arithmetic mean (*ang. arithmetic mean*) is the best known of the average measures – the sum of all the values divided by their number (formula in *For those interested*). It is used for the interval scale. For a sample it is denoted by $\overline{x}$, and for a population by μ.

The trimmed mean – computed as the arithmetic mean after removing a given percentage of the smallest and largest measurements from the sample, e.g. trimming 5% of the measurements means removing 2.5% of the largest and 2.5% of the smallest values. If the number of measurements to be removed, obtained from this calculation, is not a whole number, it is rounded down to the nearest integer.

The Winsorised mean – computed as the arithmetic mean after replacing the corresponding percentage of extreme measurements with the smallest and largest value remaining in the reduced set of values. If we decide to compute the Winsorised mean by trimming, say, 5% of the measurements, the discarded 5% is replaced by the smallest and largest value determined from the remaining 95% of the measurements. As with the trimmed mean, if converting the percentage of values to be replaced does

Copyright ©2010-2026 U-smile Stats – All rights reserved

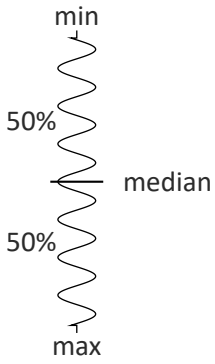
not give a whole number, it is rounded down to the nearest integer.

The geometric mean (*ang. geometric mean*) is the nth root of the product of all the values (formula in *For those interested*). It is used for the interval scale when the variable has a log-normal distribution (the logarithm of the variable has a normal distribution).

The harmonic mean (*ang. harmonic mean*) is the number of observations divided by the sum of the reciprocals of the individual values (formula in *For those interested*). It is used for the interval scale.

The median (*ang. median*)

In an ordered data set, the median is the value that divides the set into two equal parts. Half of all the observations lie below, and half above, the median.



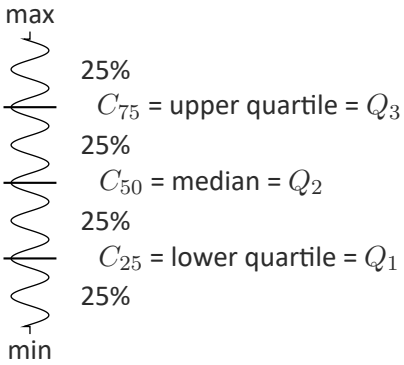
The median can be used for the interval scale and the ordinal scale.

The mode (*ang. mode*)

The mode – is the value that occurs most frequently among the measurements obtained. The mode can be used for any measurement scale.

Other measures of location

Quartiles *(ang. quartiles)*, **deciles** *(ang. deciles)*, **centiles** *(ang. centiles)* are values that divide an ordered data set into equal parts – similarly to the median, but into more than two parts.



Quartiles (Q_1, Q_2, Q_3) divide an ordered series into 4 equal parts, deciles (D_i, $i = 1, 2, ..., 9$) into 10 equal parts, and centiles (percentiles: C_i, $i = 1, 2, ..., 99$) into 100 equal parts. The second quartile, the fifth decile and the fiftieth centile are all equal to the median. These measures can be used for the interval scale and the ordinal scale.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Measures of dispersion

Knowledge of measures of central tendency is not enough to characterise the structure of a statistical collection – the group under study may show varying degrees of variability in the feature examined. Measures are therefore needed that allow us to determine the spread of the data.

ASSUMPTIONS

- measurement on the interval scale – measures of dispersion are based on distances between points, so they are not computed for the ordinal or nominal scale.

DEFINITIONS OF MEASURES

The range *(ang. range)* – the difference between the largest and smallest value in the sample (formula in *For those interested*). **The interquartile range (**IQR**)** is the analogous difference between the upper and lower quartile.

Ranges for the percentile (decile, centile) scale. Ranges between percentiles are one of the measures of dispersion – they indicate the percentage of all observations whose value lies between the selected percentiles.

Variance *(ang. variance)* — measures the degree of dispersion of the measurements around the arithmetic mean. It is calculated slightly differently for a sample and for an entire population (formulas in *For those interested*). Variance is always positive, but it is not expressed in the same units as the measurements.

Standard deviation *(ang. standard deviation)* — like variance, measures the degree of dispersion of the measurements around the arithmetic mean, but – unlike variance – is expressed in the same units as the data (it is the square root of the variance). The higher the value of the standard deviation (or the variance), the more diverse the group is with respect to the feature under study.

Note! The sample standard deviation is an approximation (estimator) of the population standard deviation. The population value of the standard deviation lies within a certain interval containing the sample standard deviation. This interval is called the **confidence interval** *(confidence interval)* for the standard deviation. **The coefficient of variation** *(ang. coefficient of variation)*

The coefficient of variation, like the standard deviation, allows us to assess the degree of homogeneity of the group under study (formula in *For those interested*). It is a dimensionless quantity, which makes it possible to compare the variability of several groups with respect to the same feature, and of the same group with respect to several different features (expressed in different units). It is generally assumed that if the coefficient V does not exceed 10%, the features show statistically insignificant variability.

Standard errors *(ang. standard errors)* — these are not measures of dispersion of the measurement results, but rather indicate the degree of precision with which we can determine the values of population parameters from their estimators calculated from the sample. The most commonly used standard error is the **standard error of the mean** *(ang. standard error of the mean, SEM)* – formula in *For those interested*.

Note! The standard error of a sample estimator can be used to determine the confidence interval for a

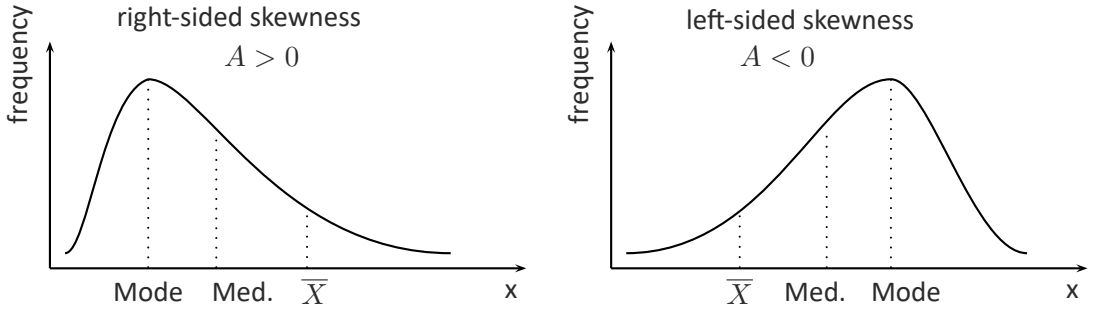
Copyright ©2010-2026 U-smile Stats – All rights reserved

population parameter.

Other distribution attributes

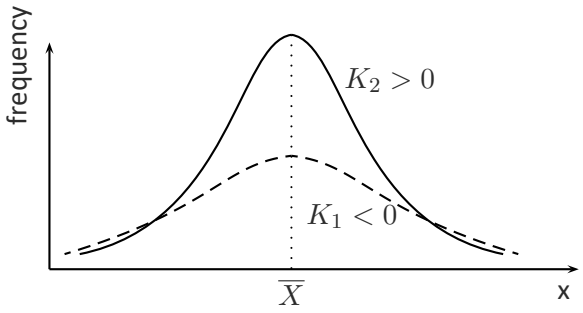
Skewness, also called the coefficient of asymmetry *(ang. skewness)*

This measure indicates how much the distribution of the data differs from a symmetric distribution. The closer the value of the asymmetry coefficient is to zero, the more symmetrically the data are distributed around the mean. This coefficient usually lies within the range [-1, 1], although for particularly high asymmetry it may fall outside this range. Positive values indicate right-sided skewness (a longer right "tail"), while negative values indicate left-sided skewness (a longer left "tail"). The formula for skewness can be found in *For those interested*.



Kurtosis, also called the coefficient of concentration *(ang. kurtosis)*

This measure indicates how closely the spread of the data around the mean resembles the spread of data in a normal distribution. The higher the value of kurtosis above zero, the more sharply peaked the studied distribution is compared with the normal distribution; the lower the value below zero, the flatter the distribution is compared with the normal distribution. The formula for kurtosis can be found in *For those interested*.



PERFORMING AND INTERPRETING THE ANALYSIS

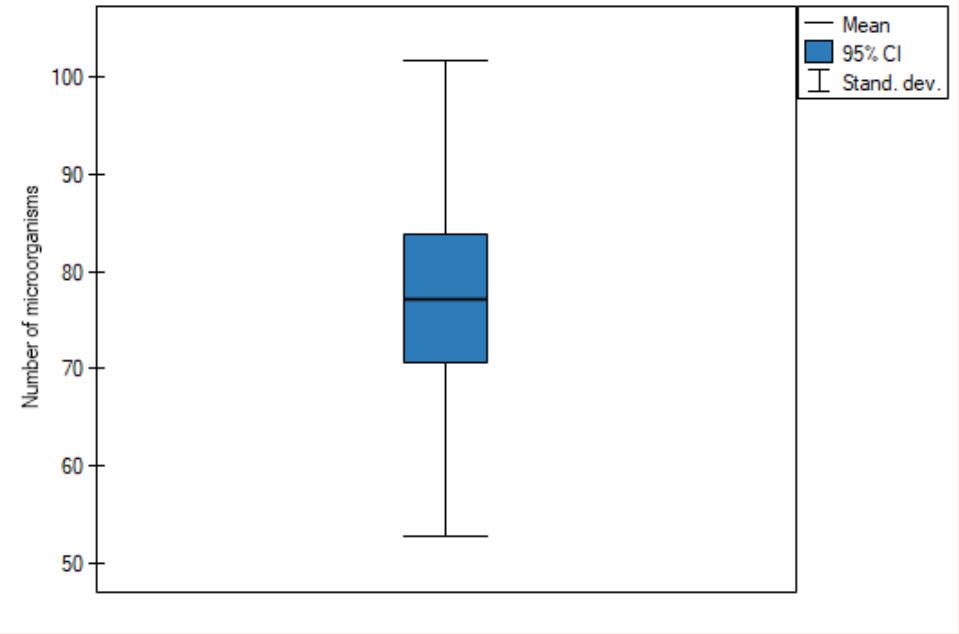
Example 11.5. (fertilisers.pqs file)
In an experiment on fertilising soil with various microbiological preparations and fertilisers, the number of microorganisms present per 1 gram of soil dry matter was calculated. We want to determine descriptive statistics for the number of actinomycetes in the sample fertilised with nitrogen, and to illustrate

Copyright ©2010-2026 U-smile Stats – All rights reserved

the results with a box-and-whisker plot. In the data sheet we select only the first 54 rows, which meet the assumptions of the analysis (these are the actinomycetes fertilised with nitrogen), and open the Descriptive statistics window.

In the descriptive statistics options window we select the variable to analyse: Number of microorganisms, and then the procedures we want to perform (e.g. the arithmetic mean together with its confidence interval, the median, the standard deviation together with its confidence interval, and information on the skewness and kurtosis of the distribution together with their errors). A message should appear at the top of the window: Data restricted by selection. To have a chart included in the report as well, we check the Add graph option and choose the box-and-whisker plot type we are interested in. We confirm the choice with the OK button and obtain the result as a report:

Descriptive statistics	
Analysed variables	Number of microorganisms
Group size	54
Number of unspecified	0
Number of missing data	0
Arithmetic mean	77.240741
Median	78.5
Standard deviation	24.425424
-95% CI for standard deviation	20.532603
+95% CI for standard deviation	30.153531
-95% CI for the group mean	70.573884
+95% CI for the group mean	83.907598
Skewness	-0.226875
Std. err. of the skewness	0.324556
Kurtosis	0.343163
Std. err. of the kurtosis	0.638893



Copyright ©2010-2026 U-smile Stats – All rights reserved

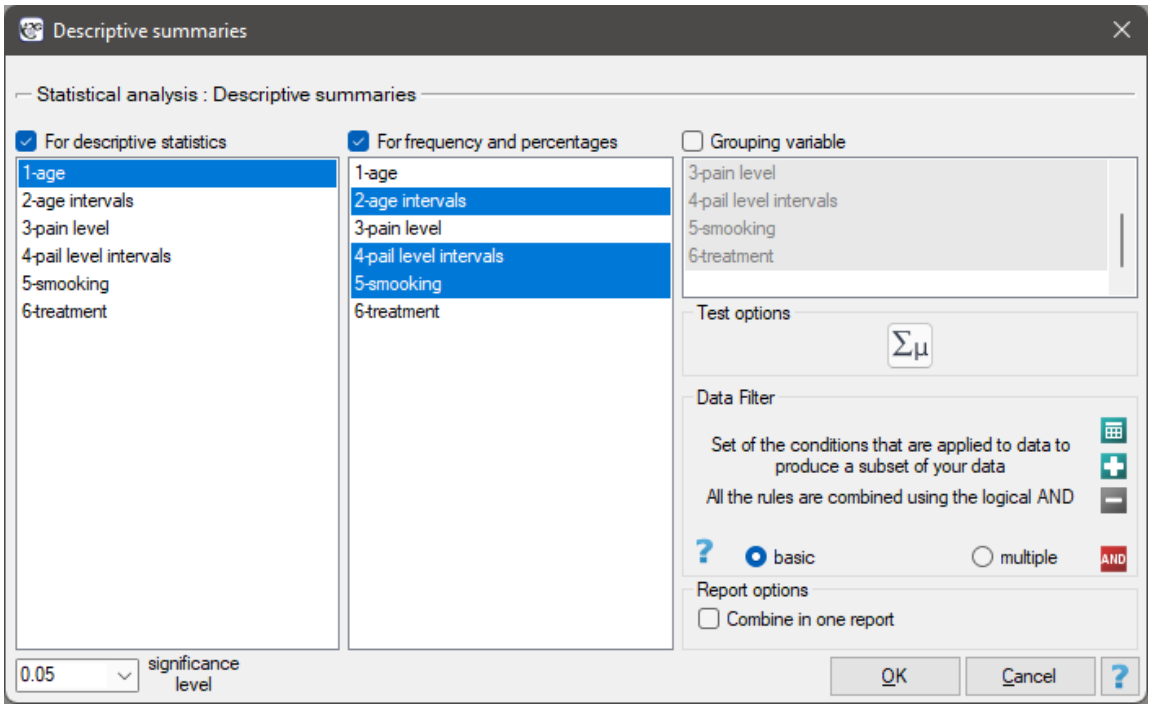
11.4 Descriptive summaries

WHEN TO USE:

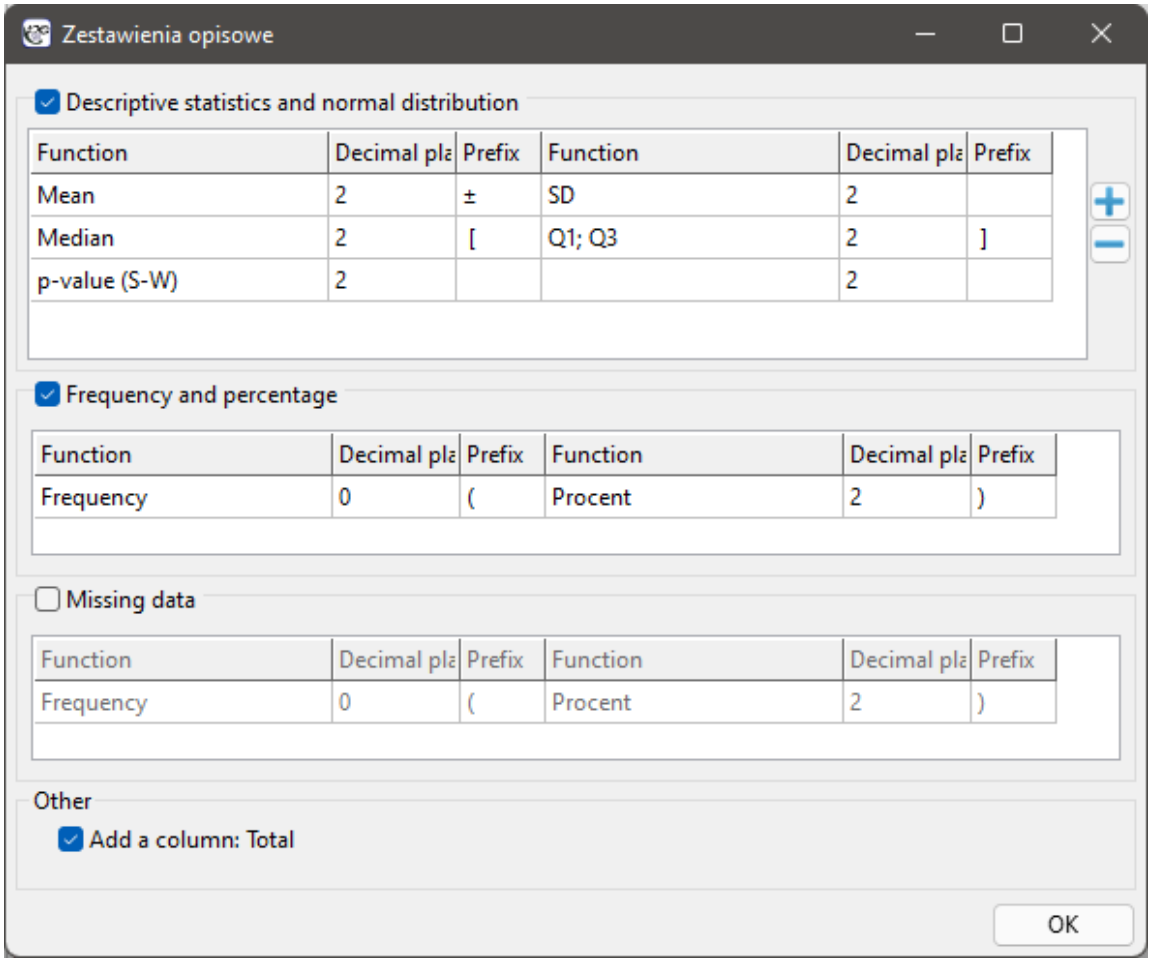
Descriptive summaries provide a quick way to prepare a report presenting a description of the data, ready to be inserted directly into a research paper. It is a tool that makes it easier to gather all the basic forms of data description in one place.

HOW TO SET UP THE ANALYSIS

Statistics → Descriptive analyses → Descriptive summaries.



Copyright ©2010-2026 U-smile Stats – All rights reserved



Depending on how the measurements were made, one of three types of measure is most often used to describe a variable:

- mean $\pm$ standard deviation, abbreviated $\overline{x} \pm sd$
- median [lower quartile; upper quartile], abbreviated $Me[Q_1; Q_3]$, or median [min; max], abbreviated $Me[min; max]$
- count (percentage) of individual categories, abbreviated $n(\%)$

Depending on the needs and whether additional assumptions are met:

- on the interval scale, data can be described using any measure – researchers are also often interested in checking whether such data are normally distributed,
- on the ordinal scale, we can use medians (with quartiles or the smallest and largest value) or the counts and percentages of individual categories,
- on the nominal scale, only the counts and percentages of individual categories are available.

PERFORMING AND INTERPRETING THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

EXAMPLE 11.6. (Descriptive summaries.pqs file)
An example way of describing data is shown by the resulting report summarising age, perceived pain level, and smoking status. These data are summarised in the first table, describing all 100 people, and in the second table, broken down by treatment method.

Descriptive summaries		
		(N=100)
age		
	Mean ± SD	66.38±16.31
	Median [Q1; Q3]	64.5 [55.75; 76]
	p-value (S-W)	0.38
age intervals		
	<=39	3(3%)
	40-49	12(12%)
	50-59	23(23%)
	60-69	22(22%)
	>=70	40(40%)
pail level intervals		
	0-3	20(20%)
	4-6	31(31%)
	7-10	49(49%)
smooking		
	no	73(73%)
	yes	27(27%)

Copyright ©2010-2026 U-smile Stats – All rights reserved

Descriptive summaries			
		treatment A(N=44)	B(N=56)
age			
	Mean ± SD	66.27±17.7	66.46±15.29
	Median [Q1; Q3]	64 [56; 74.5]	65.5 [54.75; 76.25
	p-value (S-W)	0.59	0.49
age intervals			
	<=39	2(4.55%)	1(1.79%)
	40-49	5(11.36%)	7(12.5%)
	50-59	9(20.45%)	14(25%)
	60-69	11(25%)	11(19.64%)
	>=70	17(38.64%)	23(41.07%)
pail level intervals			
	0-3	13(29.55%)	7(12.5%)
	4-6	15(34.09%)	16(28.57%)
	7-10	16(36.36%)	33(58.93%)
smooking			
	no	33(75%)	40(71.43%)
	yes	11(25%)	16(28.57%)

FOR THOSE INTERESTED

Formulas

Contingency table of expected frequencies

$$E_{11} = \frac{\sum_{i=1}^{r} O_{i1} \times \sum_{j=1}^{c} O_{1j}}{n}, \quad E_{12} = \frac{\sum_{i=1}^{r} O_{i2} \times \sum_{j=1}^{c} O_{1j}}{n}, \quad E_{1c} = \frac{\sum_{i=1}^{r} O_{ic} \times \sum_{j=1}^{c} O_{1j}}{n}$$

$$E_{21} = \frac{\sum_{i=1}^{r} O_{i1} \times \sum_{j=1}^{c} O_{2j}}{n}, \quad E_{22} = \frac{\sum_{i=1}^{r} O_{i2} \times \sum_{j=1}^{c} O_{2j}}{n}, \quad E_{2c} = \frac{\sum_{i=1}^{r} O_{ic} \times \sum_{j=1}^{c} O_{2j}}{n}$$

$$E_{r1} = \frac{\sum_{i=1}^{r} O_{i1} \times \sum_{j=1}^{c} O_{rj}}{n}, \quad E_{r2} = \frac{\sum_{i=1}^{r} O_{i2} \times \sum_{j=1}^{c} O_{rj}}{n}, \quad E_{rc} = \frac{\sum_{i=1}^{r} O_{ic} \times \sum_{j=1}^{c} O_{rj}}{n}.$$

The arithmetic mean

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

where x_i are the successive values of the variable, and n is the sample size.

The geometric mean

$$\overline{x}_G = \sqrt[n]{x_1 x_2 \ldots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

The harmonic mean

$$\overline{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

The range

$$I = \max x_i - \min x_i,$$

where x_i are the values of the variable being studied.

$$IQR = \text{interquartile range} = Q_3 - Q_1,$$

where Q_1, Q_3 are the lower and upper quartile.

Variance *(ang. variance)*

sample variance:

$$sd^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1},$$

where x_i are the successive values of the variable, $\overline{x}$ is their arithmetic mean, and n is the sample size.

population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N},$$

where x_i are the successive values of the variable, μ is their arithmetic mean, and N is the population size.

Standard deviation *(ang. standard deviation)*

sample standard deviation:

$$sd = \sqrt{sd^2},$$

population standard deviation:

$$\sigma = \sqrt{\sigma^2}.$$

The coefficient of variation

$$V = \frac{sd}{\overline{x}}100\%,$$

where sd is the standard deviation and $\overline{x}$ is the arithmetic mean.

The standard error of the mean *(ang. standard error of the mean)*

$$SEM = \text{standard error of the mean} = \frac{sd}{\sqrt{n}}.$$

Skewness

$$A = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{sd}\right)^3,$$

where:
x_i – the successive values of the variable,
$\overline{x}$, sd – the arithmetic mean and standard deviation of x_i, respectively,
n – the sample size.

Kurtosis

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{sd}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)},$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where:
$x_i -$ the successive values of the variable,
$\overline{x}$, $sd -$ the arithmetic mean and standard deviation of x_i, respectively,
$n -$ the sample size.

Copyright ©2010-2026 U-smile Stats – All rights reserved

12 PROBABILITY DISTRIBUTIONS

WHAT PROBABILITY DISTRIBUTIONS ARE

The real distribution of the data from a sample – the **empirical data distribution** – can be presented by means of Frequency tables (by selecting the Statistics→Descriptive analyses→Frequency tables menu). For example, the distribution of the number of free minutes used by the subscribers of a certain mobile network operator (*example (11.1), distribution.pqs file*) is presented by the following table:

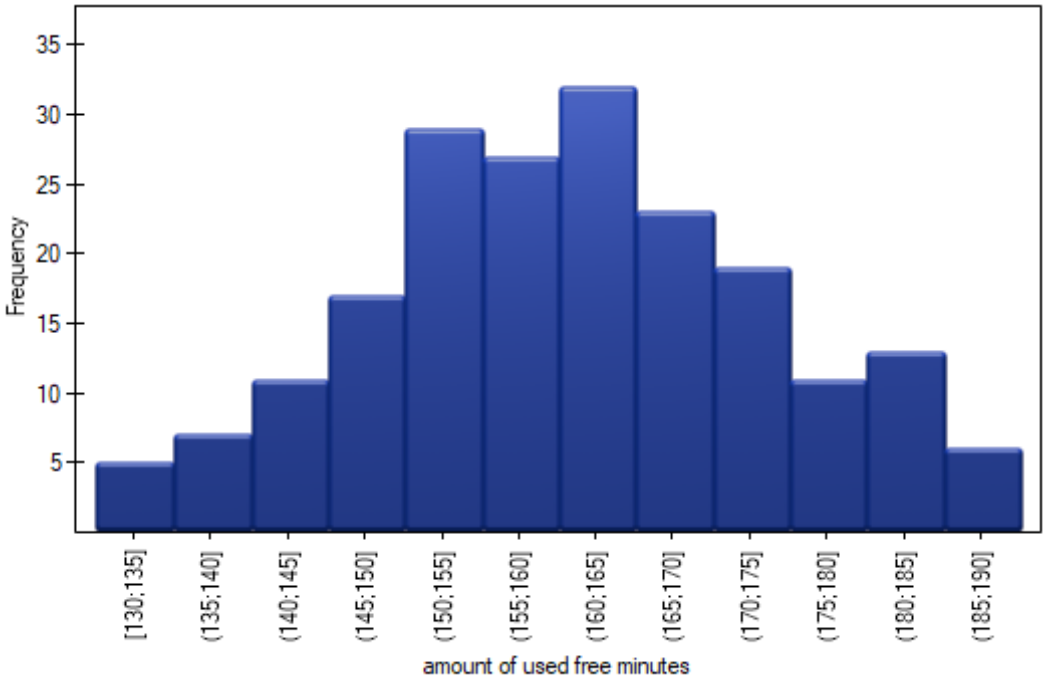
Frequency tables

Analysed variables	amount of used
Number of missing data	0
Number of deactivated cases	0

	Frequency	Percent
[130;135]	5	2.5%
(135;140]	7	3.5%
(140;145]	11	5.5%
(145;150]	17	8.5%
(150;155]	29	14.5%
(155;160]	27	13.5%
(160;165]	32	16%
(165;170]	23	11.5%
(170;175]	19	9.5%
(175;180]	11	5.5%
(180;185]	13	6.5%
(185;190]	6	3%
Total	200	100%

A graphical presentation of the results obtained in the table report is usually made by means of a histogram or a bar chart.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Such a chart can be obtained by ticking the Add chart option in the Frequency tables window.

The **theoretical distribution of the data**, also called the **probability distribution**, is usually presented graphically by means of a line chart. Such a line is described by a function (a mathematical model) called the **density function of the distribution**. The empirical distribution can be replaced by a suitable theoretical distribution – this allows the data to be described by a few parameters instead of listing all the observed values.

Note! To replace an empirical distribution with a theoretical distribution, an intuitive statement of similarity of their shapes is not enough. Specially constructed goodness-of-fit tests are used for this purpose.

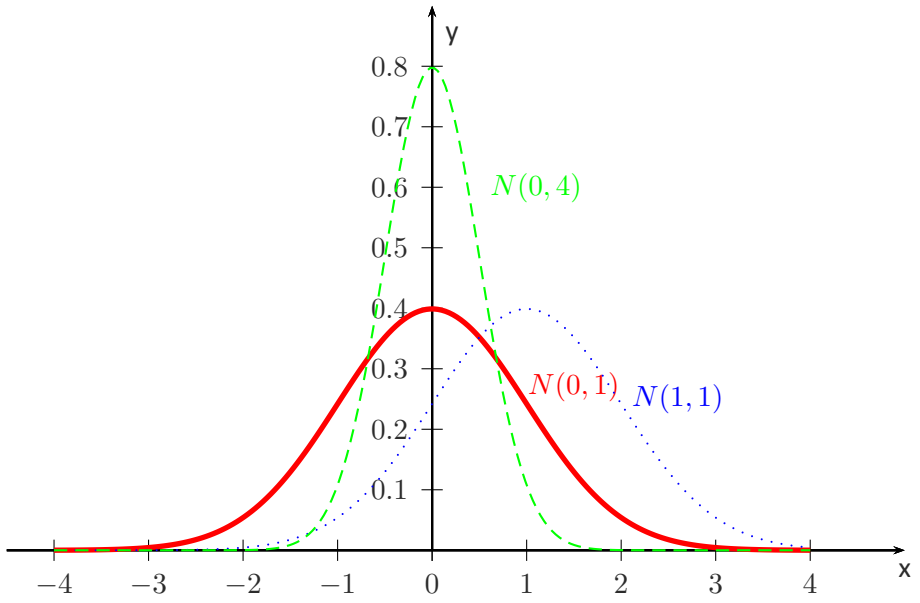
The most frequently used probability distribution is the normal distribution (the Gaussian distribution) – and it is exactly such a distribution, with a mean of 161.15 and a standard deviation of 13.03, that is followed by the data on the number of free minutes used (*example (11.1), distribution.pqs file*).

Copyright ©2010-2026 U-smile Stats – All rights reserved

12.1 CONTINUOUS PROBABILITY DISTRIBUTIONS

NORMAL DISTRIBUTION (GAUSSIAN DISTRIBUTION)

Also called the bell curve, this is one of the most important distributions in statistics – it has interesting mathematical properties and very often describes phenomena occurring in nature. It is usually denoted by the symbol $N(\mu, \sigma)$, where μ is the expected value (the mean in the population) and σ is the standard deviation – it is exactly these two numbers that fully determine the shape and location of the curve.



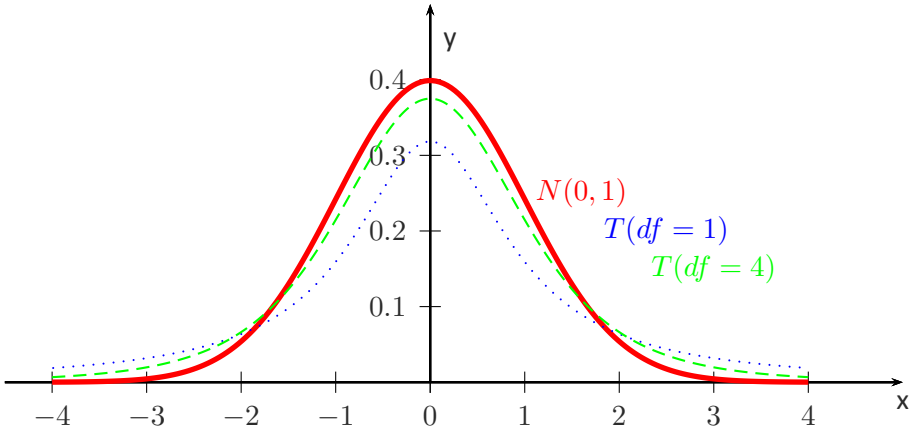
The normal distribution is a distribution symmetrical about a line perpendicular to the x-axis, passing through the point that simultaneously marks the mean, the mode and the median.

The normal distribution with a mean $\mu = 0$ and standard deviation $\sigma = 1$ ($N(0, 1)$) is the so-called **standardised normal distribution**.

THE T-STUDENT DISTRIBUTION

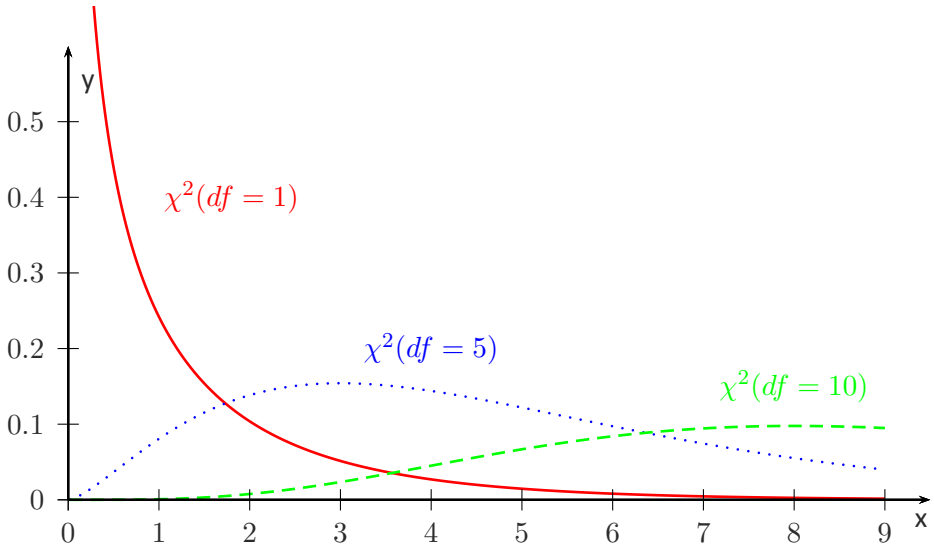
This has a shape similar to the standardised normal distribution, but has longer "tails" – meaning that extreme values occur in it slightly more often. Its shape depends on the number of degrees of freedom df: the more degrees of freedom (i.e. usually the larger the sample), the more the t-Student distribution resembles the normal distribution.

Copyright ©2010-2026 U-smile Stats – All rights reserved



THE CHI-SQUARE DISTRIBUTION (χ^2)

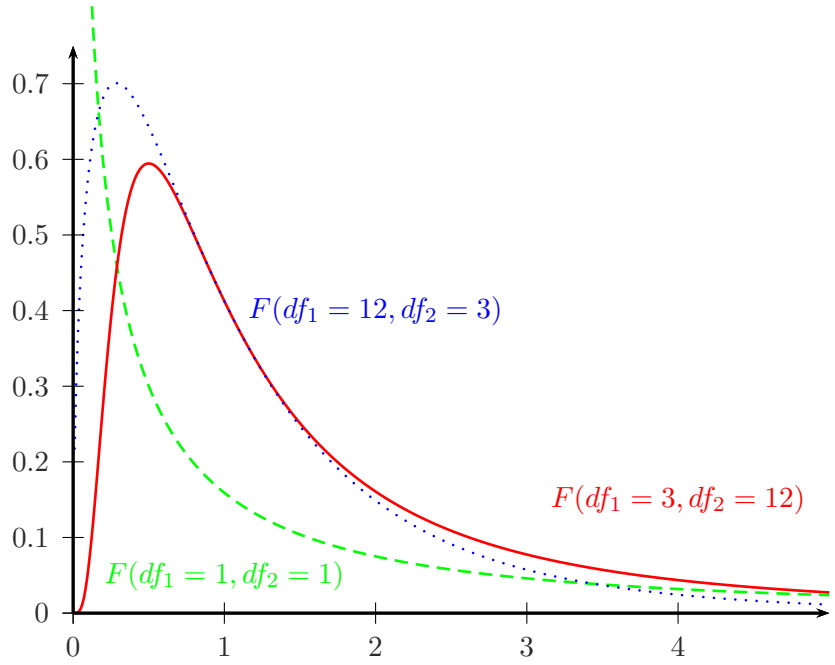
This is a right-skewed distribution (with a longer "tail" on the right side), whose shape depends on the number of degrees of freedom df. As with the t-Student distribution, the shape of the χ^2 distribution comes to resemble the shape of the normal distribution as the number of degrees of freedom increases.



THE FISHER-SNEDECOR (F) DISTRIBUTION

This has a longer right "tail", and its shape depends on two numbers of degrees of freedom: df_1 and df_2. This distribution arises as the ratio of two independent variables with a χ^2 distribution (each divided by its number of degrees of freedom), and it appears in tests comparing variances or in the analysis of variance (ANOVA).

Copyright ©2010-2026 U-smile Stats – All rights reserved



Copyright ©2010-2026 U-smile Stats – All rights reserved

12.2 PROBABILITY DISTRIBUTION CALCULATOR

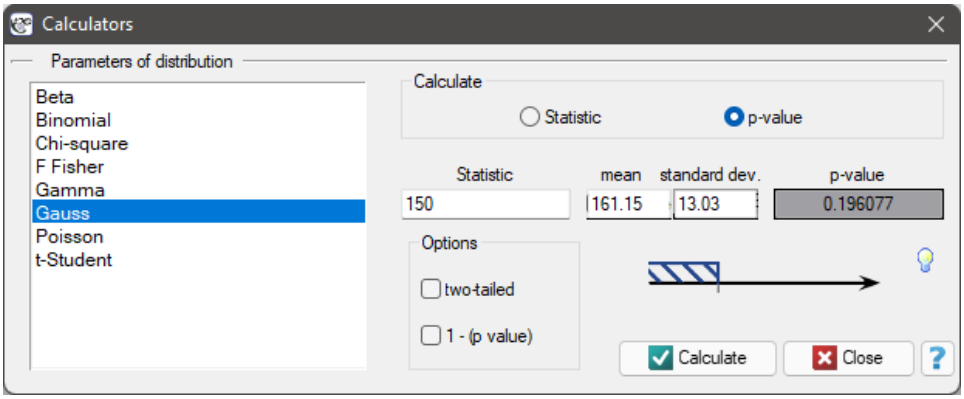
WHAT IT IS FOR



The area under the curve (the density function of the distribution) is the **probability** p of occurrence of all possible values of the studied random variable. The whole area under the curve equals $p = 1$. When we want to examine the size of only part of that area, we must give a boundary value called the **critical value** or the Statistic. For this purpose we use the Probability distribution calculator window. In this window, the value of the area under the curve (p value) of a given distribution can be calculated on the basis of the Statistic, and likewise the value of the Statistic can be determined on the basis of the p value.

HOW TO OPEN THE CALCULATOR

The Probability distribution calculator window is opened by selecting the Statistics→Calculators→Probability distribution calculator menu.



EXAMPLE

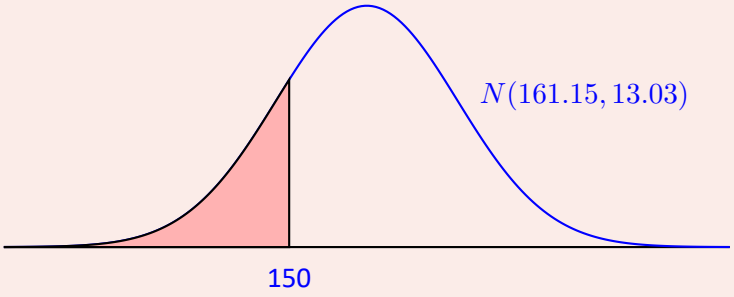
EXAMPLE 12.1. Probability distribution calculator
A certain mobile network operator carries out a series of studies on how customers use the number of "free minutes" granted under their subscription. On the basis of a sample of 200 of its customers (in which the distribution of the free minutes used takes the shape of a normal distribution), it determined the mean value $\overline{x} = 161.15\ min.$ and standard deviation $sd = 13.03\ min.$ We want to calculate the probability that a randomly drawn customer used:

1. 150 minutes or less,
2. more than 150 minutes,
3. a number of minutes from the interval $[\overline{x} - sd, \overline{x} + sd] = [148.12\ min., 174.18\ min.]$,
4. a number of minutes outside the interval $\overline{x} \pm sd$.

We open the Probability distribution calculator window, select the Gaussian distribution, and enter the mean $\overline{x} = 161.15\ min.$ and standard dev. $sd = 13.03\ min.$, also indicating that we will be calculating the p value.

Copyright ©2010-2026 U-smile Stats – All rights reserved

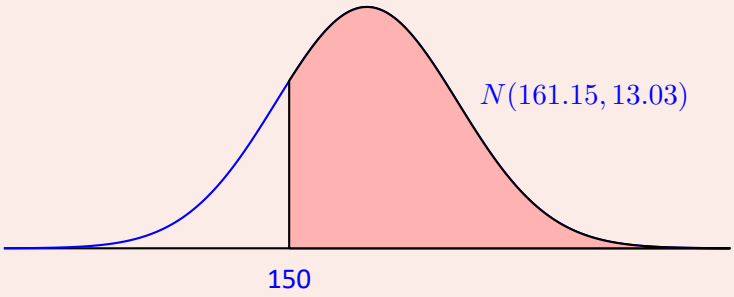
1. To calculate, on the basis of the normal (Gaussian) distribution, the probability that a randomly drawn customer used 150 free minutes or less, we enter the value 150 in the Statistic field. The chosen settings are confirmed with the Calculate button.



The obtained p value is 0.193961.

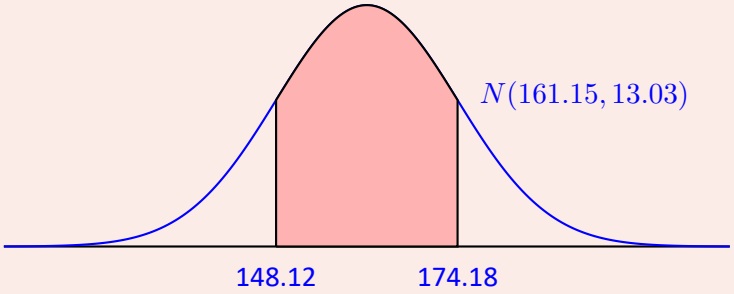
Note! Similar calculations can be performed on the basis of the empirical distribution. It is then enough to use the Frequency tables window to determine the percentage of customers using 150 minutes or less *(see example (11.1), distribution.pqs file)*. In the studied sample of 200 customers, 40 used 150 minutes or less, which is 20% of the sample, so the probability sought is $P = 0.2$.

2. To calculate, on the basis of the normal (Gaussian) distribution, the probability that a randomly drawn customer used more than 150 free minutes, we enter the value 150 in the Statistic field and tick the 1 - (p value) option. The chosen settings are confirmed with the Calculate button.



The obtained p value is 0.806039.

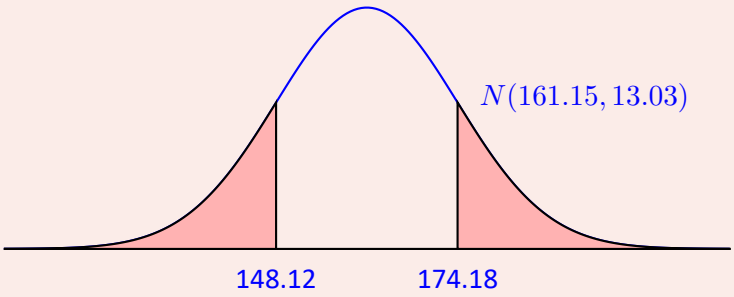
3. To calculate, on the basis of the normal (Gaussian) distribution, the probability that a randomly drawn customer used a number of minutes from the interval $[\overline{x} - sd, \overline{x} + sd] = [148.12\ min., 174.18\ min.]$, in the Statistic field we enter one of the interval's end values, and then tick the two-sided option. The chosen settings are confirmed with the Calculate button.



The obtained p value is 0.682689.

Copyright ©2010-2026 U-smile Stats – All rights reserved

4. To calculate, on the basis of the normal (Gaussian) distribution, the probability that a randomly drawn customer used a number of minutes outside the interval $[\overline{x} - sd, \overline{x} + sd] = [148.12\ min., 174.18\ min.]$, in the Statistic field we enter one of the interval's end values, and then tick the options: two-sided and 1 - (p value). The chosen settings are confirmed with the Calculate button.



The obtained p value is 0.317311.

FOR THOSE INTERESTED

Normal distribution $N(\mu, \sigma)$**. Formula:** the density function is defined as:

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right),$$

where:
$-\infty < x < +\infty$,
μ – the expected value of the population (measured by the mean),
σ – the standard deviation.

The t-Student distribution. Formula: the density function is defined as:

$$f(x, df) = \frac{\Gamma(\frac{df+1}{2})}{\Gamma(\frac{df}{2})\sqrt{df\pi}} \left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}},$$

where:
$-\infty < x < +\infty$,
df – degrees of freedom (the sample size decreased by the number of restrictions in the given calculations),
Γ is the Gamma function.

The χ^2 distribution. Formula: the density function is defined as:

$$f(x, df) = \frac{1}{2^{\frac{df}{2}}\Gamma\frac{df}{2}} x^{\frac{df}{2}-1} e^{-\frac{x}{2}},$$

where:
$x > 0$,
df – degrees of freedom (the sample size decreased by the number of restrictions in the given calculations),
Γ is the Gamma function.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Fisher-Snedecor (F) distribution. Formula: the density function is defined as:

$$F(x, df_1, df_2) = \frac{\sqrt{\frac{(df_1 x)^{df_1} d_2^{df_2}}{(df_1 x + df_2)^{df_1 + df_2}}}}{x B\left(\frac{df_1}{2}, \frac{df_2}{2}\right)},$$

where:
$x > 0$,
df_1, df_2 – degrees of freedom (it is assumed that if X and Y are independent with a χ^2 distribution with, respectively, df_1 and df_2 degrees of freedom, then $F = \frac{X/df_1}{Y/df_2}$ has the Fisher-Snedecor distribution $F(df_1, df_2)$),
B is the Beta function.

Copyright ©2010-2026 U-smile Stats – All rights reserved

13 Hypothesis testing

Generalizing the results obtained for a sample to the whole population essentially falls into 2 parts:

- **estimation** – estimating the values of population parameters on the basis of a statistical sample,
- **verification of statistical hypotheses** – checking specific assumptions formulated about the parameters of the general population, on the basis of the results obtained from a sample.

POINT AND INTERVAL ESTIMATION

In practice, we most often do not know the **parameters** (characteristics) of the whole population, and have only a sample drawn from that population. **Point estimators** are characteristics calculated on the basis of a random sample, and their accuracy is described by the **standard error**. The true value of the population parameter lies in the neighbourhood of the computed point estimator – for example, the population parameter arithmetic mean μ lies in the neighbourhood of the sample estimator $\overline{x}$.

Knowing the theoretical distribution of the estimator, we can, with a predetermined **confidence level** $(1-\alpha)$, estimate the value of the population parameter on the basis of the estimator obtained from the random sample. This procedure is called **interval estimation**, and the interval determined in this way is called the **confidence interval**; α itself is called the **significance level**.

The significance level most commonly used is 0.05, 0.01, or 0.001.

VERIFICATION OF STATISTICAL HYPOTHESES

Verifying statistical hypotheses proceeds in several steps.

Step 1. Formulating the hypotheses. These will be verified using an appropriate statistical test. Every test states a general form of the null hypothesis – $\mathcal{H}_0$ (*ang. null hypothesis*) and the alternative hypothesis – $\mathcal{H}_1$ (*ang. alternative hypothesis*):

$\mathcal{H}_0$: **there is no** statistically significant **difference** between the (means, medians, proportions, distributions, etc.) of the **populations**,

$\mathcal{H}_1$: **there is** a statistically significant **difference** between the means, medians, proportions, distributions, etc.) of the **populations**.

The researcher must formulate the hypothesis so that it is consistent both with reality and with the requirements of the chosen statistical test, e.g.:

$\mathcal{H}_0$: the percentage of women and men running a business in the population studied is the same.

If we do not know which percentage (of women or of men) might be higher in the population studied, the alternative hypothesis should be two-sided, i.e. we do not assume the direction of the difference:

$\mathcal{H}_1$: the percentage of women and men running a business in the population studied is different.

Copyright ©2010-2026 U-smile Stats – All rights reserved

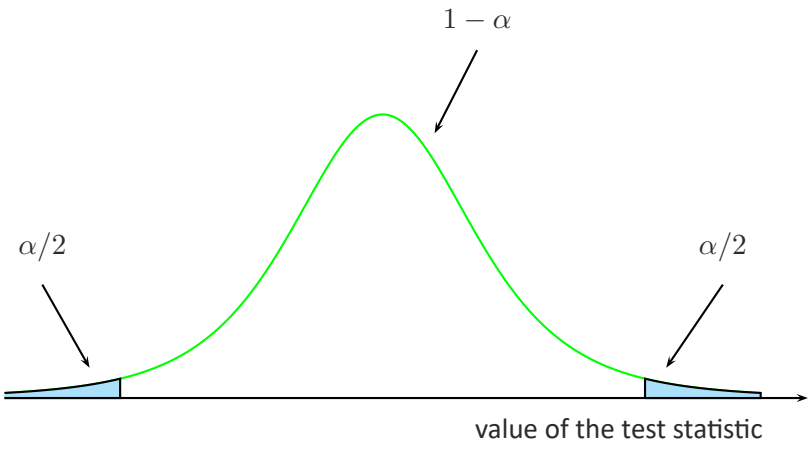
Only in rare cases, when we are certain about the direction of the expected difference, can a one-sided alternative hypothesis be used.

Step 2. Choosing the test and calculating the test statistic. We check which of the hypotheses – $\mathcal{H}_0$ or $\mathcal{H}_1$ – is more probable. Depending on the type of study and the nature of the variables, we select an appropriate statistical test.

Note! The choice of statistical test depends above all on the appropriate measurement scale (interval, ordinal, nominal) represented by the data being analysed, as well as on the study's design (dependent or independent model). Measurements of a given feature are called **dependent (paired)** when they are taken several times for the same objects. When the measurements of a given feature concern objects belonging to different groups, we speak of **independent (unpaired)** groups. Examples of studies with dependent groups: the body weight of patients before and after a weight-loss therapy, the reaction to a stimulus of the same group of people under two different conditions (e.g. at night and during the day), the assessment of creditworthiness of the same customers carried out by 2 different banks. Examples of studies with independent groups: body weight in a group of healthy patients and a group of ill patients, the effectiveness of fertilization with several different types of fertilizer, the size of GDP in different countries.

Note! The graph included in the Wizard window makes it easier to choose the appropriate statistical test.

The **test statistic** of the chosen test, calculated according to its formula, follows a theoretical distribution appropriate to it.



The program calculates the value of the test statistic and the corresponding p value (i.e. that part of the area under the curve which corresponds to the value of the test statistic). The p value lets us choose the more probable of the null and alternative hypotheses. In doing so we always assume that the null hypothesis is true, and the data collected are meant to supply sufficient evidence against this hypothesis:

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{we reject } \mathcal{H}_0 \text{ in favor of } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

A significance level of $\alpha = 0.05$ is usually adopted, accepting that in 5% of cases we will reject the null hypothesis even though it is true. In special situations a different significance level can be chosen, e.g. 0.01 or 0.001.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Step 3. Describing the results of the hypothesis verification.

TYPE I AND TYPE II ERRORS

A statistical test can fail to match reality in two ways:

		reality	
		$\mathcal{H}_0$: true	$\mathcal{H}_0$: false
test result	$\mathcal{H}_0$: true	OK	β
	$\mathcal{H}_0$: false	α	OK

We can make 2 kinds of errors:

α = Type I error – the probability of rejecting hypothesis $\mathcal{H}_0$ when it is true,

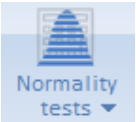
β = Type II error – the probability of accepting hypothesis $\mathcal{H}_0$ when it is false.

The power of the test is $1 - \beta$.

The values α and β are related to each other. Accepted practice is to fix the significance level α in advance and to minimize β by increasing the sample size.

Copyright ©2010-2026 U-smile Stats – All rights reserved

14 NORMALITY DISTRIBUTION TESTS



One-dimensional normality tests

WHEN TO USE:

A variety of tests may be applicable when examining the normality of a distribution, each of which pays attention to slightly different aspects of the Gaussian distribution. It is impossible to identify a single test that is good for every possible data set. Testing for normality of distribution can be done for variables or for differences determined from two variables.

ASSUMPTIONS

– measurement on an interval scale.

DEFINITIONS OF MEASURES

The Kolmogorov-Smirnov test – proposed by Kolmogorov (1933)[98], it is a relatively conservative test (it is harder to prove non-normality of the distribution using it). It is based on determining the distance between the empirical and theoretical cumulative distribution functions of the normal distribution. It is recommended for large samples, but should be used when the mean (μ) and standard deviation (σ) of the population from which the sample is drawn are known – one can then check whether the distribution matches the distribution defined by the given mean and standard deviation.

The Lilliefors test – a test proposed by Lilliefors (1967, 1969, 1973)[107][108][109]. It is a correction of the Kolmogorov-Smirnov test, used when the mean and standard deviation of the population from which the sample is drawn are not known. It is considered slightly less conservative than the Kolmogorov-Smirnov test.

The Shapiro-Wilk test – proposed by Shapiro and Wilk (1965)[153] for small samples, and later adapted for larger samples (up to 5000 objects) by Royston (1992)[146][147]. This test has relatively high power, which makes it easier to demonstrate non-normality of the distribution. The idea behind the test is illustrated by the Q-Q plot.

The D'Agostino-Pearson test – different types of statistical analysis that assume normality are sensitive to different degrees to different kinds of departure from this assumption. Tests referring to means in their hypotheses are assumed to be more sensitive to skewness, while tests comparing variances depend more heavily on kurtosis. A normal distribution should be characterised by zero skewness and zero kurtosis g_2 (or b_2 close to the value three). When the distribution is found not to be normal by the D'Agostino test (1973)[3], one can check whether this is caused by high skewness or kurtosis via the skewness test and the kurtosis test. Like the Shapiro-Wilk test, the D'Agostino test has greater power than the Kolmogorov-Smirnov and Lilliefors tests (D'Agostino, 1990[4]).

Copyright ©2010-2026 U-smile Stats – All rights reserved

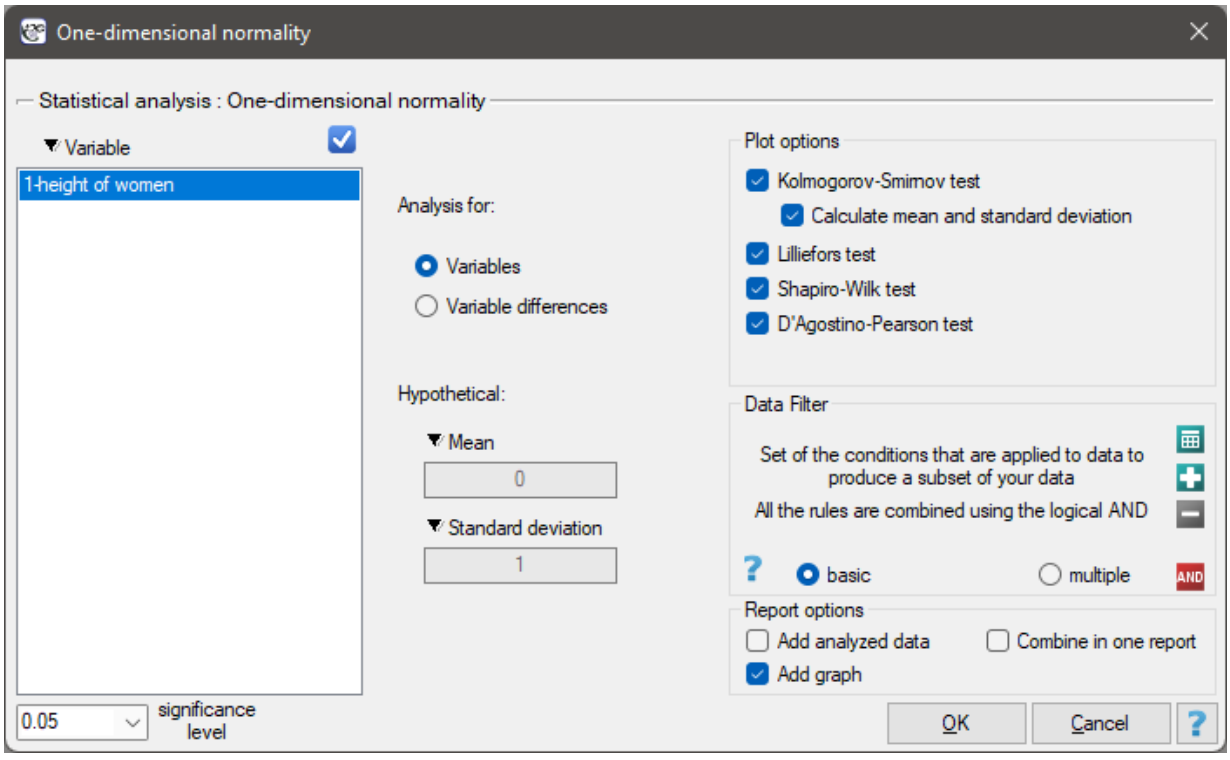
The Quantile-Quantile plot *(Q-Q plot)* – used to show the correspondence between two distributions. When checking agreement with a normal distribution, it checks the agreement of the data distribution (empirical distribution) with the theoretical Gaussian distribution. From it, one can visually check how well the normal-distribution curve fits the data. If the quantiles of the theoretical and empirical distributions match, the points fall along the line $y = x$. The horizontal axis represents the quantiles of the normal distribution, and the vertical axis the quantiles of the data distribution.

Very different deviations from the normal distribution are possible – the interpretation of a few of the most common ones is described below:

- the data lie along the line, but a few points deviate strongly from it – there are outliers in the data;
- points on the left side of the plot lie above the line, and on the right side below it – the distribution has a greater presence of values far from the mean than a normal distribution (negative kurtosis);
- points on the left side of the plot lie below the line, and on the right side above it – the distribution has a smaller presence of values far from the mean than a normal distribution (positive kurtosis);
- points on both the left and right sides of the plot lie above the line – the distribution is right-skewed (positive skewness);
- points on both the left and right sides of the plot lie below the line – the distribution is left-skewed (negative skewness).

HOW TO SET UP THE ANALYSIS

Statistics → Normality tests → One-dimensional normality (or via the Wizard).



Copyright ©2010-2026 U-smile Stats – All rights reserved

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 14.1. (Gauss.pqs file)

Women's height

Let us assume that women's height is a feature for which the average value is 168 cm. Most of the women we meet every day are of a height not significantly different from this average. Of course there are women who are quite short, and also very tall, but relatively rarely. Since very low and very high values occur rarely, and average values occur often, we can expect that the distribution of height is normal. To find out, 300 randomly selected women were measured.

Since we do not know the mean or standard deviation for women's height, but only have assumptions about these quantities, they will be determined from the sample.

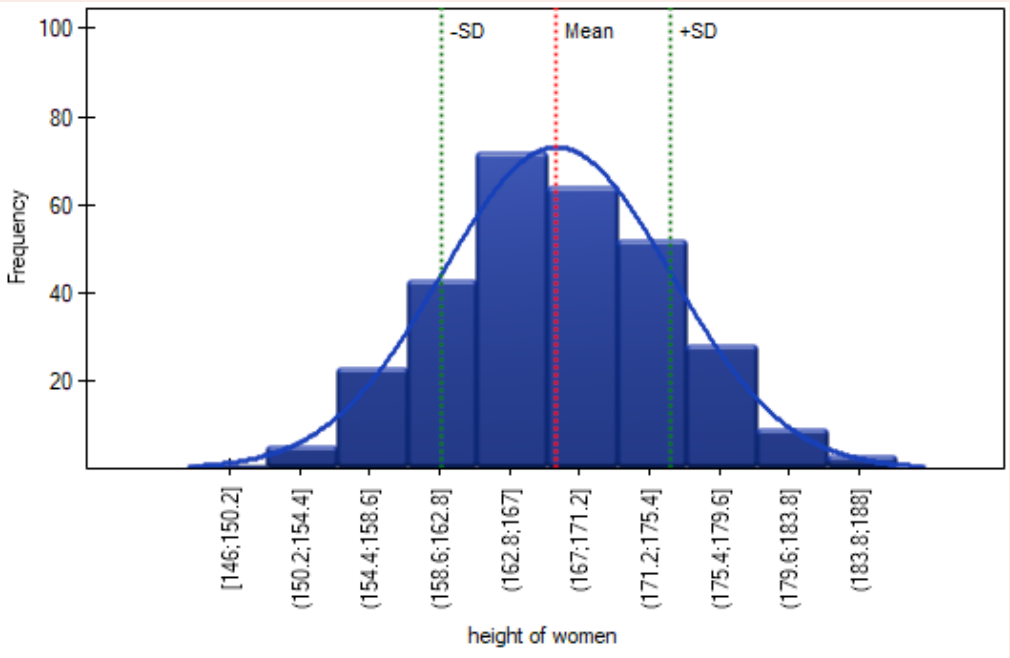
One-dimensional normality	
Analysed variables	height of women
Group size	300
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Group mean	167.726667
Group standard deviation	6.886463
Kolmogorov-Smirnov test	
D statistic	0.042332
Degrees of freedom	300
p-value	0.639671
Lilliefors test	
D statistic	0.042332
Degrees of freedom	300
p-value	0.21269
Shapiro-Wilk test	
W statistic	0.996277
Z statistic	-0.543622
p-value	0.706649
D'Agostino-Pearson test	
K-square statistic	0.284183
Degrees of freedom	2
p-value	0.867542
Skewness test	
Skewness	-0.013223
Z statistic	-0.095229
p-value	0.924133
Kurtosis test (g2)	
Kurtosis	-0.164655
Z statistic	-0.524514
p-value	0.599921

Copyright ©2010-2026 U-smile Stats – All rights reserved

all tests $p > 0.05$ (not statistically significant, $\alpha = 0.05$)

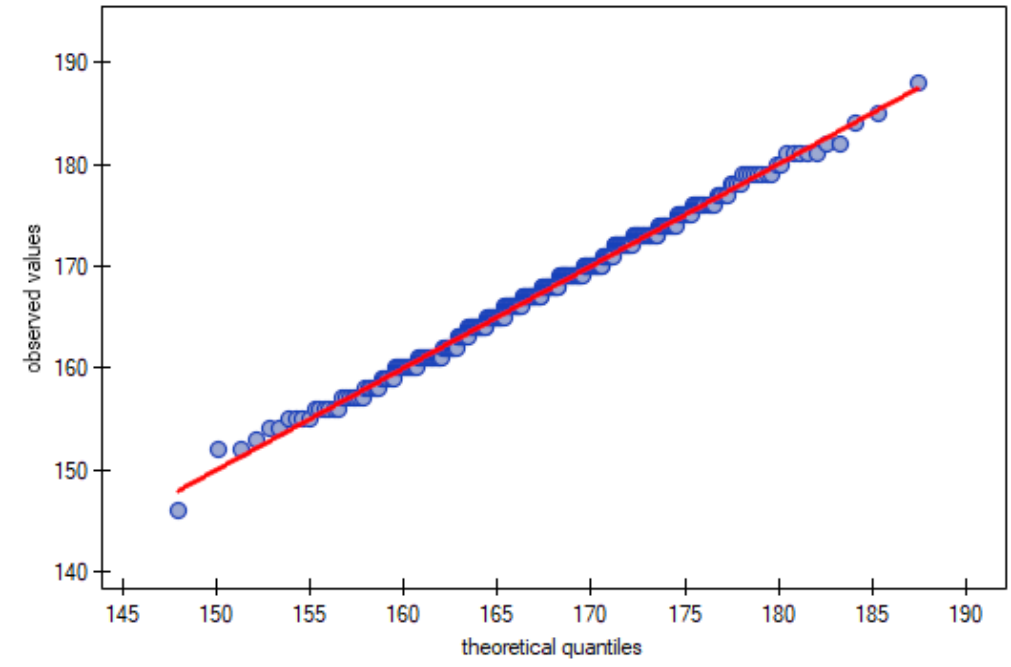
The distribution of women's height in the population studied does not deviate from a normal distribution.

The test examining skewness and kurtosis also shows no deviations. In the column chart, we presented the height distribution as 10 columns. The largest group consists of women with a height between 167 cm and 171 cm, while the smallest groups are women shorter than 150 cm or taller than 184 cm. The bell curve of the normal distribution seems to describe this distribution well.



On the quantile-quantile plot, the points lie almost perfectly on the line, which also indicates a very good fit to the normal distribution.

Copyright ©2010-2026 U-smile Stats – All rights reserved



The normal distribution can therefore be regarded as the distribution characterising women's height in the population studied.

EXAMPLE 14.2. (Gauss.pqs file)

Income

Suppose we study people's income in a certain country. We expect that most people's income will be average; however, there will be no people earning very little (below the minimum wage set by the authorities), while there will be people earning very much (company presidents), of whom there are relatively few. To check whether people's income in the country studied has a normal distribution, information on the income of 264 randomly selected people was collected.

Copyright ©2010-2026 U-smile Stats – All rights reserved

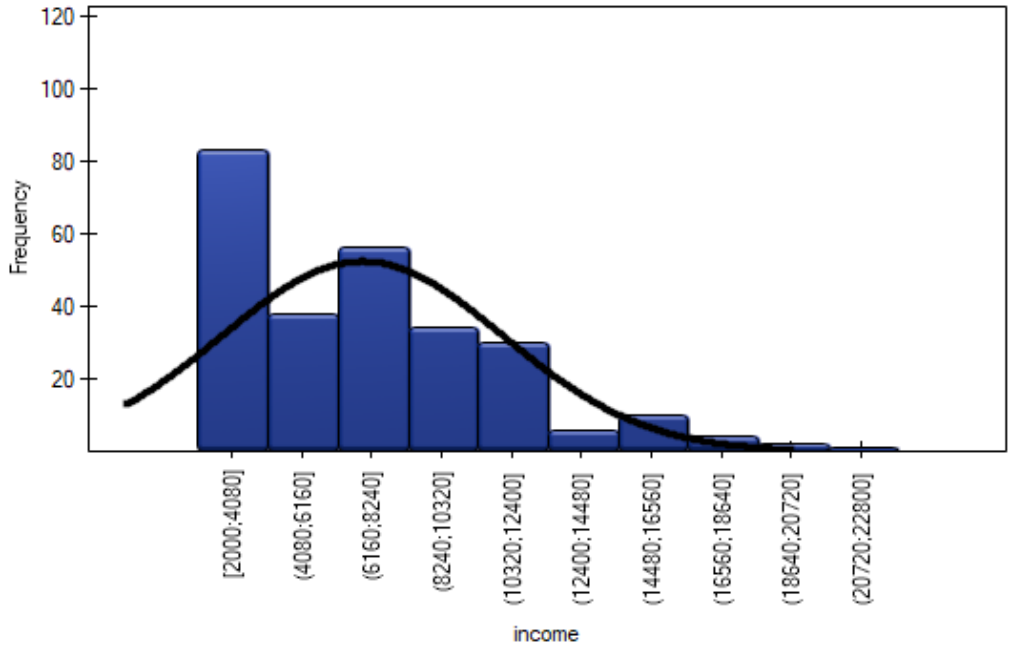
One-dimensional normality	
Analysed variables	income
Group size	264
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Group mean	6929.924242
Group standard deviation	4190.689299
Kolmogorov-Smirnov test	
D statistic	0.119718
Degrees of freedom	264
p-value	0.000938
Lilliefors test	
D statistic	0.119718
Degrees of freedom	264
p-value	<0.000001
Shapiro-Wilk test	
W statistic	0.920662
Z statistic	6.33239
p-value	<0.000001
D'Agostino-Pearson test	
K-square statistic	28.15808
Degrees of freedom	2
p-value	0.000001
Skewness test	
Skewness	0.825115
Z statistic	4.959279
p-value	0.000001
Kurtosis test (g2)	
Kurtosis	0.656828
Z statistic	1.887759
p-value	0.059058

all tests $p < \alpha$ (statistically significant, $\alpha = 0.05$)

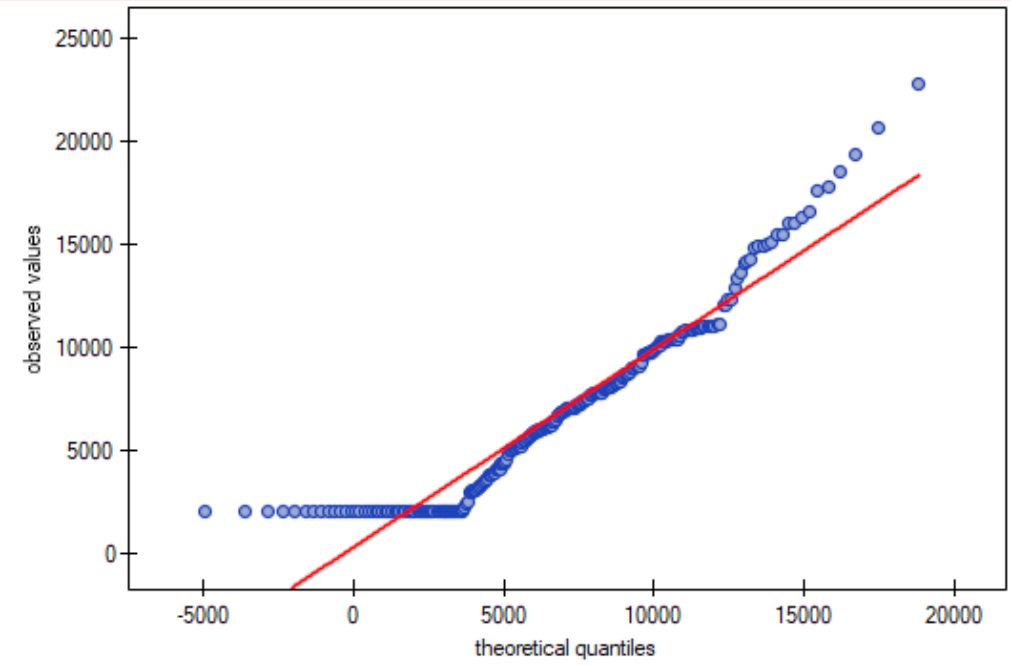
The distribution of income in the population studied is not a normal distribution.

The positive and statistically significant ($p < \alpha$) skewness value indicates that the right tail of the function is too long. The distribution is also more slender than the normal distribution, but this difference is not statistically significant (kurtosis test).

Copyright ©2010-2026 U-smile Stats – All rights reserved



On the quantile-quantile plot, the deviation from the normal distribution is shown by right-hand skewness, i.e. the position of the initial and final points of the plot well above the line.



As a result, the data collected do not support agreement between the income distribution and a normal distribution.

Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Hypotheses for normality-of-distribution tests:

$\mathcal{H}_0$: the distribution of the feature studied in the population is normal,
$\mathcal{H}_1$: the distribution of the feature studied in the population differs from normal.

Interpretation: the p value, determined on the basis of the test statistic of the chosen test, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The Kolmogorov-Smirnov test
Formula: on the basis of the sample data collected in a cumulative frequency distribution, and the corresponding values of the area under the theoretical normal-distribution curve, we determine the value of the test statistic D:

$$D = \sup_x |F_n(x) - F(x)|,$$

where:
$F_n(x)$ – the empirical cumulative distribution function calculated at successive points of the distribution, for an n-element sample,
$F(x)$ – the theoretical cumulative distribution function of the normal distribution.
The test statistic has the Kolmogorov-Smirnov distribution.

The Lilliefors test
Explanation: the test statistic D is calculated using the same formula as the Kolmogorov-Smirnov test, but has the Lilliefors distribution.

The Shapiro-Wilk test
Formulas: the Shapiro-Wilk test statistic is defined by:

$$W = \frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

where:
a_i – coefficients determined on the basis of the expected values of the ordered statistics, assigned weights, and the covariance matrix,
$\overline{x}$ – the mean of the sample data.

This statistic is transformed into a statistic with a normal distribution:

$$Z = \frac{g(W) - \mu}{\sigma},$$

where $g(W)$, μ, and σ depend on the sample size:

- for small sample sizes $n \in \langle 4; 12)$:

  $g(W) = -\ln(\gamma - \ln(1 - W))$,
  $\gamma = 0.459n - 2.273$,
  $\mu = -0.0006714n^3 + 0.025054n^2 - 0.39978n + 0.5440$,
  $\sigma = \exp(-0.0020322n^3 + 0.062767n^2 - 0.77857n + 1.3822)$;

- for large sample sizes $n \in \langle 12; 5000\rangle$:

Copyright ©2010-2026 U-smile Stats – All rights reserved

$g(W) = \ln(1 - W)$,
$\mu = 0.0038915u^3 - 0.083751u^2 - 0.31082u - 1.5851$,
$\sigma = \exp(0.0030302u^2 - 0.082676u - 0.4803)$,
$u = \ln(n)$.

The D'Agostino-Pearson test
Formulas: the test statistic is defined by:

$$K^2 = Z_A^2 + Z_K^2,$$

where:
Z_A^2 – the test statistic of the skewness test,
Z_K^2 – the test statistic of the kurtosis test.
This statistic asymptotically has the χ^2 distribution with two degrees of freedom.

The D'Agostino skewness test
Hypotheses:

$\mathcal{H}_0$: the distribution is not skewed (skewness in the population equals zero),
$\mathcal{H}_1$: the distribution is skewed (skewness in the population deviates from zero).

Formulas: the test statistic is defined by:

$$Z_A = \delta \ln\left(\frac{Y}{\alpha} + \sqrt{\frac{Y^2}{\alpha^2} + 1}\right),$$

where:
$Y = \sqrt{(b_1)}\sqrt{\frac{(n+1)(n+3)}{6(n-2)}}$,
$\sqrt{(b_1)} = \frac{m_3}{m_2^{(3/2)}}$,
$m_k = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^k}{n}$,
$\beta(\sqrt{(b_1)}) = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$,
$W^2 = -1 + \sqrt{2(\beta(\sqrt{(b_1)}) - 1}$,
$\delta = \frac{1}{\sqrt{\ln W}}$,
$\alpha = \sqrt{\frac{2}{W^2-1}}$.
The statistic Z asymptotically (for large sample sizes) has the normal distribution.

The D'Agostino kurtosis test
Hypotheses:

$\mathcal{H}_0$: kurtosis in the population matches the kurtosis of a normal distribution,
$\mathcal{H}_1$: kurtosis in the population differs from the kurtosis of a normal distribution.

Formulas: the test statistic is defined by:

$$Z_K = \frac{(1 - \frac{2}{9H}) - \left(\frac{1-\frac{2}{A}}{1+x\sqrt{\frac{2}{H-4}}}\right)^{1/3}}{\sqrt{\frac{2}{9H}}},$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where:
$E(b_2) = \frac{3(n-1)}{n+1}$,
$b_2 = \frac{m_4}{m_2^2}$,
$var(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$,
$x = \frac{b_2 - E(b_2)}{\sqrt{var(b_2)}}$,
$\sqrt{\beta(b_2)} = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-5)}}$,
$H = 6 + \frac{8}{\sqrt{\beta(b_2)}}\left(\frac{2}{\sqrt{\beta(b_2)}} + \sqrt{1 + \frac{4}{\beta(b_2)}}\right)$.
The statistic Z asymptotically (for large sample sizes) has the normal distribution.

Multivariate normality tests

WHEN TO USE:

Many methods of multivariate analysis, including MANOVA, Hotelling's tests, and regression models, are based on the assumption of multivariate normality. If a set of variables is characterised by a multivariate normal distribution, then each variable has a normal distribution. However, when all the individual variables are characterised by a normal distribution, their set need not have a multivariate normal distribution – therefore, testing the univariate normality of each variable may be useful, but cannot be assumed to be sufficient.

ASSUMPTIONS

- measurement on an interval scale for all analysed variables.

DEFINITIONS OF MEASURES

Different types of statistical analysis that assume normality are sensitive to different degrees to different kinds of departure from this assumption. Tests referring to means in their hypotheses are assumed to be more sensitive to skewness, while tests comparing covariances depend more heavily on kurtosis.

The Mardia test – proposed by Mardia in 1970[119] and modified in 1974[120], this test examines the normality of the distribution by analysing separately the magnitude of multivariate skewness and multivariate kurtosis.

The Jarque-Bera test – Jarque and Bera[91] proposed combining the two Mardia measures (skewness and kurtosis) into a single test.

The Hanusz-Tarasinska test – a similar way of combining skewness and kurtosis information into a single test is provided by the method of Hanusz and Tarasinska (2014)[79].

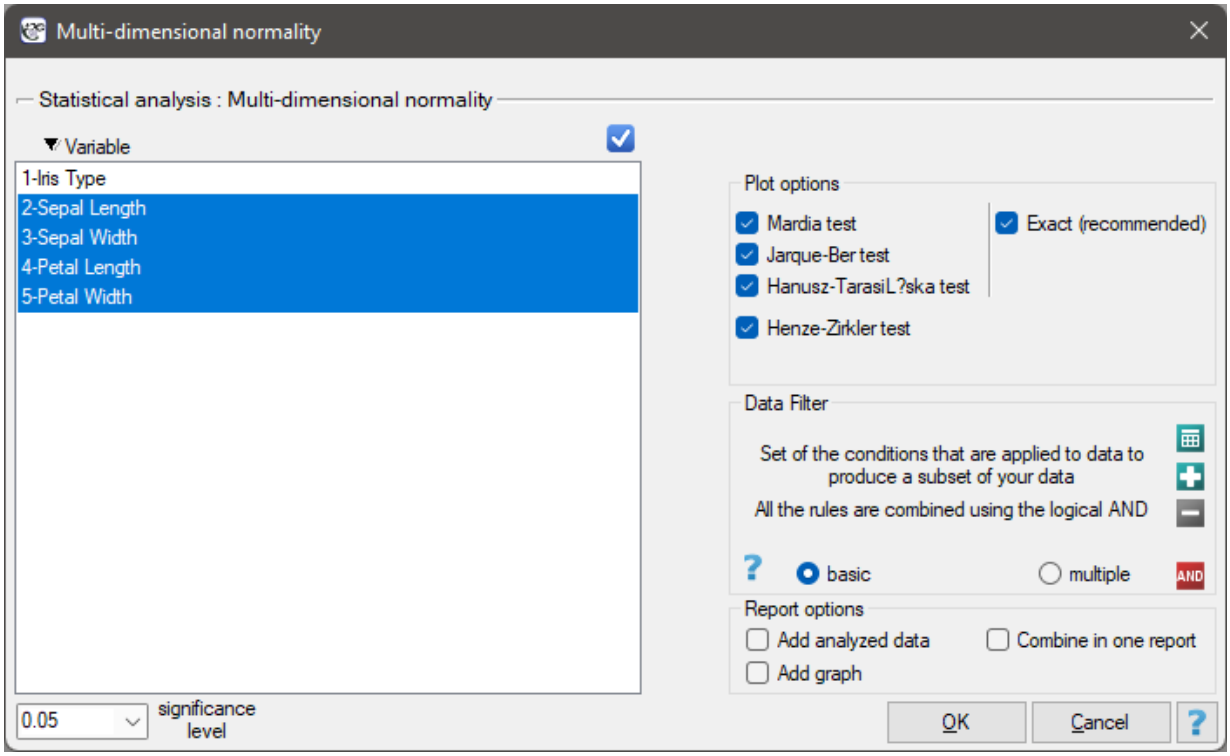
The Henze-Zirkler test – Henze and Zirkler (1990)[82] proposed a test examining multivariate normality of the distribution, extending the work of Baringhaus and Henze on the empirical characteristic

Copyright ©2010-2026 U-smile Stats – All rights reserved

function[56]. In the literature, this test is considered one of the strongest tests dedicated to the multivariate normal distribution (Thode, 2002)[167].

HOW TO SET UP THE ANALYSIS

Statistics → Normality tests → Multi-dimensional normality.



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 14.3. (Iris.pqs file)
We examine the normality of the distribution for the classic data set of R.A. Fisher from 1936[60]. The file is included in the help supplied with the program, and contains measurements of the length and width of the petals and calyx sepals for 3 varieties of iris. The analysis will be performed separately for each variety.

In the analysis window, we select all the tests and the graph, and to repeat the analysis for each iris variety, we set a multiple filter. All the results will be returned to the same worksheet, so we select the Combine into one report option.

Copyright ©2010-2026 U-smile Stats – All rights reserved

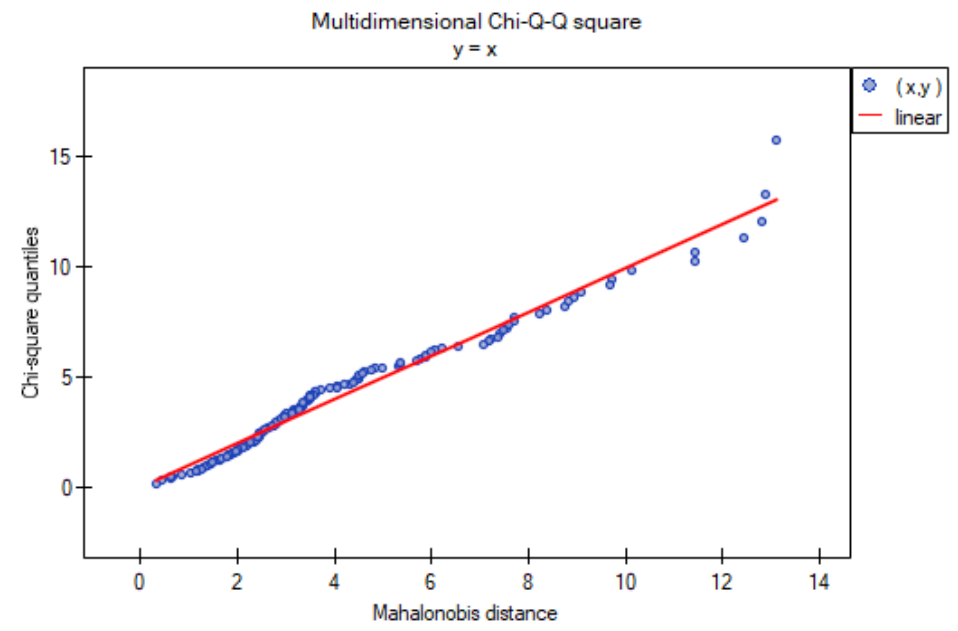
Multi-dimensional normality	ID1	ID2	ID3
Analysed variables	Sepal Length Sepal Width Petal Length Petal Width	Sepal Length Sepal Width Petal Length Petal Width	Sepal Length Sepal Width Petal Length Petal Width
Data Filter	Iris Type=seto	Iris Type=vers	Iris Type=virgi
Number of unspecified	0	0	0
Number of missing data	0	0	0
Significance level	0.05	0.05	0.05
Size	50	50	50
Number of variables in the model	4	4	4
Exact moments	Yes	Yes	Yes
Mardia skewness test			
Skewness	3.079721	3.022201	3.152472
Chi-square statistic	27.859728	27.339392	28.517842
Degrees of freedom	20	20	20
p-value	0.112762	0.125983	0.097696
Mardia kurtosis test			
Kurtosis	26.537656	22.879375	24.299061
Z statistic	2.192645	-0.113103	0.781699
p-value	0.028333	0.909949	0.434391
Jarque-Ber test			
Chi-square statistic	32.667419	27.352184	29.128896
Degrees of freedom	21	21	21
p-value	0.050038	0.159498	0.110941
Hanusz-TarasiL?ska test			
t-statistic	1.857782	-0.096737	0.654631
Degrees of freedom	20	20	20
p-value	0.077985	0.923898	0.520164
Henze-Zirkler test			
Statistic	0.948845	0.838801	0.75701
p-value	0.049954	0.226199	0.497024

all tests $p > 0.05$ for the versicolor and virginica varieties; borderline results for the setosa variety

The data distribution does not deviate from the multivariate normal distribution for the versicolor and virginica varieties; for the setosa variety, the results are inconclusive.

For the setosa variety, the Mardia kurtosis test and the Henze-Zirkler test indicate deviations from the multivariate normal distribution. We can also observe such deviations on the graph below, where as the Mahalanobis distance increases, the points lie further and further from the straight line.

Copyright ©2010-2026 U-smile Stats – All rights reserved



FOR THOSE INTERESTED

Hypotheses (common to all tests):

$\mathcal{H}_0$: the distribution of the data in the population $=$ a multivariate normal distribution,
$\mathcal{H}_1$: the distribution of the data in the population $\neq$ a multivariate normal distribution.

The Mardia test
Formulas: Mardia defined multivariate skewness and kurtosis as follows:

$$skew = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}^3, \quad kurt = \frac{1}{n}\sum_{i=1}^{n} m_{ii}^2$$

where:
$m_{ij} = \left(X_i - \bar{X}\right)^T S^{-1}\left(X_j - \bar{X}\right)$,
$S = \frac{1}{n}\sum_{j=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T$,
$\bar{X}$ – the mean, S – the covariance matrix.
For data coming from a sample rather than a population, the formulas for skewness and kurtosis are multiplied respectively: skewness by $(n/(n-1))^3$, and kurtosis by $(n/(n-1))^2$.

Skewness test: when the sample is drawn from a population with a multivariate normal distribution (null hypothesis), the test statistic is defined by (Mardia, 1970):

$$\chi^2(M) = \frac{n}{6} skew$$

or, with a correction for exact moments for groups with smaller sample sizes (< 20) (Mardia, 1974):

$$\chi_c^2(M) = \frac{n}{6}\frac{(n+1)(n+3)(k+1)}{n((n+1)(k+1)-6)} skew$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with $df = f = \frac{k(k+1)(k+2)}{6}$ degrees of freedom.

Kurtosis test: when the sample is drawn from a population with a multivariate normal distribution (null hypothesis), the test statistic is defined by (Mardia, 1974):

$$Z(M) = \frac{kurt - k(k+2)}{\sqrt{\frac{8k(k+2)}{n}}}$$

or, with a correction (Mardia, 1974):

$$Z_c(M) = \frac{(n+1)kurt - k(k+2)(n-1)}{\sqrt{\frac{8k(k+2)(n-3)(n-k-1)(n-k+1)}{(n+3)(n+5)}}}$$

This statistic asymptotically (for large sample sizes) has the normal distribution.

Interpretation: the p value, determined on the basis of the test statistic, is compared with the significance level α for both tests (the skewness test and the kurtosis test): if $p \leq \alpha$ for at least one test, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$ for both tests used, there is no reason to reject $\mathcal{H}_0$.

The Jarque-Bera test
Formulas: this test is based on the skewness and kurtosis statistics of the Mardia test. The test statistic is defined by:

$$\chi^2(JB) = \chi^2(M) + (Z(M))^2$$

or, with a correction (Mardia, 1974):

$$\chi_c^2(JB) = \chi_c^2(M) + (Z_c(M))^2$$

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with $df = f + 1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The Hanusz-Tarasinska test
Formulas: this test is based on the skewness and kurtosis statistics of the Mardia test. The test statistic is defined by:

$$t_c(HT) = \frac{Z_c(M)}{\sqrt{\frac{\chi_c^2(M)}{f}}}$$

The test statistic has the t-Student distribution with $df = f$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The Henze-Zirkler test
Formulas: the test statistic is defined by:

$$Z(HZ)_\beta = n\left(4I_E + D_{n,\beta}I_{E^c}\right)$$

where:
I_E and I_{E^c} – indicator functions depending on the singularity of the covariance matrix,

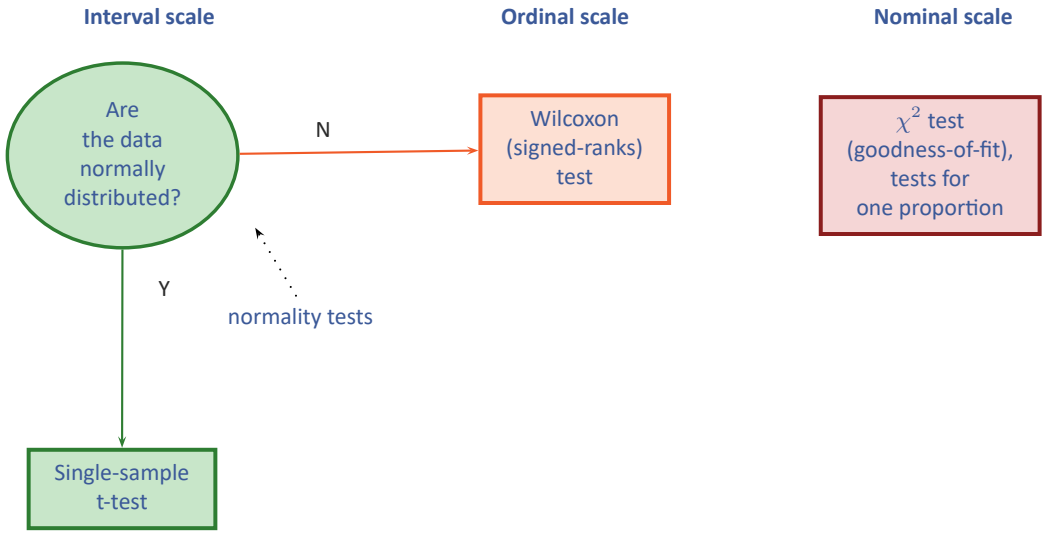
Copyright ©2010-2026 U-smile Stats – All rights reserved

$D_{n,\beta} = \frac{1}{n^2}\sum \exp\left(\frac{-\beta^2||Y_j - Y_k||^2}{2}\right) + (1+2\beta^2)^{-p/2} - 2(1+\beta^2)^{-p/2}\sum \exp\left(\frac{-\beta^2||Y_j||^2}{2(1+\beta^2)}\right)$,
$Y_i = S^{1/2}(X_i - \bar{X})$,
$\beta* = 2^{-1/2}\left(\frac{n(2k+1)}{4}\right)^{1/(k+4)}$ – the optimal value of the parameter β.

The statistic $Z(HZ)_\beta$ asymptotically (for large sample sizes) has the normal distribution, based on the mean and variance described by Henze and Zirkler, and is read one-sided.

The one-sided p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

15 COMPARISON - 1 GROUP



Copyright ©2010-2026 U-smile Stats – All rights reserved

15.1 PARAMETRIC TESTS

The single-sample t-test

WHEN TO USE THIS TEST:

The single-sample t-test *(single-sample t test)* lets you verify the hypothesis that the sample studied, with mean $\overline{x}$, comes from a population for which the mean μ is a given value.

ASSUMPTIONS

- measurement on an interval scale,
- normality of distribution of the feature studied.

Note! When the sample is large and the population standard deviation is known, the test statistic can be calculated using the formula $t = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$, which then has the normal distribution. As $n \to \infty$, the t-Student distribution converges to the normal distribution $N(0,1)$ – in practice, it is assumed that for $n > 30$ the t-Student distribution can be approximated by the normal distribution. Calculations can be based on raw records or on averaged data, i.e. the arithmetic mean, standard deviation, and sample size.

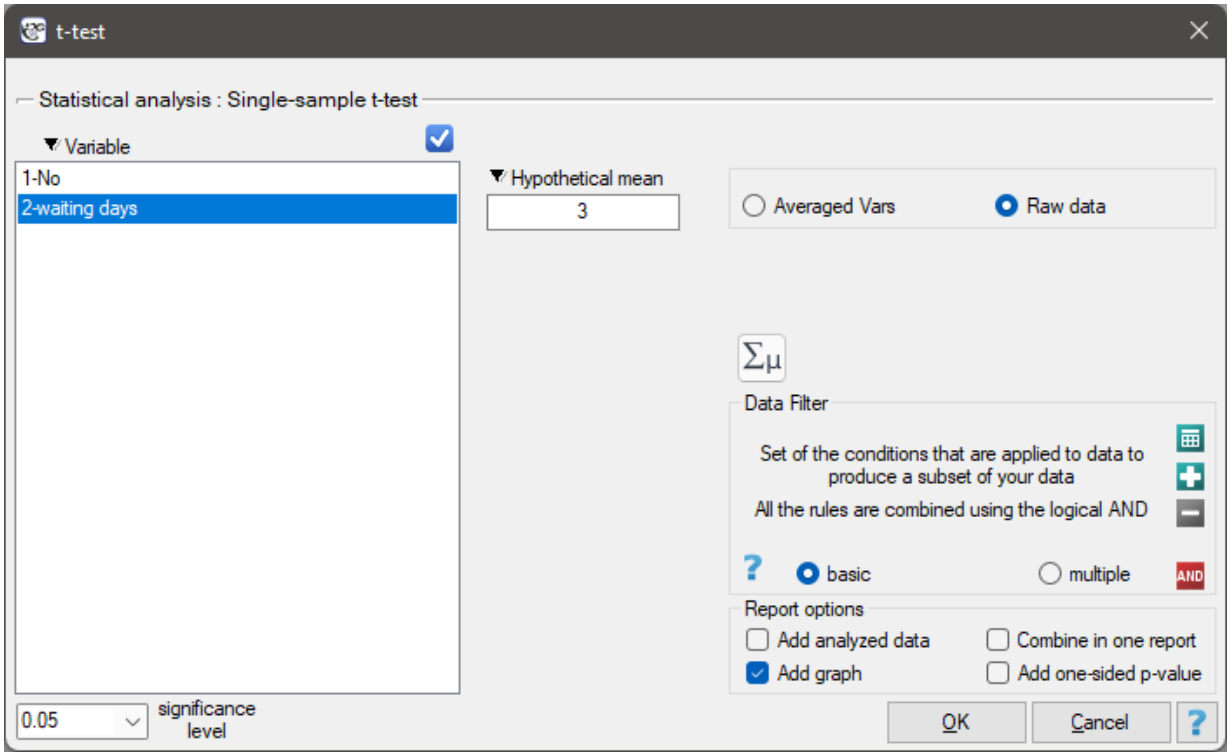
DEFINITIONS OF MEASURES

Cohen's d – a standardised effect size that determines how large a part of the variability present is accounted for by the difference between the means. When interpreting an effect, researchers often use general guidelines defined by Cohen[47], defining a small (0.2), medium (0.5), and large (0.8) effect size.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → t-Student (or via the Wizard).

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

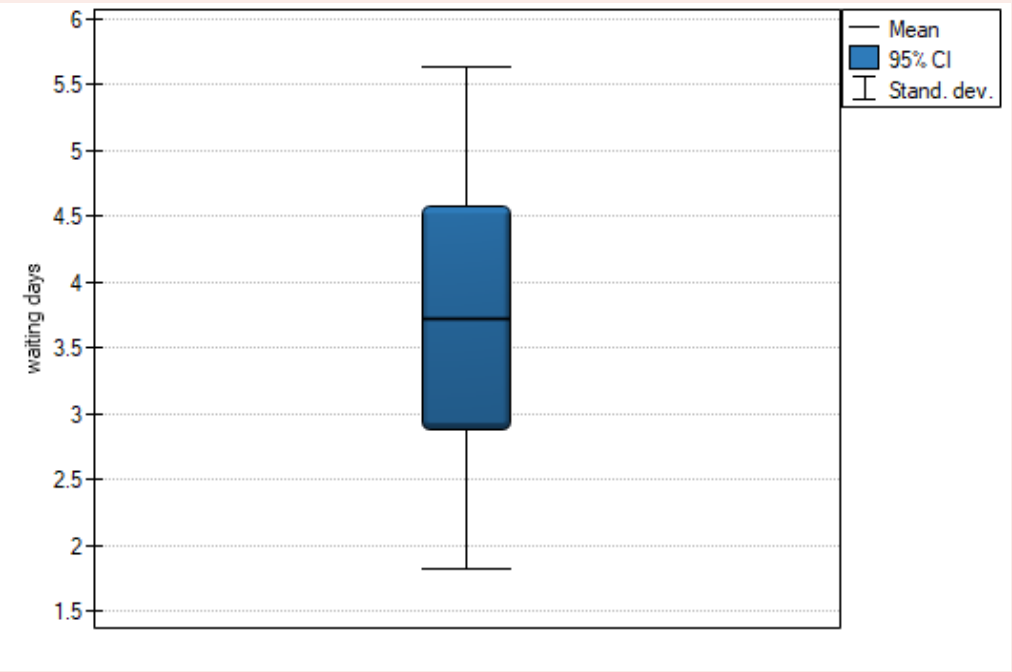
EXAMPLE 15.1. (courier.pqs file)
We want to check whether the waiting time for a delivery by a certain courier company is, on average, 3 days ($\mu_0 = 3$). To do this, a sample of 22 people was drawn from the population of the company's customers, and the number of days that elapsed between sending the parcel and its delivery was recorded for each. The following values were obtained: (1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7).

The number of days of waiting for a delivery in the population studied satisfies the assumption of normality of distribution.

Single-sample t-test	
Analysed variables	waiting days
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Group size	22
Hypothetical mean	3
Difference of the means	0.7273
-95% CI for the difference	-0.1182
+95% CI for the difference	1.5728
t-statistic	1.7889
Degrees of freedom	21
Two sided p-value	0.0881

Copyright ©2010-2026 U-smile Stats – All rights reserved

Summary	
Group	waiting days
Sample size	22
Arithmetic mean	3.7273
Standard error of the mean	0.4066
Standard deviation	1.9069
-95% CI for the group mean	2.8818
+95% CI for the group mean	4.5728



$p = 0.088074$ **(not statistically significant,** $\alpha = 0.05$**)**

There is no reason to reject the null hypothesis that the mean number of days of waiting for a delivery by the courier company analysed is 3.

For the sample studied, the mean is $\overline{x} = 3.727$, and the standard deviation is $sd = 1.907$.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu = \mu_0, \\ \mathcal{H}_1: &\quad \mu \neq \mu_0, \end{aligned}$$

where:
μ – the mean of the feature in the population represented by the sample,
μ_0 – a given value.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Formulas:

The test statistic is defined by:

$$t = \frac{\overline{x} - \mu_0}{sd}\sqrt{n},$$

where:
sd – the sample standard deviation,
n – the sample size.

The test statistic has the t-Student distribution with $n - 1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Cohen's d

$$d = \left|\frac{\overline{x} - \mu_0}{sd}\right|$$

The single-sample chi-square test for a population variance

WHEN TO USE THIS TEST:

The χ^2 test for the variance of a single sample *(The Single-Sample Chi-square Test for a Population Variance)* lets you verify the hypothesis that the sample studied comes from a population for which the variance (or standard deviation σ) is a given value. Hypotheses can refer either to the variance or, equivalently, to the standard deviation.

ASSUMPTIONS

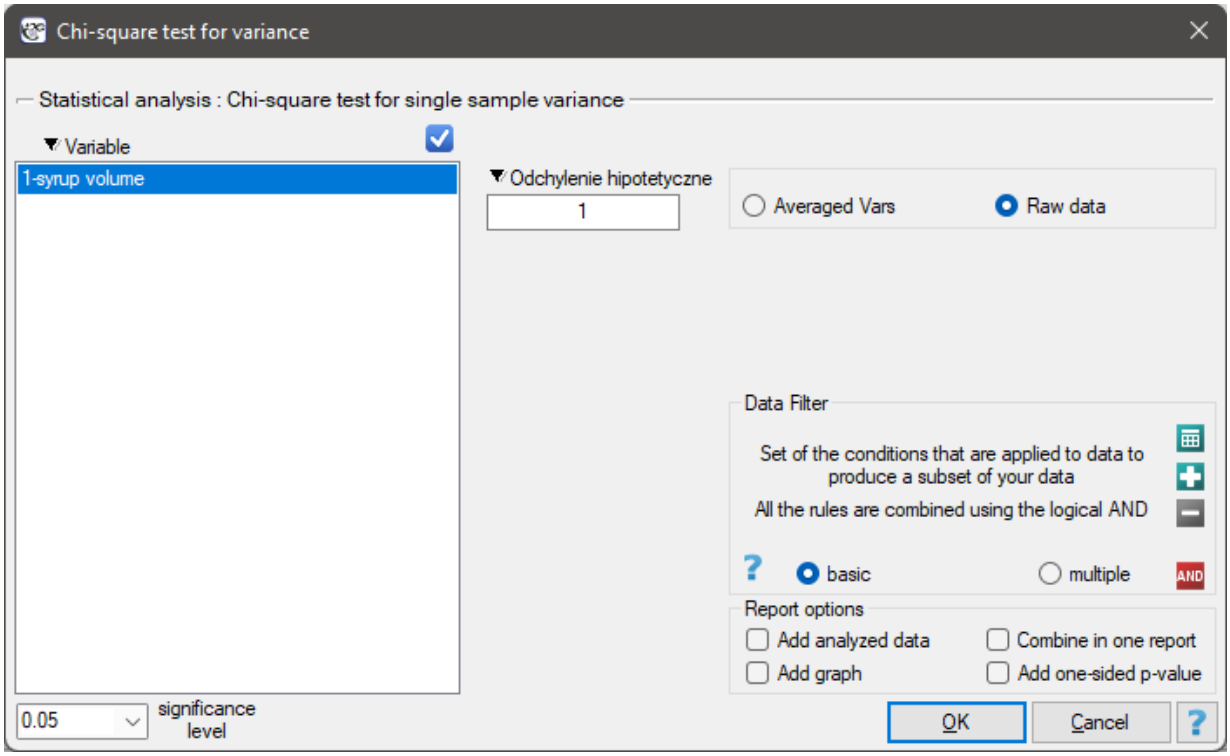
- measurement on an interval scale,
- normality of distribution of the feature studied.

Note! Calculations can be based on raw records or on averaged data, i.e. the standard deviation and sample size.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → Chi-square test for variance.

Copyright ©2010-2026 U-smile Stats – All rights reserved



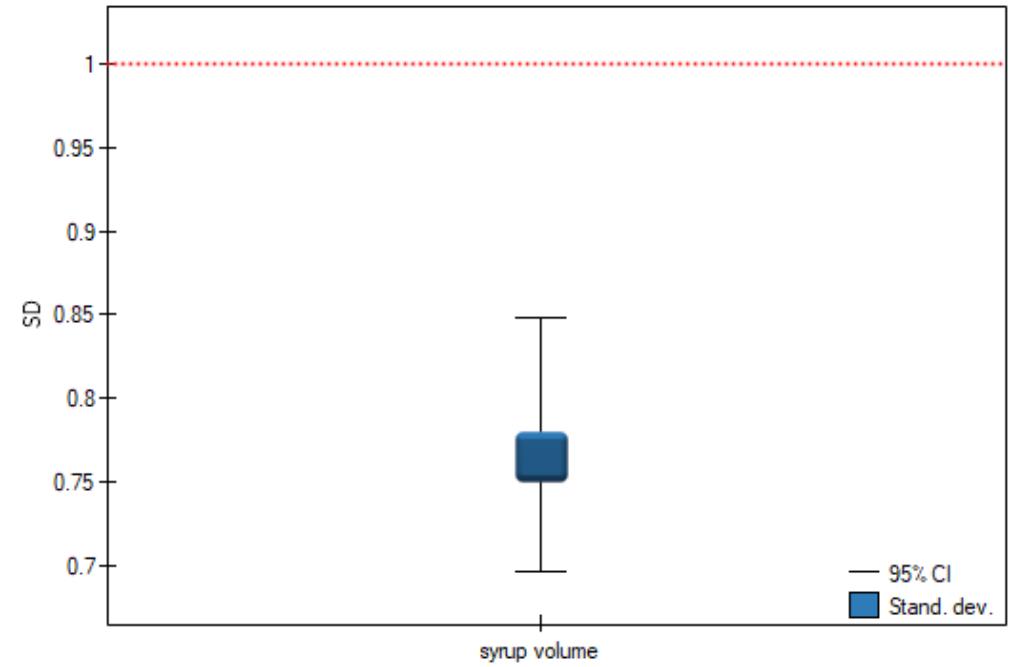
PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 15.2. (dispenser.pqs file)
Before starting production of another batch of a certain cough syrup, control measurements were made of the volume of syrup poured into bottles. The technical documentation of the dosing device shows that the permissible variation in syrup volume, measured by the standard deviation, is 1 ml. We need to check whether the tested device is working properly.

The normality of the distribution of the volume of syrup poured into bottles was checked (with the Lilliefors test), obtaining a result consistent with this normality. The analysis concerning the standard deviation can therefore be performed with the chi-square test for variance.

Chi-square test for single sample variance	
Analysed variables	syrup volume
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Group size	200
Group mean	199.9848
Odchylenie hipotetyczne	1
Group standard deviation	0.7646
-95% CI for standard deviation	0.6963
+95% CI for standard deviation	0.8479
Chi-square statistic	116.3476
Degrees of freedom	199
Two sided p-value	<0.0001

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p = 0.000001$ **(statistically significant,** $\alpha = 0.05$**)**

The scatter of the dosing device is different from 1 ml.

We can, however, consider the device's performance to be correct, because the sample standard deviation is 0.76, which is significantly smaller than the permissible value from the technical documentation.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \sigma = \sigma_0, \\ \mathcal{H}_1: &\quad \sigma \neq \sigma_0, \end{aligned}$$

where:
σ – the standard deviation of the feature in the population represented by the sample,
σ_0 – a given value.

Formulas: The test statistic is defined by:

$$\chi^2 = \frac{(n-1)sd^2}{\sigma_0^2},$$

where:
sd – the sample standard deviation,
n – the sample size.

Interpretation: The test statistic has the χ^2 distribution with the number of degrees of freedom calculated using the formula $df = n - 1$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The p value, determined on the basis of the test statistic and the χ^2 distribution, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

If the standard deviation value is smaller than the given value, the p value is calculated as twice the area under the chi-square distribution curve to the left of the corresponding critical value; if it is larger than the given value, it is twice the corresponding area to the right.

Copyright ©2010-2026 U-smile Stats – All rights reserved

15.2 NON-PARAMETRIC TESTS

The Wilcoxon test (signed-ranks)

WHEN TO USE THIS TEST:

The Wilcoxon signed-ranks test *(Wilcoxon signed-ranks test)*, also known as the single-sample Wilcoxon test (Wilcoxon, 1945, 1949[176]), lets you verify the hypothesis that the sample studied comes from a population for which the median (θ) is a known value.

ASSUMPTIONS

- measurement on an ordinal scale or an interval scale.

Note! Depending on the sample size, the test statistic takes a different form – the statistic T for a small sample size, or the statistic Z for a large sample size (see *For those interested*).

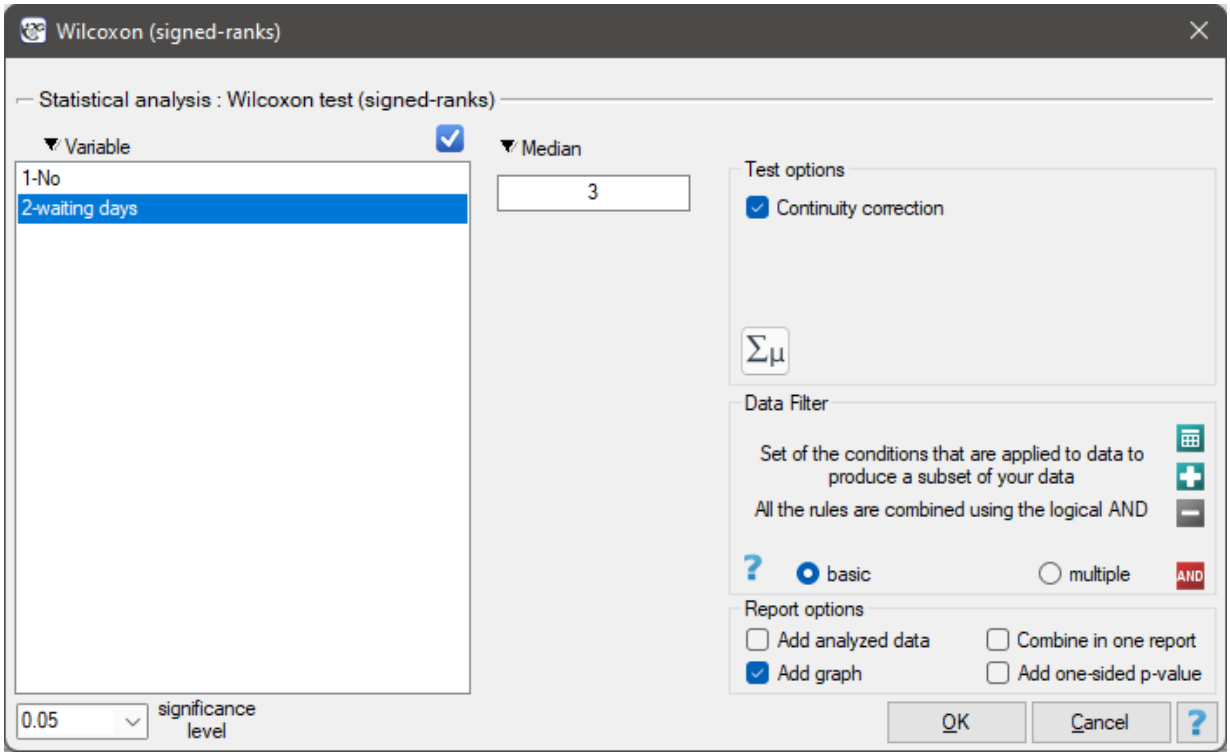
DEFINITIONS OF MEASURES

Cohen's d – a standardised effect size. The distribution of the Wilcoxon test statistic is approximated by the normal distribution, which can be converted into the effect size $r = |Z/n|$[68], from which Cohen's d is then obtained using the standard conversion used in meta-analyses. When interpreting an effect, researchers often use general guidelines defined by Cohen[47], defining a small (0.2), medium (0.5), and large (0.8) effect size.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Wilcoxon (signed-ranks) (or via the Wizard).

Copyright ©2010-2026 U-smile Stats – All rights reserved



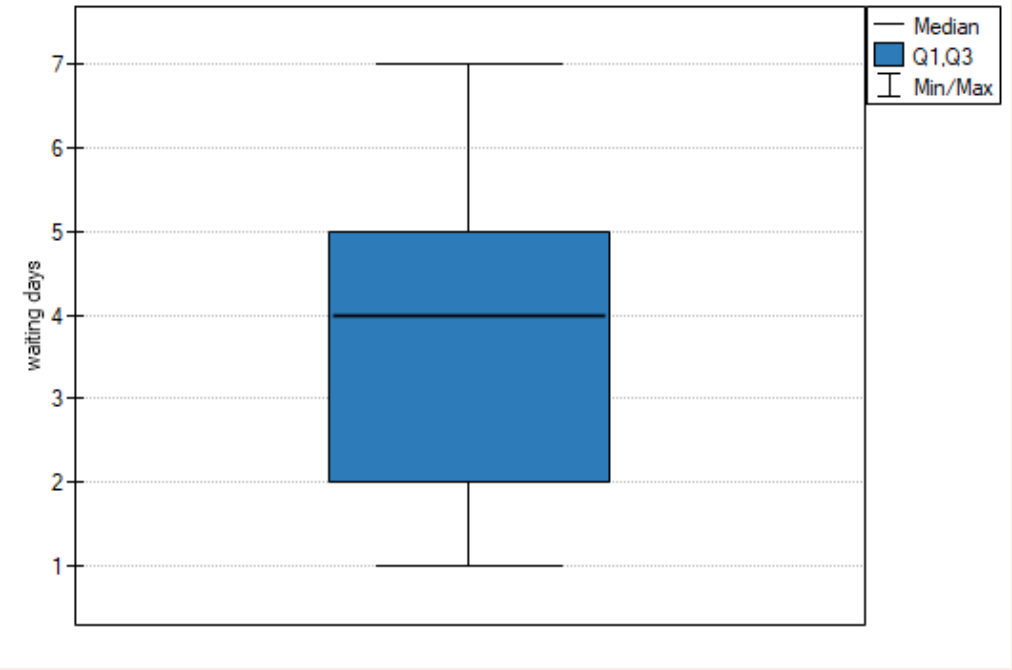
PERFORMING AND INTERPRETING THE ANALYSIS

Continuation of Example 15.1 (courier.pqs file): this time, instead of the mean, we will check the median number of days of waiting for a delivery.

Wilcoxon test (signed-ranks)	
Analysed variables	waiting days
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Group size	22
Number of omitted values (equal median)	3
Hypothetical median	3
T statistic	56
Two sided p-value (exact)	0.1232
Z statistic (adjusted for ties)	1.5726
Two sided p-value (asymptotic)	0.1158

Summary	
Group	waiting days
Sample size	22
Median	4
Lower quartile	2
Upper quartile	5

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p = 0.123212$ for the T statistic (not statistically significant, $\alpha = 0.05$)

There is no reason to reject the null hypothesis that the median number of days of waiting for a delivery by the courier company analysed is 3.

We would reach the same conclusion on the basis of the value $p = 0.111161$ or $p = 0.115817$ of the Wilcoxon test based on the Z statistic, or Z with the continuity correction.

FOR THOSE INTERESTED

Hypotheses:
Hypotheses relate to the equality of the sum of positive and negative ranks, or are simplified to medians:

$$\begin{aligned} \mathcal{H}_0 &: \quad \theta = \theta_0, \\ \mathcal{H}_1 &: \quad \theta \neq \theta_0. \end{aligned}$$

where:
θ – the median of the feature studied in the population represented by the sample studied,
θ_0 – a given value.

We determine the value of the test statistic Z (T – for a small sample size), and on this basis the p value, which we compare with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

For a small sample size

$$T = \min\left(\sum R_-, \sum R_+\right),$$

where $\sum R_+$ and $\sum R_-$ are, respectively, the sum of positive ranks and the sum of negative ranks. This statistic has the Wilcoxon distribution.

Copyright ©2010-2026 U-smile Stats – All rights reserved

For a large sample size

$$Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}},$$

where:
n – the number of ranked signs (the number of ranks),
t – the number of cases included in a tied rank.
The formula for the test statistic Z includes a correction for tied ranks. This correction should be used when tied ranks are present (when there are no tied ranks, the correction is not calculated, since then $\left(\sum t^3 - \sum t\right)/48 = 0$). The Z statistic asymptotically (for large sample sizes) has the normal distribution.

Continuity correction for the Wilcoxon test (Marascuilo and McSweeney, 1977[118])
A continuity correction is used to allow the test statistic to take on all values of real numbers, in line with the assumption of a normal distribution. The formula for the test statistic with the continuity correction is then:

$$Z = \frac{\left|T - \frac{n(n+1)}{4}\right| - 0.5}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}}.$$

Cohen's d

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

The chi-square goodness-of-fit test

WHEN TO USE THIS TEST:

The χ^2 goodness-of-fit test *(Chi-square goodness-of-fit test)*, also called the one-sample χ^2 test, is designed to test the agreement of the values observed for r ($r \geq 2$) categories $X_1, X_2, ..., X_r$ of a single feature X with hypothetical expected values for that feature. The values of all n measurements should be gathered into a table consisting of r rows (categories: $X_1, X_2, ..., X_r$). For each category X_i, we record the frequency of its occurrence O_i, and either the frequency expected for it E_i, or the probability of its occurrence p_i. The expected frequency is calculated as the product $E_i = np_i$. The resulting table can take one of the forms below.

ASSUMPTIONS

- measurement on a nominal scale – any ordering of the categories is not taken into account,
- large expected frequencies, according to the Cochran interpretation (1952)[42],
- the sum of the observed frequencies should equal the sum of the expected frequencies, and the sum of all probabilities p_i should equal 1.

Copyright ©2010-2026 U-smile Stats – All rights reserved

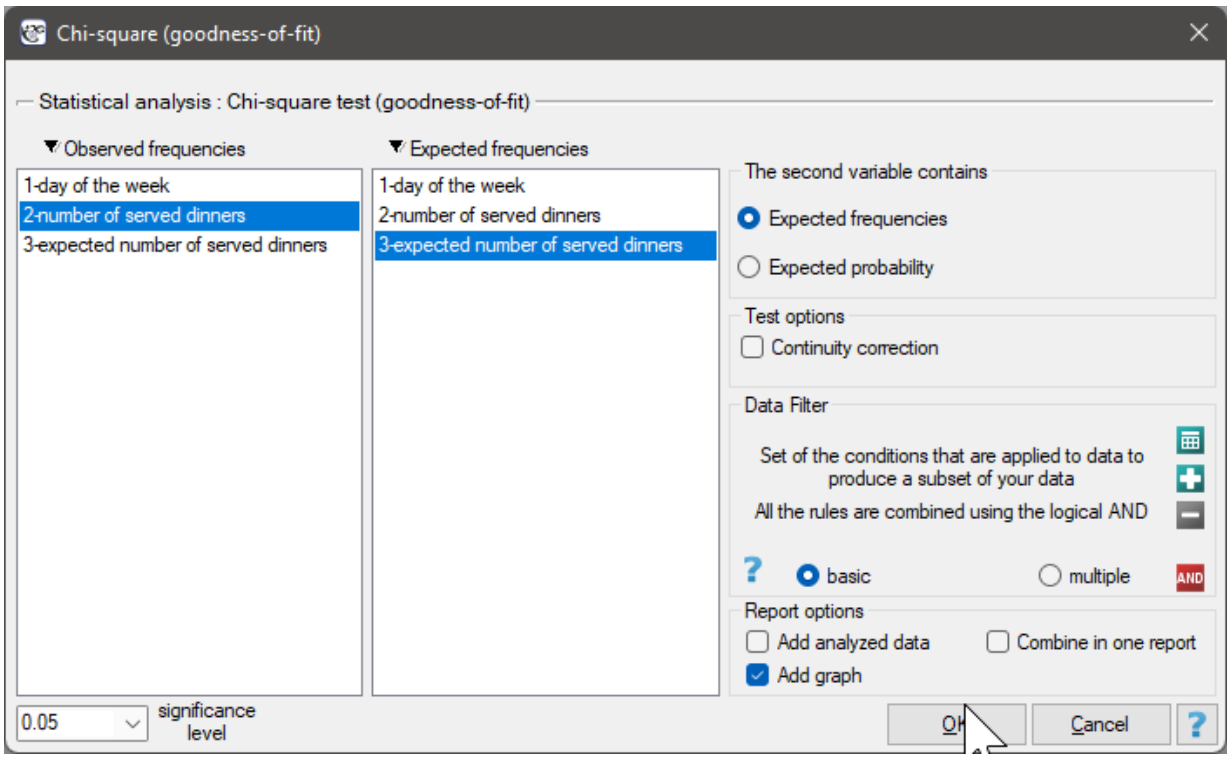
Note! The table of data can take one of the following forms:

Category X_i	O_i	E_i
X_1	O_1	E_i
X_2	O_2	E_2
...	...	...
X_r	O_r	E_r

Category X_i	O_i	p_i
X_1	O_1	p_1
X_2	O_2	p_2
...	...	...
X_r	O_r	p_r

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Chi-square (or via the Wizard).



PERFORMING AND INTERPRETING THE ANALYSIS

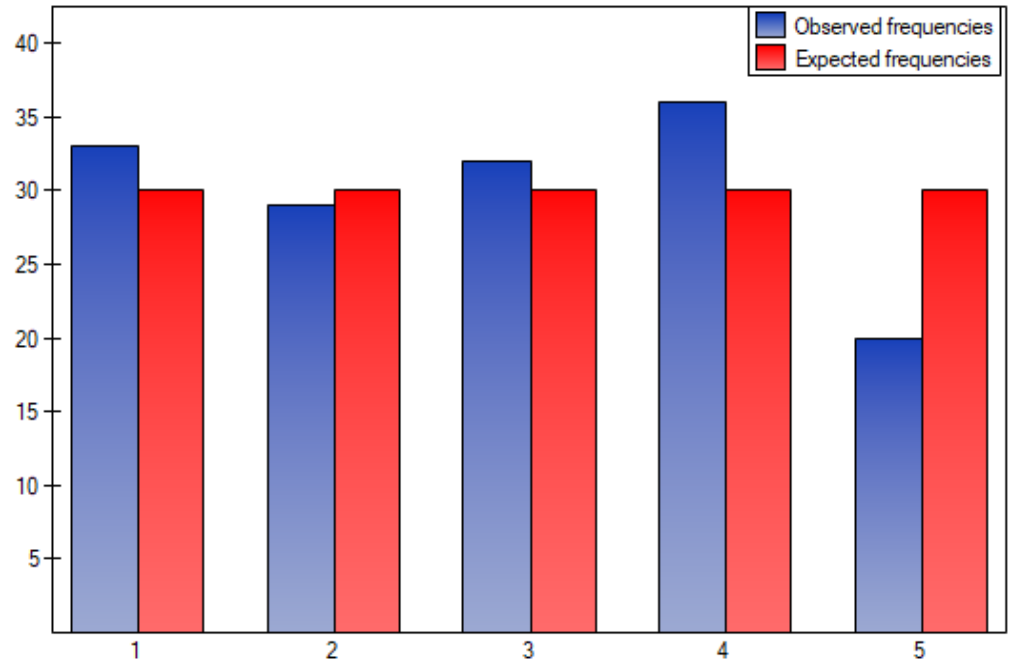
Example 15.3. (dinners.pqs file)
We would like to find out whether the number of dinners served on successive days of the week (Monday to Friday) in a certain school canteen is statistically the same. To do this, a one-week sample was taken, and the number of dinners served was recorded for each day: Monday – 33, Tuesday – 29, Wednesday – 32, Thursday – 36, Friday – 20.

In total, 150 dinners were served over the whole week (5 days). We assume that the probability of a dinner being served is the same on each day, i.e. it equals $\frac{1}{5}$. The expected number of dinners served for each of the five days of the week is therefore $E_i = 150 \cdot \frac{1}{5} = 30$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

day of the week	number of served dinners	expected number of served dinners
Monday	33	30
Tuesday	29	30
Wednesday	32	30
Thursday	36	30
Friday	20	30

Chi-square test (goodness-of-fit)	
Analysed variables	number of served dinners
	expected number of served
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	No
Size	150
Chi-square statistic	5
Degrees of freedom	4
p-value	0.2873



$p = 0.287297$ **for** $df = 4$ **(not statistically significant,** $\alpha = 0.05$**)**

There is no reason to reject the null hypothesis that the number of dinners served agrees with the expected number of dinners served on each day.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! If we wanted to make a larger number of comparisons within a single study, we could apply the **Bonferroni correction**[2] or another of the corrections described in the Multiple comparisons section. This correction is used to limit the size of the Type I error committed when we compare the observed and expected values between selected days, e.g.: Fri $\Longleftrightarrow$ Mon, Fri $\Longleftrightarrow$ Tue, Fri $\Longleftrightarrow$ Wed, Fri $\Longleftrightarrow$ Thu – assuming the comparisons are performed independently. The significance level α for each comparison is determined according to this correction using the formula $\alpha = \frac{0.05}{r}$, where r is the number of comparisons performed – for our example, the significance level for a single comparison under the Bonferroni correction is $\alpha = \frac{0.05}{4} = 0.0125$. It should be remembered, however, that by reducing α for each comparison, we also reduce the power of the test.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad O_i = E_i \text{ for all categories,} \\ \mathcal{H}_1: &\quad O_i \neq E_i \text{ for at least one category.} \end{aligned}$$

Formulas:

The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}.$$

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with the number of degrees of freedom calculated using the formula $df = (r-1)$.

The p value, determined on the basis of the test statistic and the χ^2 distribution, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Tests for one proportion

WHEN TO USE THIS TEST:

Tests for one proportion are used when there are two possible outcomes to obtain (one of which is a distinguished outcome, with count m), and we know how often these outcomes occur in the sample (we know the proportion p). These tests are used to verify the hypothesis that the proportion in the population from which the sample is drawn is a given value. Depending on the sample size n, we can choose between the Z **test for one proportion** – for large samples, and the **exact binomial test** – for small samples.

ASSUMPTIONS

– measurement on a nominal scale – any ordering of the categories is not taken into account.

Note! Additional condition for the Z test for one proportion: a large sample size (according to the

Copyright ©2010-2026 U-smile Stats – All rights reserved

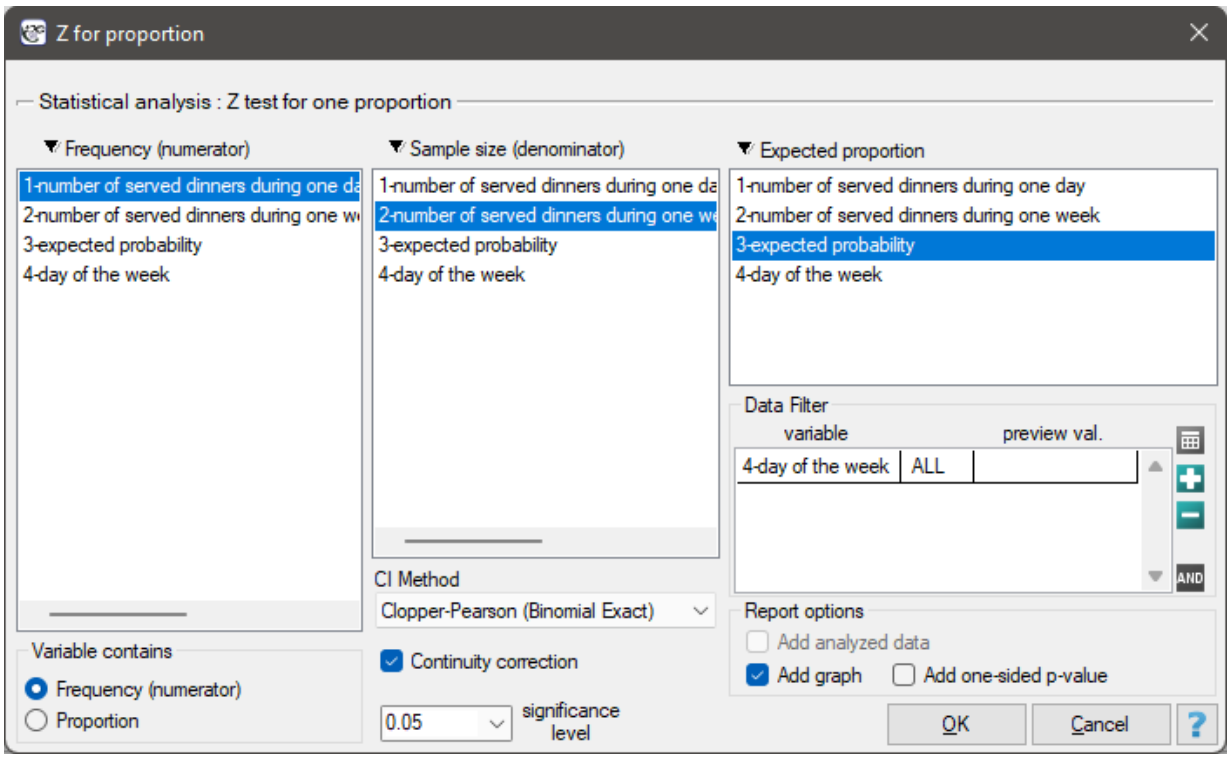
interpretation of Marascuilo and McSweeney (1977)[118], each of the values $np > 5$ and $n(1-p) > 5$). When this condition is not met, it is recommended to use the exact binomial test instead.

DEFINITIONS OF MEASURES

Confidence interval for a proportion – for the sample estimator, which in this case is the proportion value p, a confidence interval is calculated. For large samples, one can rely on intervals based on the normal distribution – the so-called Wald intervals. More universal, however, are the intervals proposed by Wilson (1927)[179], as well as by Agresti and Coull (1998)[5]. The Clopper-Pearson intervals (1934)[40] are more accurate for smaller samples. A comparison of methods for constructing confidence intervals for proportions can be found in the work of Brown L.D. and others (2001)[31].

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Z for proportion.



PERFORMING AND INTERPRETING THE ANALYSIS

Continuation of Example 15.3 (dinners.pqs file): suppose we would like to check whether $\frac{1}{5}$ of all the dinners served in the school canteen over the week are served on Friday. For the sample drawn, $m = 20$, $n = 150$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

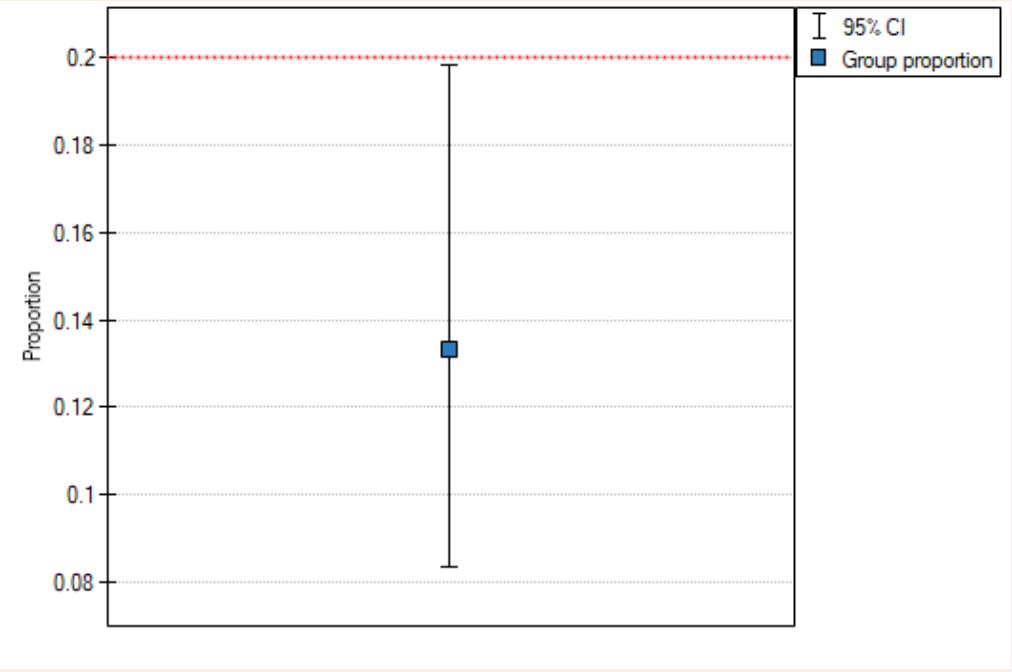
number of served dinners during one day	number of served dinners during one week	expected probability	day of the week
33	150	0.2	Monday
29	150	0.2	Tuesday
32	150	0.2	Wednesday
36	150	0.2	Thursday
20	150	0.2	Friday

When setting the analysis options, we enable a filter, selecting the appropriate day of the week – Friday. Not setting the filter does not generate an error, but simply results in statistics being calculated for each successive day of the week.

Z test for one proportion	
Analysed variables	number of served dinners du
	number of served dinners du
	expected probability
Data Filter	day of the week=Friday
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
1	
Group proportion	0.1333
Clopper-Pearson (Binomial Exact)	
-95% CI for the proportion	0.0834
+95% CI for the proportion	0.1984
Z statistic	1.9392
Two sided p-value (asymptotic)	0.0525
Two sided p-value (exact)	0.0447

Data :

number of serve	number of s	expected pro
20	150	0.2



Copyright ©2010-2026 U-smile Stats – All rights reserved

$p = 0.041227$ for the Z test without the continuity correction, $p = 0.044711$ for the exact binomial test (statistically significant, $\alpha = 0.05$); $p = 0.052479$ for the Z test with the continuity correction (not statistically significant)

The proportion of dinners served on Friday is statistically lower than $\frac{1}{5}$ of the dinners served over the whole week – except for the result of the Z test with the continuity correction.

The proportion of the distinguished value in the sample is $p = \frac{m}{n} = 0.133$, and the 95% Clopper-Pearson confidence interval for this fraction, $(0.083, 0.198)$, does not include the hypothetical value of 0.2.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad p = p_0, \\ \mathcal{H}_1: &\quad p \neq p_0, \end{aligned}$$

where:
p – the probability (distinguished proportion) in the population,
p_0 – the expected probability (expected proportion).

The Z test for one proportion *(Z test for one proportion)*
Formulas: The test statistic is defined by:

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where:
$p = \frac{m}{n}$ – the proportion for the sample from this population,
m – the count of the distinguished values in the sample,
n – the sample size.
The test statistic modified with the continuity correction is defined by:

$$Z = \frac{|p - p_0| - \frac{1}{2n}}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

The Z statistic, both without and with this correction, asymptotically (for large sample sizes) has the normal distribution.

The binomial test *(Binomial test for one proportion)*
Definition: This test directly uses the **binomial distribution**, also called the Bernoulli distribution, which belongs to the group of discrete distributions (i.e. distributions in which the analysed variable takes a finite number of values). The analysed variable can take $k = 2$ values, the first usually referred to as a success and the second as a failure. The probability of a success is p_0, and of a failure $1 - p_0$.

Formulas: The probability for a specific point in this distribution is calculated using the formula:

$$P(m) = \binom{n}{m} p_0^m (1 - p_0)^{n-m},$$

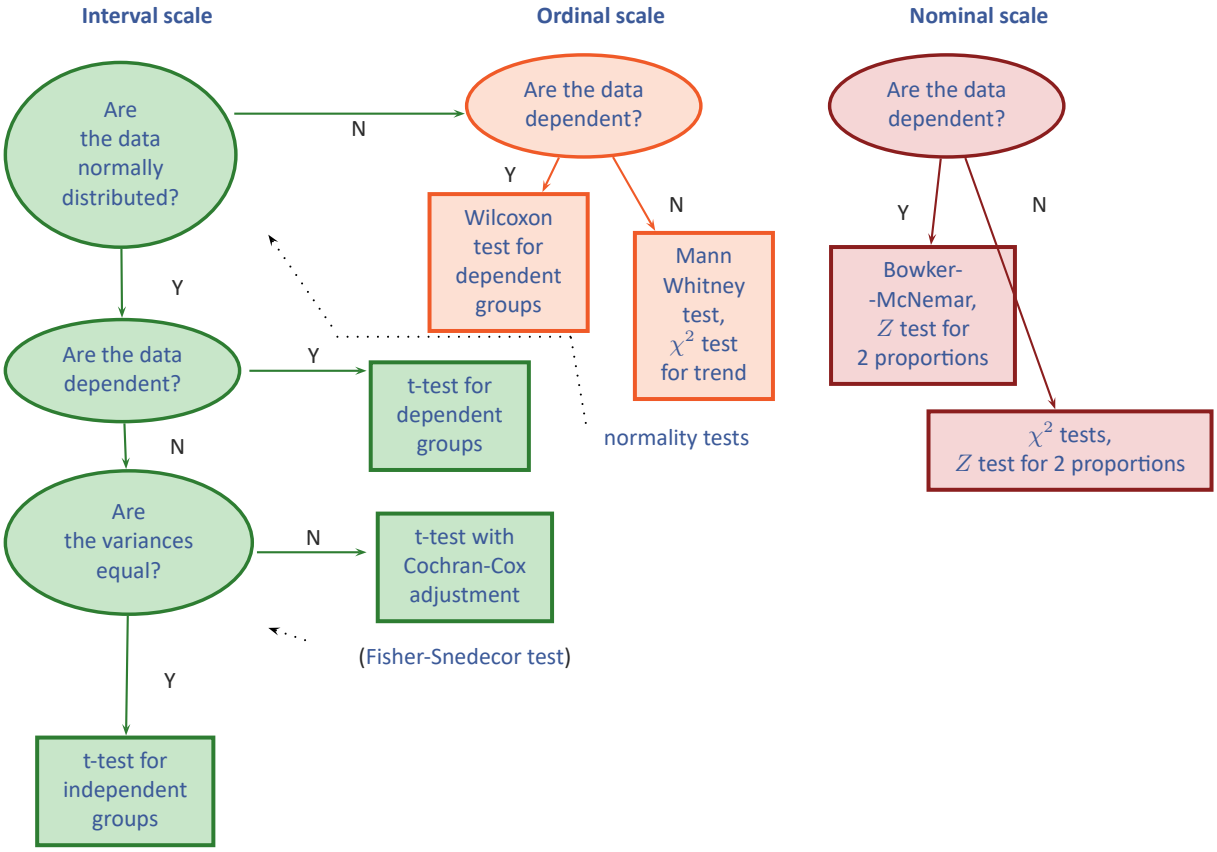
Copyright ©2010-2026 U-smile Stats – All rights reserved

where:
$\binom{n}{m} = \frac{n!}{m!(n-m)!}$,
m – the count of the distinguished values in the sample,
n – the sample size.

Interpretation: On the basis of the sum of the appropriate probabilities P, a one-sided and a two-sided p value are determined, with the two-sided p value defined as twice the smaller of the one-sided probabilities. The p value is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

16 COMPARISON - TWO GROUPS



Copyright ©2010-2026 U-smile Stats – All rights reserved

16.1 PARAMETRIC TESTS

The Fisher-Snedecor test

WHEN TO USE THIS TEST:

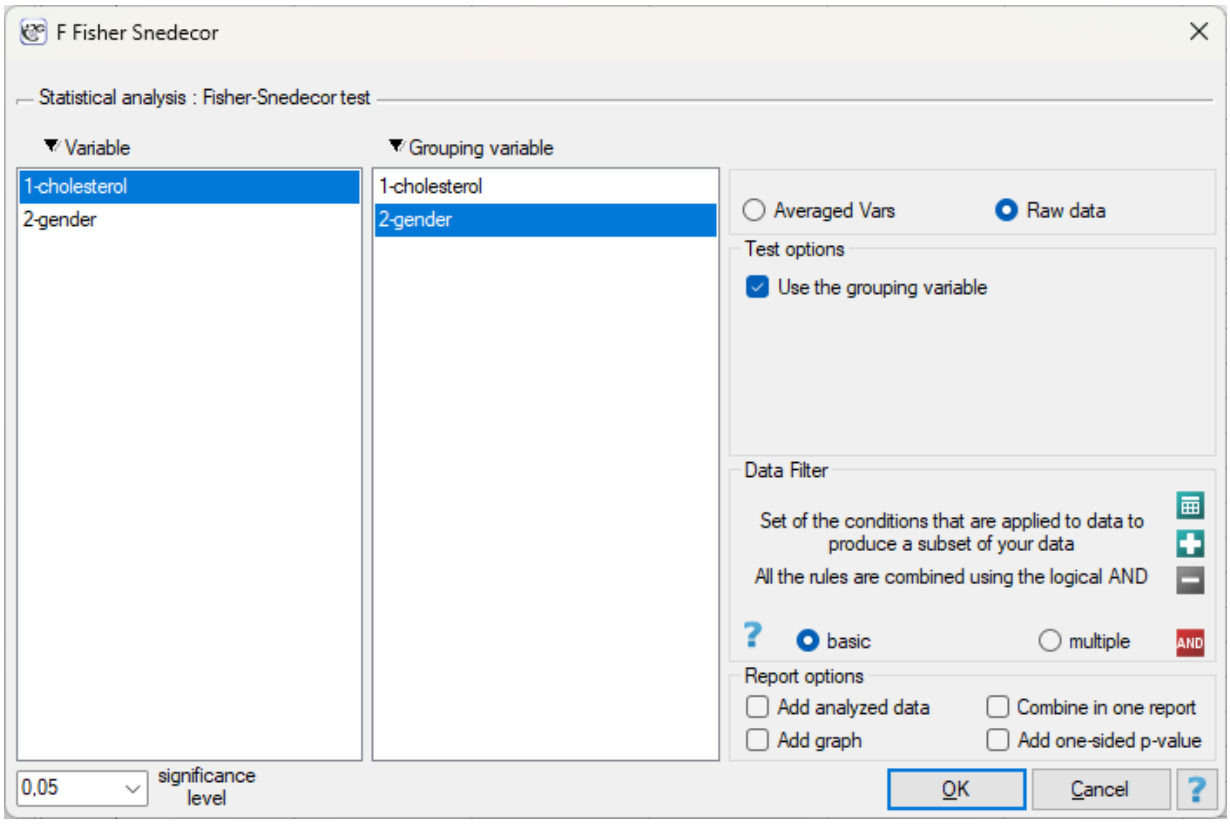
The Fisher-Snedecor test checks whether two compared groups have similarly varied results (similar variability). It is usually used as a preliminary step before the t-test for independent groups – its result indicates whether to use the basic version of the t-test, or its variant with the Cochran-Cox correction.

ASSUMPTIONS

- measurement on an interval scale,
- normality of distribution of the analysed variable in both populations,
- an independent model.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → F Fisher Snedecor.



Copyright ©2010-2026 U-smile Stats – All rights reserved

Step 1. Indicate the grouping variable (e.g. sex) and the outcome variable (e.g. total cholesterol level).

Step 2. Run the analysis.

Note! Calculations can be based on raw data or on averaged data, i.e. standard deviations and sample sizes.

PERFORMANCE AND INTERPRETATION

Example 16.1. (cholesterol.pqs file)
A sample of 500 women and a sample of 500 men, living in Poland and aged over 40, were randomly selected from their respective populations. The study concerned the assessment of cardiovascular disease risk. Total cholesterol level was among the analysed parameters. Before comparing the mean values of this parameter between women and men, it is first checked whether the variances of total cholesterol level in the two populations are equal.

The distribution of total cholesterol level in both groups is a normal distribution (this was checked with the Lilliefors test). The standard deviation in the group of men was $sd_1 = 47.6$ ($n_1 = 500$), and in the group of women $sd_2 = 43.5$ ($n_2 = 500$).

Fisher-Snedecor test	
Analysed variables	cholesterol
	gender
Number of unspecified	0
Number of missing data	0
Significance level	0,05
Grouping variable	gender
Group name	man
Group size	500
Group mean	191,456
Group standard deviation	43,514631
Group name	woman
Group size	500
Group mean	201,078
Group standard deviation	47,638711
F statistic	0,834354
Degrees of freedom	499 / 499
Two sided p-value	0,043351

$p = 0.0434$ **(statistically significant,** $\alpha = 0.05$**)**

The variances of total cholesterol level for women and men living in Poland differ significantly, although the difference is small. In the subsequent comparison of means, the t-test with the Cochran-Cox correction should therefore be used, rather than its basic version.

The p value, calculated on the basis of the test statistic, is always compared with the assumed significance level α: when $p \leq \alpha$, the difference in variances is statistically significant; when $p > \alpha$, there is no reason to claim that the variances differ. This decision determines which version of the t-test for independent groups should be used in the next step of the analysis.

Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \sigma_1^2 = \sigma_2^2, \\ \mathcal{H}_1: &\quad \sigma_1^2 \neq \sigma_2^2, \end{aligned}$$

where σ_1^2, σ_2^2 – variances of an analysed variable of the 1st and the 2nd population.

Formulas: The test statistic is defined by:

$$F = \frac{sd_1^2}{sd_2^2},$$

where sd_1^2, sd_2^2 – variances of an analysed variable of the samples chosen randomly from the 1st and the 2nd population. The test statistic has the F Snedecor distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

The t-test for independent groups

WHEN TO USE THIS TEST:

The t-test for independent groups is used to verify the hypothesis about the equality of means of an analysed variable in two populations – e.g. when comparing the value of some feature between two independent, unrelated groups (women and men, a treated group and a control group).

ASSUMPTIONS

- measurement on an interval scale,
- normality of distribution of the analysed variable in both populations,
- an independent model,

You do not need to check the equality of variances in both groups yourself – PQStat does this automatically and selects the appropriate version of the test (see *How to set up the analysis*).

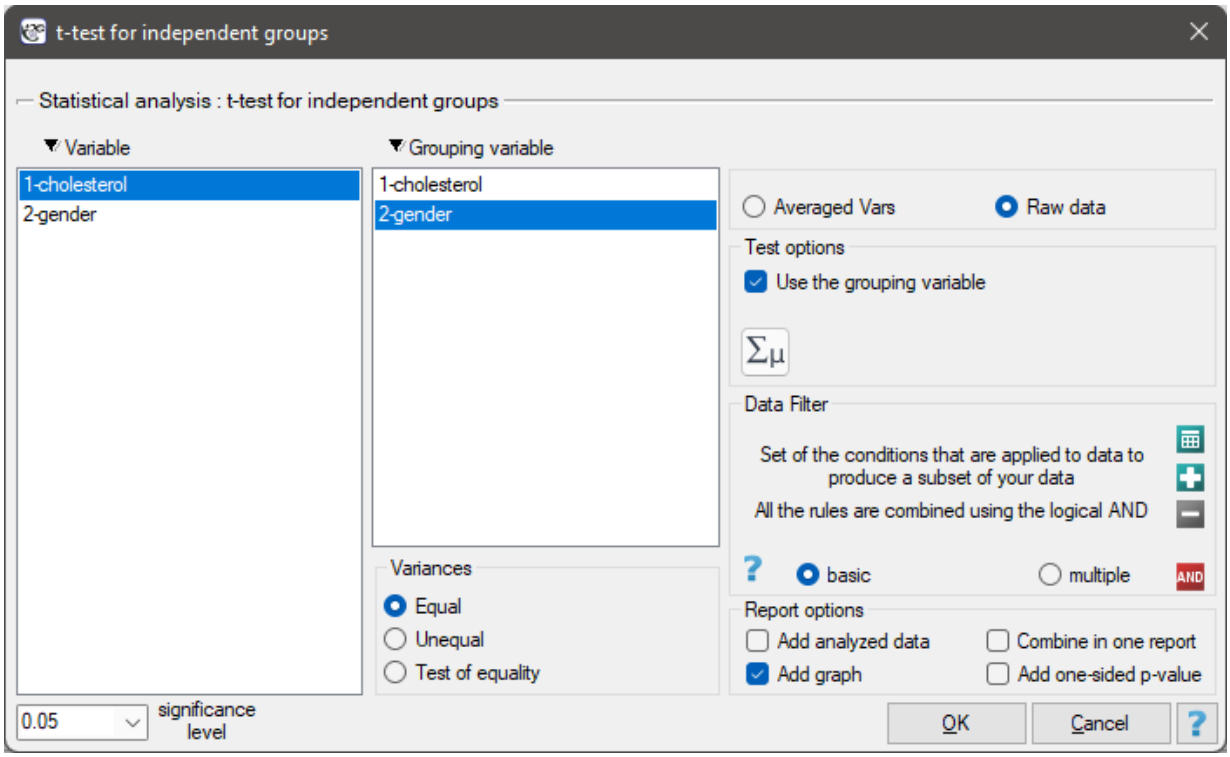
The t-test with the Cochran-Cox adjustment

The Cochran-Cox adjustment *(1957)*[44] is a variant of the t-test for independent groups, applied automatically when the variances of the analysed variable in the two populations differ significantly. The assumptions and purpose of the test are the same as for the basic variant – only the way the test statistic and the degrees of freedom are calculated is different (see *For those interested*).

HOW TO SET UP THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

Statistics → Parametric tests → t-test for independent groups (or via the Wizard).



Step 1. Indicate the grouping variable (e.g. sex) and the outcome variable (e.g. total cholesterol concentration).

Step 2. In the option concerning variances, choose one of the settings:

- equal – the t-test for independent groups will be calculated,
- different – the t-test with the Cochran-Cox adjustment will be calculated,
- check equality – the program will calculate the Fisher-Snedecor test itself and, on this basis, automatically select the appropriate variant of the test.

Step 3. Run the analysis.

Note! Calculations can be based on raw data or on averaged data, i.e. arithmetic means, standard deviations and sample sizes.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 16.2. (cholesterol.pqs file)
Five hundred subjects each were drawn from a population of women and a population of men living in Poland, over 40 years of age. The study concerned the assessment of cardiovascular disease risk. Among the parameters studied is the total cholesterol concentration. The purpose of this study is to

Copyright ©2010-2026 U-smile Stats – All rights reserved

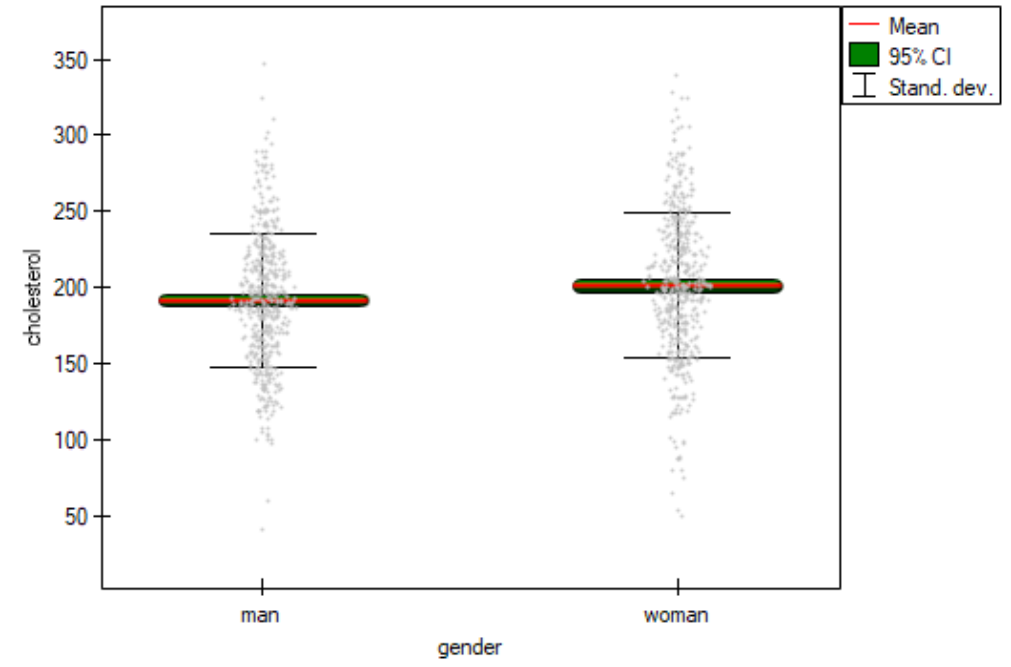
compare men and women with respect to this value – we want to show that these populations already differ at the level of total cholesterol concentration, and not only in the individual cholesterol fractions.

The distribution of age in both groups is a normal distribution (this was checked with the Lilliefors test). The mean total cholesterol concentration in the male group was $\overline{x}_1 = 201.1$, and the standard deviation $sd_1 = 47.6$; in the female group $\overline{x}_2 = 191.5$ and $sd_2 = 43.5$, respectively. The Fisher-Snedecor test indicates small, but statistically significant ($p = 0.0434$), differences in variances. The Student's t-test with the Cochran-Cox adjustment will be used in the analysis.

t-test for independent groups	
Analysed variables	cholesterol
	gender
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Correction for different variances	Yes
Grouping variable	gender
Difference of the means	-9.622
-95% CI for the difference	-15.2843
+95% CI for the difference	-3.9597
Standard error of the difference	2.8855
Pooled standard deviation	NA
t-statistic	-3.3346
Degrees of freedom	989.9277
Two sided p-value (Cochran-Cox)	0.0009
Fisher-Snedecor test	
Variance ratio F	0.8344
p-value	0.0434

Summary		
Group	man	woman
Sample size	500	500
Arithmetic mean	191.456	201.078
Standard error of the mean	1.946	2.1305
Standard deviation	43.5146	47.6387
-95% CI for the group mean	187.6326	196.8922
+95% CI for the group mean	195.2794	205.2638

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p = 0.0009$ **(statistically significant,** $\alpha = 0.05$**)**

Women and men living in Poland differ, in a statistically significant way, in their total cholesterol concentration. The average Polish man over 40 years of age has a higher total cholesterol concentration than the average Polish woman – the difference is almost 10 units.

The p value, calculated on the basis of the test statistic, is always compared with the assumed significance level α: if $p \leq \alpha$, the difference between the means is statistically significant; if $p > \alpha$, there are no grounds to claim that the groups differ. It is also worth reporting the **effect size** (Cohen's d) – statistical significance alone does not indicate how large the difference is; with large samples, even small differences can turn out to be statistically significant.

DEFINITIONS OF MEASURES

Cohen's d – a standardized effect-size measure that determines how large a part of the observed variation in the data is the observed difference between the means. Unlike the p value, it does not depend on sample size, which makes it possible to assess the practical significance of a difference, not just its statistical significance.

FOR THOSE INTERESTED

The t-test for independent groups

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2, \\ \mathcal{H}_1: &\quad \mu_1 \neq \mu_2. \end{aligned}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where μ_1, μ_2 – means of the analysed variable in the first and the second population.

Formulas:

The test statistic is defined by:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1-1)sd_1^2 + (n_2-1)sd_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

where $\overline{x}_1, \overline{x}_2$ – means of the 1st and the 2nd sample, n_1, n_2 – the 1st and the 2nd sample size, sd_1^2, sd_2^2 – variances of the 1st and the 2nd sample. The test statistic has the t-Student distribution with $df = n_1 + n_2 - 2$ degrees of freedom.

Pooled standard deviation:

$$SD_p = \sqrt{\frac{(n_1-1)sd_1^2 + (n_2-1)sd_2^2}{n_1+n_2-2}},$$

Standard error of the difference of means:

$$SE_{\overline{x}_1-\overline{x}_2} = \sqrt{\frac{(n_1-1)sd_1^2 + (n_2-1)sd_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

Cohen's d formula:

Cohen's d is calculated using the formula:

$$d = \left|\frac{\overline{x}_1 - \overline{x}_2}{SD_p}\right|$$

When interpreting an effect, researchers often use general guidelines proposed by Cohen [47], defining a small (0.2), medium (0.5) and large (0.8) effect size.

The t-test with the Cochran-Cox adjustment

Formulas:

The test statistic is defined by:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}}.$$

The test statistic has the t-Student distribution with the degrees of freedom proposed by Satterthwaite (1946)[149] and calculated using the formula:

$$df = \frac{\left(\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}\right)^2}{\left(\frac{sd_1^2}{n_1}\right)^2 \cdot \frac{1}{(n_1-1)} + \left(\frac{sd_2^2}{n_2}\right)^2 \cdot \frac{1}{(n_2-1)}}.$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

The t-test for dependent groups

WHEN TO USE THIS TEST:

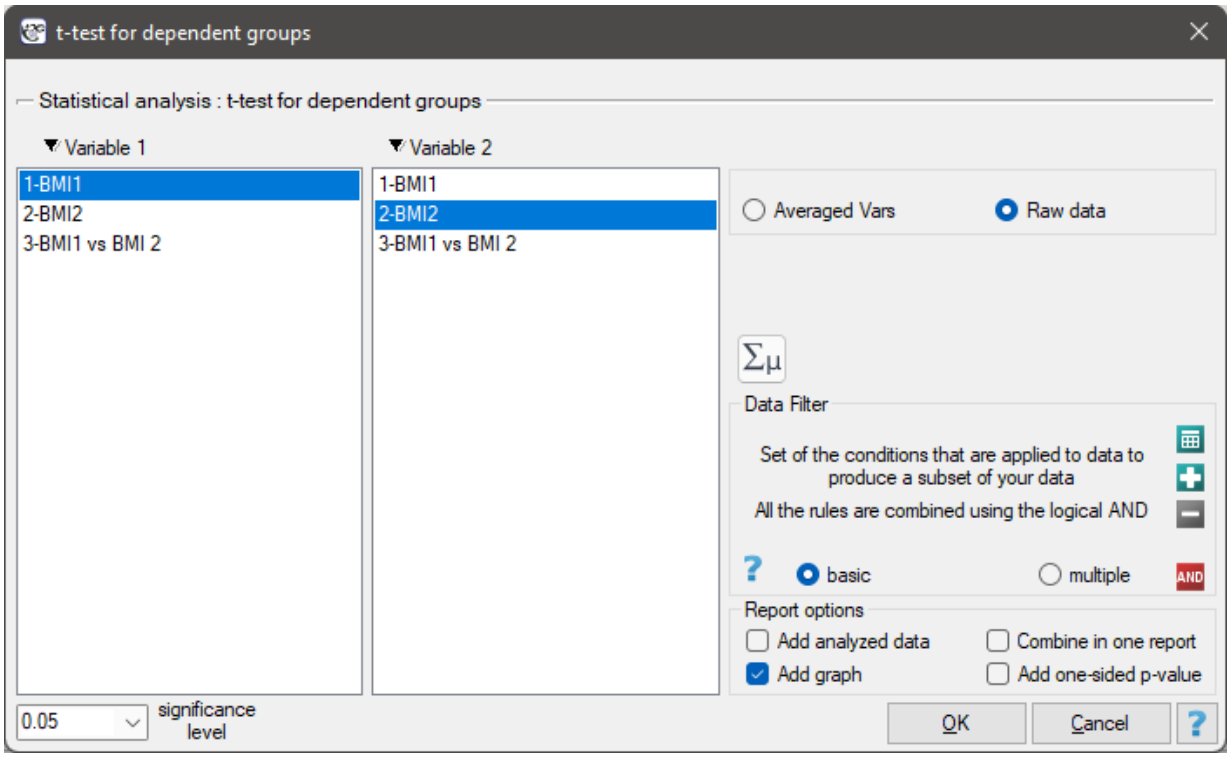
The t-test for dependent groups is used when the same characteristic is measured twice for the same subjects, under different conditions (e.g. before and after a diet), and we want to check whether the measurement changed.

ASSUMPTIONS

- measurement on an interval scale,
- normality of distribution of the differences d_i,
- a dependent model.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → t-test for dependent groups (or via the Wizard).



Step 1. Indicate the two variables corresponding to the first and second measurement (e.g. BMI before the diet and BMI after the diet).

Step 2. Run the analysis.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! Calculations can be based on raw data or on averaged data, i.e. the mean of the differences, the standard deviation of the differences, and the sample size.

PERFORMING AND INTERPRETING THE ANALYSIS

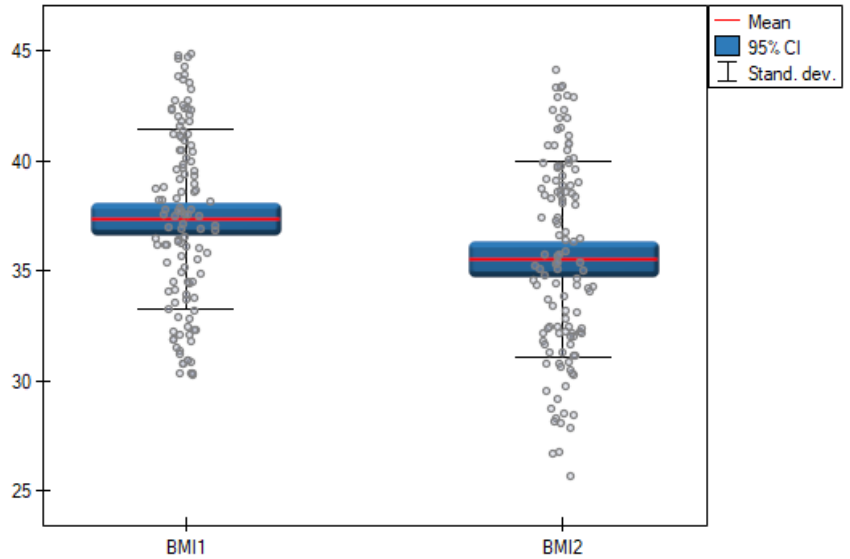
EXAMPLE 16.3. (BMI.pqs file)
A clinic treating eating disorders studied the effect of a recommended "diet A" on body weight change. A sample of 120 obese patients were put on the diet. Their BMI levels were measured twice: before the diet and after 180 days of following it. The aim of the study was to compare the obtained BMI measurements in order to assess the effectiveness of the diet.

The distribution of the difference between the pairs of measurements was a normal distribution (Lilliefors test, $p = 0.0837$), which allowed the t-test for dependent groups to be applied.

t-test for dependent groups	
Analysed variables	BMI1
	BMI2
Number of unspecified	0
Number of missing data (pairs)	1
Significance level	0.05
Number of pairs	120
Mean of the difference	1.8219
-95% CI for the mean difference	1.4136
+95% CI for the mean difference	2.2303
Standard error of the mean difference	0.2062
SD of difference	2.2592
t-statistic	8.8342
Degrees of freedom	119
Two sided p-value	<0.0001

Summary		
Group	BMI1	BMI2
Sample size	120	120
Arithmetic mean	37.3532	35.5313
Standard error of the mean	0.3703	0.4083
Standard deviation	4.0567	4.4727
-95% CI for the group mean	36.6199	34.7228
+95% CI for the group mean	38.0864	36.3397

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p < 0.0001$ **(statistically significant,** $\alpha = 0.05$**)**

The mean BMI of the studied patients changed significantly under the influence of the diet – before the diet it was higher on average by less than 2 units.

The p value, calculated on the basis of the test statistic, is always compared with the assumed significance level α: when $p \leq \alpha$, the change is statistically significant; when $p > \alpha$, there is no reason to claim that a change occurred. It is also worth reporting the **effect size** (Cohen's d) – statistical significance alone does not indicate how large the change is; with large samples, even small differences may turn out to be statistically significant.

DEFINITIONS OF MEASURES

Cohen's d – a standardized effect-size measure that determines how large a part of the observed variability in the data is the observed difference between the means. Unlike the p value, it does not depend on sample size, which makes it possible to assess the practical significance of a difference, not just its statistical significance.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_0 = 0, \\ \mathcal{H}_1: &\quad \mu_0 \neq 0, \end{aligned}$$

where μ_0 – mean of the differences d_i in the population.

Formulas:

The test statistic is defined by:

$$t = \frac{\overline{d}}{sd_d}\sqrt{n},$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where $\overline{d}$ – mean of the differences d_i in the sample, sd_d – standard deviation of the differences d_i in the sample, n – number of differences d_i in the sample. The test statistic has the t-Student distribution with $n-1$ degrees of freedom.

Standard deviation of the difference:

$$sd_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}},$$

Standard error of the mean of differences:

$$SEM_d = \frac{SD_d}{\sqrt{n}}.$$

Cohen's d formula:

Cohen's d is calculated using the formula:

$$d = \left|\frac{\overline{d}}{sd_d}\right|$$

When interpreting an effect, researchers often use general guidelines proposed by Cohen [47] defining small (0.2), medium (0.5) and large (0.8) effect sizes.

Copyright ©2010-2026 U-smile Stats – All rights reserved

16.2 NON-PARAMETRIC TESTS

The Mann-Whitney U test

WHEN TO USE THIS TEST:

The Mann-Whitney U test is used when comparing two independent groups on the same characteristic, and the data do not meet the assumptions of the t-test for independent groups (e.g. they are not normally distributed). It checks whether the results in the two groups differ (assuming that the distributions of the variable in both groups are fairly similar – the comparison of rank variances can be checked with the Conover rank test).

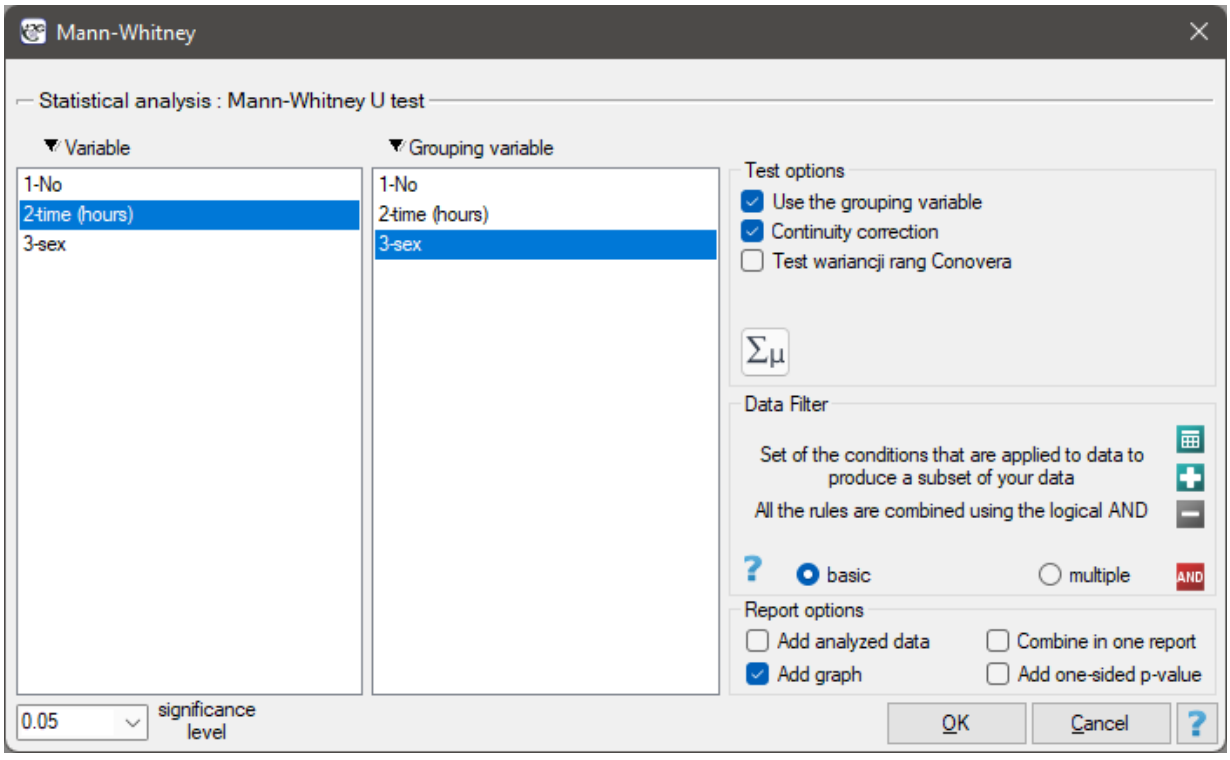
The test is also known as the Wilcoxon Mann-Whitney test (Mann and Whitney (1947)[114] and Wilcoxon (1949)[178]).

ASSUMPTIONS

- measurement on an ordinal scale or an interval scale,
- an independent model.

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Mann-Whitney (or via the Wizard).



Copyright ©2010-2026 U-smile Stats – All rights reserved

Step 1. Indicate the grouping variable (e.g. sex) and the outcome variable (e.g. time spent in front of the computer).

Step 2. Run the analysis.

PERFORMANCE AND INTERPRETATION

Example 16.4. (computer.pqs file)
It was hypothesised that at a certain university, male math students spend statistically more time in front of a computer screen than female math students. To verify this, a sample of 54 people (25 women and 29 men) was drawn from the population of people studying math at this university. These people were asked how many hours a day they spend in front of the computer screen. The following results were obtained: (time, sex): (2, k) (2, m) (2, m) (3, k) (3, k) (3, k) (3, k) (3, m) (3, m) (4, k) (4, k) (4, k) (4, k) (4, m) (4, m) (5, k) (5, k) (5, k) (5, k) (5, k) (5, k) (5, k) (5, k) (5, m) (5, m) (5, m) (5, m) (6, k) (6, k) (6, k) (6, k) (6, k) (6, m) (6, m) (6, m) (6, m) (6, m) (6, m) (6, m) (7, k) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (8, k) (8, m) (8, m).

Mann-Whitney U test	
Analysed variables	time (hours)
	sex
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Grouping variable	sex
Mann-Whitney	
U statistic	225.5
U statistic'	499.5
Two sided p-value (exact)	0.0149
Z statistic (adjusted for ties)	2.413
Two sided p-value (asymptotic)	0.0158

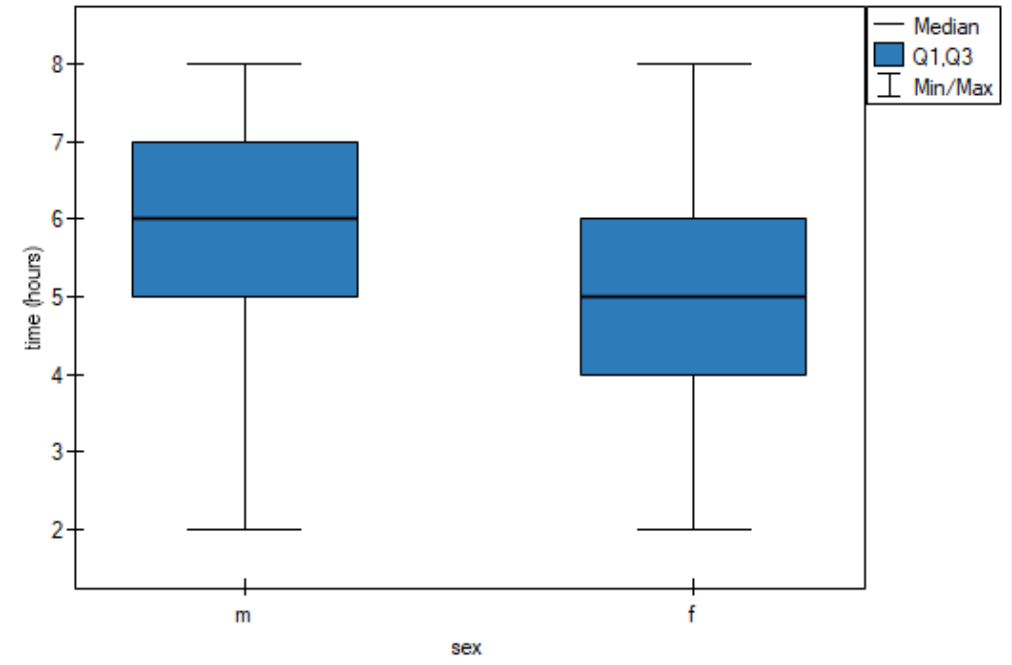
Summary		
Group	m	f
Sample size	29	25
Median	6	5
Minimum	2	2
Maximum	8	8
Lower quartile	5	4
Upper quartile	7	6

$p = 0.0149$ (exact U statistic, statistically significant, $\alpha = 0.05$)

There are statistically significant differences between female and male math students in the amount of time spent in front of the computer – female students spend less time in front of the computer than male students.

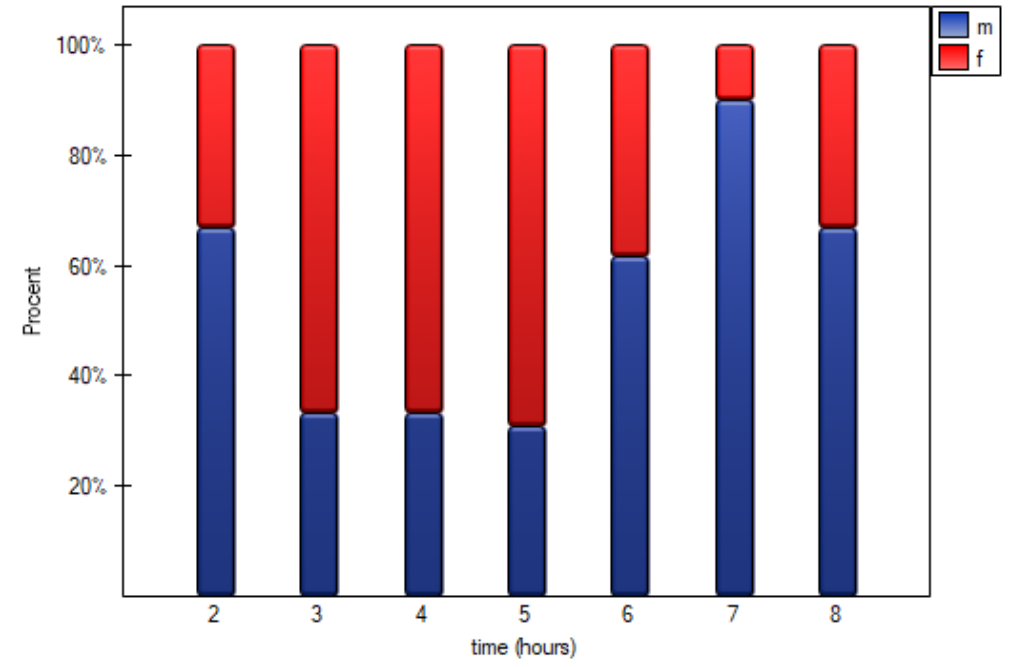
Copyright ©2010-2026 U-smile Stats – All rights reserved

The p value, calculated on the basis of the test statistic, is always compared with the assumed significance level α: when $p \leq \alpha$, the difference is statistically significant; when $p > \alpha$, there is no reason to claim that the groups differ. The same decision would be reached on the basis of the Z statistic without the continuity correction ($p = 0.0154$) and with it ($p = 0.0158$). The difference can be described using the median, quartiles, and the largest and smallest values, seen in a box-and-whisker plot, or using a table of counts and percentages (run in the analysis window by setting descriptive statistics $\Sigma\mu$) or a column plot. It is also worth reporting the **effect size** (Cohen's d) – statistical significance alone does not indicate how large the difference is.



Frequency(Percent)		
Group	m	f
2	2 (6.897%)	1 (4%)
3	2 (6.897%)	4 (16%)
4	2 (6.897%)	4 (16%)
5	4 (13.793%)	9 (36%)
6	8 (27.586%)	5 (20%)
7	9 (31.034%)	1 (4%)
8	2 (6.897%)	1 (4%)
summary	29	25

Copyright ©2010-2026 U-smile Stats – All rights reserved



DEFINITIONS OF MEASURES

r (effect size) – a measure of effect strength based on a transformation of the test statistic Z; it ranges from 0 (no effect) to 1 (maximum effect).

Cohen's d – a standardized effect-size measure, calculated here from r following the standard conversion used in meta-analyses. It makes it possible to assess the practical significance of a difference regardless of sample size.

FOR THOSE INTERESTED

Hypotheses:
Hypotheses concern the equality of mean ranks for the compared populations, or are simplified to medians:

$$\begin{aligned} \mathcal{H}_0: &\quad \phi_1 = \phi_2, \\ \mathcal{H}_1: &\quad \phi_1 \neq \phi_2, \end{aligned}$$

where ϕ_1, ϕ_2 – distributions of an analysed variable in the 1st and the 2nd population.

Formulas:
Depending on the sample size, the test statistic is calculated using different formulas. **For a small sample size:**

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,$$

or

$$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2,$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where n_1, n_2 are sample sizes, R_1, R_2 are rank sums for the samples. Standardly, the smaller of U or U' is interpreted. This statistic has the Mann-Whitney distribution and does not contain any correction for ties. The value of the exact probability of the Mann-Whitney distribution is calculated with the accuracy up to the hundredth place of the fraction. **For a large sample size:**

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n1+n_2+1)}{12} - \frac{n_1 n_2 \sum(t^3-t)}{12(n_1+n_2)(n_1+n_2-1)}}},$$

where U can be replaced with U', t – number of cases included in a tie. The formula for the Z statistic includes the correction for ties, used when ties occur (if there are none, the correction is not calculated, since $\frac{n_1 n_2 \sum(t^3-t)}{12(n_1+n_2)(n_1+n_2-1)} = 0$). The Z statistic asymptotically (for large sample sizes) has the normal distribution. **The Mann-Whitney test with the continuity correction** (Marascuilo and McSweeney (1977)[118]) – used to guarantee that the test statistic can take all real-number values, in line with the normal-distribution assumption:

$$Z = \frac{\left|U - \frac{n_1 n_2}{2}\right| - 0.5}{\sqrt{\frac{n_1 n_2 (n1+n_2+1)}{12} - \frac{n_1 n_2 \sum(t^3-t)}{12(n_1+n_2)(n_1+n_2-1)}}}.$$

Formulas (effect size measures):
r is calculated using the formula $r = |Z/(n_1 + n_2)|$ [68], and Cohen's d:

$$d = \frac{2r}{\sqrt{1-r^2}}$$

When interpreting an effect, researchers often use general guidelines proposed by Cohen [47] defining small (0.2), medium (0.5) and large (0.8) effect sizes.

The Wilcoxon test (matched-pairs)

WHEN TO USE THIS TEST:

The Wilcoxon matched-pairs test is used when the same characteristic is measured twice for the same subjects, under different conditions (e.g. before and after treatment), and the data do not meet the assumptions of the t-test for dependent groups (e.g. they are not normally distributed). It checks whether the measurement changed.

The test is also called the Wilcoxon test for dependent groups (Wilcoxon 1945[**?**], 1949[**?**]) – it is the extension for two dependent samples of the Wilcoxon test (signed-ranks), designed for one sample.

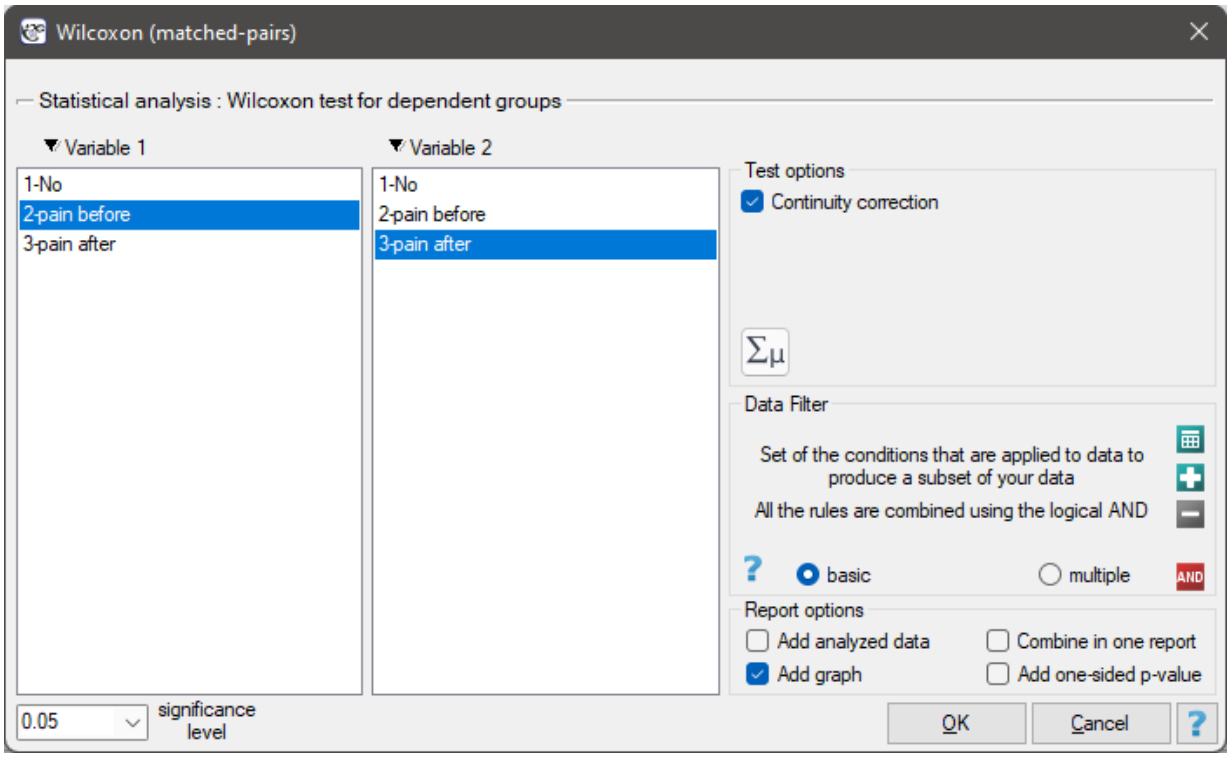
ASSUMPTIONS

- measurement on an ordinal scale or an interval scale,
- a dependent model.

Copyright ©2010-2026 U-smile Stats – All rights reserved

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Wilcoxon (matched-pairs) (or via the Wizard).



Step 1. Indicate the two variables corresponding to the first and second measurement (e.g. pain level before and after treatment).

Step 2. Run the analysis.

Note! From version 1.8.0, the median calculated for the difference column includes all pairs of results except those with a difference of 0.

PERFORMANCE AND INTERPRETATION

Example 16.5. (pain.pqs file)
A sample of 22 patients suffering from cancer was drawn. They were examined to check their level of felt pain (1–10 scale, where 1 means no pain and 10 means the greatest pain). This examination was repeated after a month of treatment with a new medicine, meant to lower the level of felt pain. The following results were obtained: (pain before, pain after): (2, 2) (2, 3) (3, 1) (3,1) (3, 2) (3, 2) (3, 3) (4, 1) (4, 3) (4, 4) (5, 1) (5, 1) (5, 2) (5, 4) (5, 4) (6, 1) (6, 3) (7, 2) (7, 4) (7, 4) (8, 1) (8, 3). We want to check whether the treatment had an effect on the level of felt pain in the population from which the sample was drawn.

Copyright ©2010-2026 U-smile Stats – All rights reserved

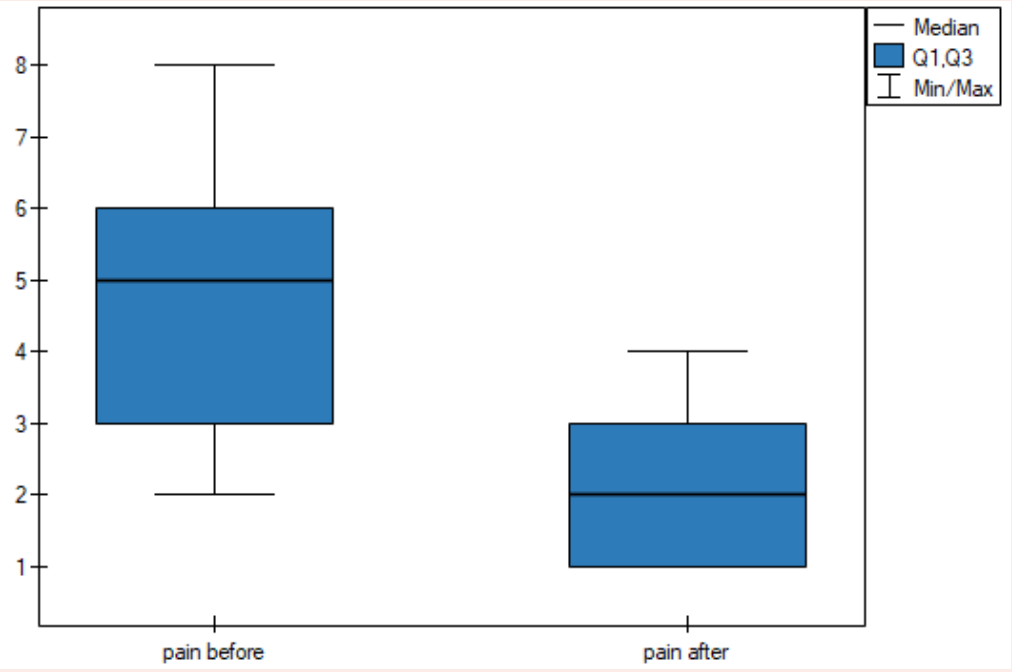
Wilcoxon test for dependent groups	
Analysed variables	pain before
	pain after
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Number of pairs	22
Number of omitted pairs (equal values)	3
Median of the difference	3
T statistic	3.5
Two sided p-value (exact)	0.0001
Z statistic (adjusted for ties)	3.6849
Two sided p-value (asymptotic)	0.0002

Summary		
Group	pain before	pain after
Sample size	22	22
Median	5	2
Minimum	2	1
Maximum	8	4
Lower quartile	3	1
Upper quartile	6	3
Sum of positive ranks	3.5	
Sum of negative ranks	186.5	

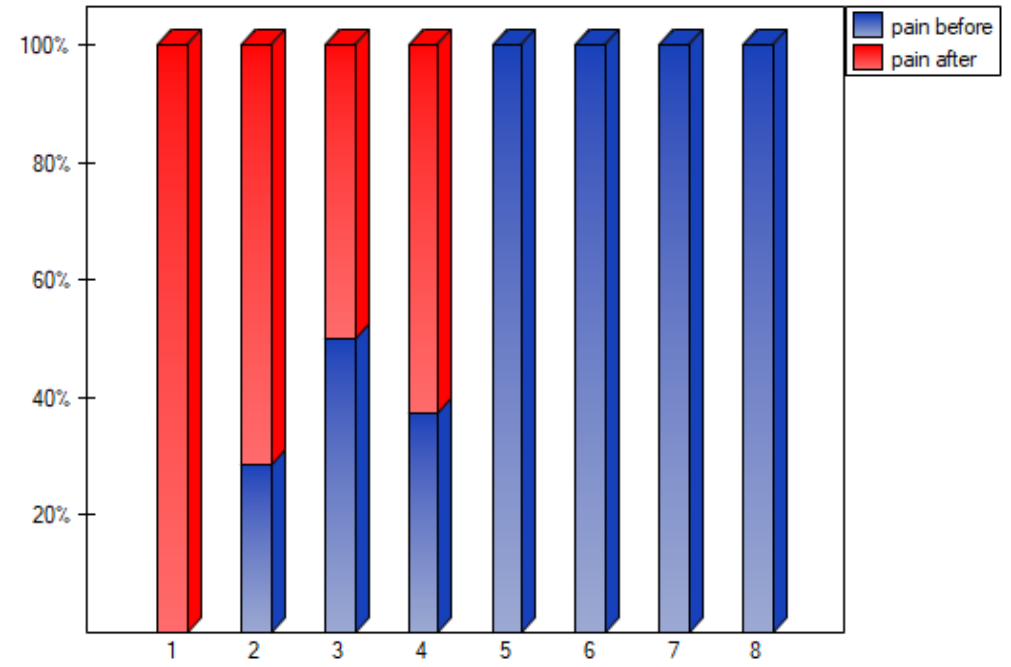
$p = 0.000211$ **(statistically significant,** $\alpha = 0.05$**)**

In the analysed population, the level of felt pain changed significantly after a month of treatment – it decreased (the sum of the negative ranks is significantly greater than the sum of the positive ranks).

The p value, calculated on the basis of the test statistic, is always compared with the assumed significance level α: when $p \leq \alpha$, the change is statistically significant; when $p > \alpha$, there is no reason to claim that a change occurred. The same decision would be reached on the basis of $p = 0.00021$ or $p = 0.000229$ of the Wilcoxon test based on the Z statistic (without and with the continuity correction, respectively). The difference can be seen in a box-and-whisker plot or a column plot. It is also worth reporting the **effect size** (Cohen's d) – statistical significance alone does not indicate how large the change is.



Copyright ©2010-2026 U-smile Stats – All rights reserved



DEFINITIONS OF MEASURES

r (effect size) – a measure of effect strength based on a transformation of the test statistic Z; it ranges from 0 (no effect) to 1 (maximum effect).

Cohen's d – a standardized effect-size measure, calculated here from r following the standard conversion used in meta-analyses. It makes it possible to assess the practical significance of a difference regardless of sample size.

FOR THOSE INTERESTED

Hypotheses: concern the equality of the sums of positive and negative ranks, or are simplified to medians:

$$\begin{aligned} \mathcal{H}_0: &\quad \theta_0 = 0, \\ \mathcal{H}_1: &\quad \theta_0 \neq 0, \end{aligned}$$

where θ_0 – median of the differences d_i in a population.

Formulas: depending on the sample size, the test statistic is calculated using different formulas. **For a small sample size:**

$$T = \min\left(\sum R_-, \sum R_+\right),$$

where $\sum R_+$ – sum of positive ranks, $\sum R_-$ – sum of negative ranks. This statistic has the Wilcoxon distribution and does not contain any correction for ties. **For a large sample size:**

$$Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}},$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where n – number of ranked signs (number of ranks), t – number of cases included in a tie. The formula for the Z statistic includes the correction for ties, used when ties occur (if there are none, the correction is not calculated, since $\frac{\sum t^3 - \sum t}{48} = 0$). The Z statistic asymptotically (for large sample sizes) has the normal distribution. **The Wilcoxon test with the continuity correction** (Marascuilo and McSweeney (1977)[118]) – used to guarantee that the test statistic can take all real-number values, in line with the normal-distribution assumption:

$$Z = \frac{\left|T - \frac{n(n+1)}{4}\right| - 0.5}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}}.$$

r is calculated using the formula $r = |Z/n|$ [68], and Cohen's d:

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

When interpreting an effect, researchers often use general guidelines proposed by Cohen [47] defining small (0.2), medium (0.5) and large (0.8) effect sizes.

The Chi-square tests

WHEN TO USE THESE TESTS:

The Chi-square tests check whether there is a dependence between two features gathered in a contingency table (e.g. whether sex is related to exam passing). Depending on the size of the table, we speak of the 2×2 chi-square test (two categories per feature) or the $R \times C$ chi-square test (more categories):

- chi-square test for small tables (2×2),
- chi-square test for large tables ($R \times C$).

ASSUMPTIONS

- measurement on a nominal scale – any ordering of categories is not taken into account,
- an independent model,
- large expected frequencies according to the Cochran interpretation (1952)[42].

Note! The p value of the test is always compared with the assumed significance level α: when $p \leq \alpha$, we conclude that the analysed features are related; when $p > \alpha$, there is no reason to claim that such a dependence exists.

ADDITIONALLY

Copyright ©2010-2026 U-smile Stats – All rights reserved

- When Cochran's condition is not satisfied, instead of the chi-square test one can use: Fisher's exact test for $R \times C$ tables, Fisher's exact test for 2×2 tables, the chi-square test with Yates correction, the mid-p test for 2×2 tables.
- When we obtain an $R \times 2$ table and the R categories can be ordered, it is possible to check for a trend: the chi-square test for trend.
- When, for a table larger than 2×2, a significant dependence or difference is detected and we want to locate it, multiple comparisons with an appropriate correction can be performed – in the test option window, select Multiple column comparisons (RxC).
- To describe the strength of the relationship between the analysed features, one can determine: measures of dependence.
- To describe, for 2×2 tables, the size of a risk factor's impact, one can determine: Odds Ratio (OR) and Relative Risk (RR).

FOR THOSE INTERESTED

Hypotheses (general form):

$$\begin{aligned} \mathcal{H}_0: &\quad O_{ij} = E_{ij} \text{ for all categories,} \\ \mathcal{H}_1: &\quad O_{ij} \neq E_{ij} \text{ for at least one category,} \end{aligned}$$

where O_{ij} – observed frequencies in a contingency table, E_{ij} – expected frequencies in a contingency table.

Hypotheses (independence test form):

$\mathcal{H}_0$: there is no dependence between the analysed features of the population (both classifications, according to feature X and feature Y, are statistically independent),
$\mathcal{H}_1$: there is a dependence between the analysed features of the population.

The Chi-square test for large tables

WHEN TO USE THIS TEST:

The χ^2 test for $r \times c$ tables, also known as Pearson's Chi-square test (Karl Pearson 1900), is used when we want to check whether two features with any number of categories (e.g. country of residence and education) are related. It is an extension of the χ^2 (goodness-of-fit) test to two features.

ASSUMPTIONS

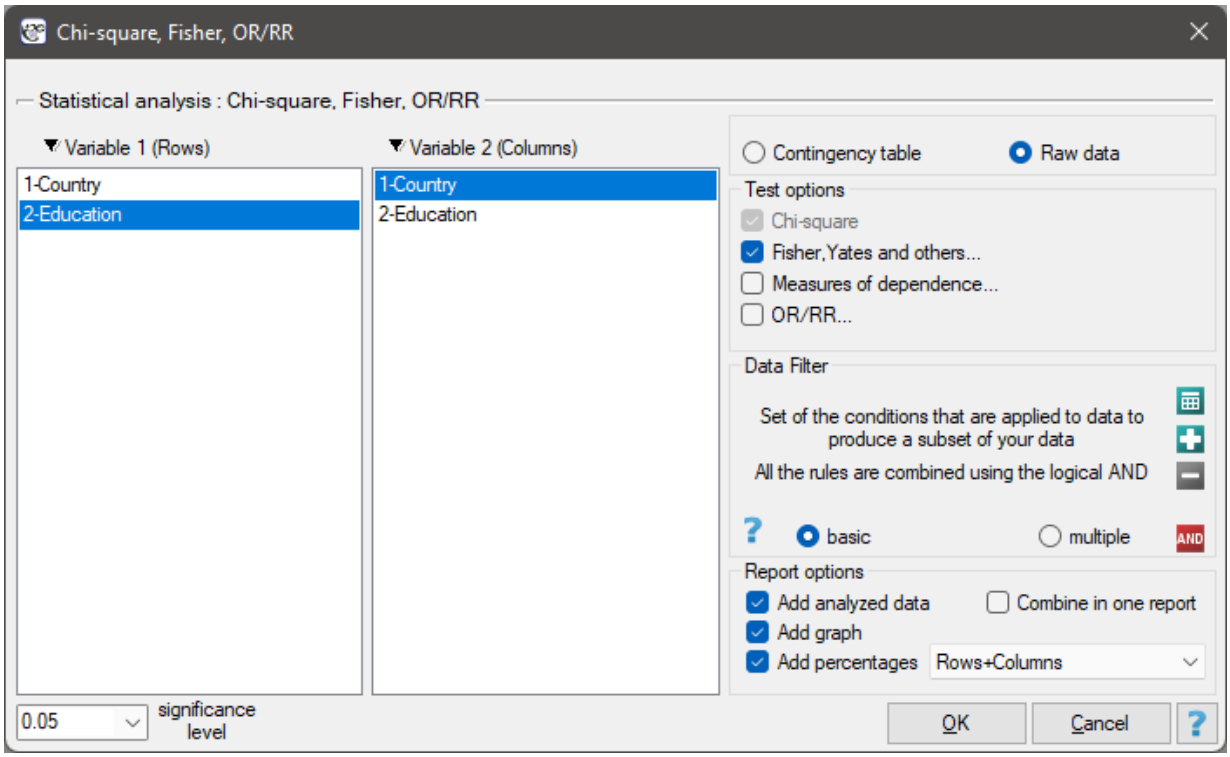
- measurement on a nominal scale,
- an independent model,

Copyright ©2010-2026 U-smile Stats – All rights reserved

– large expected frequencies according to the Cochran interpretation (1952)[42].

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Chi-square, Fisher, OR/RR (or via the Wizard).



Step 1. Indicate the two features whose dependence you want to check.

Step 2. Run the analysis.

PERFORMANCE AND INTERPRETATION

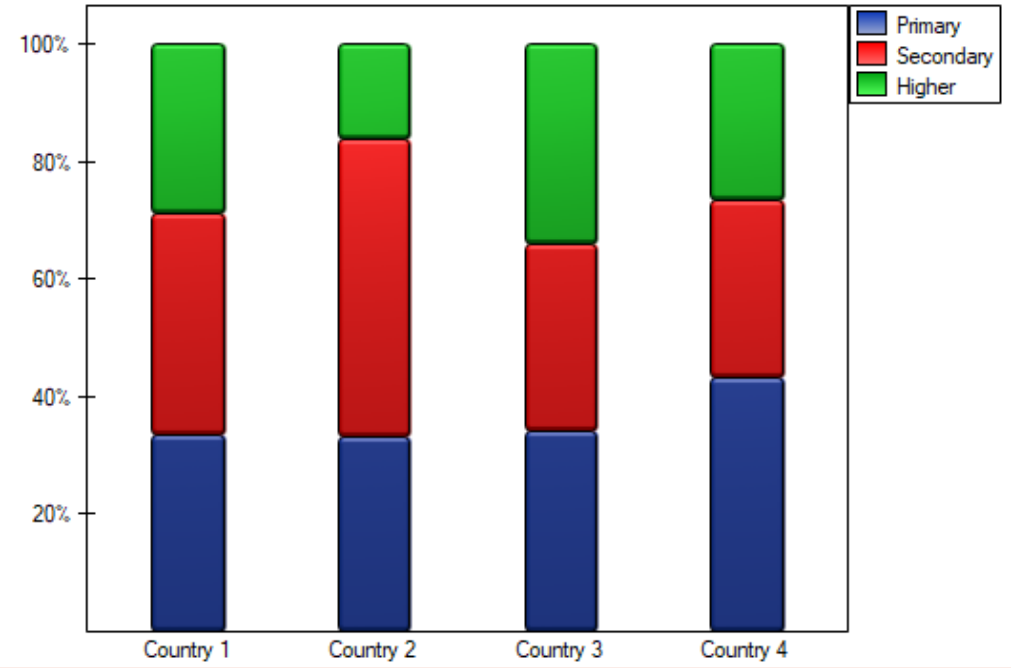
EXAMPLE 16.6. (country-education.pqs file)
There is a sample of 605 people ($n = 605$), for whom 2 features are analysed (X=country of residence, Y=education). The first feature occurs in 4 categories, and the second in 3 categories (X_1=Country 1, X_2=Country 2, X_3=Country 3, X_4=Country 4, Y_1=primary, Y_2=secondary, Y_3=higher). The data distribution is shown in the contingency table below:

Data :				✓Country	
↓Education	Country 1	Country 2	Country 3	Country 4	Summary
Primary	50	51	52	65	218
Secondary	56	78	48	45	227
Higher	43	25	52	40	160
Summary	149	154	152	150	

Copyright ©2010-2026 U-smile Stats – All rights reserved

Based on this sample, we would like to find out whether there is a dependence between education and country of residence in the analysed population.

Chi-square, Fisher, OR/RR	
Analysed variables	Country
	Education
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	605
Cochran condition	fulfilled
Pearson's chi-square statistic	23.6104
Degrees of freedom	6
p-value	0.0006



$p = 0.0006$ **(statistically significant,** $\alpha = 0.05$**)**

There is a dependence between country of residence and education in the analysed population.

Cochran's condition is satisfied. If we are interested in more precise information about the detected dependence, we can obtain it by determining multiple comparisons through the Fisher, Yates and others... option, then Multiple column comparisons (RxC) and one of the corrections, e.g. Benjamini-Hochberg.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Multiple comp. Benjamini-Hochberg : p-value, Chi^2, *Cochran not.				
Country	Country 1	Country 2	Country 3	Country 4
Country 1		0.0314	0.4777	0.2555
Country 2	0.0314		0.0015	0.0028
Country 3	0.4777	0.0015		0.2555
Country 4	0.2555	0.0028	0.2555	

A closer look reveals that only the second country differs from the other countries in educational attainment in a statistically significant way.

FOR THOSE INTERESTED

Hypotheses:

$\mathcal{H}_0$: there is no dependence between the analysed features in the population,
$\mathcal{H}_1$: there is a dependence between the analysed features in the population.

Formulas: The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}.$$

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with a number of degrees of freedom calculated using the formula: $df = (r-1)(c-1)$.

The Fisher test for large tables

WHEN TO USE THIS TEST:

The Fisher test for $r \times c$ tables, also called the Fisher-Freeman-Halton test *(ang. Fisher-Freeman-Halton test)*, Freeman G.H., Halton J.H. (1951)[65], is used instead of the χ^2 test for $r \times c$ tables when the expected frequencies in the table are too small (Cochran's condition is not satisfied). It is an extension to $r \times c$ tables of the Fisher exact test, determining the exact probability of the occurrence of a particular distribution of numbers in the table, given a known n and fixed marginal sums.

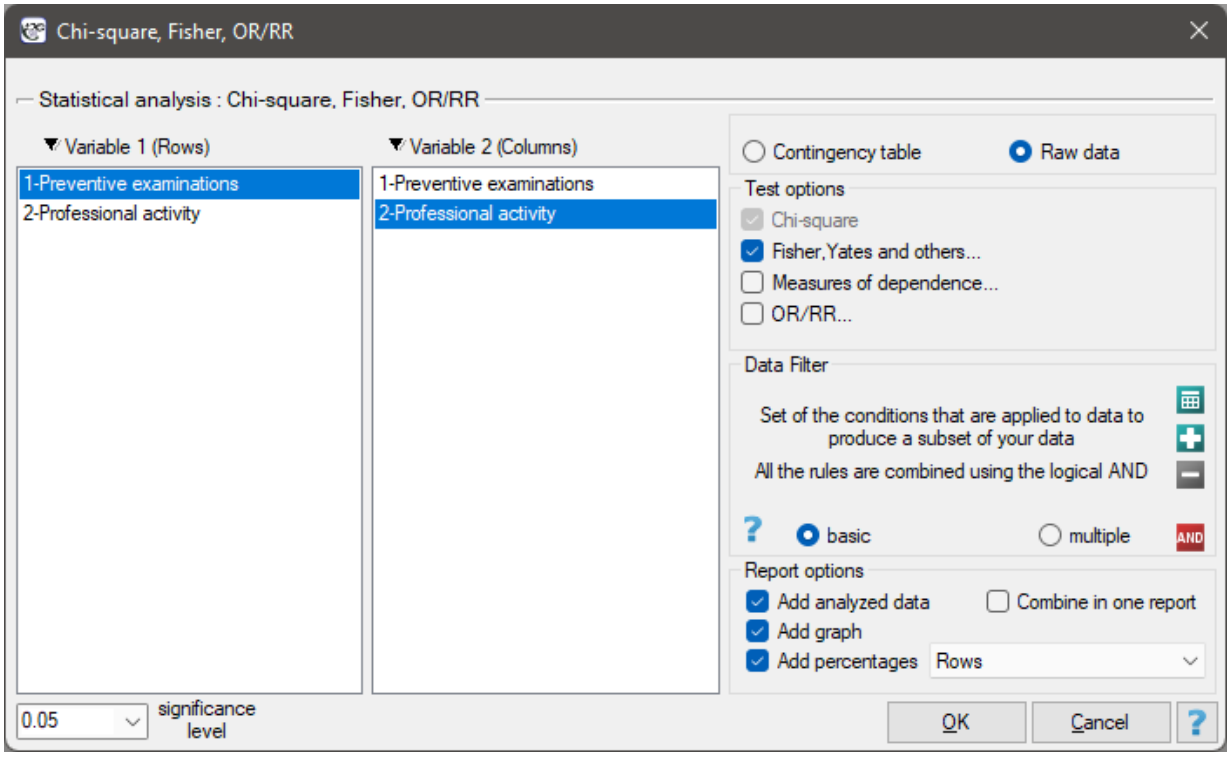
ASSUMPTIONS

- measurement on a nominal scale,
- an independent model,
- small expected frequencies (the χ^2 test is not advisable).

Copyright ©2010-2026 U-smile Stats – All rights reserved

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Chi-square, Fisher, OR/RR (or via the Wizard).



Step 1. Indicate the two features whose dependence you want to check.

Step 2. Run the analysis.

Note! The procedure for calculating the p value for this test is based on the algorithm published by Mehta (1986)[123].

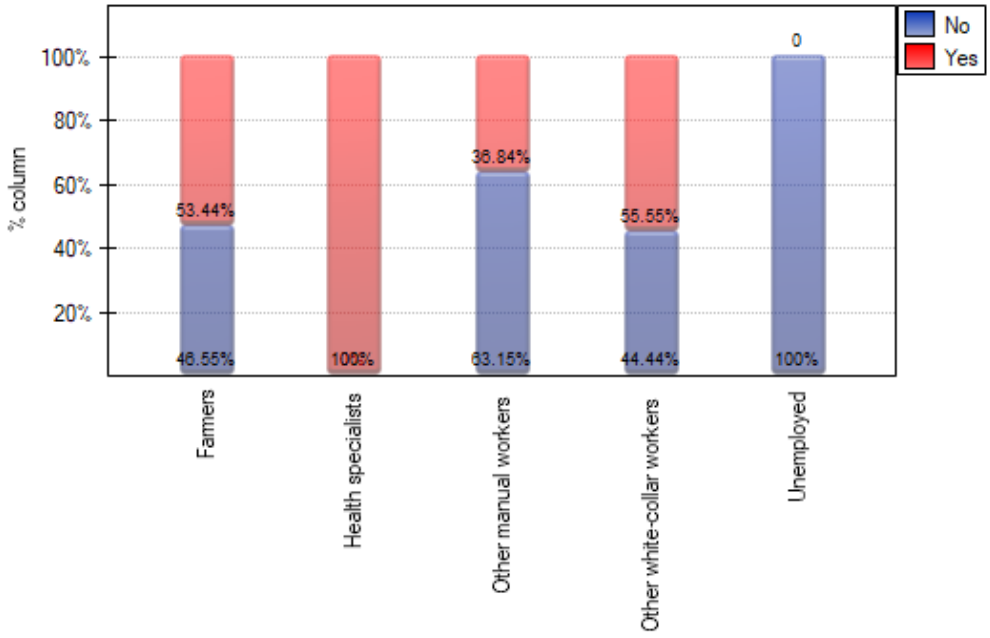
PERFORMANCE AND INTERPRETATION

EXAMPLE 16.7. (job-prevention.pqs file)
In the population of people living in the rural areas of Komorniki municipality, it was examined whether performing preventive health examinations depends on the type of occupational activity of the residents. A random sample of 120 people was collected and asked about their education and whether they perform preventive examinations. Complete answers were obtained from 113 people.

Data :					↙Professional activity	
↓Preventive examinations	Farmers	Health specialists	Other manual workers	Other white-collar workers	Unemployed	Summary
No	27	0	12	4	17	60
Yes	31	10	7	5	0	53
Summary	58	10	19	9	17	

Copyright ©2010-2026 U-smile Stats – All rights reserved

Chi-square, Fisher, OR/RR	
Analysed variables	Preventive examinations
	Professional activity
Number of unspecified	0
Number of missing data	7
Significance level	0.05
Size	113
Cochran condition	not fulfilled
Pearson's chi-square statistic	28.378
Degrees of freedom	4
p-value	<0.0001
Fisher exact test	
One sided p-value	NA
Two sided p-value	<0.0001



$p < 0.0001$ **(Fisher test, statistically significant,** $\alpha = 0.05$**)**

There is a dependence between performing preventive examinations and the type of work performed, in the analysed population.

Cochran's condition is not satisfied, so we should not use the chi-square test – hence the choice of the Fisher test. If we are interested in more precise information about the detected dependence, we can obtain it by determining multiple comparisons through the Fisher, Yates and others... option, then Multiple column comparisons (RxC) and one of the corrections, e.g. Benjamini-Hochberg.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Multiple comp. Benjamini-Hochberg : p-value, Chi^2/*Fisher(Cochran not.)					
Professional act	Farmers	Health speci	Other manu	Other white-	Unemployed
Farmers		0.0091*	0.2611	1*	0.0004
Health specialist	0.0091*		0.0039*	0.0464*	<0.0001*
Other manual w	0.2611	0.0039*		0.4799*	0.0139*
Other white-coll	1*	0.0464*	0.4799*		0.0048*
Unemployed	0.0004	<0.0001*	0.0139*	0.0048*	

A closer analysis allows us to conclude that health professionals perform preventive examinations significantly more often than the other groups (100% of people in this group performed examinations), and the unemployed significantly less often (no one in this group performed an examination). Farmers, other manual workers, and other white-collar workers take preventive examinations in about 50% of cases, which means these three groups do not differ from each other in a statistically significant way. p values in the table marked with an asterisk come from the Fisher exact test with the Benjamini-Hochberg correction, while unmarked values come from the chi-square test with the same correction (where Cochran's condition was satisfied).

FOR THOSE INTERESTED

Hypotheses:

$\mathcal{H}_0$: there is no dependence between the analysed features in the population,
$\mathcal{H}_1$: there is a dependence between the analysed features in the population.

Formulas: Marginal row sums are defined as:

$$W_i = \sum_{j=1}^{c} O_{ij},$$

where O_{ij} – observed frequencies in a contingency table, and marginal column sums as:

$$K_i = \sum_{i=1}^{r} O_{ij}.$$

Given fixed marginal sums, for various configurations of observed frequencies denoted as U_{ij}, we calculate the probabilities P:

$$P = \frac{D^{-1} \prod_{j=1}^{c} K_j!}{U_{1j}!U_{2j}! \ldots U_{rj}},$$

where

$$D = \frac{(W_1 + W_2 + \ldots + W_r)!}{W_1!W_2! \ldots W_r!}.$$

Interpretation: The exact significance level p is the sum of those P probabilities (calculated for new values U_{ij}) that are less than or equal to the P probability of the table with the initial values O_{ij}.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Chi-square test for small tables

WHEN TO USE THIS TEST:

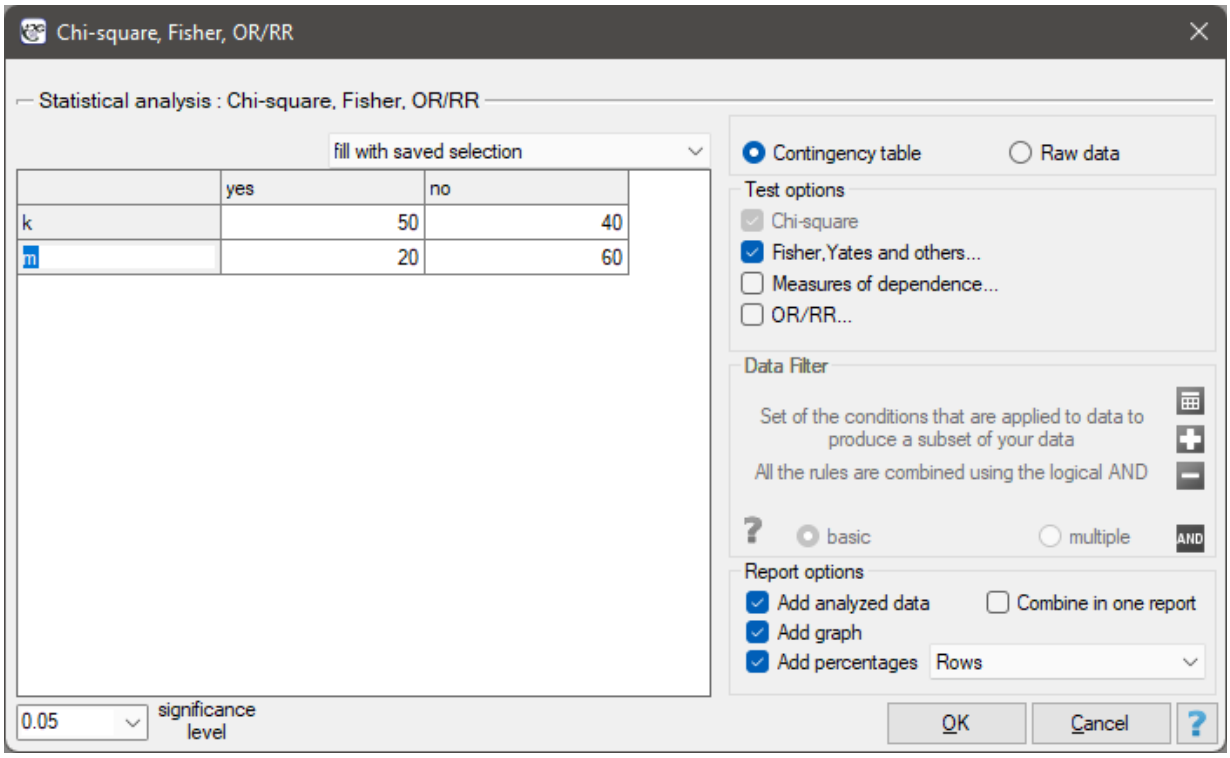
The χ^2 test for 2×2 tables (also known as Pearson's Chi-square test, Karl Pearson 1900) is used when we want to check whether two features with two categories each (e.g. sex and exam passing) are related. It is a special case of the χ^2 test for $r \times c$ tables for the smallest possible table.

ASSUMPTIONS

- measurement on a nominal scale,
- an independent model,
- a 2×2 contingency table (see table (11.1)).

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Chi-square, Fisher, OR/RR (or via the Wizard).



Step 1. Indicate the two features whose dependence you want to check.

Step 2. Run the analysis.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! When the expected frequencies in the table are small (below 5), it is better to base the conclusion on the test with Yates correction, Fisher's exact test, or the mid-p test, rather than the basic χ^2 test – all are calculated automatically together with the basic test.

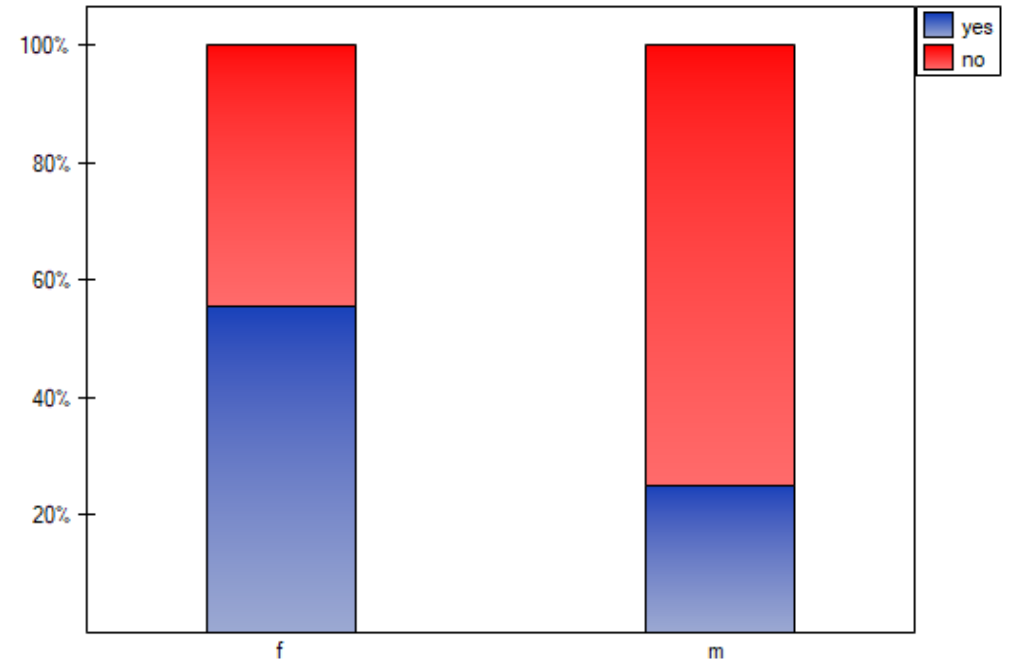
PERFORMANCE AND INTERPRETATION

EXAMPLE 16.8. (sex-exam.pqs file)
There is a sample of 170 people ($n = 170$), for whom 2 features are analysed (X=sex, Y=exam passing). Each of these features occurs in two categories (X_1=f, X_2=m, Y_1=yes, Y_2=no). Based on this sample, we would like to find out whether there is a dependence between sex and exam passing in the analysed population. The data distribution is shown in the contingency table below:

Observed frequencies O_{ij}		exam passing		
		yes	no	total
sex	f	50	40	90
	m	20	60	80
	total	70	100	170

Chi-square, Fisher, OR/RR	
Analysed variables	Contingency table
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	170
Cochran condition	fulfilled
Pearson's chi-square statistic	16.325396825
Degrees of freedom	1
p-value	0.000053344
Fisher exact test	
One sided p-value	0.000043395
Two sided p-value	0.00008324
Chi-square statistic with Yates correction	15.088258929
Degrees of freedom	1
p-value	0.000102599
mid-p test	
2 * one sided p-value	0.000054186

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p = 0.000053$ **(chi-square test, statistically significant, $\alpha = 0.05$)**

There is a dependence between sex and exam passing in the analysed population.

The table of expected frequencies contains no values below 5, so Cochran's condition is satisfied – the basic χ^2 test could be applied directly. The other test variants lead to the same conclusion: the test with Yates correction ($p = 0.000103$), Fisher's exact test ($p = 0.000083$), and the mid-p test ($p = 0.000054$).

FOR THOSE INTERESTED

Hypotheses:

$\mathcal{H}_0$: there is no dependence between the analysed features in the population,
$\mathcal{H}_1$: there is a dependence between the analysed features in the population.

Formulas: The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with 1 degree of freedom.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Chi-square test corrections for small tables

WHEN TO USE THESE VARIANTS:

When the expected frequencies in a 2×2 table are small (Cochran's condition is not satisfied), the basic χ^2 test may give an overly optimistic result. In such a situation it is better to base the conclusion on one of the variants below, more suitable for small frequencies. All of them are calculated automatically together with the basic test.

ASSUMPTIONS

- measurement on a nominal scale,
- an independent model,
- a 2×2 contingency table (see table (11.1)).

HOW TO SET UP THE ANALYSIS

Statistics $\rightarrow$ NonParametric tests $\rightarrow$ Chi-square, Fisher, OR/RR (or via the Wizard) – all variants appear in one report together with the basic test.

THE CHI-SQUARE TEST WITH THE YATES CORRECTION FOR CONTINUITY

The χ^2 test with the Yates correction *(ang. Chi-square test with Yates correction)*, Frank Yates (1934)[181], is more conservative than the χ^2 test – it rejects the null hypothesis more rarely than the χ^2 test. The correction for continuity guarantees that the test statistic can take all real-number values, in line with the χ^2 distribution assumption.

THE FISHER TEST FOR 2×2 TABLES

The Fisher test for 2×2 tables, also called the Fisher exact test *(ang. Fisher exact test)*, R. A. Fisher (1934[58], 1935[59]), determines the exact probability of the occurrence of a particular distribution of numbers in the table, given a known n and fixed marginal sums.

THE MID-P TEST

The mid-p test is a correction of the Fisher exact test, recommended by many statisticians (Lancaster 1961[101], Anscombe 1981[9], Pratt and Gibbons 1981[139], Plackett 1984[138], Miettinen 1985[124], Barnard 1989[15], Rothman 2008[144]) as a way of reducing the conservatism of the Fisher exact test – with the mid-p test we reject the null hypothesis sooner than with the exact Fisher test. For large samples, the p value of the χ^2 test with the Yates correction and of the Fisher test are similar, while the p value of the uncorrected χ^2 test corresponds to the mid-p value.

Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

The Chi-square test with the Yates correction – the test statistic is defined by:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}.$$

The Fisher test – given known marginal sums, for various configurations of observed frequencies we calculate the probabilities P:

$$P = \frac{\binom{O_{11}+O_{21}}{O_{11}}\binom{O_{12}+O_{22}}{O_{12}}}{\binom{O_{11}+O_{12}+O_{21}+O_{22}}{O_{11}+O_{12}}}.$$

The exact significance level p is the sum of those probabilities that are less than or equal to the analysed probability.

The mid-p test – the one-sided p value is calculated by transforming the probability value of the Fisher exact test:

$$p_{I(mid-p)} = p_{I(Fisher)} - 0.5 \cdot P_{punktu(tabeli \quad zadanej)},$$

where $p_{I(mid-p)}$ – one-sided p value of the mid-p test, $p_{I(Fisher)}$ – one-sided p value of the Fisher exact test. The two-sided p value is defined as twice the smaller of the one-sided probabilities:

$$p_{II(mid-p)} = 2p_{I(mid-p)}.$$

The Chi-square test for trend

WHEN TO USE THIS TEST:

The χ^2 test for trend, also called the Cochran-Armitage trend test *(ang. Cochran-Armitage test for trend)*[43][10], is used when we have ordered categories of one feature (e.g. education level) and two groups of the other feature (e.g. smoker/non-smoker), and we want to check whether the proportions across the ordered categories increase or decrease in a systematic way – i.e. whether a trend is visible.

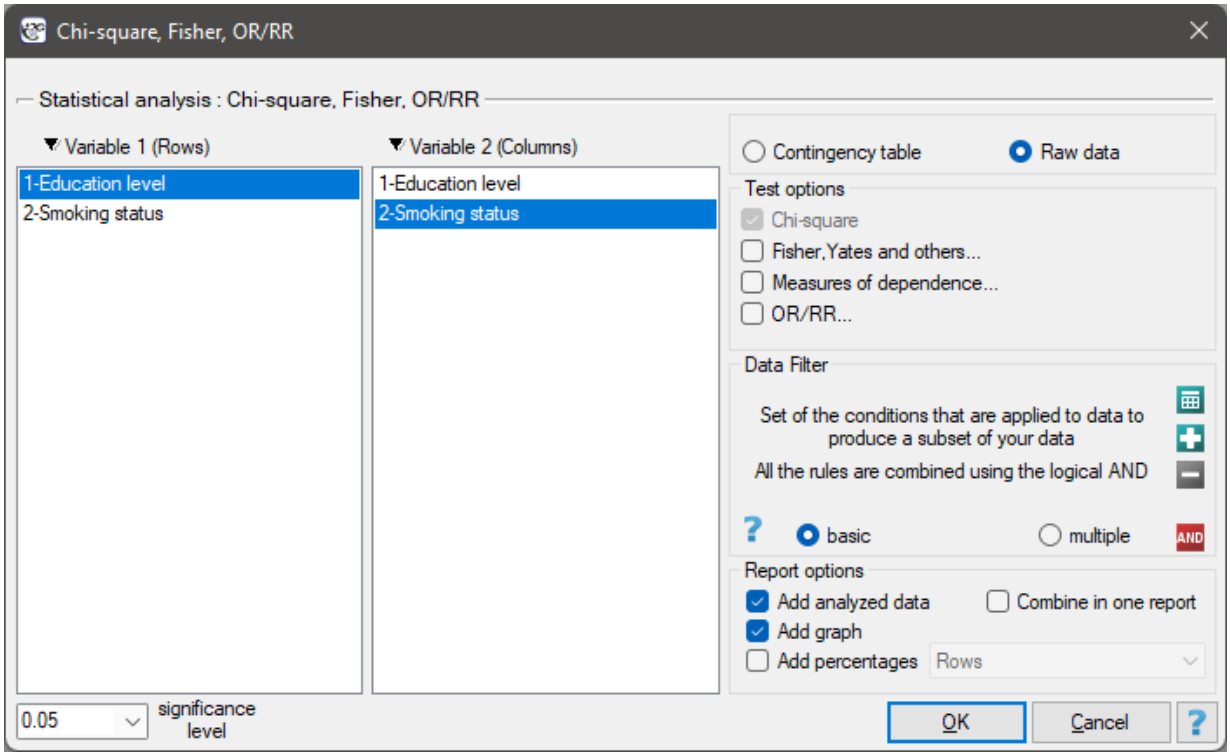
ASSUMPTIONS

- measurement on an ordinal scale or an interval scale (the categories must be orderable),
- an independent model (the second feature – 2 independent groups).

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Chi-square, Fisher, OR/RR, option Fisher, Yates and others... → Chi-square for trend.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Step 1. Indicate the ordered feature (e.g. education) and the two-category feature (e.g. smoking).

Step 2. Run the analysis.

Note! Before the analysis, you need to indicate the order in which the categories of the ordered feature should appear. To do this, from the variable's properties select Codes/Labels/Format... and assign the order by specifying consecutive natural numbers, then assign labels to them.

PERFORMANCE AND INTERPRETATION

EXAMPLE 16.9. (smoking-education.pqs file)
We check whether cigarette smoking is related to the education of residents of a village. A sample of 122 people was drawn. We assume that the relationship can be of two types: the more educated people are, the more often they smoke, or the more educated people are, the less often they smoke – so we are looking for an increasing or decreasing trend.

Codes and labels for the values

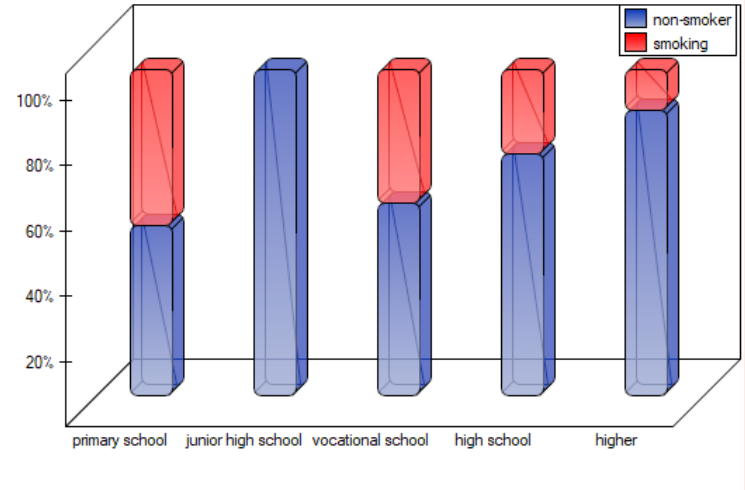
Value	Current value	Label
high school	4	high school
higher	5	higher
junior high school	2	junior high school
primary school	1	primary school
vocational school	3	vocational school

Chi-square statistic for trend	4.808
Degrees of freedom	1
p-value	0.0283

Copyright ©2010-2026 U-smile Stats – All rights reserved

$p = 0.0091$ **(statistically significant,** $\alpha = 0.05$**)**

There is a trend in the frequency of smoking with increasing education in the analysed population.



As the graph shows, the more educated people are, the less often they smoke. However, the result obtained for people with lower secondary education deviates from this trend – since this category has only two people, it did not have much effect on the overall trend. Due to the very small size of this group, the analysis was repeated for the combined primary and lower secondary education categories.

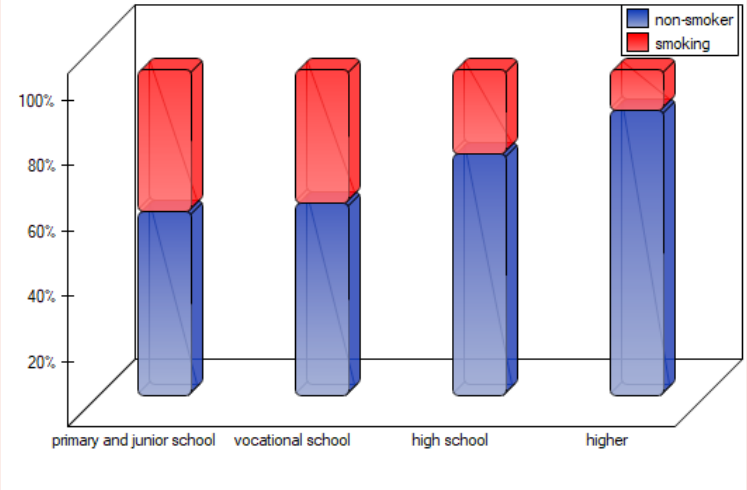
Codes and labels for the values

Value	Current value	Label
high school	2	high school
higher	3	higher
junior high school	0	primary and junior school
primary school	0	primary and junior school
vocational school	1	vocational school

$p = 0.0078$ **(after combining categories, statistically significant,** $\alpha = 0.05$**)**

The trend in smoking frequency with increasing education is confirmed again after combining the least numerous categories.

Copyright ©2010-2026 U-smile Stats – All rights reserved



EXAMPLE 16.10. (viewers.pqs file)
Because of a decrease in the audience of a particular soap opera, a viewer opinion survey was carried out. 100 people who had recently started watching the soap opera and 300 who had watched it regularly from the beginning were surveyed. Among other things, they were asked to rate their level of preoccupation with the characters' fate. The results are shown in the table below:

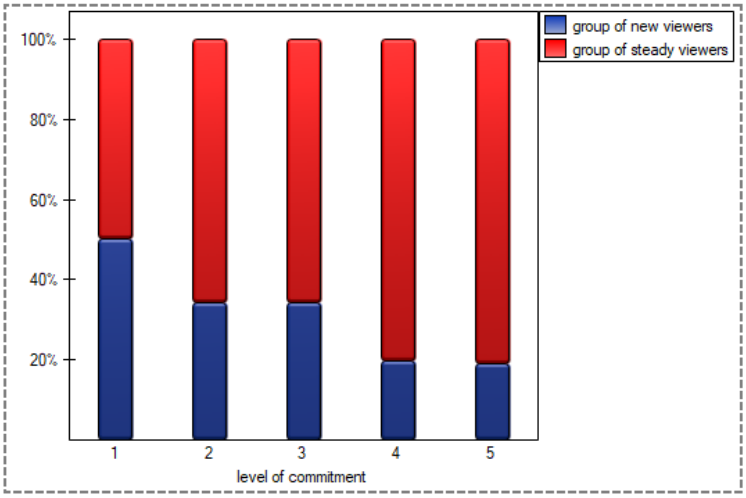
Level of commitment	group		
	group of new viewers	group of steady viewers	total
rather small	7	7	14
average	13	25	38
rather high	30	58	88
high	24	99	123
very high	26	111	137
total	100	300	400

The new viewers make up 25% of all analysed viewers. This proportion is not the same for every level of commitment, however – it looks as follows:

Level of commitment	group		
	group of new viewers	group of steady viewers	total
rather small	p_1=50.00%	50.00%	100%
average	p_2=34.21%	65.79%	100%
rather high	p_3=34.09%	65.91%	100%
high	p_4=19.51%	80.49%	100%
very high	p_5=18.98%	81.02%	100%
total	**25.00%**	**75.00%**	**100%**

Copyright ©2010-2026 U-smile Stats – All rights reserved

Chi-square, Fisher, OR/RR	
Analysed variables	level of commitment group
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	400
Cochran condition	fulfilled
Pearson's chi-square statistic	14.89
Degrees of freedom	4
p-value	0.0049
Chi-square statistic for trend	12.3703
Degrees of freedom	1
p-value	0.0004



Expected:		✓group
↓level of commit	group of nev	group of ste
1	3.5	10.5
2	9.5	28.5
3	22	66
4	30.75	92.25
5	34.25	102.75

Data :		✓group	
↓level of commit	group of nev	group of ste	Summary
1	7	7	14
2	13	25	38
3	30	58	88
4	24	99	123
5	26	111	137
Summary	100	300	88

% of row :		✓group
↓level of commit	group of nev	group of ste
1	50%	50%
2	34.211%	65.789%
3	34.091%	65.909%
4	19.512%	80.488%
5	18.978%	81.022%

$p = 0.000436$ **(statistically significant, $\alpha = 0.05$)**

There is a trend in the proportion of new viewers in the population of soap opera viewers: the more interested a group is in the characters' fate, the smaller share of it is made up of new viewers (a decreasing trend).

FOR THOSE INTERESTED

Notation: the $r \times 2$ contingency table of observed frequencies:

Observed frequencies O_{ij}		Feature 2 (group) G_1	Feature 2 (group) G_2	Total
Feature 1 (feature X)	X_1	O_{11}	O_{12}	$W_1 = O_{11} + O_{12}$
	X_2	O_{21}	O_{22}	$W_2 = O_{21} + O_{22}$
	...	...	...	...
	X_r	O_{r1}	O_{r2}	$W_r = O_{r1} + O_{r2}$
	Total	$C_1 = \sum_{i=1}^{r} O_{i1}$	$C_2 = \sum_{i=1}^{r} O_{i2}$	$n = C_1 + C_2$

Let $p_1, p_2, ..., p_r$ denote the proportions $p_1 = \frac{O_{11}}{W_1}$, $p_2 = \frac{O_{21}}{W_2}$,..., $p_r = \frac{O_{r1}}{W_r}$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Hypotheses:

$\mathcal{H}_0$: in the analysed population, there is no trend in the proportions $p_1, p_2, ..., p_r$,
$\mathcal{H}_1$: in the analysed population, there is a trend in the proportions $p_1, p_2, ..., p_r$.

Formulas: The test statistic is defined by:

$$\chi^2 = \frac{\left[\left(\sum_{i=1}^{r} i \cdot O_{i1}\right) - C_1 \left(\sum_{i=1}^{r} \frac{i \cdot W_i}{n}\right)\right]^2}{\frac{C_1}{n}\left(1 - \frac{C_1}{n}\right)\left[\left(\sum_{i=1}^{n} i^2 W_i\right) - n\left(\sum_{i=1}^{n} \frac{i \cdot W_i}{n}\right)^2\right]}.$$

This statistic has the χ^2 distribution with 1 degree of freedom.

The Relative Risk and the Odds Ratio

WHEN TO USE THESE MEASURES:

The Odds Ratio and the Relative Risk are used when we want to check how exposure to some factor (e.g. cigarette smoking) is related to the occurrence of an analysed phenomenon (e.g. a disease) – in other words, how much greater (or smaller) the odds or risk of the phenomenon occurring is in the exposed group compared to the unexposed group. Both measures are calculated on the basis of data collected in a 2×2 contingency table.

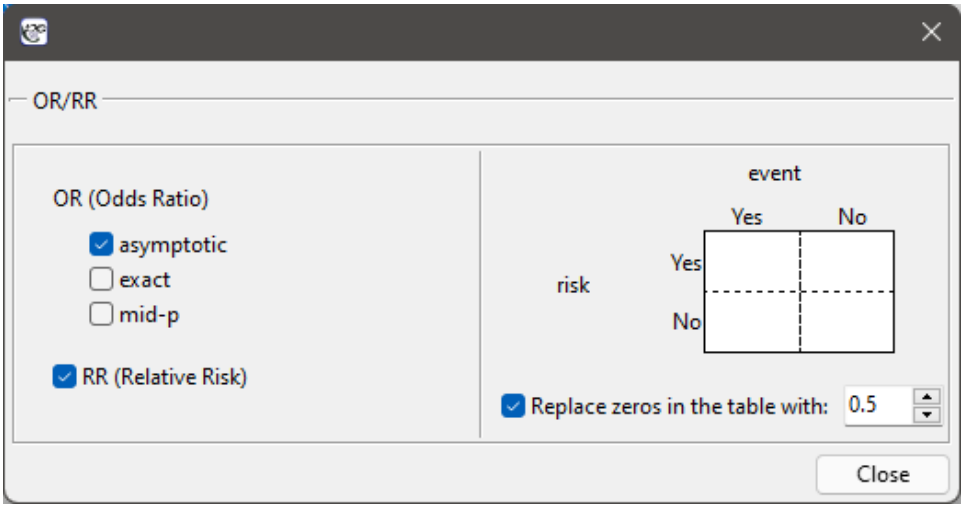
	Ill	Healthy
Smokers	O_{11}	O_{12}
Non-smokers	O_{21}	O_{22}

If the study is a **case-control** study (usually retrospective – the researcher decides on the sample size in which the phenomenon occurs, and on the control sample, free of the phenomenon), the **Odds Ratio** is calculated. If the study is a **cohort** study (usually prospective – the researcher ensures that the structure of the phenomenon's occurrence in the sample is close to its structure in the population), the **Relative Risk** is calculated.

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Chi-square, Fisher, OR/RR, option OR/RR... (or via the Wizard).

Copyright ©2010-2026 U-smile Stats – All rights reserved



Note! When the data table contains zeros, it may not be possible to calculate the odds ratio or relative risk. To address this, in the analysis window you can select the Replace 0 in the table with: option (continuity correction) – this adds the specified value to all cells of the table.

DEFINITIONS OF MEASURES

Odds Ratio (OR) (*ang. Odds Ratio*) – indicates how many times greater (or smaller) the odds of the analysed phenomenon occurring in the group exposed to the risk factor are compared to the unexposed group. It is calculated for case-control studies.

Relative Risk (RR) (*ang. Relative Risk*) – indicates how many times greater (or smaller) the risk of the analysed phenomenon occurring in the group exposed to the risk factor is compared to the unexposed group. It is calculated for cohort studies.

When interpreting the significance of either measure, we often use its confidence interval – we then check whether the interval contains the value 1 (no difference between groups). We can also compare the p value of the significance test (see "For those interested") with the assumed significance level α: when $p \leq \alpha$, the difference is statistically significant.

FOR THOSE INTERESTED

The odds ratio (2×2 table)

Formulas:
The odds of being a case in the exposed group and in the unexposed group are calculated according to the formulas:

$$odds_{exposed} = \frac{O_{11}/(O_{11}+O_{12})}{O_{12}/(O_{11}+O_{12})} = \frac{O_{11}}{O_{12}},$$

$$odds_{unexposed} = \frac{O_{21}/(O_{21}+O_{22})}{O_{22}/(O_{21}+O_{22})} = \frac{O_{21}}{O_{22}}.$$

The Odds Ratio:

$$OR = \frac{O_{11}/O_{12}}{O_{21}/O_{22}} = \frac{O_{11}O_{22}}{O_{12}O_{21}}.$$

The test of significance for the OR verifies the hypothesis that the odds of the analysed phenomenon

Copyright ©2010-2026 U-smile Stats – All rights reserved

occurring are the same in the exposed and unexposed groups.
Hypotheses:

$$\begin{aligned}\mathcal{H}_0: &\quad OR = 1,\\ \mathcal{H}_1: &\quad OR \neq 1.\end{aligned}$$

Formulas:
The test statistic is defined by:

$$z = \frac{\ln(OR)}{SE},$$

where $SE = \sqrt{\frac{1}{O_{11}} + \frac{1}{O_{12}} + \frac{1}{O_{21}} + \frac{1}{O_{22}}}$ – standard error of $\ln(OR)$. This statistic asymptotically (for large sample sizes) has the normal distribution.

The relative risk (2×2 table)

Formulas:
The estimated risk of the analysed phenomenon occurring is defined by the formula $R = \frac{O_{11}+O_{21}}{n}$, and the Relative Risk:

$$RR = \frac{O_{11}/(O_{11}+O_{12})}{O_{21}/(O_{21}+O_{22})}.$$

The test of significance for the RR verifies the hypothesis that the risk of the analysed phenomenon occurring is the same in the exposed and unexposed groups.
Hypotheses:

$$\begin{aligned}\mathcal{H}_0: &\quad RR = 1,\\ \mathcal{H}_1: &\quad RR \neq 1.\end{aligned}$$

Formulas:
The test statistic is defined by:

$$z = \frac{\ln(RR)}{SE},$$

where $SE = \sqrt{\frac{1}{O_{11}} - \frac{1}{O_{11}+O_{12}} + \frac{1}{O_{21}} - \frac{1}{O_{21}+O_{22}}}$ – standard error of $\ln(RR)$. This statistic asymptotically (for large sample sizes) has the normal distribution.

The Z test for two independent proportions

WHEN TO USE THIS TEST:

The Z test for two independent proportions is used in similar situations as the χ^2 test (2×2): when we have 2 independent samples in which the analysed feature takes one of two possible outcomes, and we want to check whether the share (proportion) of the distinguished outcome is the same in the two populations from which the samples were drawn.

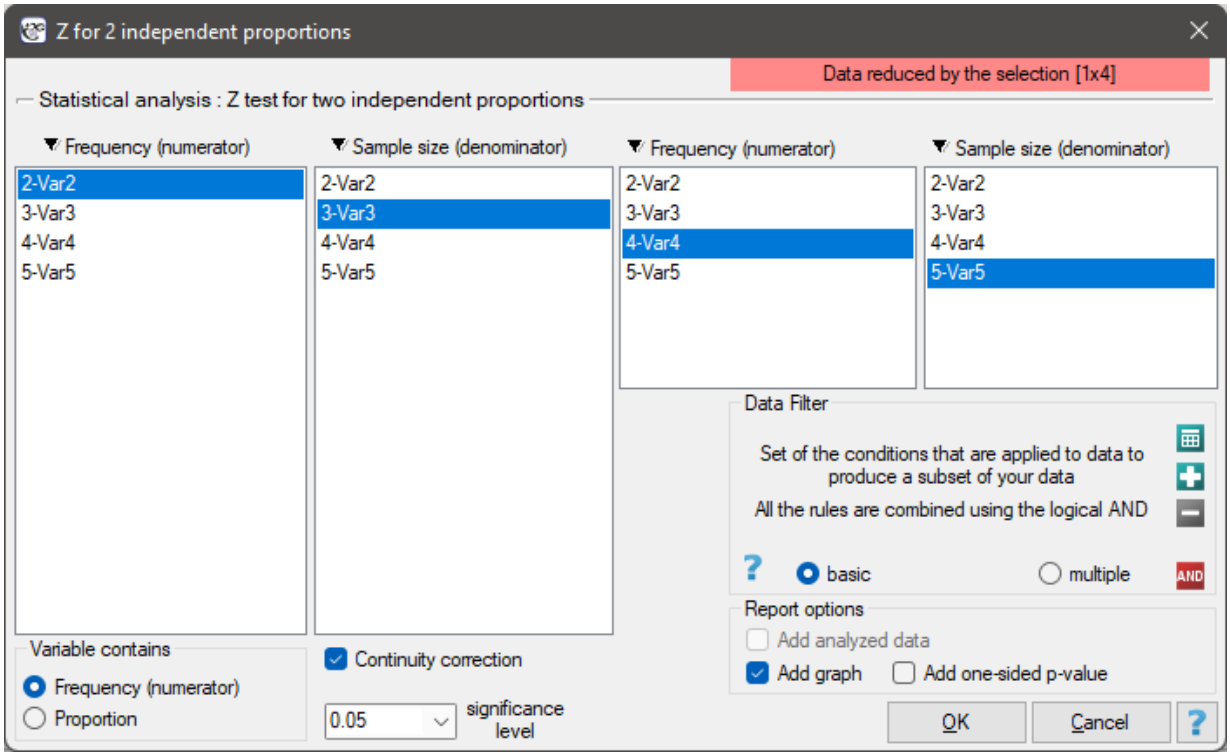
ASSUMPTIONS

Copyright ©2010-2026 U-smile Stats – All rights reserved

- measurement on a nominal scale – any ordering of categories is not taken into account,
- an independent model,
- large sample size.

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Z for 2 independent proportions.



Step 1. Indicate the data for both samples – as the number of successes and sample size, or as a proportion and sample size.

Step 2. Run the analysis.

Note! When the data sheet contains more information, before starting the analysis you should select the relevant area (data without headers). In the test window, also select the option describing the content of the variable (frequency (numerator) or proportion).

Note! Besides the difference between proportions, the program also calculates the NNT and/or NNH indicators. Confidence intervals for the difference between two independent proportions (from PQStat version 1.3.0) are estimated using the Newcombe-Wilson method[18][128][179], and the confidence interval for NNT/NNH – as the inverse of the interval for the proportion, using Altman's method[6].

PERFORMANCE AND INTERPRETATION

Copyright ©2010-2026 U-smile Stats – All rights reserved

EXAMPLE 16.11. **EXAMPLE (??) cont.** *(sex-exam.pqs file)*
We know that $\frac{50}{90} = 55.56\%$ of all women in the sample passed the exam, and $\frac{20}{80} = 25.00\%$ of all men in the sample passed the exam. We want to check whether the pass-rate proportion differs between the sexes in the analysed population.

The data can be written in two ways – as a numerator and a denominator for each sample, or as a proportion and a denominator for each sample:

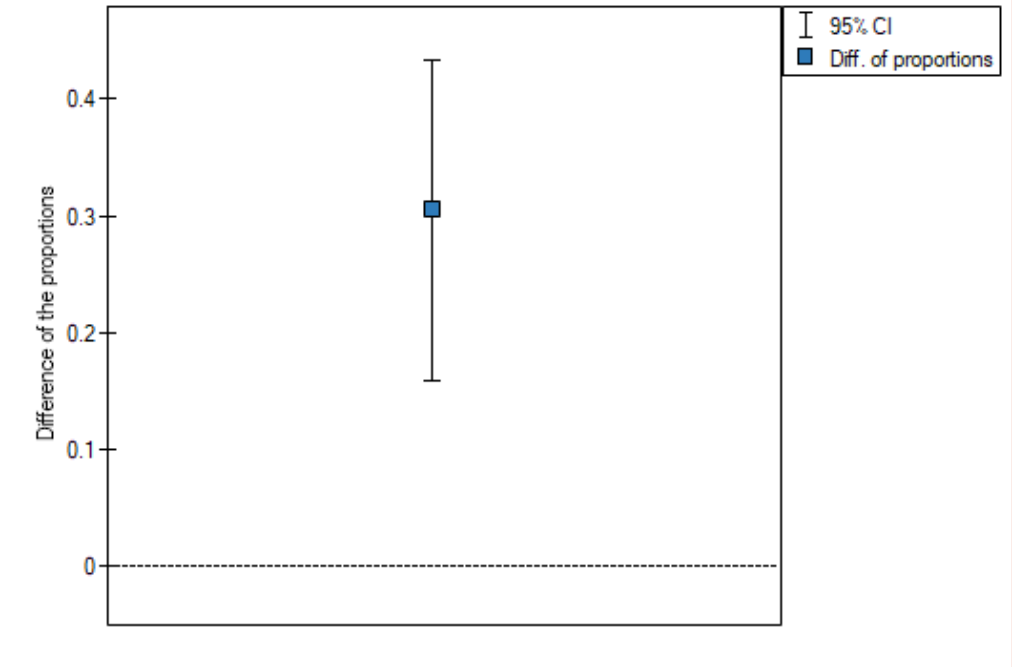
for Z test (frequency)	numerator - women	denominator - wome	numerator - men	denominator - men
	50	90	20	80
for Z test (proportion)	proportion - women	denominator - wome	proportion - men	denominator - men
	0.555555555556	90	0.25	80

Z test for two independent proportions	
Analysed variables	Var2
	Var3
	Var4
	Var5
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Difference of the proportions	0.3056
-95% CI for the difference of the proportions	0.1587
+95% CI for the difference of the proportions	0.4335
NNT	3.2727
-95% CI NNT	2.3067
+95% CI NNT	6.3014
Z statistic	3.8844
Two sided p-value (asymptotic)	0.0001

Data :

Var2	Var3	Var4	Var5
50	90	20	80

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p = 0.000053$ **(without continuity correction, statistically significant, $\alpha = 0.05$)**

The proportion of men who passed the exam differs from the proportion of women who passed the exam, in the analysed population.

The difference between the proportions distinguished in the sample is 30.56%, and its 95% confidence interval $(15.90\%, 43.35\%)$ does not contain 0. The same conclusion is reached with the Z test with the continuity correction ($p = 0.0001$) – similarly to the previously performed tests: the exact Fisher test, mid-p, chi-square, and chi-square with the Yates correction. The exam was passed significantly more often by women (55.56% of the sample) than by men (25.00% of the sample).

Example 16.12. Let us assume that a disease has a 100% mortality rate without treatment, and therapy reduces the mortality rate to 50% – this is the result of a 20-year study. We want to know how many people would need to be treated to prevent 1 death over 20 years. For this purpose, two 100-person samples were drawn from the population of diseased people: in the untreated sample all 100 patients die, in the treated sample 50 out of 100 patients survive.

Patients – not treated		Patients – treated	
sample numerator	sample size (denominator)	sample numerator	sample size (denominator)
100	100	50	100

Copyright ©2010-2026 U-smile Stats – All rights reserved

Z test for two independent proportions	
Analysed variables	Var2
	Var3
	Var4
	Var5
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Difference of the proportions	0.5
-95% CI for the difference of the proportions	0.397
+95% CI for the difference of the proportions	0.5962
NNT	2
-95% CI NNT	1.6774
+95% CI NNT	2.5191
Z statistic	8.0017
Two sided p-value (asymptotic)	<0.0001

Data :

Var2	Var3	Var4	Var5
100	100	50	100

$p < 0.000001$, $NNT = 2$ **(statistically significant,** $\alpha = 0.05$**)**

The difference between the proportions is statistically significant, and the NNT indicator is 2 – meaning treating 2 patients for 20 years will prevent 1 death.

The calculated 95% confidence interval for NNT should be rounded to whole numbers, giving an NNT of 2 to 3 patients.

Example 16.13. A certain proportion difference in a study comparing the effectiveness of drug 1 vs drug 2 was (95%CI)$= -0.08$ (-0.27 to 0.11). This negative proportion difference suggests that drug 1 was less effective than drug 2, and its use exposed patients to risk.

NNH(95%CI)$= 2.5$ **(**3.7 **to** ∞ **to NNT** 9.1**)**

The result is not statistically significant – because the confidence interval for NNH contains infinity and crosses from NNH to NNT, there are no grounds to claim a significant increase in risk[6].

DEFINITIONS OF MEASURES

NNT (*ang. number needed to treat*) – an indicator used in medicine, denoting the number of patients who need to be treated for a specified time in order to cure one person who would not otherwise have recovered.

NNH (*ang. number needed to harm*) – an indicator used in medicine, denoting the number of patients whose exposure to a risk over a specified time results in harm to one person who would not otherwise have been harmed.

FOR THOSE INTERESTED

Copyright ©2010-2026 U-smile Stats – All rights reserved

Definition: Let $p_1 = \frac{m_1}{n_1}$ and $p_2 = \frac{m_2}{n_2}$ denote the distinguished proportions in samples of sizes n_1, n_2, where m_1, m_2 – the frequencies of the distinguished outcome in the first and second sample.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad P_1 = P_2, \\ \mathcal{H}_1: &\quad P_1 \neq P_2, \end{aligned}$$

where P_1, P_2 – distinguished proportions in the first and second population.

Formulas: The test statistic is defined by:

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $p = \frac{m_1+m_2}{n_1+n_2}$.

The test statistic modified by the continuity correction is defined by:

$$Z = \frac{p_1 - p_2 - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

The statistic, with and without the continuity correction, asymptotically (for large sample sizes) has the normal distribution.

Formulas – NNT/NNH: NNT is calculated using the formula:

$$NNT = \frac{1}{|p_1 - p_2|},$$

and is quoted when the difference $p_1 - p_2$ is positive. NNH is calculated the same way as NNT, but is quoted when the difference $p_1 - p_2$ is negative.

Interpretation: The narrower the confidence interval, the more precise the estimate. If the confidence interval for the difference in proportions contains 0 (and for NNT/NNH – contains ∞), this is an indication that the result should be treated as not statistically significant.

Confidence interval estimation method – confidence intervals for the difference between two independent proportions, from PQStat version 1.3.0, are estimated using the Newcombe-Wilson method (Bender 2001[18], Newcombe 1998[128], Wilson 1927[179]); previously the Wald method was used. The Wald method is suitable for large samples and proportion differences far from 0 or 1 – for small samples and values close to these extremes it can lead to unreliable results (Newcombe 1998[128], Miettinen 1985[125], Beal 1987[17], Wallenstein 1997[170]). The Newcombe-Wilson method is also suitable for extreme values of proportions.

The McNemar test, the Bowker test of internal symmetry

WHEN TO USE THESE TESTS:

Copyright ©2010-2026 U-smile Stats – All rights reserved

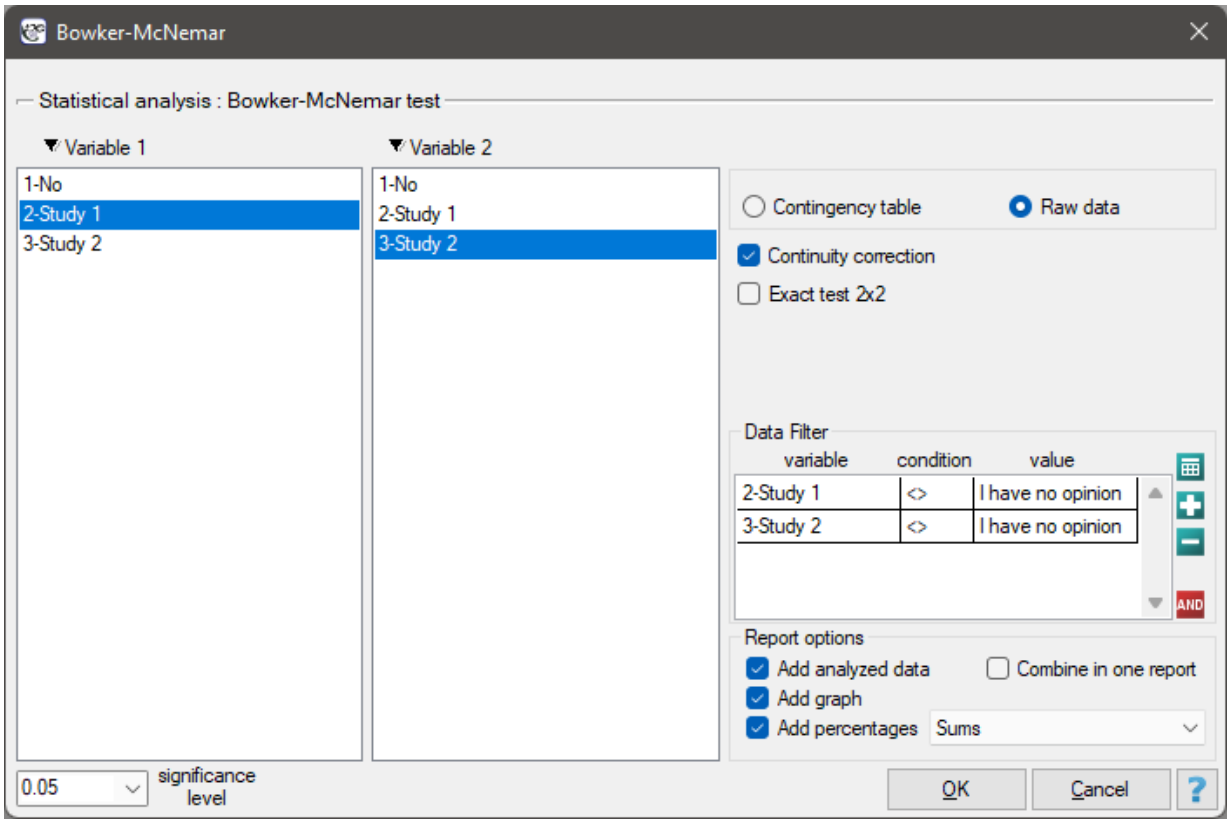
These tests are used when the same feature is measured twice for the same subjects (e.g. an opinion before and after some event), and we want to check whether the results changed. **The McNemar test** *(ang. McNemar test)*, McNemar (1947)[122], is used when the analysed feature has only 2 categories (denoted **(+)** and **(–)**). **The Bowker test of internal symmetry** *(ang. Bowker test of internal symmetry)*, Bowker (1948)[24], is its extension to features with more than 2 categories ($c > 2$). Both tests can be calculated on the basis of raw data or a ready-made contingency table.

ASSUMPTIONS

- measurement on a nominal scale – any ordering of categories is not taken into account,
- a dependent model.

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Bowker-McNemar (or via the Wizard).



Step 1. Indicate the two variables corresponding to the first and second measurement of the same feature.

Step 2. Run the analysis.

Note! A common general rule for the validity of the asymptotic McNemar test is the **Rufibach condition**: the number of discordant pairs should be greater than 10 ($O_{12} + O_{21} \geq 10$)[148]. When this condition

Copyright ©2010-2026 U-smile Stats – All rights reserved

is not satisfied, it is better to base the conclusion on the exact probability value of the test (based on the binomial distribution) or on its less conservative version, the mid-p McNemar test[57] – both are calculated automatically together with the basic test.

PERFORMANCE AND INTERPRETATION

EXAMPLE 16.14. (opinion.pqs file)
Two surveys of student opinion about a particular academic professor were carried out. Both surveys allowed students to rate the professor negatively, positively, or choose a neutral answer – no opinion. Both surveys were carried out on the same sample of 250 students, with the first survey done the day before the exam in the course taught by the rated professor, and the second the day after the exam. Below is a fragment of the raw data and the complete data as a contingency table. We want to check whether the two surveys give similar results.

Study 1	Study 2
I have no opinion	positive
positive	negative
positive	negative
positive	positive

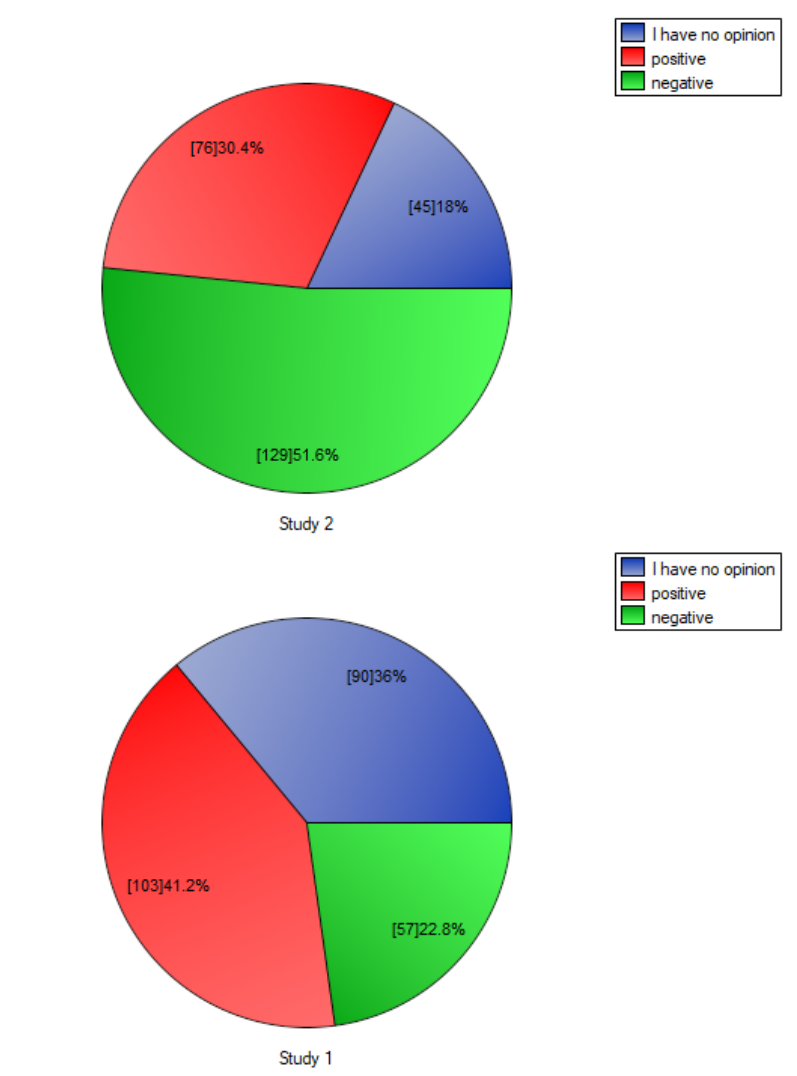
	negative	positive	I have no opinion
negative	50	4	3
positive	44	54	5
I have no opinion	35	18	37

Bowker-McNemar test	
Analysed variables	Study 1
	Study 2
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Number of pairs	250
Odds Ratio	NA
-95% CI for the Odds Ratio	NA
+95% CI for the Odds Ratio	NA
Bowker test	
Chi-square statistic	63.2378
Degrees of freedom	3
p-value	<0.0001

Data :			✓Study 2
↓Study 1	I have no op	positive	negative
I have no opinic	37	18	35
positive	5	54	44
negative	3	4	50
Summary	45	76	129

% of sum :			✓Study 2
↓Study 1	I have no op	positive	negative
I have no opinic	14.8%	7.2%	14%
positive	2%	21.6%	17.6%
negative	1.2%	1.6%	20%

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p < 0.0001$ **(Bowker test, statistically significant,** $\alpha = 0.05$**)**

Students' opinions about the professor changed between the two surveys.

The table shows that significantly more students changed their opinion to negative after the exam than to positive – many students who rated the professor positively before the exam no longer expressed that opinion afterwards.

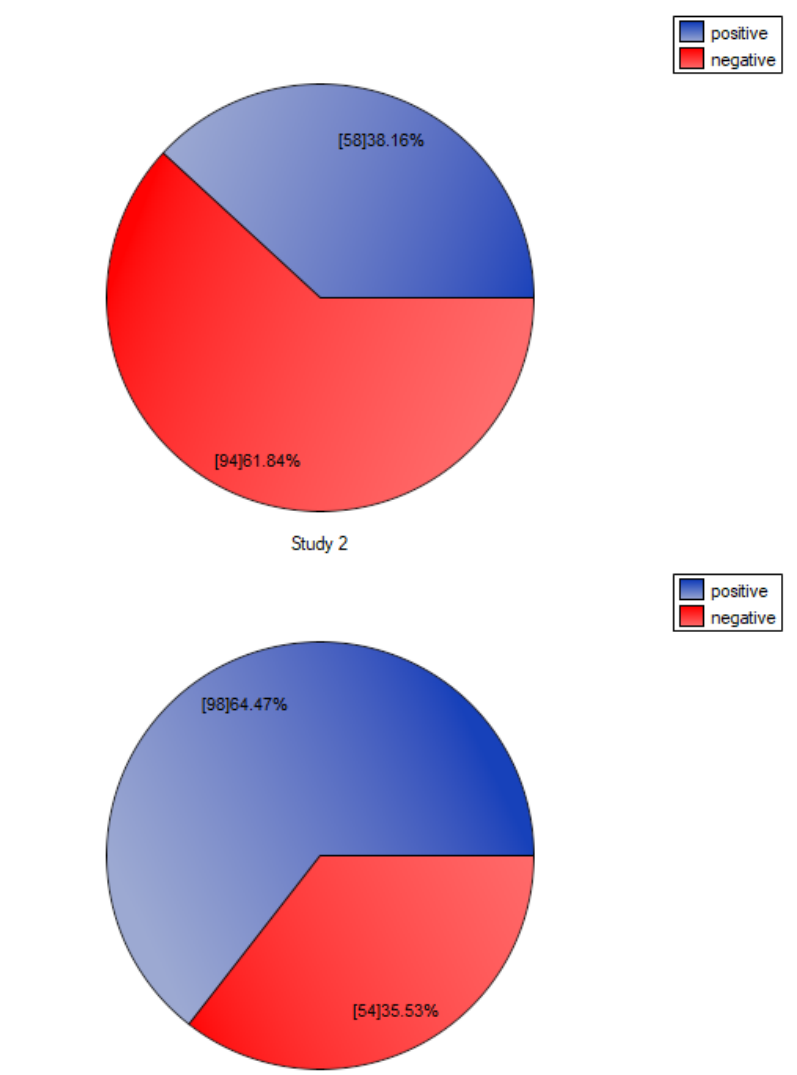
If we limited the study to people with a defined opinion about the professor (rating him only positively or negatively), we could use the simpler McNemar test:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Bowker-McNemar test	
Analysed variables	Study 1
	Study 2
Data Filter	Study 1<>I have no opinion
	and Study 2<>I have no opi
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Number of pairs	152
Odds Ratio	11
-95% CI for the Odds Ratio	3.9525
+95% CI for the Odds Ratio	30.6139
McNemar test	
Chi-square statistic	31.6875
Degrees of freedom	1
p-value	<0.0001
Rufibach condition	fulfilled

Data :		✓Study 2	
ıStudy 1	positive	negative	Summary
positive	54	44	98
negative	4	50	54
Summary	58	94	152

% of sum :		✓Study 2
ıStudy 1	positive	negative
positive	35.526%	28.947%
negative	2.632%	32.895%



Copyright ©2010-2026 U-smile Stats – All rights reserved

$p < 0.0001$ **(McNemar test, statistically significant, $\alpha = 0.05$)**

Among students with a clear-cut opinion, its distribution (positive/negative) changed between the two surveys.

Significantly more students changed their opinion to negative after the exam than to positive. The odds of changing an opinion from positive (before the exam) to negative (after the exam) are eleven times greater ($\frac{44}{4}$) than the odds of changing in the opposite direction ($\frac{4}{44}$).

DEFINITIONS OF MEASURES

Odds ratio (OR) (*ang. odds ratio*) – if a study of the same feature on the same objects was carried out twice, the odds ratio describes how many times more (or less) likely a change of result in one direction (e.g. from **(+)** to **(–)**) is than a change in the opposite direction.

FOR THOSE INTERESTED

Notation:
Notation for the 2×2 contingency table for McNemar's test:

Observed frequencies O_{ij}		$X^{(2)}$		
		(+)	**(–)**	**Total**
$X^{(1)}$	**(+)**	O_{11}	O_{12}	$O_{11} + O_{12}$
	(–)	O_{21}	O_{22}	$O_{21} + O_{22}$
	Total	$O_{11} + O_{21}$	$O_{12} + O_{22}$	$n = O_{11} + O_{12} + O_{21} + O_{22}$

The McNemar test – hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \quad O_{12} = O_{21}, \\ \mathcal{H}_1 &: \quad O_{12} \neq O_{21}. \end{aligned}$$

Formulas:
The test statistic is defined by:

$$\chi^2 = \frac{(O_{12} - O_{21})^2}{O_{12} + O_{21}}.$$

This statistic asymptotically (for large frequencies) has the χ^2 distribution with 1 degree of freedom.

Continuity correction for the McNemar test is more conservative than the basic test (it rejects the null hypothesis more rarely). The test statistic with the correction is defined by:

$$\chi^2 = \frac{(|O_{12} - O_{21}| - 1)^2}{O_{12} + O_{21}}.$$

Odds ratio of a result change – if the same feature was measured twice on the same objects, for such a table we calculate the odds ratio for a result change (from **(+)** to **(–)** and vice versa). The odds of a change from **(+)** to **(–)** is O_{12}, and from **(–)** to **(+)** is O_{21}. The odds ratio is calculated using the formula:

$$OR = \frac{O_{12}}{O_{21}}.$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

The confidence interval for the odds ratio is built on the basis of the standard error:

$$SE = \sqrt{\frac{1}{O_{12}} + \frac{1}{O_{21}}}.$$

For samples with small frequencies, the exact range of the confidence interval for the odds ratio can additionally be determined[106].

The Bowker test of internal symmetry extends the McNemar test to $c \times c$ tables ($c > 2$ categories):

Observed frequencies O_{ij}		$X^{(2)}$				
		$X_1^{(2)}$	$X_2^{(2)}$	...	$X_c^{(2)}$	Total
$X^{(1)}$	$X_1^{(1)}$	O_{11}	O_{12}	...	O_{1c}	$\sum_{j=1}^{c} O_{1j}$
	$X_2^{(1)}$	O_{21}	O_{22}	...	O_{2c}	$\sum_{j=1}^{c} O_{2j}$
	...	...	...	...	...	...
	$X_c^{(1)}$	O_{c1}	O_{c2}	...	O_{cc}	$\sum_{j=1}^{c} O_{cj}$
	Total	$\sum_{i=1}^{c} O_{i1}$	$\sum_{i=1}^{c} O_{i2}$	...	$\sum_{i=1}^{c} O_{ic}$	$n = \sum_{i=1}^{c}\sum_{j=1}^{c} O_{ij}$

Hypotheses (where $j \neq i$, $j, i \in \{1, 2, ..., c\}$, and O_{ij} and O_{ji} are the frequencies of symmetric pairs):

$$\begin{aligned} \mathcal{H}_0 &: \ O_{ij} = O_{ji}, \\ \mathcal{H}_1 &: \ O_{ij} \neq O_{ji} \text{ for at least one pair } O_{ij}, O_{ji}. \end{aligned}$$

Formulas:
The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j>i} \frac{(O_{ij} - O_{ji})^2}{O_{ij} + O_{ji}}.$$

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with a number of degrees of freedom calculated using the formula $df = \frac{c(c-1)}{2}$.

The Z test for two dependent proportions

WHEN TO USE THIS TEST:

The Z test for two dependent proportions is used when the same feature with two possible outcomes (e.g. yes/no, agree/disagree) is measured twice for the same subjects, and we want to check whether the share of one of the outcomes changed between the measurements (e.g. before and after some event). This is a similar situation to the McNemar test.

ASSUMPTIONS

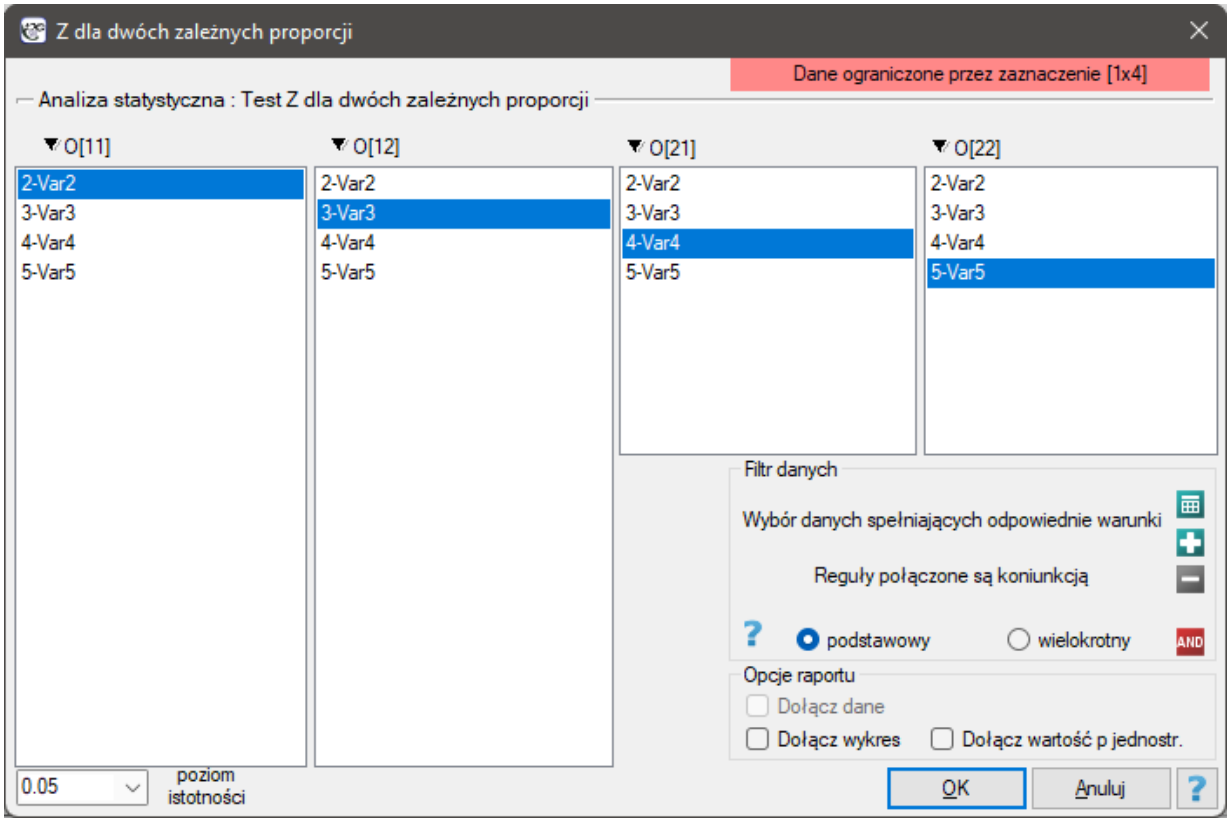
- measurement on a nominal scale – any ordering of categories is not taken into account,
- a dependent model,

Copyright ©2010-2026 U-smile Stats – All rights reserved

– large sample size.

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests → Z for 2 dependent proportions.



Step 1. Indicate the two variables corresponding to the first and second measurement of the same feature.

Step 2. Run the analysis.

Note! The confidence interval for the difference between two dependent proportions is estimated using the Newcombe-Wilson method.

PERFORMANCE AND INTERPRETATION

Example 16.15. **Example (16.14) cont.** *(opinion.pqs file)*
If we limit the study to people with a defined opinion about the professor (rating him only positively or negatively), we obtain 152 such students. Among them: 50 students rated the professor positively both before and after the exam; 4 students rated him well before the exam, but poorly after the exam; 44 students rated him poorly before the exam, but well after the exam; 54 students rated him negatively both before and after the exam. So we know that $\frac{50+4}{152} = 35.53\%$ of students expressed a negative

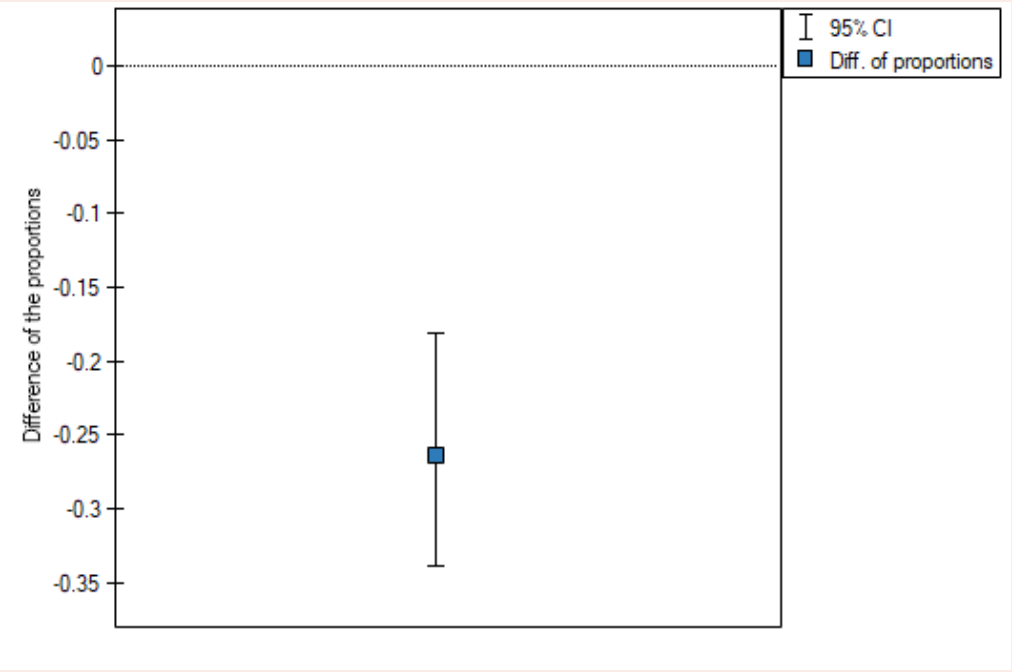
Copyright ©2010-2026 U-smile Stats – All rights reserved

opinion before the exam, and after the exam the corresponding percentage was $\frac{50+44}{152} = 61.84\%$. We want to check whether the proportion of negative ratings changed between the measurements.

Z test for two dependent proportions	
Analysed variables	Var3
	Var4
	Var5
	Var6
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	No
O[11]+O[12]	54
O[11]+O[21]	94
Proportion 1	0.3553
Proportion 2	0.6184
Difference of the proportions	-0.2632
-95% CI for the difference of the proportions	-0.3388
+95% CI for the difference of the proportions	-0.1807
Z statistic	-5.7735
Two sided p-value (asymptotic)	<0.0001

Data :

Var3	Var4	Var5	Var6
50	4	44	54



$p = 0.0001$ **(statistically significant, $\alpha = 0.05$)**

The proportion of negative ratings of the professor before the exam differs, in a statistically significant way, from the proportion of negative ratings after the exam.

The difference between the proportions distinguished in the sample is 26.32%, and its 95% confidence interval $(18.07\%, 33.88\%)$ does not contain 0 – the same conclusion is reached by the McNemar test. After the exam, the professor is rated negatively significantly more often.

Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Definition: Notation for the contingency table of two dependent measurements of the same feature X (first measurement $X^{(1)}$, second measurement $X^{(2)}$), each with two possible outcomes **(+)** and **(–)**:

Observed frequencies O_{ij}		$X^{(2)}$		
		(+)	**(–)**	**Total**
$X^{(1)}$	**(+)**	O_{11}	O_{12}	$O_{11}+O_{12}$
	(–)	O_{21}	O_{22}	$O_{21}+O_{22}$
	Total	$O_{11}+O_{21}$	$O_{12}+O_{22}$	$n=O_{11}+O_{12}+O_{21}+O_{22}$

where O_{11} – the number of people with a **(+)** outcome in both measurements, O_{12} – the number of people who changed outcome from **(+)** to **(–)**, O_{21} – the number of people who changed outcome from **(–)** to **(+)**, O_{22} – the number of people with a **(–)** outcome in both measurements.

For these data we calculate the distinguished proportions $p_1=\frac{O_{11}+O_{12}}{n}$ and $p_2=\frac{O_{11}+O_{21}}{n}$.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0:&\quad P_1-P_2=0,\\ \mathcal{H}_1:&\quad P_1-P_2\neq 0,\end{aligned}$$

where P_1, P_2 – distinguished proportions in the population, for the first and second measurement.

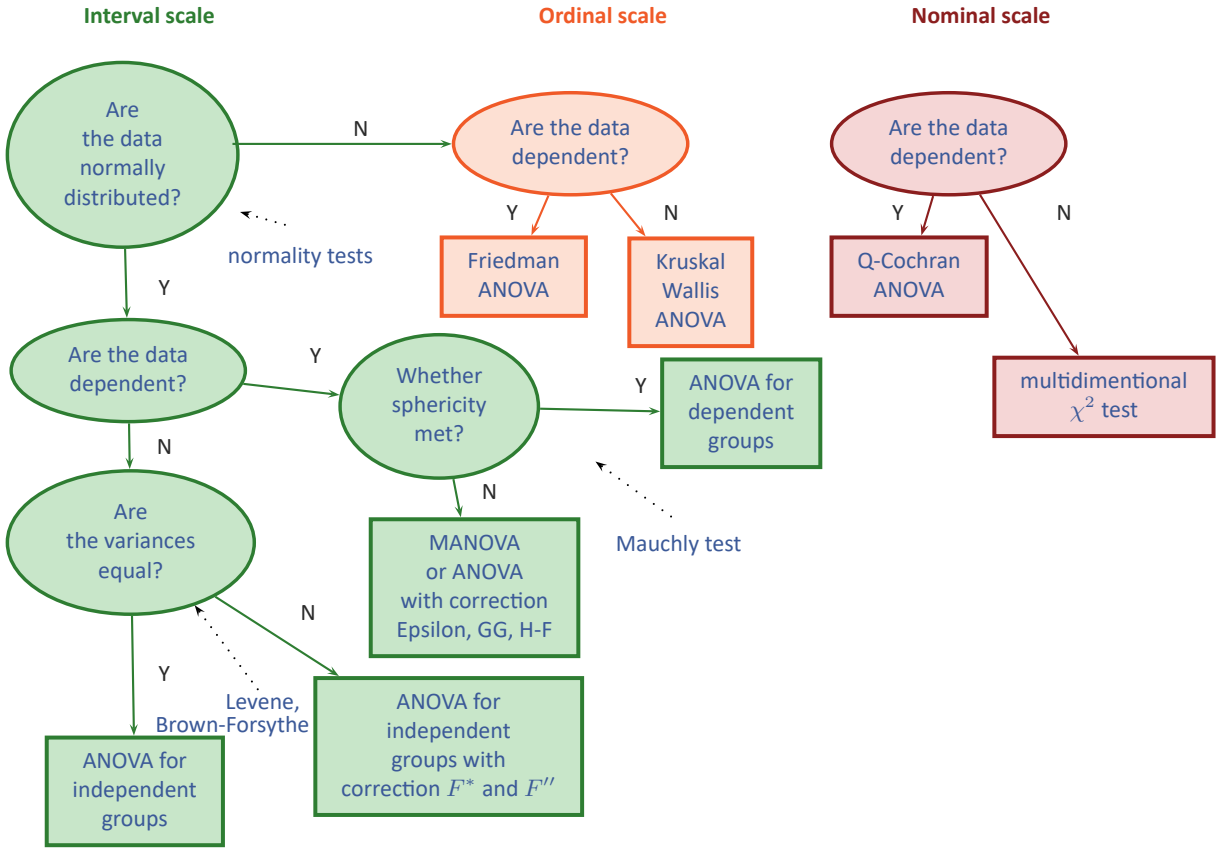
Formulas: The test statistic is defined by:

$$Z=\frac{p_1-p_2}{\sqrt{O_{21}+O_{12}}}\cdot n,$$

and asymptotically (for large sample sizes) has the normal distribution.

Copyright ©2010-2026 U-smile Stats – All rights reserved

17 COMPARISON - MORE THAN TWO GROUPS



Note

The proposed test selection scheme for the multiple group comparison is not the only possible scheme and does not include all the tests proposed in the software for this comparison.

Note

Note, that simultaneous comparison of more than two groups can NOT be replaced with multiple performance the tests for the comparison of two groups. It is the result of the necessity of controlling the I type error α. Choosing the α and using the k-fold selected test for the comparison of 2 groups, we could make the assumed level much higher α. It is possible to avoid this error using the ANOVA test (Analysis of Variance) and contrasts or the POST-HOC tests dedicated to them.

Copyright ©2010-2026 U-smile Stats – All rights reserved

17.1 PARAMETRIC TESTS

The ANOVA for independent groups

WHEN TO USE THIS TEST:

The one-way analysis of variance (ANOVA) for independent groups, proposed by Ronald Fisher, lets you check whether the mean value of an analysed feature is the same in at least three independent populations (e.g. groups of patients, product variants, companies). The test only indicates whether such differences exist at all – it does not say between which groups they occur; that is the role of the POST-HOC tests, described in the next section.

ASSUMPTIONS

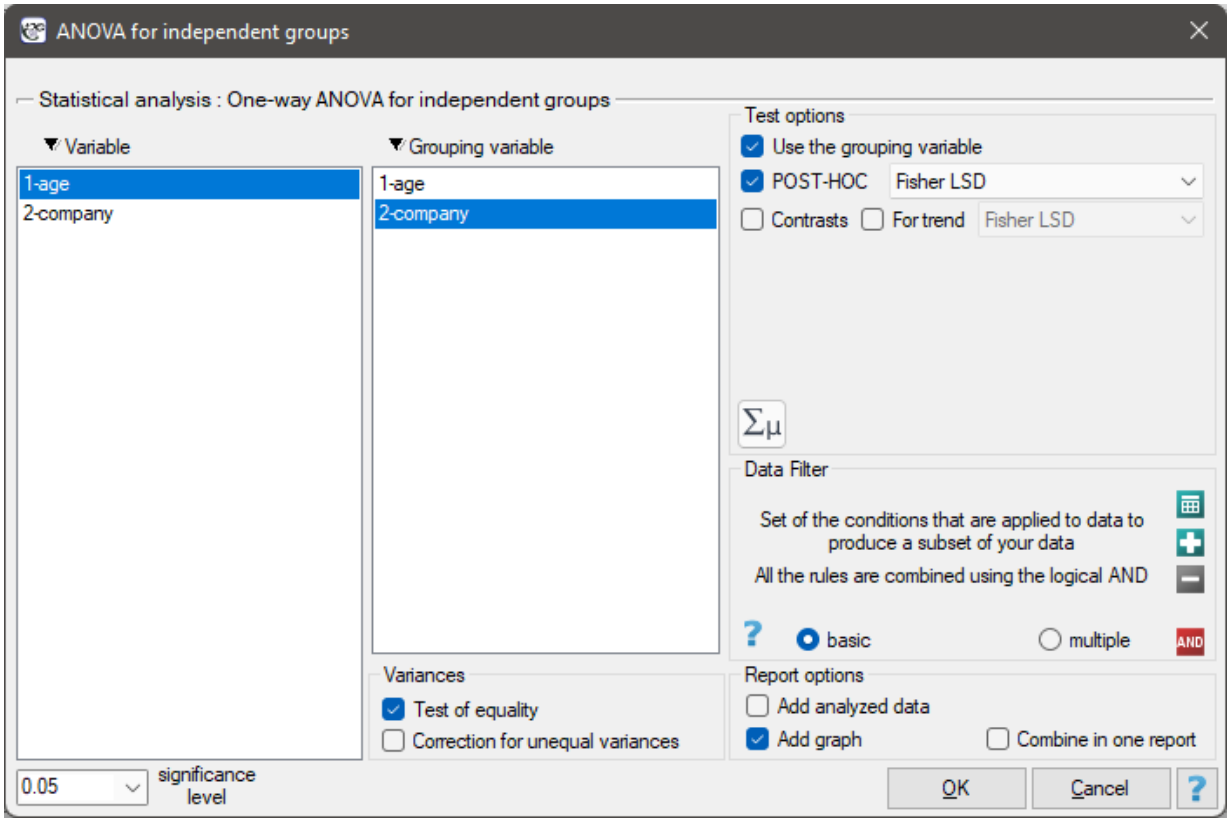
- measurement on an interval scale,
- normality of distribution of the analysed variable in each population,
- an independent model,
- equality of variances of the analysed variable across all populations.

When the assumption of equality of variances is not met, use the ANOVA for independent groups with F^* and F'' corrections.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → ANOVA for independent groups (or via the Wizard).

Copyright ©2010-2026 U-smile Stats – All rights reserved



Step 1. Select the grouping variable (e.g. company name) and the outcome variable (e.g. worker's age).

Step 2. Run the analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

Example 17.1. (age ANOVA.pqs file)
There are 150 people chosen at random from the population of workers of 3 different transport companies. From each company, 50 people are drawn into the sample. Before the experiment begins, we need to check whether the average age of the workers of these companies is similar, since the next stage of the experiment depends on it. The age of each participant is recorded in years.
Age (company 1): 27, 33, 25, 32, 34, 38, 31, 34, 20, 30, 30, 27, 34, 32, 33, 25, 40, 35, 29, 20, 18, 28, 26, 22, 24, 24, 25, 28, 32, 32, 33, 32, 34, 27, 34, 27, 35, 28, 35, 34, 28, 29, 38, 26, 36, 31, 25, 35, 41, 37
Age (company 2): 38, 34, 33, 27, 36, 20, 37, 40, 27, 26, 40, 44, 36, 32, 26, 34, 27, 31, 36, 36, 25, 40, 27, 30, 36, 29, 32, 41, 49, 24, 36, 38, 18, 33, 30, 28, 27, 26, 42, 34, 24, 32, 36, 30, 37, 34, 33, 30, 44, 29
Age (company 3): 34, 36, 31, 37, 45, 39, 36, 34, 39, 27, 35, 33, 36, 28, 38, 25, 29, 26, 45, 28, 27, 32, 33, 30, 39, 40, 36, 33, 28, 32, 36, 39, 32, 39, 37, 35, 44, 34, 21, 42, 40, 32, 30, 23, 32, 34, 27, 39, 37, 35

Before proceeding with the ANOVA analysis, normality of the data distribution was confirmed. The analysis window also tested the assumption of equality of variance, obtaining $p > 0.05$ in both tests.

Copyright ©2010-2026 U-smile Stats – All rights reserved

One-way ANOVA for independent groups	
Analysed variables	age
	company
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	company
ANOVA for independent groups	
Eta-square	0.0692
Total sum of squares (SS[T])	5151.8933
Between-groups sum of squares (SS[BG])	356.4133
Within-groups sum of squares (SS[WG])	4795.48
Mean square between-groups (MS[BG])	178.2067
Mean square within-groups (MS[WG])	32.6223
Between-groups degrees of freedom (df[BG])	2
Within-groups degrees of freedom (df[WG])	147
Total degrees of freedom (df[T])	149
F statistic	5.4627
p-value	0.0051
Equality of variances - Brown-Forsythe	
F statistic	0.8484
p-value	0.4302
Equality of variances - Levene	
F statistic	0.8671
p-value	0.4223

$p = 0.0051$ **(statistically significant,** $\alpha = 0.05$**)**

The average age of the workers of the three transport companies differs statistically significantly.

Based on the ANOVA result alone, we do not yet know which companies differ from one another in terms of workers' age – the answer to that question is provided by the POST-HOC analysis, described in the next section.

We can provide a detailed description of the data by selecting Descriptive statistics in the analysis window Σμ

Summary			
Group	1	2	3
Sample size	50	50	50
Arithmetic mean	30.26	32.68	33.98
Standard error of the mean	0.74	0.8992	0.7754
Standard deviation	5.2326	6.3582	5.4828
-95% CI for the group mean	28.7729	30.873	32.4218
+95% CI for the group mean	31.7471	34.487	35.5382

Copyright ©2010-2026 U-smile Stats – All rights reserved

DEFINITIONS OF MEASURES

Partial η^2 – an effect-size measure indicating what proportion of the variability of the analysed feature results from belonging to the compared groups (the factor), and what proportion remains unexplained. The higher the value, the stronger the association between group membership and the outcome of the analysed feature.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2 = ... = \mu_k, \\ \mathcal{H}_1: &\quad \text{not all } \mu_j \text{ are equal } (j = 1, 2, ..., k), \end{aligned}$$

where μ_1,μ_2,...,μ_k – means of the analysed variable in the populations from which the samples were drawn.

Formulas: the test statistic is defined by:

$$F = \frac{MS_{BG}}{MS_{WG}},$$

where:

$MS_{BG} = \dfrac{SS_{BG}}{df_{BG}}$ – mean square between-groups,

$MS_{WG} = \dfrac{SS_{WG}}{df_{WG}}$ – mean square within-groups,

$SS_{BG} = \displaystyle\sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N}$ – between-groups sum of squares,

$SS_{WG} = SS_T - SS_{BG}$ – within-groups sum of squares,

$SS_T = \left(\displaystyle\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2\right) - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N}$ – total sum of squares,

$df_{BG} = k - 1$ – degrees of freedom (between-groups),

$df_{WG} = df_T - df_{BG}$ – degrees of freedom (within-groups),

$df_T = N - 1$ – total degrees of freedom,

$N = \sum_{j=1}^{k} n_j$,

n_j – sample sizes for $(j = 1, 2, ...k)$,

x_{ij} – values of the variable in the samples for $(i = 1, 2, ...n_j)$, $(j = 1, 2, ...k)$.

The test statistic has the F Snedecor distribution with df_{BG} and df_{WG} degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Partial η^2

$$\eta^2 = \frac{SS_{BG}}{SS_{BG} + SS_{res}}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

Contrasts and POST-HOC tests

WHEN TO USE THIS TEST:

The ANOVA result only tells us whether there are statistically significant differences among the populations under study – it does not indicate which groups differ. Multiple comparisons answer that question: **contrasts** (when comparisons are planned in advance, usually only selected ones, so-called a'priori) or **POST-HOC tests** (when we look for differences only after the ANOVA has been performed, usually between all possible pairs of groups).

ASSUMPTIONS

- a previously performed ANOVA for independent groups showed a statistically significant difference,
- the choice of method depends on whether the assumption of equality of variances is met: the LSD Fisher, Scheffe and Tukey tests are used when the variances do not differ significantly.

For groups with significantly different variances, use the methods described in the ANOVA for independent groups with F^* and F'' corrections chapter.

HOW TO SET UP THE ANALYSIS

The POST-HOC tests are set up in the same options window used to run the ANOVA for independent groups.

Step 1. Resume the analysis using the  button.

Step 2. In the test options window, select a multiple-comparison method, e.g. Tukey HSD.

Step 3. Select the option to add a graph.

Step 4. Run the analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

Homogeneous groups

For each post-hoc test, homogeneous groups are constructed. Each homogeneous group represents a set of groups that are not statistically significantly different from each other. For example, suppose we divided subjects into six groups regarding smoking status: Nonsmokers (NS), Passive smokers (PS), Noninhaling smokers (NI), Light smokers (LS), Moderate smokers (MS), Heavy smokers (HS), and we examine their expiratory parameters. In an ANOVA we obtained statistically significant differences in expiratory parameters between the studied groups. To indicate which groups differ significantly and which do not, we perform post-hoc tests. As a result, in addition to the table with the results of each pair of comparisons and the statistical significance expressed as a p value:

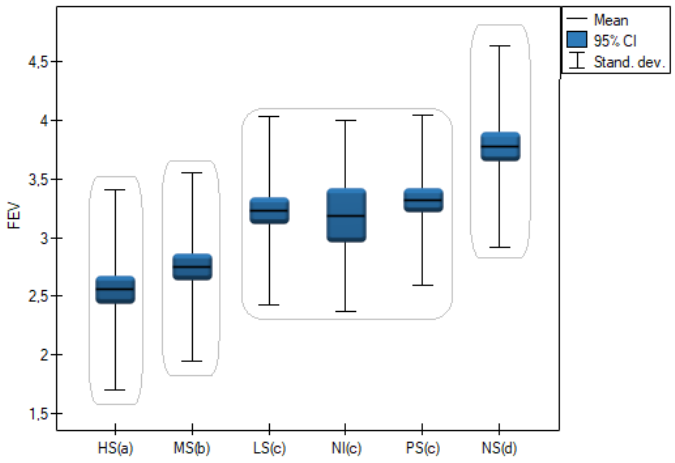
Copyright ©2010-2026 U-smile Stats – All rights reserved

Statystyk	HS	MS	LS	NI	PS	NS
HS		2,434752	8,276928	4,920204	9,41745	15,066567
MS	2,434752		5,842176	3,380331	6,982697	12,631814
LS	8,276928	5,842176		0,314585	1,140522	6,789639
NI	4,920204	3,380331	0,314585		1,035914	4,60873
PS	9,41745	6,982697	1,140522	1,035914		5,649117
NS	15,066567	12,631814	6,789639	4,60873	5,649117	
Wartość p	HS	MS	LS	NI	PS	NS

we obtain a division into homogeneous groups:

Homoge	HS(a)	MS(b)	LS(c)	NI(c)	PS(c)	NS(d)
A	*					
B		*				
C			*	*	*	
D						*

In this case we obtained 4 homogeneous groups, i.e. A, B, C and D, which indicates the possibility of conducting the study based on a smaller division – i.e. instead of the six groups studied originally, further analyses can be conducted based on the four homogeneous groups determined here. The order of the groups was established based on weighted means calculated for each homogeneous group, such that letter A was assigned to the group with the lowest weighted mean, and further letters of the alphabet to groups with progressively higher means.



Continuation of Example 17.1 (age ANOVA.pqs file): the ANOVA showed a statistically significant difference in the average age of workers of the three transport companies ($p = 0.0051$). We now check which companies differ from one another in workers' age by performing the Tukey POST-HOC test.

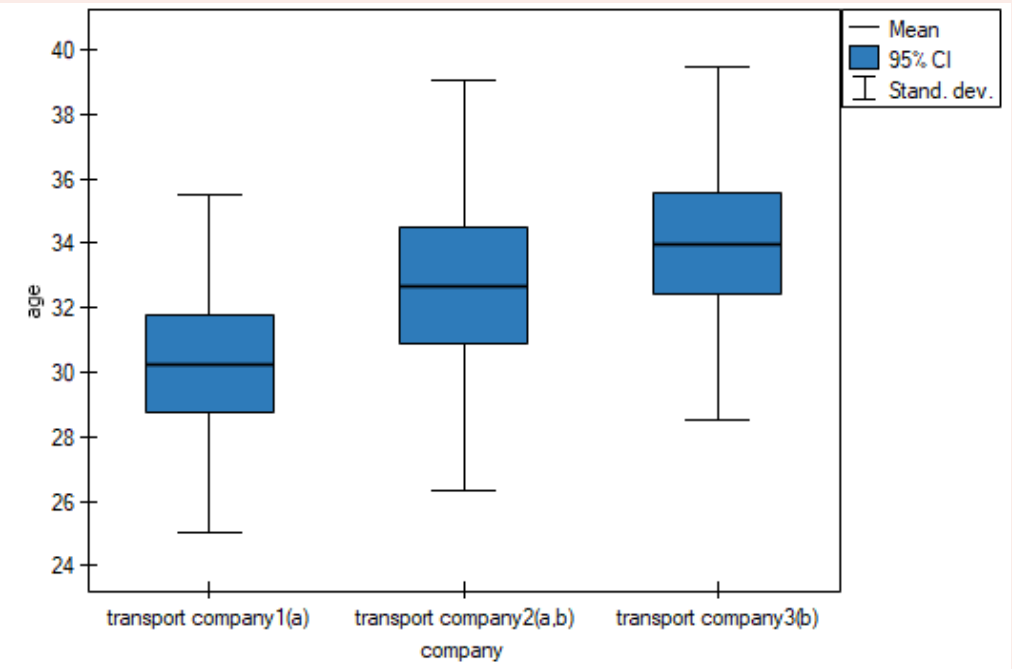
Copyright ©2010-2026 U-smile Stats – All rights reserved

POST-HOC (Tukey HSD)			
	transport company1	transport company2	transport company3
Difference of the r			
transport company1		2.42	3.72
transport company2	2.42		1.3
transport company3	3.72	1.3	
CD	transport company1	transport company2	transport company3
transport company1		2.7309	2.7309
transport company2	2.7309		2.7309
transport company3	2.7309	2.7309	
Statistic q	transport company1	transport company2	transport company3
transport company1		2.996	4.6054
transport company2	2.996		1.6094
transport company3	4.6054	1.6094	
p-value	transport company1	transport company2	transport company3
transport company1		0.0897	0.004
transport company2	0.0897		0.4923
transport company3	0.004	0.4923	
Homogeneous	transport company1(a)	transport company2(a,b)	transport company3(b)
A	*	*	
B		*	*

$p < 0.05$ **(statistically significant only for the comparison of company 1 and 3)**

A significant difference in average age applies only to the workers of the first and third transport companies.

The critical difference (CD) determined for each pair of comparisons is the same (because the group sizes are equal) and equals 2.7309. Comparing the CD value with the value of the mean difference shows that significant differences occur only between the mean age of workers of the first and third transport companies (only for this pair is the CD value smaller than the difference of means). We reach the same conclusion by comparing the p value of the POST-HOC test with the significance level $\alpha = 0.05$. Workers of the first company are, on average, just over 3 years younger than workers of the third company. Two overlapping homogeneous groups were obtained, which are also marked on the graph.



Copyright ©2010-2026 U-smile Stats – All rights reserved

Test for trend. If, besides the mere existence of differences between groups, we are interested in whether the means change according to a specific, ordered trend (e.g. increase or decrease across successive groups), we can use a variant of the POST-HOC tests based on contrasts – the so-called test for trend. It is run in the same options window as the other POST-HOC tests, by indicating the expected order of the groups. The construction of trend contrasts and the formula for the test statistic are described in the "For those interested" section.

FOR THOSE INTERESTED

Formulas: the number of all possible simple comparisons is calculated using the formula:

$$c = \binom{k}{2} = \frac{k(k-1)}{2}$$

Contrasts

Comparisons within contrasts and POST-HOC tests are written in the form of hypotheses.

The first example – **simple comparisons** (comparison of 2 selected means):

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2, \\ \mathcal{H}_1: &\quad \mu_1 \neq \mu_2. \end{aligned}$$

The second example – **complex comparisons** (comparison of a combination of selected means):

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \tfrac{\mu_2+\mu_3}{2}, \\ \mathcal{H}_1: &\quad \mu_1 \neq \tfrac{\mu_2+\mu_3}{2}. \end{aligned}$$

To define the selected hypotheses, a contrast value c_j, $(j = 1, 2, ...k)$ must be assigned to each mean. The c_j values are chosen so that the sums for the compared sides are opposite numbers, and their value for means not taking part in the comparison equals 0.

The first example: $c_1 = 1$, $c_2 = -1$, $c_3 = 0, ...c_k = 0$.

The second example: $c_1 = 2$, $c_2 = -1$, $c_3 = -1$, $c_4 = 0$,..., $c_k = 0$.

We can choose the correct hypothesis by:

(i) Comparing the difference between the selected means with the **critical difference (CD)** determined by the appropriate POST-HOC test: if the difference of means $\geq CD$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if the difference of means $< CD$, there is no reason to reject $\mathcal{H}_0$.

(ii) Comparing the p value, determined on the basis of the test statistic of the appropriate POST-HOC test, with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The LSD Fisher test
For simple and complex comparisons, equal-size groups as well as unequal-size groups, when the variances do not differ significantly.

Copyright ©2010-2026 U-smile Stats – All rights reserved

(i) The critical difference is calculated using the formula:

$$CD = \sqrt{F_{\alpha,1,df_{WG}}} \cdot \sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n_j}\right) MS_{WG}},$$

where $F_{\alpha,1,df_{WG}}$ is the critical value (statistic) of the F Snedecor distribution for a given significance level α and degrees of freedom 1 and df_{WG}, respectively.

(ii) The test statistic is defined by:

$$t = \frac{\sum_{j=1}^{k} c_j \overline{x}_j}{\sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n_j}\right) MS_{WG}}}.$$

The test statistic has the t-Student distribution with df_{WG} degrees of freedom.

The Scheffe test

For simple comparisons, equal-size groups as well as unequal-size groups, when the variances do not differ significantly.

(i) The critical difference is calculated using the formula:

$$CD = \sqrt{F_{\alpha,df_{BG},df_{WG}}} \cdot \sqrt{(k-1)\left(\sum_{j=1}^{k} \frac{c_j^2}{n_j}\right) MS_{WG}},$$

where $F_{\alpha,df_{BG},df_{WG}}$ is the critical value (statistic) of the F Snedecor distribution for a given significance level α and df_{BG} and df_{WG} degrees of freedom.

(ii) The test statistic is defined by:

$$F = \frac{\left(\sum_{j=1}^{k} c_j \overline{x}_j\right)^2}{(k-1)\left(\sum_{j=1}^{k} \frac{c_j^2}{n_j}\right) MS_{WG}}.$$

The test statistic has the F Snedecor distribution with df_{BG} and df_{WG} degrees of freedom.

The Tukey test

For simple comparisons, equal-size groups as well as unequal-size groups, when the variances do not differ significantly.

(i) The critical difference is calculated using the formula:

$$CD = \frac{\sqrt{2} \cdot q_{\alpha,df_{WG},k} \cdot \sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n_j}\right) MS_{WG}}}{2},$$

where $q_{\alpha,df_{WG},k}$ is the critical value (statistic) of the studentized range distribution for a given significance level α and df_{WG} and k degrees of freedom.

Copyright ©2010-2026 U-smile Stats – All rights reserved

(ii) The test statistic is defined by:

$$q = \sqrt{2}\frac{\sum_{j=1}^{k} c_j \overline{x}_j}{\sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n_j}\right) MS_{WG}}}.$$

The test statistic has the studentized range distribution with df_{WG} and k degrees of freedom.

Info. The algorithm for calculating the p value and the statistic of the studentized range distribution in PQStat is based on Lund's work (1983)[111]. Other applications or web pages may calculate slightly different values than PQStat, since they may be based on less precise or more restrictive algorithms (Copenhaver and Holland (1988), Gleason (1999)).

Test for trend

The test examining the existence of a trend can be calculated in the same situation as the ANOVA for independent variables, since it is based on the same assumptions, but it captures the alternative hypothesis differently – indicating the existence of a trend in the mean values across successive populations. The analysis of the trend in the arrangement of means is based on LSD Fisher contrasts. By building appropriate contrasts, you can study any type of trend, e.g. linear, quadratic, cubic, etc. Below is a table of sample contrast values for selected trends.

		Contrast									
Number of groups	Trend	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
3	linear	-1	0	1							
	quadratic	1	-2	1							
4	linear	-3	-1	1	3						
	quadratic	1	-1	-1	1						
	cubic	-1	3	-3	1						
5	linear	-2	-1	0	1	2					
	quadratic	2	-1	-2	-1	2					
	cubic	-1	2	0	-2	1					
6	linear	-5	-3	-1	1	3	5				
	quadratic	5	-1	-4	-4	-1	5				
	cubic	-5	7	4	-4	-7	5				
7	linear	-3	-2	-1	0	1	2	3			
	quadratic	5	0	-3	-4	-3	0	5			
	cubic	-1	1	1	0	-1	-1	1			
8	linear	-7	-5	-3	-1	1	3	5	7		
	quadratic	7	1	-3	-5	-5	-3	1	7		
	cubic	-7	5	7	3	-3	-7	-5	7		
9	linear	-4	-3	-2	-1	0	1	2	3	4	
	quadratic	28	7	-8	-17	-20	-17	-8	7	28	
	cubic	-14	7	13	9	0	-9	-13	-7	14	
10	linear	-9	-7	-5	-3	-1	1	3	5	7	9
	quadratic	6	2	-1	-3	-4	-4	-3	-1	2	6
	cubic	-42	14	35	31	12	-12	-31	-35	-14	42

Linear trend

A linear trend, like other trends, can be analyzed by entering the appropriate contrast values. However,

Copyright ©2010-2026 U-smile Stats – All rights reserved

if the direction of the linear trend is known, simply use the Linear trend option and indicate the expected order of the populations by assigning them consecutive natural numbers. The analysis is performed based on the linear contrast, i.e. the groups indicated according to the natural order are assigned appropriate contrast values, and the LSD Fisher statistic is calculated.

With the expected direction of the trend known, the alternative hypothesis is one-sided and the one-sided p value is interpreted. Interpreting the two-sided p value means that the researcher does not know (does not assume) the direction of the possible trend.

The p value, determined on the basis of the test statistic, is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The ANOVA for independent groups with F^* and F'' corrections

WHEN TO USE THIS TEST:

The F^* (Brown-Forsythe, 1974[34]) and F'' (Welch, 1951[173]) corrections are variants of the ANOVA for independent groups, used when the variances of the analysed variable differ significantly between the compared populations. The purpose of the test is the same as for the basic variant – only the way of calculating the test statistic and degrees of freedom differs (see *For those interested*).

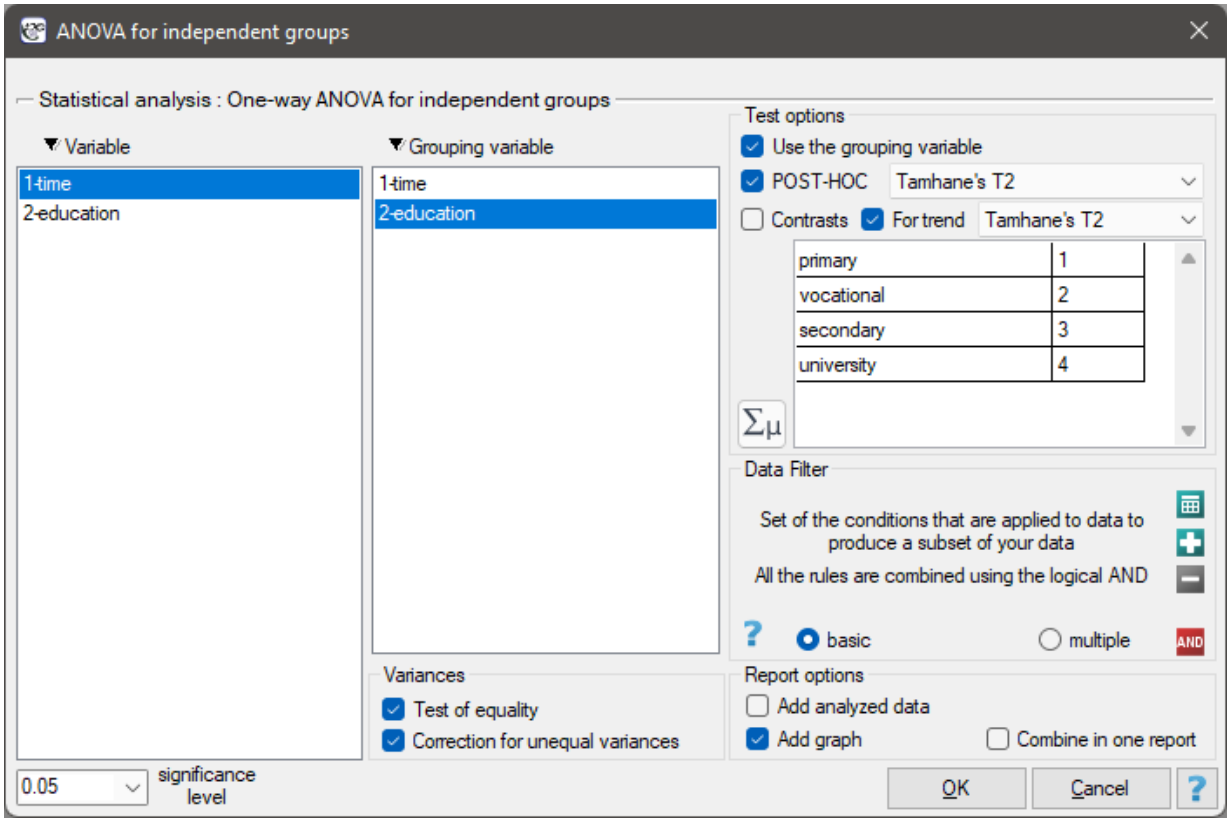
ASSUMPTIONS

- measurement on an interval scale,
- normality of distribution of the analysed variable in each population,
- an independent model,
- significantly different variances of the analysed variable between populations (otherwise use the ANOVA for independent groups without correction).

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → ANOVA for independent groups (or via the Wizard).

Copyright ©2010-2026 U-smile Stats – All rights reserved



Step 1. Select the grouping variable and the outcome variable.

Step 2. Check the assumption of equality of variances in the analysis window (Levene's test or the Brown-Forsythe test).

Step 3. If the variances differ significantly, run the analysis with the F^*/F'' correction and select an appropriate POST-HOC test (e.g. Tamhane's T2) for groups with unequal variances.

Step 4. Run the analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

Example 17.2. (unemployment.pqs file)
There are many factors that affect how long it takes to find a job during an economic crisis. One of the most important may be the level of education. Sample data on education and time (in months) spent unemployed are gathered in the file. We want to check whether there are differences in average job-search time between education categories.

Due to differences in variance between the populations (for Levene's test $p = 0.0001$, and for the Brown-Forsythe test $p = 0.0002$):

Copyright ©2010-2026 U-smile Stats – All rights reserved

Equality of variances - Brown-Forsythe	
F statistic	7.0103
p-value	0.0002
Equality of variances - Levene	
F statistic	7.4708
p-value	0.0001

the analysis is performed with the correction for unequal variances enabled. The resulting adjusted F statistic is shown below.

One-way ANOVA for independent groups	
Analysed variables	time
	education
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	education
ANOVA for independent groups	
Eta-square	0.1001
Total sum of squares (SS[T])	4226.3607
Between-groups sum of squares (SS[BG])	423.2131
Within-groups sum of squares (SS[WG])	3803.1475
Mean square between-groups (MS[BG])	141.071
Mean square within-groups (MS[WG])	15.8464
Between-groups degrees of freedom (df[BG])	3
Within-groups degrees of freedom (df[WG])	240
Total degrees of freedom (df[T])	243
F statistic	8.9024
p-value	<0.0001
Correction for unequal variances - F* (Brown-Forsyth	
F statistic	8.9024
Degrees of freedom	3 / 203.3014
p-value	<0.0001
Correction for unequal variances - F'' (Welch)	
F statistic	8.6631
Degrees of freedom	3 / 130.5548
p-value	<0.0001

$p < 0.0001$ for the F^* test and $p < 0.0001$ for the F'' test, both statistically significant, $\alpha = 0.05$

Average job-search time differs depending on the education one has.

By performing one of the POST-HOC tests dedicated to comparing groups with unequal variances (see *For those interested*), we find out which education categories the differences apply to:

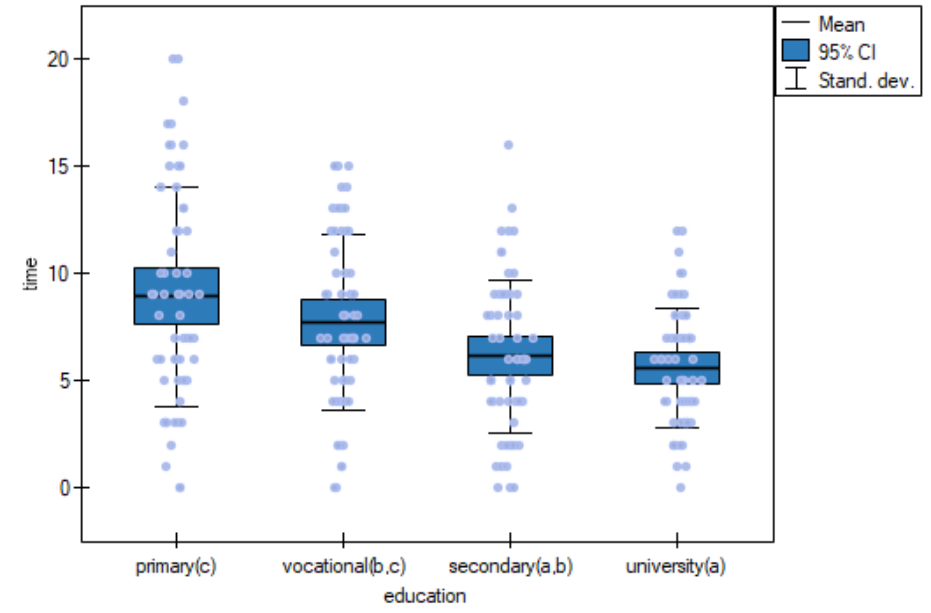
Copyright ©2010-2026 U-smile Stats – All rights reserved

POST-HOC (Tamhane's T2)

	primary	vocational	secondary	university
Difference of				
primary		1.2131	2.7869	3.3443
vocational	1.2131		1.5738	2.1311
secondary	2.7869	1.5738		0.5574
university	3.3443	2.1311	0.5574	
CD	primary	vocational	secondary	university
primary		2.2457	2.1376	2.0038
vocational	2.2457		1.8608	1.7024
secondary	2.1376	1.8608		1.551
university	2.0038	1.7024	1.551	
Statistic T	primary	vocational	secondary	university
primary		-1.4462	-3.4947	-4.4866
vocational	-1.4462		-2.2633	-3.3564
secondary	-3.4947	-2.2633		-0.9623
university	-4.4866	-3.3564	-0.9623	
p-value	primary	vocational	secondary	university
primary		0.6251	0.0041	0.0001
vocational	0.6251		0.1433	0.0066
secondary	0.0041	0.1433		0.9158
university	0.0001	0.0066	0.9158	
Homogeneous	primary(c)	vocational(b	secondary(a	university(a)
A			*	*
B		*	*	
C	*	*		

The critical difference (CD) determined for each pair of comparisons is not the same (even though the group sizes are equal), because the variances are not equal. Relating the CD value to the resulting differences in mean values gives the same result as comparing the p value with the significance level $\alpha = 0.05$. The differences concern primary and higher education, primary and secondary education, and vocational and higher education. The resulting homogeneous groups overlap. Overall, however, looking at the graph, we can expect that the more educated a person is, the less time it takes them to find a job.

Copyright ©2010-2026 U-smile Stats – All rights reserved



To test this hypothesis, we perform a trend analysis. To do this, we resume the analysis with the button and, in the test options window, select the Tamhane's T2 method together with the For trend option, indicating the order of the education categories with consecutive natural numbers.

Linear trend (Fisher LSD)	
The ordinal number for [primary]	1
The ordinal number for [vocational]	2
The ordinal number for [secondary]	3
The ordinal number for [university]	4
F statistic	25.9283
Two sided p-value	<0.0001
One sided p-value	<0.0001

Depending on whether the direction of the relationship between education and job-search time is known to us, we use the one-sided or two-sided p value. Both of these values are smaller than the assumed significance level – the trend we expected is confirmed, i.e. at the significance level $\alpha = 0.05$ we can say that this trend really does exist in the population from which the sample was drawn.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2 = ... = \mu_k, \\ \mathcal{H}_1: &\quad \text{not all } \mu_j \text{ are equal } (j = 1, 2, ..., k). \end{aligned}$$

Formulas: the test statistic is defined by:

$$F^* = \frac{SS_{BG}}{\sum_{j=1}^{k} \left(1 - \frac{n_j}{n} sd_j^2\right)},$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

$$F'' = \frac{\frac{\sum_{j=1}^{k} w_j(\overline{x}_j - \widetilde{x})}{k-1}}{1 + \frac{2(k-2)}{k^2-1}\sum_{j=1}^{k} h_j},$$

where:
sd_j – standard deviation of group j,
$w_j = \frac{n_j}{sd_j^2}$ – weight of group j,
$\widetilde{x}$ – weighted mean,
$h_j = \frac{\left(1 - \frac{w_j}{\sum_{j=1}^{k} w_j}\right)^2}{n_j - 1}$.

The test statistic has the F Snedecor distribution with $k-1$ and adjusted df_{WG_k} degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

POST-HOC tests

An introduction to contrasts and POST-HOC tests is provided in section 17.1.2, which concerns the one-way analysis of variance. The methods below, unlike the LSD Fisher, Scheffe and Tukey tests, are used when the variances of the analysed variable differ significantly.

The T2 Tamhane test
For simple and complex comparisons, equal-size groups as well as unequal-size groups (Tamhane A. C., 1977[161]).

(i) The critical difference is calculated using the formula:

$$CD = \sqrt{F_{\alpha_{Sidak},1,df_v}} \cdot \sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2 sd_j^2}{n_j}\right)},$$

where $F_{\alpha_{Sidak},1,df_v}$ is the critical value (statistic) of the F Snedecor distribution for the modified significance level α_{Sidak} and for degrees of freedom 1 and df_v, respectively,
$\alpha_{Sidak} = 1 - (1-\alpha)^{(1/k)}$,

$$df_v = \frac{\left(\sum_{j=1}^{k} \frac{c_j^2 sd_j^2}{n_j}\right)^2}{\sum_{j=1}^{k} \frac{c_j^4 sd_j^4}{n_j^2(n_j-1)}}$$

(ii) The test statistic is defined by:

$$t = \frac{\left(\sum_{j=1}^{k} c_j \overline{x}_j\right)^2}{\sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2 sd_j^2}{n_j}\right)}}.$$

The test statistic has the t-Student distribution with df_v degrees of freedom, and the p value is adjusted for the number of possible simple comparisons.

The BF test (Brown-Forsythe)
For simple comparisons, equal-size groups as well as unequal-size groups (Brown M. B. and Forsythe A. B., 1974[33]).

Copyright ©2010-2026 U-smile Stats – All rights reserved

(i) The critical difference is calculated using the formula:

$$CD = \sqrt{F_{\alpha,k-1,df_v}} \cdot \sqrt{(k-1)\left(\sum_{j=1}^{k} \frac{c_j^2 sd_j^2}{n_j}\right)},$$

where $F_{\alpha,k-1,df_v}$ is the critical value (statistic) of the F Snedecor distribution for a given significance level α and $k-1$ and df_v degrees of freedom.

(ii) The test statistic is defined by:

$$F = \frac{\left(\sum_{j=1}^{k} c_j \overline{x}_j\right)^2}{(k-1)\left(\sum_{j=1}^{k} \frac{c_j^2 sd_j^2}{n_j}\right)}.$$

The test statistic has the F Snedecor distribution with $k-1$ and df_v degrees of freedom.

The GH test (Games-Howell)

For simple comparisons, equal-size groups as well as unequal-size groups (Games P. A. and Howell J. F., 1976[69]).

(i) The critical difference is calculated using the formula:

$$CD = \frac{q_{\alpha,k,df_v} \cdot \sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2 sd_j^2}{n_j}\right)}}{\sqrt{2}},$$

where q_{α,k,df_v} is the critical value (statistic) of the studentized range distribution for a given significance level α and k and df_v degrees of freedom.

(ii) The test statistic is defined by:

$$q = \sqrt{2}\frac{\sum_{j=1}^{k} c_j \overline{x}_j}{\sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2 sd_j^2}{n_j}\right)}}.$$

The test statistic has the studentized range distribution with k and df_v degrees of freedom.

Test for trend

The test examining the existence of a trend can be calculated in the same situation as the ANOVA for independent groups with F^* and F'' corrections, since it is based on the same assumptions, but it captures the alternative hypothesis differently – indicating the existence of a trend in the mean values across successive populations. The analysis of the trend in the arrangement of means is based on T2 Tamhane contrasts. A table of sample contrast values for selected trends can be found in the description of the test for trend for the ANOVA without correction for unequal variances.

A linear trend is analyzed based on the linear contrast, assigning appropriate contrast values to the groups ordered naturally and calculating the T2 Tamhane statistic. With the expected direction of the trend known, the alternative hypothesis is one-sided and the one-sided p value is interpreted; interpreting the two-sided p value means that the researcher does not know (does not assume) the direction of the possible trend. The determined p value is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Brown-Forsythe test and the Levene test

WHEN TO USE THIS TEST:

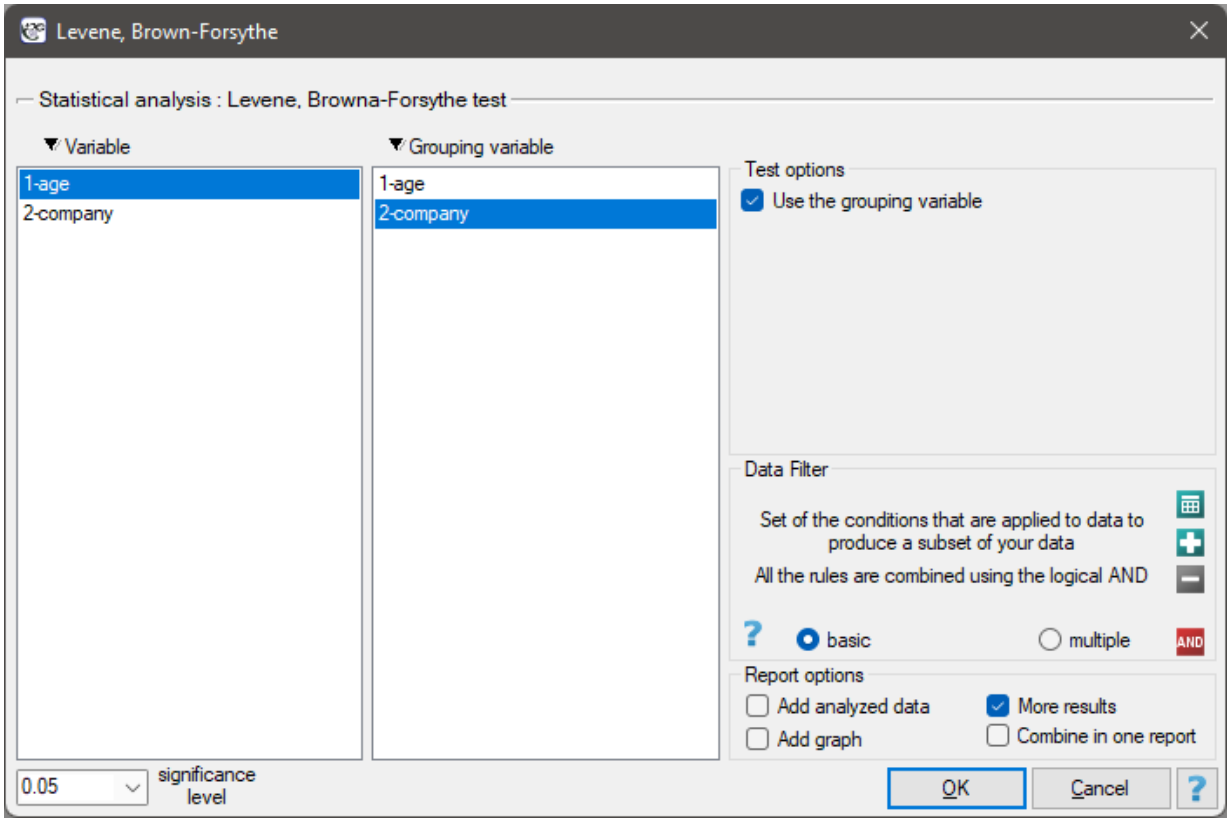
The Levene test (Levene, 1960[105]) and the Brown-Forsythe test (Brown and Forsythe, 1974[32]) let you check whether the spread (variance) of an analysed feature is the same in at least two populations. The result of this test is usually checked as a preliminary condition before performing the ANOVA for independent groups – it indicates whether the variant with the correction for unequal variances should be used.

ASSUMPTIONS

- measurement on an interval scale,
- normality of distribution of the analysed variable in each population,
- an independent model.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → Levene, Brown-Forsythe.



Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! The Brown-Forsythe test is less sensitive than the Levene test to a violation of the assumption of distribution normality.

PERFORMING AND INTERPRETING THE ANALYSIS

In practice, these tests are performed as part of a broader analysis – an example of their use and interpretation can be found in the ANOVA for independent groups with F^* and F'' corrections chapter (Example 17.2).

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2, \\ \mathcal{H}_1: &\quad \text{not all } \sigma_j^2 \text{ are equal } (j = 1, 2, ..., k), \end{aligned}$$

where σ_1^2,σ_2^2,...,σ_k^2 – variances of the analysed variable in the populations from which the samples were drawn.

Procedure: The analysis is based on calculating the absolute deviation of the measurement results from the mean (in the Levene test) or from the median (in the Brown-Forsythe test) in each of the analysed groups. This absolute deviation is the set of data which undergoes exactly the same procedure as the analysis of variance for independent groups.

Formulas: the test statistic is defined by:

$$F = \frac{MS_{BG}}{MS_{WG}},$$

The test statistic has the F Snedecor distribution with df_{BG} and df_{WG} degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The ANOVA for dependent groups

WHEN TO USE THIS TEST:

The single-factor repeated-measures analysis of variance, i.e. the ANOVA for dependent groups, lets you check whether the mean value of an analysed feature is the same across at least three measurements taken on the same objects under different conditions (e.g. before treatment, during treatment, and after treatment). The test only indicates whether such differences exist at all – it does not say between which measurements they occur; that is the role of the POST-HOC tests, described below.

ASSUMPTIONS

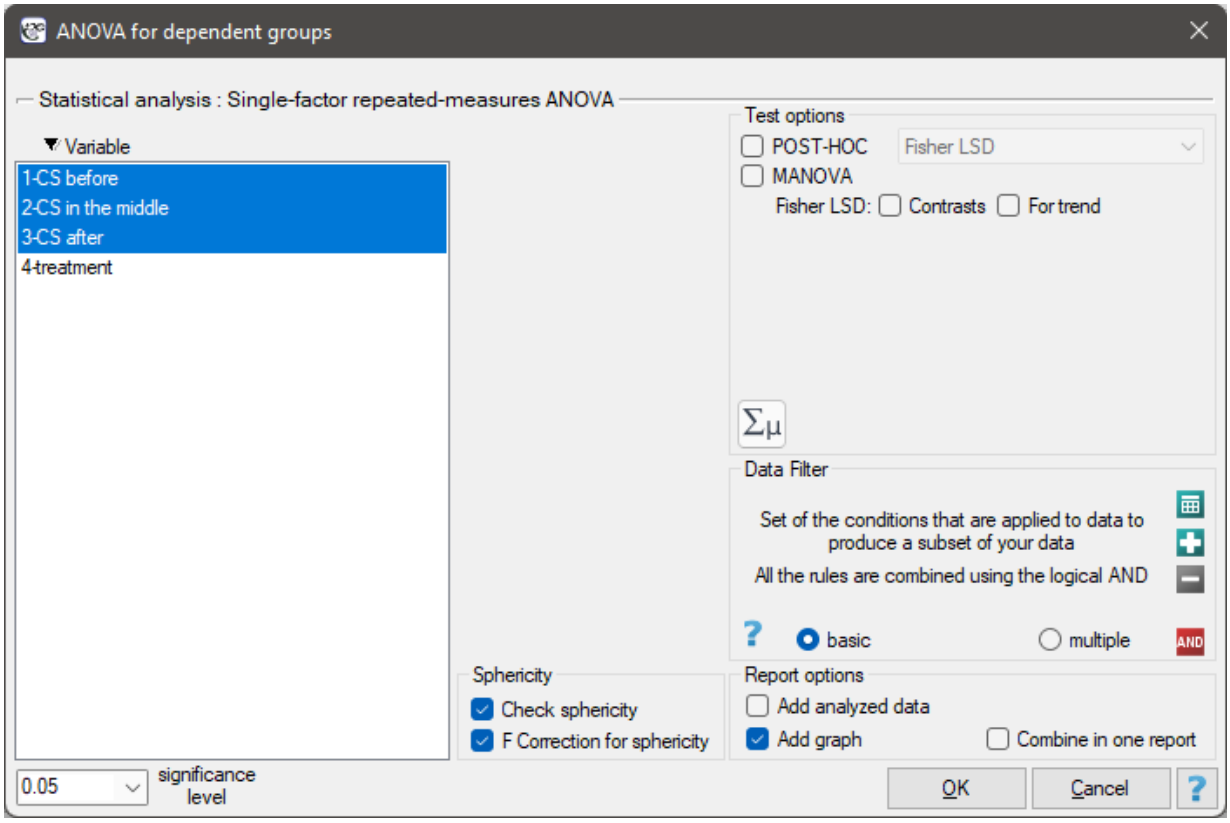
Copyright ©2010-2026 U-smile Stats – All rights reserved

- measurement on an interval scale,
- the normal distribution for all variables that are the differences of measurement pairs (or normality of the analysed variable for each measurement),
- a dependent model,
- sphericity: the variances of the differences between all pairs of measurements should be close to each other.

The sphericity assumption is checked using Mauchly's test. When it is not met, use the ANOVA for dependent groups with Epsilon correction and MANOVA.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → ANOVA for dependent groups (or via the Wizard).



Step 1. Select the variables corresponding to successive measurements.

Step 2. Check the sphericity assumption (Mauchly's test).

Step 3. Run the analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

Since evaluating the result of the ANOVA for dependent groups is inseparable from first checking the sphericity assumption, a full example of performing and interpreting this test – including a situation where the assumption is violated – can be found in the ANOVA for dependent groups with Epsilon correction and MANOVA chapter (see Example 17.3).

DEFINITIONS OF MEASURES

Partial η^2 – an effect-size measure indicating what proportion of the within-subject variability of the outcomes results from the repeated measurements of the analysed variable, and what proportion remains unexplained. The higher the value, the stronger the association between the measurement and the outcome of the analysed feature.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2 = ... = \mu_k, \\ \mathcal{H}_1: &\quad \text{not all } \mu_j \text{ are equal } (j = 1, 2, ..., k), \end{aligned}$$

where μ_1,μ_2,...,μ_k – means of the analysed feature at successive measurements in the population studied.

Formulas: the test statistic is defined by:

$$F = \frac{MS_{BC}}{MS_{res}},$$

where:

$MS_{BC} = \dfrac{SS_{BC}}{df_{BC}}$ – mean square between-conditions,

$MS_{res} = \dfrac{SS_{res}}{df_{res}}$ – mean square residual,

$SS_{BC} = \sum_{j=1}^{k} \left(\dfrac{(\sum_{i=1}^{n} x_{ij})^2}{n} \right) - \dfrac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}\right)^2}{N}$ – between-conditions sum of squares,

$SS_{res} = SS_T - SS_{BS} - SS_{BC}$ – residual sum of squares,

$SS_T = \left(\sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}^2 \right) - \dfrac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}\right)^2}{N}$ – total sum of squares,

$SS_{BS} = \sum_{i=1}^{n} \left(\dfrac{\left(\sum_{j=1}^{k} x_{ij}\right)^2}{k} \right) - \dfrac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}\right)^2}{N}$ – between-subjects sum of squares,

$df_{BC} = k - 1$ – degrees of freedom (between-conditions),

$df_{res} = df_T - df_{BC} - df_{BS}$ – degrees of freedom (residual),

$df_T = N - 1$ – total degrees of freedom,

$df_{BS} = n - 1$ – degrees of freedom (between-subjects),

$N = nk$,

n – sample size,

Copyright ©2010-2026 U-smile Stats – All rights reserved

x_{ij} – values of the variable for i subjects $(i = 1, 2, ...n)$ in j conditions $(j = 1, 2, ...k)$.

The test statistic has the F Snedecor distribution with df_{BC} and df_{res} degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Partial η^2

$$\eta^2 = \frac{SS_{BC}}{SS_{BC} + SS_{res}}$$

POST-HOC tests

An introduction to contrasts and POST-HOC tests is provided in section 17.1.2, which concerns the one-way analysis of variance for independent groups. The methods below apply analogously to repeated measures (the counts in each measurement are always the same).

The LSD Fisher test
For simple and complex comparisons.

Example – **simple comparisons** (comparison of 2 selected means):

$$\begin{aligned} \mathcal{H}_0 &: \quad \mu_j = \mu_{j+1}, \\ \mathcal{H}_1 &: \quad \mu_j \neq \mu_{j+1}. \end{aligned}$$

(i) The critical difference is calculated using the formula:

$$CD = \sqrt{F_{\alpha,1,df_{res}}} \cdot \sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n}\right) MS_{res}},$$

where $F_{\alpha,1,df_{res}}$ is the critical value (statistic) of the F Snedecor distribution for a given significance level α and degrees of freedom 1 and df_{res}, respectively.

(ii) The test statistic is defined by:

$$t = \frac{\sum_{j=1}^{k} c_j \overline{x}_j}{\sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n}\right) MS_{res}}}.$$

The test statistic has the t-Student distribution with df_{res} degrees of freedom. For contrasts, $\sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n}\right) MS_{res}}$ is replaced by the contrast standard error $SE_{contrast}$, and the degrees of freedom become df_{BS}.

The Scheffe test
For simple comparisons.

(i) The critical difference is calculated using the formula:

$$CD = \sqrt{F_{\alpha,df_{BC},df_{res}}} \cdot \sqrt{(k-1)\left(\sum_{j=1}^{k} \frac{c_j^2}{n}\right) MS_{res}},$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where $F_{\alpha,df_{BC},df_{res}}$ is the critical value (statistic) of the F Snedecor distribution for a given significance level α and df_{BC} and df_{res} degrees of freedom.

(ii) The test statistic is defined by:

$$F = \frac{\left(\sum_{j=1}^{k} c_j \overline{x}_j\right)^2}{(k-1)\left(\sum_{j=1}^{k} \frac{c_j^2}{n}\right) MS_{res}}.$$

The test statistic has the F Snedecor distribution with df_{BC} and df_{res} degrees of freedom.

The Tukey test
For simple comparisons.

(i) The critical difference is calculated using the formula:

$$CD = \frac{\sqrt{2} \cdot q_{\alpha,df_{res},k} \cdot \sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n}\right) MS_{res}}}{2},$$

where $q_{\alpha,df_{res},k}$ is the critical value (statistic) of the studentized range distribution for a given significance level α and df_{res} and k degrees of freedom.

(ii) The test statistic is defined by:

$$q = \sqrt{2} \frac{\sum_{j=1}^{k} c_j \overline{x}_j}{\sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n}\right) MS_{res}}}.$$

The test statistic has the studentized range distribution with df_{res} and k degrees of freedom.

Info. The algorithm for calculating the p value and the statistic of the studentized range distribution in PQStat is based on Lund's work (1983)[111]. Other applications or web pages may calculate slightly different values than PQStat, since they may be based on less precise or more restrictive algorithms (Copenhaver and Holland (1988), Gleason (1999)).

Test for trend
The test examining the existence of a trend can be calculated in the same situation as the ANOVA for dependent variables, since it is based on the same assumptions, but it captures the alternative hypothesis differently – indicating the existence of a trend in the mean values across successive measurements. The analysis of the trend in the arrangement of means is based on LSD Fisher contrasts. A table of sample contrast values for selected trends can be found in the description of the test for trend for the ANOVA of independent variables.

A linear trend is analyzed based on the linear contrast, assigning appropriate contrast values to the measurements ordered naturally and calculating the LSD Fisher statistic. With the expected direction of the trend known, the alternative hypothesis is one-sided and the one-sided p value is interpreted; interpreting the two-sided p value means that the researcher does not know (does not assume) the direction of the possible trend. The determined p value is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Mauchly's sphericity

WHEN TO USE THIS TEST:

The Mauchly test (1940[121]) checks the sphericity assumption – a condition similar to, but stronger than, equality of variances, required before performing the ANOVA for dependent groups. This assumption is met if the variances of the differences between all pairs of repeated measurements are close to each other. A significant test result ($p \leq \alpha$) means the sphericity assumption is violated and an appropriate correction must be used (see ANOVA for dependent groups with Epsilon correction and MANOVA).

ASSUMPTIONS

- measurement on an interval scale,
- multivariate normal distribution, or normality of distribution of each analysed variable,
- a dependent model.

Instead of the sphericity assumption, a simpler but more stringent condition is sometimes considered – compound symmetry – which assumes symmetry of the covariance matrix: equal variances of the variables and equal covariances between them. Meeting the compound symmetry condition entails meeting the sphericity assumption.

PERFORMING AND INTERPRETING THE ANALYSIS

In practice, this test is performed as part of a broader analysis – an example of its use and interpretation can be found in the ANOVA for dependent groups with Epsilon correction and MANOVA chapter (Example 17.3).

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \sigma_{diff(1)} = \sigma_{diff(2)} = ... = \sigma_{diff(s)}, \\ \mathcal{H}_1: &\quad \text{not all } \sigma_{diff(i)} \text{ are equal } (i = 1, 2, ..., s), \end{aligned}$$

where $\sigma_{diff(i)}$ – population variance of the differences between the i-th pair of repeated measurements, s – number of pairs.

Formulas: Mauchly's W value is defined as follows:

$$W = \frac{\prod_{j=1}^{k-1} \lambda_j}{\left(\frac{1}{k-1}\sum_{j=1}^{k-1} \lambda_j\right)^{k-1}}.$$

The test statistic is defined by:

$$\chi^2 = (f-1)(n-1)\ln W,$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where:
$f = \frac{2(k-1)^2+(k-1)+2}{6(k-1)(n-1)}$,
λ_j – eigenvalue of the expected covariance matrix,
k – number of variables analyzed.

Interpretation: this statistic asymptotically (for large sample sizes) has the χ^2 distribution with $df = \frac{k(k-1)}{2} - 1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

A value of $W \approx 1$ indicates that the sphericity assumption is met. When interpreting the results of this test, however, it is important to remember that it is sensitive to violations of the normality assumption.

The ANOVA for dependent groups with Epsilon correction and MANOVA

WHEN TO USE THIS TEST:

The Epsilon corrections and the multivariate MANOVA approach apply to the ANOVA for dependent groups and are used when the sphericity assumption is not met, i.e. when the variances of the differences between all pairs of measurements are not close to each other.

ASSUMPTIONS

- measurement on an interval scale,
- the normal distribution for all variables that are the differences of measurement pairs,
- a dependent model,
- lack of sphericity (otherwise the ANOVA for dependent groups without correction is sufficient).

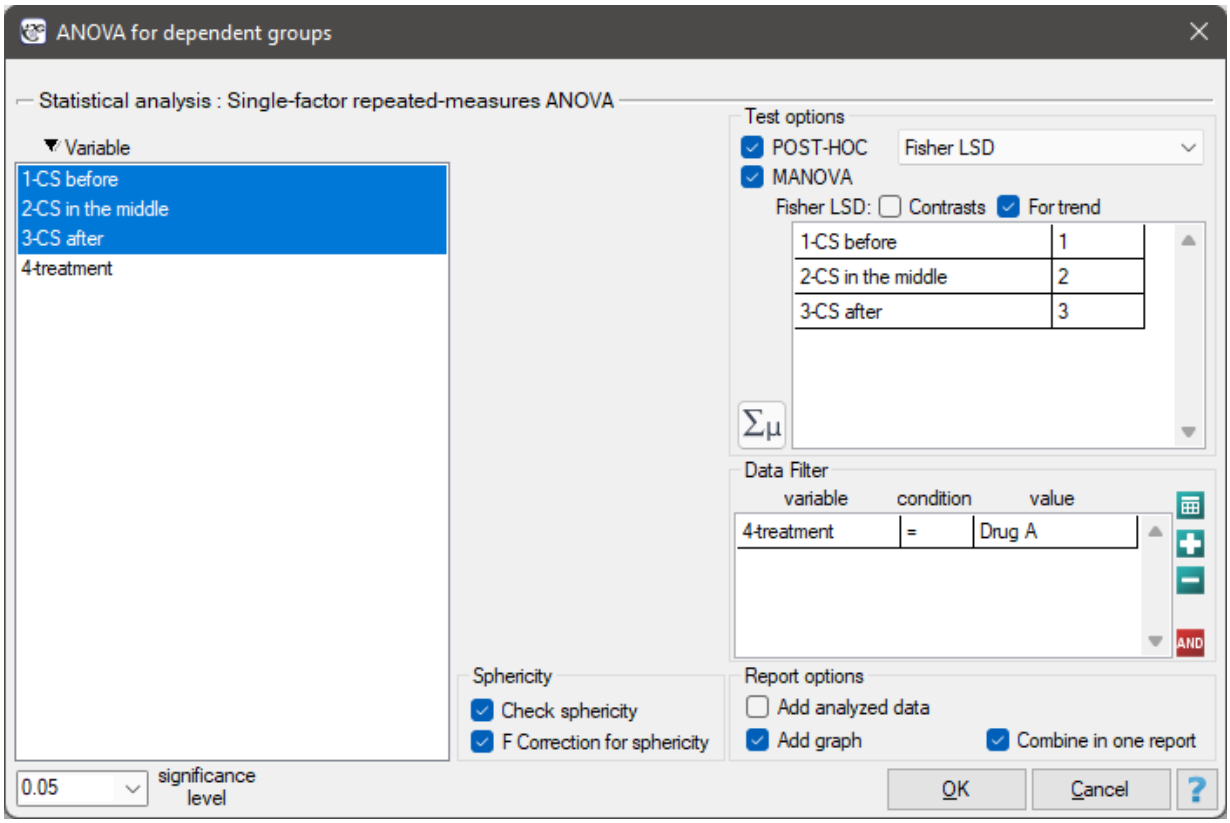
The degree to which sphericity is met is represented by the value of W in Mauchly's test, as well as by the **Epsilon** (ε) values calculated together with the corrections. $\varepsilon = 1$ indicates strict adherence to the sphericity condition – the smaller the value of Epsilon is below 1, the more the sphericity assumption is violated (the lower limit Epsilon can reach is $\frac{1}{k-1}$).

Three corrections are available for adjusting the degrees of freedom of the F test: the simplest but weakest **Epsilon lower-bound correction**; the somewhat stronger, though still conservative, **Greenhouse-Geisser correction** (1959)[73]; and the strongest **Huynh-Feldt correction** (1976)[89]. When sphericity is significantly violated, however, it is most advisable to perform an analysis that does not require this assumption, namely **MANOVA** (*multivariate analysis of variance*) – applied to repeated measures (without an independent-groups factor), it tests the same hypothesis as the ANOVA for dependent variables, based on the differences between adjacent measurements. A broader discussion of this method can be found in the one-way MANOVA chapter.

HOW TO SET UP THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

Statistics → Parametric tests → ANOVA for dependent groups (or via the Wizard).



Step 1. Check the sphericity assumption with Mauchly's test (ideally filtering the data by group, if the study has more than one group).

Step 2. If sphericity is violated, select an appropriate Epsilon correction or the MANOVA statistic in the options window.

Step 3. Run the analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 17.3. (pressure.pqs file)
The effectiveness of two treatments for hypertension was studied. A sample of 56 patients was collected and randomly assigned to two groups: a group treated with drug A and a group treated with drug B. Systolic blood pressure was measured three times in each group: before treatment, during treatment, and after 6 months of treatment. We will check whether, in each group, the mean systolic blood pressure changes across the treatment stages.

Since the data are normally distributed, we begin the analysis by checking the sphericity assumption – testing is performed for each group separately, using a multiple filter. Failure to meet the sphericity assumption in the group treated with drug B is indicated both by the observed values of the covariance and correlation matrices and by the result of Mauchly's test ($W = 0.68$, $p = 0.0063$).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Covariance matrix			
	CS before	CS in the mi	CS after
CS before	181.9153	81.8942	-23.7619
CS in the middle	81.8942	172.6177	81.7143
CS after	-23.7619	81.7143	270.9206

Correlation matrix			
	CS before	CS in the mi	CS after
CS before	1	0.4621	-0.107
CS in the middle	0.4621	1	0.3779
CS after	-0.107	0.3779	1

SPHERICITY ASSUMPTION	
Mauchly sphericity	
W	0.677
Chi-square statistic	10.1439
Degrees of freedom	2
p-value	0.0063

We resume the analysis and, in the test options window, select the primary filter to perform a repeated-measures ANOVA for those treated with drug A, followed by a correction of this analysis and the MANOVA statistic for those treated with drug B.

Results for those treated with drug A:

Copyright ©2010-2026 U-smile Stats – All rights reserved

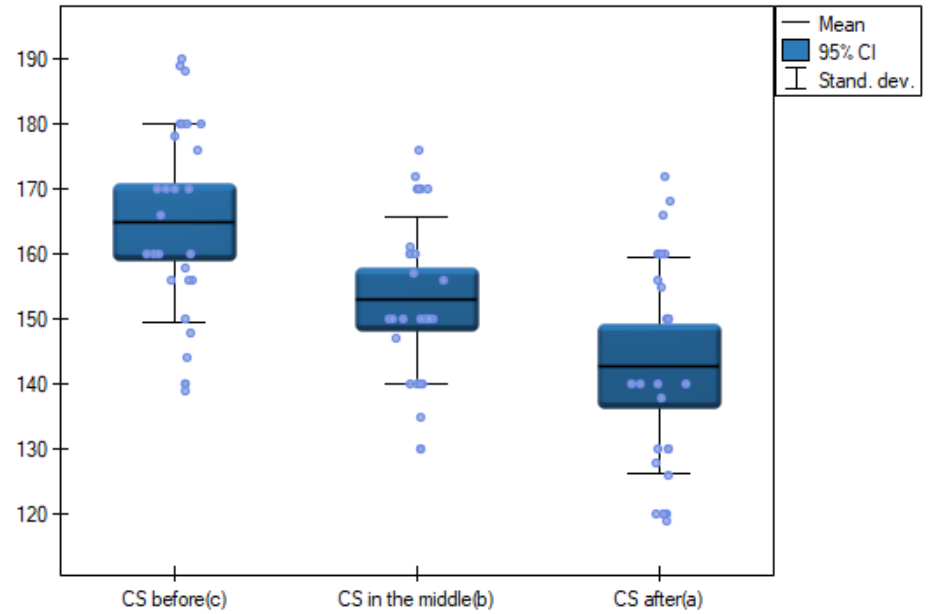
Single-factor repeated-measures ANOVA	
Analysed variables	CS before
	CS in the middle
	CS after
Data Filter	treatment=Drug A
Number of unspecified	0
Number of missing data (rows)	0
Significance level	0.05
ANOVA for dependent groups	
Eta-square	0.6639
Total sum of squares (SS[T])	24960.9524
Between-conditions sum of squares (SS[BC])	6787.5238
Between-subjects sum of squares (SS[BS])	14736.9524
Residual sum of squares (SS[RES])	3436.4762
Between-conditions degrees of freedom (df[BC])	2
Between-subjects degrees of freedom (df[BS])	27
Residual degrees of freedom (df[RES])	54
Total degrees of freedom (df[T])	83
Mean square between-conditions (MS[BC])	3393.7619
Mean square between-subjects (MS[BS])	545.8131
Mean square residual (MS[RES])	63.6384
F statistic	53.3288
p-value	<0.0001

Copyright ©2010-2026 U-smile Stats – All rights reserved

POST-HOC (Fisher LSD)			
	CS before	CS in the mi	CS after
Difference of t			
CS before		11.7857	22
CS in the middle	11.7857		10.2143
CS after	22	10.2143	
CD	CS before	CS in the mi	CS after
CS before		4.2745	4.2745
CS in the middle	4.2745		4.2745
CS after	4.2745	4.2745	
Statistic t	CS before	CS in the mi	CS after
CS before		5.5279	10.3187
CS in the middle	5.5279		4.7908
CS after	10.3187	4.7908	
p-value	CS before	CS in the mi	CS after
CS before		<0.0001	<0.0001
CS in the middle	<0.0001		<0.0001
CS after	<0.0001	<0.0001	
Jednorodne	CS before(c	CS in the mi	CS after(a)
A			*
B		*	
C	*		

Linear trend	
The ordinal number for [CS before]	1
The ordinal number for [CS in the middle]	2
The ordinal number for [CS after]	3
F statistic	106.4765
Two sided p-value	<0.0001
One sided p-value	<0.0001

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p < 0.0001$ for those treated with drug A (statistically significant, $\alpha = 0.05$)

Systolic blood pressure differs significantly between the treatment stages for drug A.

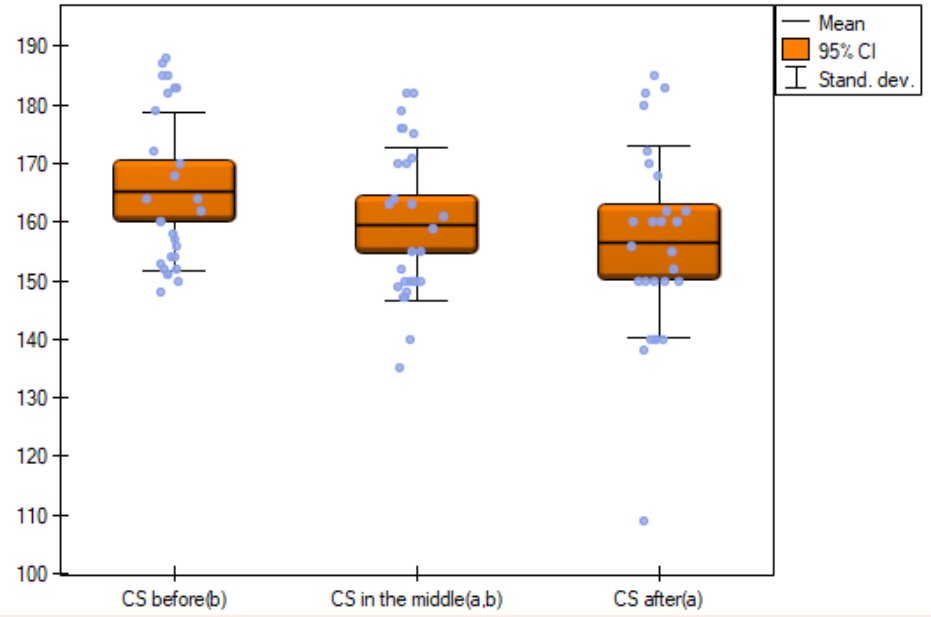
More than 66% of the within-subject variation in outcomes can be explained by the use of drug A ($\eta^2 = 0.66$). The differences apply to all treatment stages compared (POST-HOC result). The decreasing trend in systolic blood pressure due to treatment is also significant ($p < 0.0001$) – so drug A can be considered effective.

Results for those treated with drug B:

Correction - GG (Greenhouse-Geisser)	
GG Epsilon	0.7558
df1	1.5117
df2	40.8149
p-value	0.0561
Correction - HF (Huyhn-Feldt)	
HF Epsilon	0.7911
df1	1.5822
df2	42.7183
p-value	0.0536
Correction – lower-bound Epsilon	
Lower-bound Epsilon	0.5
df1	1
df2	27
p-value	0.0768

Copyright ©2010-2026 U-smile Stats – All rights reserved

MANOVA	
Wilk's Lambda	0.8331
Eta-square	0.1669
df1	2
df2	26
F statistic	2.6049
p-value	0.0931
Hotelling-Lawley Trace	0.2004
Eta-square	0.1669
df1	2
df2	26
F statistic	2.6049
p-value	0.0931
Pillai-Bartlett Trace	0.1669
Eta-square	0.1669
df1	2
df2	26
F statistic	2.6049
p-value	0.0931



$p > 0.05$ for those treated with drug B, both for the epsilon corrections and for Wilks' Lambda (not statistically significant, $\alpha = 0.05$)

No significant differences in systolic blood pressure between treatment stages for drug B.

Barely under 17% of the within-subject variation in outcomes can be explained by the use of drug B ($\eta^2 = 0.17$).

Copyright ©2010-2026 U-smile Stats – All rights reserved

17.2 NON-PARAMETRIC TESTS

The Kruskal-Wallis ANOVA

WHEN TO USE THIS TEST:

The Kruskal-Wallis one-way analysis of variance by ranks[99][100] is the non-parametric equivalent of the ANOVA for independent groups, extending the Mann-Whitney U test to more than two populations. It lets you check whether the distribution of the analysed feature (most often: its median) is the same in at least three independent populations – it is used instead of the ANOVA for independent groups when that test's assumptions are not met.

ASSUMPTIONS

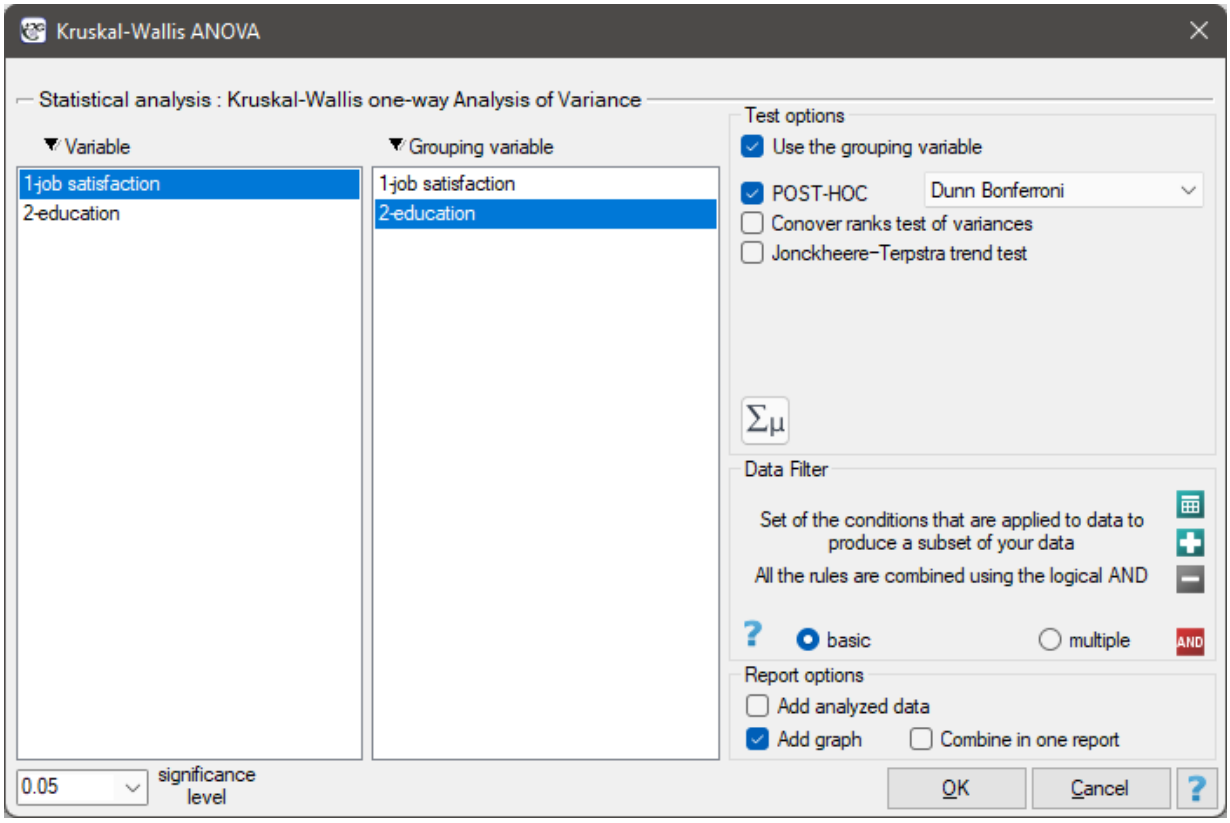
- measurement on an ordinal scale or an interval scale,
- an independent model.

The test assumes that the distributions of the analysed variable in the compared populations have a similar shape – equality of rank variances can be checked with Conover's rank test. If we are interested in a trend in the arrangement of the studied groups, we can additionally perform the Jonckheere-Terpstra test for trend.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Kruskal-Wallis ANOVA (or via the Wizard).

Copyright ©2010-2026 U-smile Stats – All rights reserved



Step 1. Select the grouping variable and the outcome variable.

Step 2. If the test shows a significant difference, select a POST-HOC test, e.g. Dunn or Conover-Iman, together with an appropriate correction (e.g. Bonferroni).

Step 3. Run the analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 17.4. (jobSatisfaction.pqs file)
A group of 120 people was surveyed, for whom the job they currently hold is their first job obtained after completing their education. The respondents rated their job satisfaction on a five-point scale, where: 1 – unsatisfying job, 2 – job giving little satisfaction, 3 – job giving an average level of satisfaction, 4 – job giving a fairly high level of satisfaction, 5 – very satisfying job. We will check whether the level of reported job satisfaction differs between education categories.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Kruskal-Wallis one-way Analysis of Variance	
Analysed variables	job satisfaction
	education
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	education
Kruskal-Wallis ANOVA	
Degrees of freedom	3
H statistic (adjusted for ties)	16.1954
p-value	0.001

$p = 0.001034$ **(statistically significant,** $\alpha = 0.05$**)**

The level of job satisfaction differs statistically significantly between education categories.

The Dunn POST-HOC analysis with the Bonferroni correction shows that significant differences occur between people with primary and secondary education, and between people with primary and tertiary education. Slightly more differences can be confirmed by choosing the stronger Conover-Iman POST-HOC test.

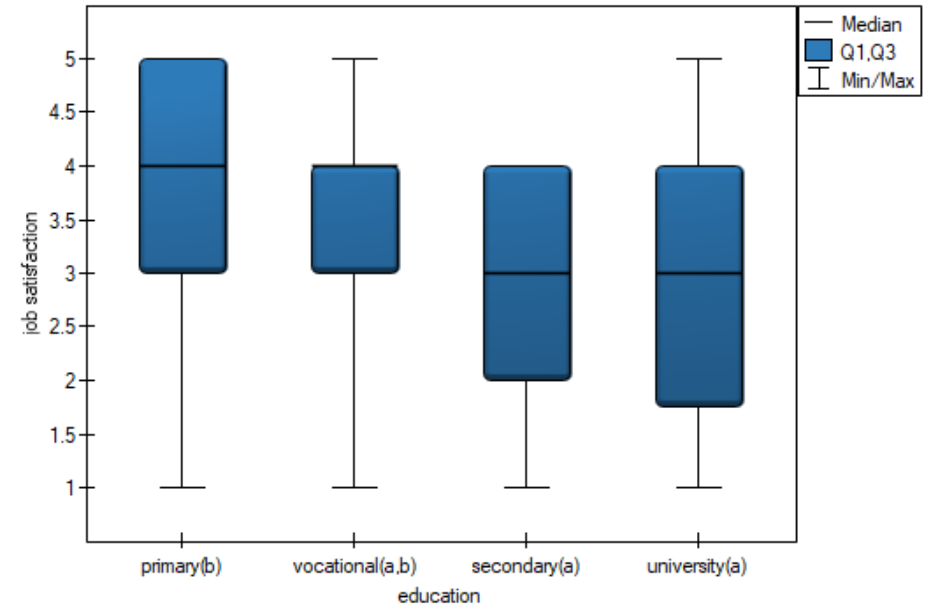
POST-HOC (Dunn Bonferroni)				
	primary	vocational	secondary	university
The mean ran				
primary		11.5833	30.4306	30.5208
vocational	11.5833		18.8472	18.9375
secondary	30.4306	18.8472		0.0903
university	30.5208	18.9375	0.0903	
CD	primary	vocational	secondary	university
primary		23.5125	23.5125	25.7566
vocational	23.5125		21.0302	23.5125
secondary	23.5125	21.0302		23.5125
university	25.7566	23.5125	23.5125	
Statistic	primary	vocational	secondary	university
primary		1.2997	3.4145	3.1263
vocational	1.2997		2.3644	2.1249
secondary	3.4145	2.3644		0.0101
university	3.1263	2.1249	0.0101	
p-value	primary	vocational	secondary	university
primary		1	0.0038	0.0106
vocational	1		0.1084	0.2016
secondary	0.0038	0.1084		1
university	0.0106	0.2016	1	
Homogeneous	primary(b)	vocational(a	secondary(a	university(a)
A		*	*	*
B	*	*		

Copyright ©2010-2026 U-smile Stats – All rights reserved

POST-HOC (Dunn Benjamini-Hochberg)				
	primary	vocational	secondary	university
The mean ran				
primary		11.5833	30.4306	30.5208
vocational	11.5833		18.8472	18.9375
secondary	30.4306	18.8472		0.0903
university	30.5208	18.9375	0.0903	
CD	primary	vocational	secondary	university
primary		NA	NA	NA
vocational	NA		NA	NA
secondary	NA	NA		NA
university	NA	NA	NA	
Statistic	primary	vocational	secondary	university
primary		1.2997	3.4145	3.1263
vocational	1.2997		2.3644	2.1249
secondary	3.4145	2.3644		0.0101
university	3.1263	2.1249	0.0101	
p-value	primary	vocational	secondary	university
primary		0.2324	0.0038	0.0053
vocational	0.2324		0.0361	0.0504
secondary	0.0038	0.0361		0.9919
university	0.0053	0.0504	0.9919	
Homogeneous	primary(c)	vocational(b	secondary(a	university(a,
A			*	*
B		*		*
C	*	*		

The graph showing medians and quartiles reveals the homogeneous groups determined by the POST-HOC test. For the Dunn result with the Bonferroni correction, two homogeneous groups are visible which are not fully distinct: group (a) – people who rate job satisfaction lower, and group (b) – people who rate it higher. Vocational education belongs to both of these groups, meaning that people with this education rate job satisfaction quite variably.

Copyright ©2010-2026 U-smile Stats – All rights reserved



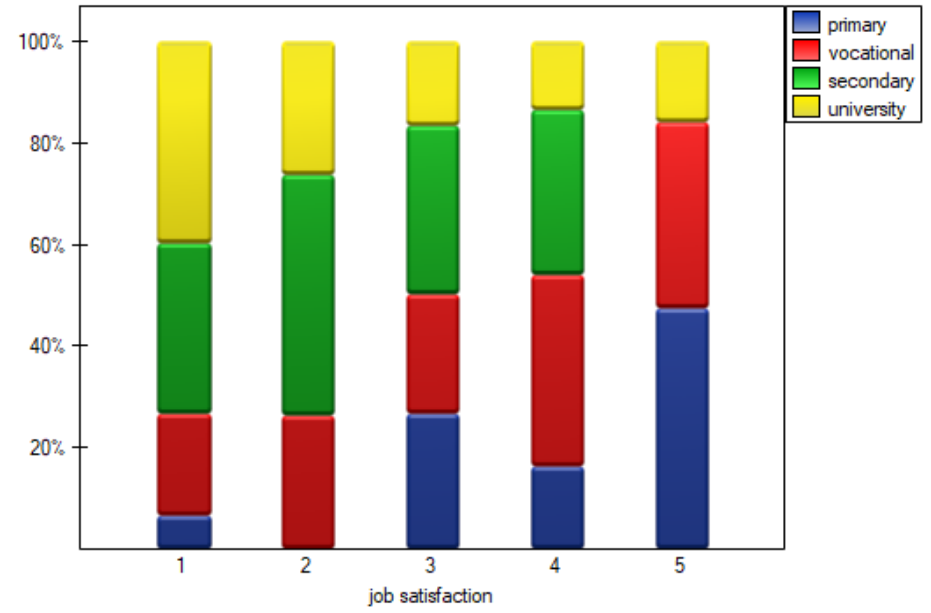
We can provide a detailed description of the data by selecting Descriptive statistics in the analysis window Σμ and choosing to add counts and percentages to the description.

Summary				
Group	primary	vocational	secondary	university
Sample size	24	36	36	24
Median	4	4	3	3
Lower quartile	3	3	2	1.75
Upper quartile	5	4	4	4

Frequency(Percent)				
Group	primary	vocational	secondary	university
1	1 (4.167%)	3 (8.333%)	5 (13.889%)	6 (25%)
2	0 (0%)	5 (13.889%)	9 (25%)	5 (20.833%)
3	8 (33.333%)	7 (19.444%)	10 (27.778%	5 (20.833%)
4	6 (25%)	14 (38.889%	12 (33.333%	5 (20.833%)
5	9 (37.5%)	7 (19.444%)	0 (0%)	3 (12.5%)
summary	24	36	36	24

We can also show the distribution of responses in a column chart.

Copyright ©2010-2026 U-smile Stats – All rights reserved



FOR THOSE INTERESTED

Hypotheses: they relate to the equality of mean ranks for successive populations or are simplified to medians:

$$\begin{aligned} \mathcal{H}_0: &\quad \phi_1 = \phi_2 = ... = \phi_k, \\ \mathcal{H}_1: &\quad \text{not all } \phi_j \text{ are equal } (j = 1, 2, ..., k), \end{aligned}$$

where $\phi_1, \phi_2, ...\phi_k$ are the distributions of the analysed variable in the populations from which the samples were drawn.

Formulas: the test statistic is defined by:

$$H = \frac{1}{C}\left(\frac{12}{N(N+1)}\sum_{j=1}^{k}\left(\frac{\left(\sum_{i=1}^{n_j} R_{ij}\right)^2}{n_j}\right) - 3(N+1)\right),$$

where:
$N = \sum_{j=1}^{k} n_j$,
n_j – sample sizes for $(j = 1, 2, ...k)$,
R_{ij} – ranks assigned to the values of the variable, for $(i = 1, 2, ...n_j)$, $(j = 1, 2, ...k)$,
$C = 1 - \frac{\sum(t^3 - t)}{N^3 - N}$ – correction for ties,
t – number of cases included in a tie. The correction C is used when ties occur (otherwise $C = 1$).

The H statistic asymptotically (for large sample sizes) has the χ^2 distribution with $df = (k-1)$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

POST-HOC tests

An introduction to contrasts and POST-HOC tests is provided in section 17.1.2, which concerns the one-way analysis of variance.

The Dunn test

For simple comparisons, equal-size groups as well as unequal-size groups (Dunn, 1964[53]). It includes a correction for tied ranks (Zar, 2010[187]) and is corrected for multiple testing – most often with the Bonferroni or Sidak correction (other, newer corrections are described in more detail in the Multiple comparisons section).

(i) The critical difference is calculated using the formula:

$$CD = Z_{\alpha(corrected)} \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum(t^3-t)}{12(N-1)}\right) \sum_{j=1}^{k} \frac{c_j^2}{n_j}},$$

where t is the number of cases included in a tie, and $Z_{\alpha(corrected)}$ is the critical value (statistic) of the normal distribution for the significance level α adjusted for the number of possible simple comparisons c according to the selected correction.

(ii) The test statistic is defined by:

$$Z = \frac{\sum_{j=1}^{k} c_j \overline{R}_j}{\sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum(t^3-t)}{12(N-1)}\right) \sum_{j=1}^{k} \frac{c_j^2}{n_j}}},$$

where $\overline{R}_j$ – mean rank of group j, for $(j = 1, 2, ...k)$. The test statistic asymptotically (for large sample sizes) has the normal distribution, and the p value is adjusted for the number of possible simple comparisons c according to the selected correction.

The Conover-Inman test

The non-parametric equivalent of LSD Fisher[48], used for simple comparisons of both equal-size and unequal-size groups.

(i) The critical difference is calculated using the formula:

$$CD = \sqrt{F_{\alpha,1,N-k}} \cdot \sqrt{S^2 \frac{N-1-H}{N-k} \sum_{j=1}^{k} \frac{c_j^2}{n_j}},$$

where $S^2 = \frac{1}{N-1}\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} R_{ij}^2 - N\frac{(N+1)^2}{4}\right)$, and $F_{\alpha,1,N-k}$ is the critical value (statistic) of the F Snedecor distribution for a given significance level α and for degrees of freedom 1 and $N-k$, respectively.

(ii) The test statistic is defined by:

$$t = \frac{\sum_{j=1}^{k} c_j \overline{R}_j}{\sqrt{S^2 \frac{N-1-H}{N-k} \sum_{j=1}^{k} \frac{c_j^2}{n_j}}},$$

where $\overline{R}_j$ – mean rank of group j, for $(j = 1, 2, ...k)$. The test statistic has the t-Student distribution with $N-k$ degrees of freedom.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Jonckheere-Terpstra test for trend

WHEN TO USE THIS TEST:

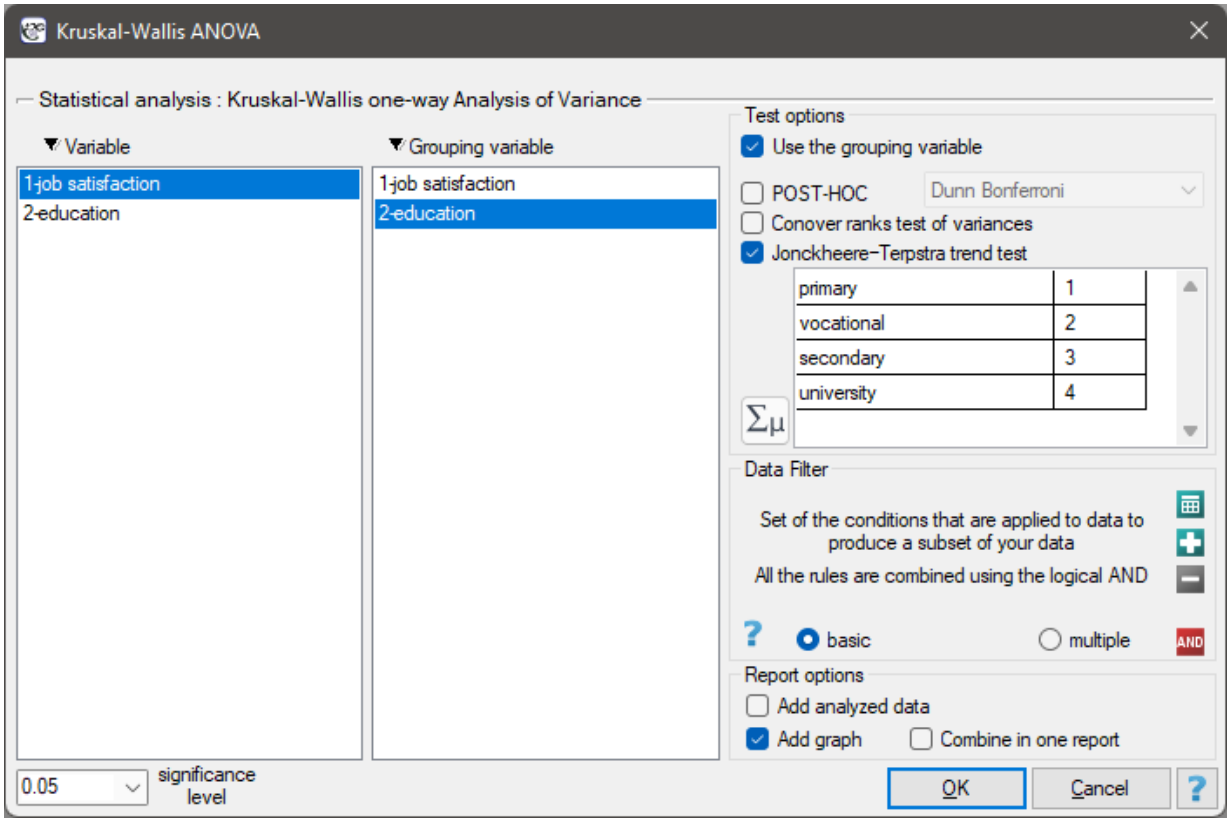
The Jonckheere-Terpstra test for trend, described independently by Jonckheere (1954)[93] and Terpstra (1952)[164], can be performed in the same situation as the Kruskal-Wallis ANOVA, since it is based on the same assumptions. Unlike the Kruskal-Wallis ANOVA, however, it does not check for arbitrary differences between groups, but whether the medians of successive, ordered populations increase or decrease in the expected direction.

ASSUMPTIONS

- measurement on an ordinal scale or an interval scale,
- an independent model,
- a known or assumed order of the compared populations.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Kruskal-Wallis ANOVA (or via the Wizard).



Copyright ©2010-2026 U-smile Stats – All rights reserved

Step 1. Resume the analysis using the button.

Step 2. Select the Jonckheere-Terpstra trend test option.

Step 3. Indicate the expected order of the populations by assigning them consecutive natural numbers.

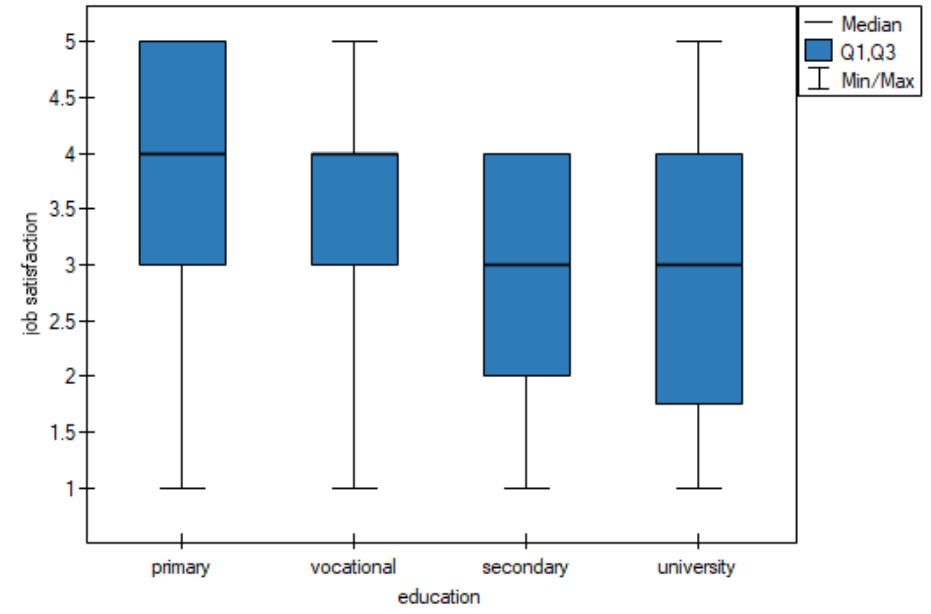
Step 4. Run the analysis.

Note! The phrase "with at least one strict inequality", used in the alternative hypothesis of this test, means that the median of at least one population should be greater than the median of another population, in the expected order.

PERFORMING AND INTERPRETING THE ANALYSIS

Continuation of Example 17.4 (jobSatisfaction.pqs file): it is suspected that better-educated people have higher job demands, which may reduce satisfaction with their first job, which often does not meet such demands. We will therefore check whether the level of satisfaction with the first job changes in the expected direction with the level of education.

Jonckheere–Terpstra trend test	
The ordinal number for [primary]	1
The ordinal number for [vocational]	2
The ordinal number for [secondary]	3
The ordinal number for [university]	4
L statistic	1856
Z statistic	3.9157
One sided p-value	<0.0001
Two sided p-value	0.0001

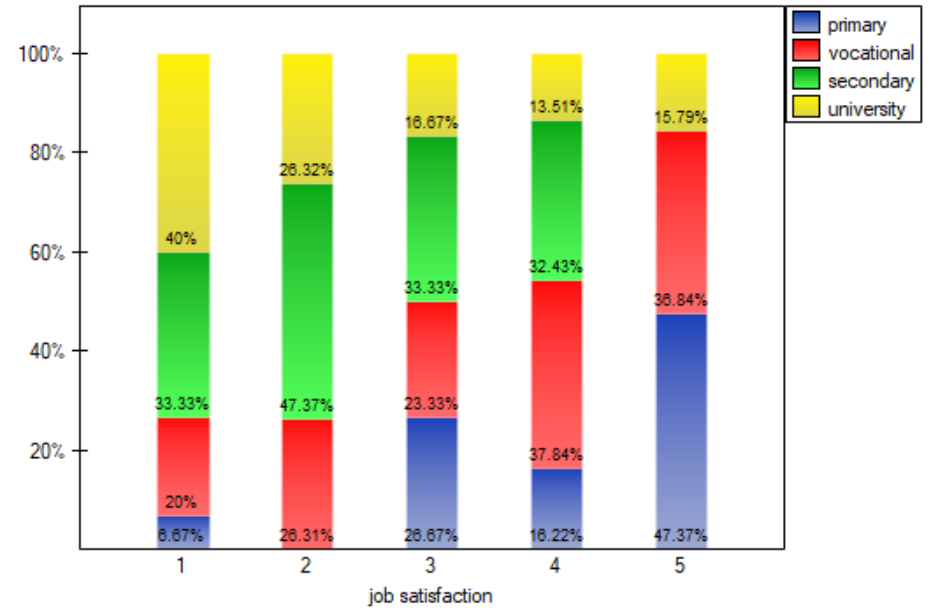


$p = 0.000045$ **(one-sided, statistically significant,** $\alpha = 0.05$**)**

The level of satisfaction with the first job really does change in the expected direction of the trend.

Copyright ©2010-2026 U-smile Stats – All rights reserved

We can also confirm the existence of this trend by showing the percentage distribution of the responses obtained.



FOR THOSE INTERESTED

Hypotheses (simplified to medians):

$$\begin{aligned}\mathcal{H}_0: &\quad \theta_1 = \theta_2 = ... = \theta_k,\\ \mathcal{H}_1: &\quad \theta_1 \geq \theta_2 \geq ... \geq \theta_k, \text{ with at least one strict inequality}\end{aligned}$$

Formulas: The test statistic is defined by:

$$Z = \frac{L - \left[\frac{N^2 - \sum_{j=1}^{k} n_j^2}{4}\right]}{SE}$$

where:
L – sum of the l_{ij} values obtained for each pair of compared populations,
l_{ij} – number of results higher than a given value in the next group,
$SE = \sqrt{\frac{A}{72} + \frac{B}{36N(N-1)(N-2)} + \frac{C}{8N(N-1)}}$,
$A = N(N-1)(2N+5) - \sum_{j=1}^{k} n_j(n_j-1)(2n_j+5) - \sum_{l=1}^{g} t_l(t_l-1)(2t_l+5)$,
$B = \sum_{j=1}^{k} n_j(n_j-1)(n_j-2) \cdot \sum_{l=1}^{g} t_l(t_l-1)(t_l-2)$,
$C = \sum_{j=1}^{k} n_j(n_j-1) \cdot \sum_{l=1}^{g} t_l(t_l-1)$,
g – number of groups of different tied ranks,
t_l – number of cases included in a tie,
$N = \sum_{j=1}^{k} n_j$,

Copyright ©2010-2026 U-smile Stats – All rights reserved

n_j – sample sizes for $(j = 1, 2, ...k)$.

Interpretation: The formula for the test statistic Z includes a correction for tied ranks. This correction is applied when tied ranks are present (when there are no tied ranks, the formula reduces to the original Jonckheere-Terpstra formula without this correction). The Z statistic asymptotically (for large sample sizes) has the normal distribution.

With the expected direction of the trend known, the alternative hypothesis is one-sided and the one-sided p value is interpreted. Interpreting the two-sided p value means that the researcher does not know (does not assume) the direction of the possible trend.

The p value, determined on the basis of the test statistic, is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The Conover ranks test of variance

WHEN TO USE THIS TEST:

The Conover squared ranks test lets you check whether the dispersion (variability) of an analysed feature is similar across several populations – similarly to the Fisher-Snedecor test (for two populations) or the Levene and Brown-Forsythe tests (for at least two populations). It is their non-parametric counterpart: it does not assume normality of the data distribution and is based on ranks[48]. It is usually used as a complement to the Kruskal-Wallis ANOVA, since this test examines the variability of the data (distances from the mean) rather than its level.

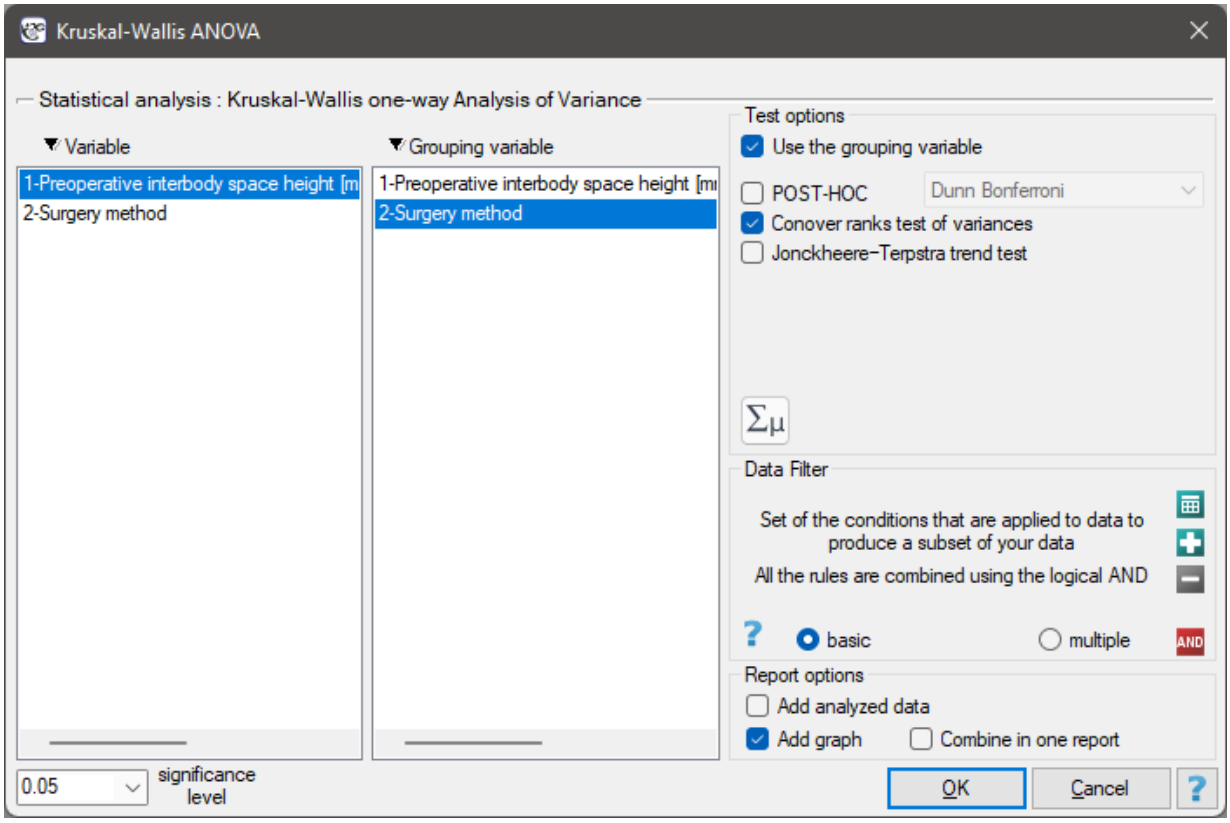
ASSUMPTIONS

– measurement on an interval scale.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Kruskal-Wallis ANOVA, option Conover ranks test of variance, or Statistics → Nonparametric tests → Mann-Whitney, option Conover ranks test of variance.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 17.5. (surgeryMethod.pqs file)
Patients have been prepared for spinal surgery and will be operated on using one of three methods. Each patient has been provisionally assigned to a surgery type. At a later stage we intend to compare patients' condition after surgery, so we want the groups of patients to be comparable – they should be similar, among other things, in terms of the height of the interbody space (WPMT) before surgery. The similarity should concern not only the average values but also the variability of the groups.

The distribution of the data was checked.

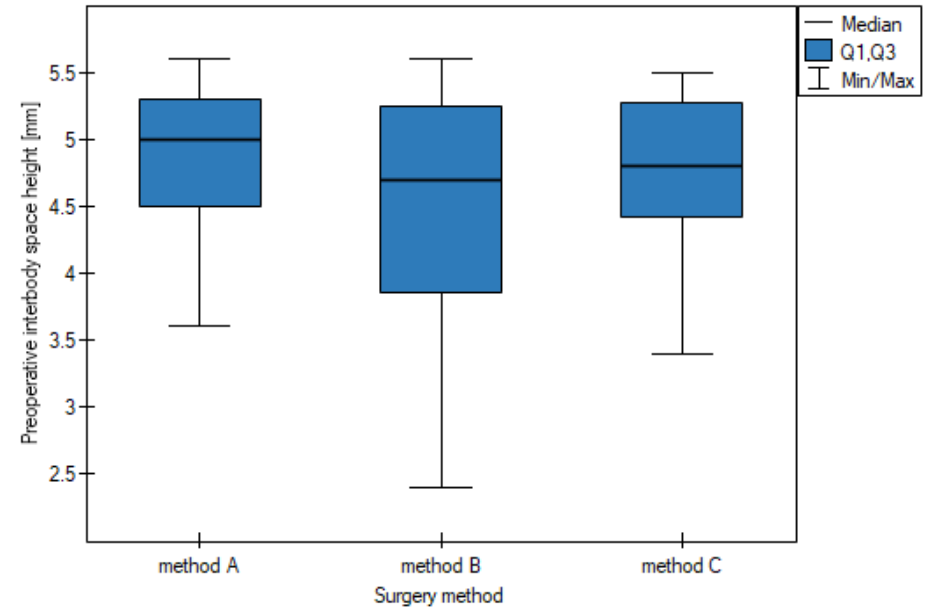
Copyright ©2010-2026 U-smile Stats – All rights reserved

One-dimensional normality	ID1	ID2	ID3
Analysed variables	Preoperative ir	Preoperative ir	Preoperative ir
Data Filter	Surgery methc	Surgery methc	Surgery methc
Group size	21	55	46
Number of unspecified	0	0	0
Number of missing data	0	0	0
Significance level	0.05	0.05	0.05
Group mean	4.8762	4.5218	4.7957
Group standard deviation	0.5718	0.8315	0.5296
Shapiro-Wilk test			
W statistic	0.9178	0.931	0.9309
Z statistic	1.4157	2.6875	2.3618
p-value	0.0784	0.0036	0.0091
D'Agostino-Pearson test			
K-square statistic	2.9229	4.1294	3.6689
Degrees of freedom	2	2	2
p-value	0.2319	0.1269	0.1597
Skewness test			
Skewness	-0.8514	-0.6489	-0.6467
Z statistic	-1.7051	-1.9901	-1.8381
p-value	0.0882	0.0466	0.0661
Kurtosis test (g2)			
Kurtosis	-0.0954	-0.3252	0.2169
Z statistic	0.124	-0.4109	0.5389
p-value	0.9013	0.6811	0.59

It turns out that for two of the surgery methods, WPMT shows deviations from normality, largely caused by the skewness of the data. We will conduct further comparative analysis using the Kruskal-Wallis test, to check whether the WPMT level differs between the methods, and the Conover test, to indicate whether the spread of WPMT scores is similar for each method.

Kruskal-Wallis one-way Analysis of Variance	
Analysed variables	Preoperative interbody space
	Surgery method
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	Surgery method
Kruskal-Wallis ANOVA	
Degrees of freedom	2
H statistic (adjusted for ties)	3.1627
p-value	0.2057
Conover ranks test of variances	
Chi-square statistic	12.2712
Degrees of freedom	2
p-value	0.0022

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p = 0.0022$ for the Conover variance test (statistically significant, $\alpha = 0.05$), $p = 0.2057$ for the Kruskal-Wallis test (not statistically significant)

The WPMT ranges differ significantly between surgery methods, while the WPMT level does not.

We first interpret the result of the Conover variance test, which indicates statistically significant differences in the ranges of the compared groups. The graph suggests that the differences mainly concern group 3. Since differences in WPMT range were detected, the interpretation of the Kruskal-Wallis test result (comparing WPMT levels for these methods) should be cautious, since this test is sensitive to heterogeneity of variance. Although the Kruskal-Wallis test showed no significant differences, it is recommended that patients with low WPMT values (mostly assigned to surgery method B) be distributed more evenly – i.e. that we check whether they could instead be offered surgery method A or C. After reassigning patients, the analysis should be repeated.

FOR THOSE INTERESTED

Hypotheses:

$\mathcal{H}_0$: the dispersion of the data in the compared populations is the same,
$\mathcal{H}_1$: at least two populations differ in the amount of data dispersion.

Formulas: The test statistic is defined by:

$$\chi^2 = \frac{1}{D^2}\left(\sum_{j=1}^{k}\frac{S_j^2}{n_j} - N\overline{S}^2\right)$$

where:
$N = n_1 + n_2 + ... + n_k$,

Copyright ©2010-2026 U-smile Stats – All rights reserved

n_j – size of the individual groups,
S_j – sum of squared ranks in group j,
$\overline{S} = \frac{1}{N}\sum_{j=1}^{k} S_j$ – mean of all squared ranks,
$D^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} R_i^4 - N\overline{S}^2\right)$,
R_i – ranks for values representing the distance of the measurement from the mean of a given group.

This statistic has the χ^2 distribution with $k-1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The Friedman ANOVA

WHEN TO USE THIS TEST:

The Friedman repeated-measures analysis of variance by ranks (Friedman, 1937[67]) is the non-parametric equivalent of the ANOVA for dependent groups. It is used when measurements of the analysed feature are taken several times on the same objects, and the assumptions of the parametric test are not met – also when we have rankings from different judges and want to assess their agreement. The Iman-Davenport (1980[90]) modification of this statistic is now recommended in place of the traditional Friedman statistic, which tends to be overly conservative.

ASSUMPTIONS

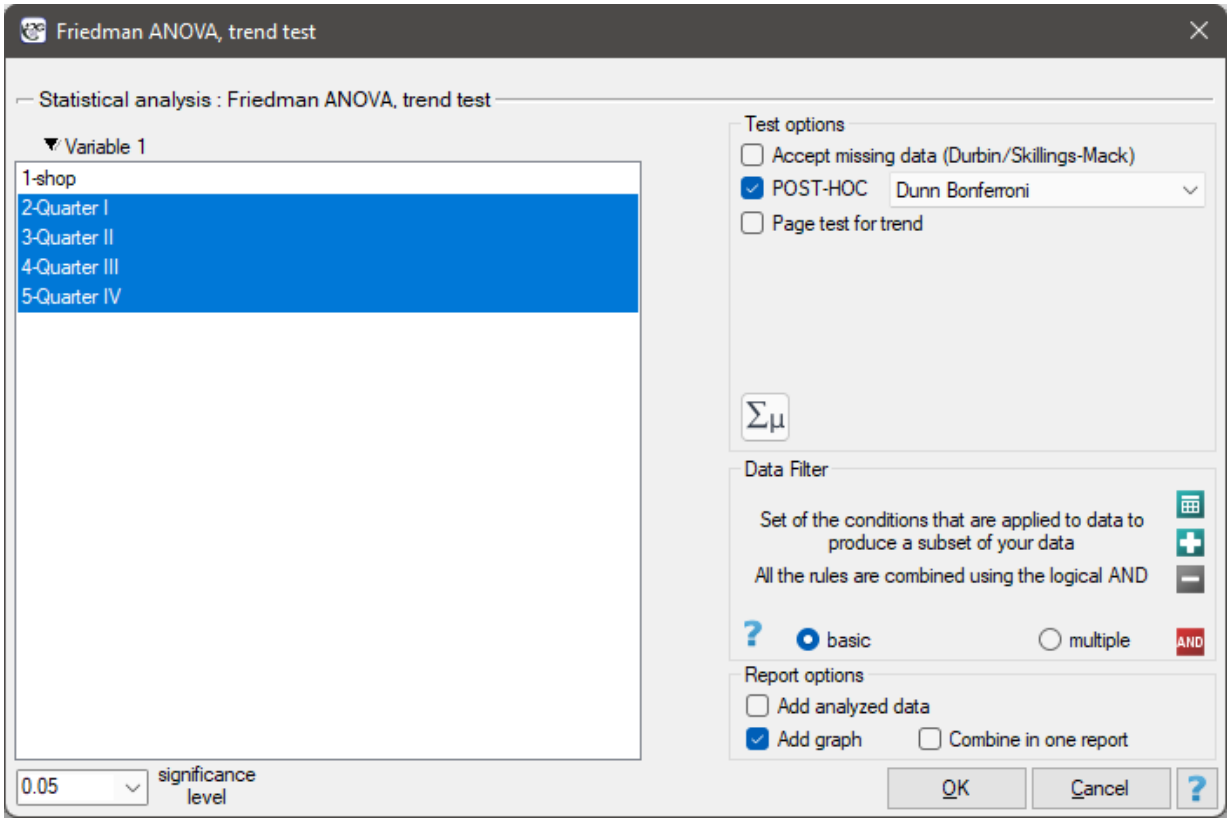
- measurement on an ordinal scale or an interval scale,
- a dependent model.

When there are gaps in the data, you can instead calculate the Durbin ANOVA (balanced gaps) or the Skillings-Mack ANOVA (gaps in any arrangement). If we are interested in a trend in the arrangement of the studied groups, we can additionally perform the Page test for trend.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Friedman ANOVA (missing data allowed) (or via the Wizard).

Copyright ©2010-2026 U-smile Stats – All rights reserved



Step 1. Select the variables corresponding to successive measurements.

Step 2. If the test shows a significant difference, select a POST-HOC test, e.g. Dunn or Conover-Iman, together with an appropriate correction (e.g. Bonferroni).

Step 3. Run the analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 17.6. (chocolate bar.pqs file)
Quarterly sales of a certain chocolate bar were measured in 14 randomly chosen supermarkets. The study began in January and ended in December. During the second quarter, an intensive billboard advertising campaign for the product was underway. We will check whether the campaign had an effect on the sales of the advertised chocolate bar.

Shop	Quarter I	Quarter II	Quarter III	Quarter IV
SK1	3415	4556	5772	5432
SK2	1593	1937	2242	2794
SK3	1976	2056	2240	2085
SK4	1526	1594	1644	1705
SK5	1538	1634	1866	1769
SK6	983	1086	1135	1177
SK7	1050	1209	1245	977
SK8	1861	2087	2054	2018
SK9	1714	2415	2361	2424
SK10	1320	1621	1624	1551
SK11	1276	1377	1522	1412
SK12	1263	1279	1350	1490
SK13	1271	1417	1583	1513
SK14	1436	1310	1357	1468

Copyright © 2010-2026 Usmile Stats - All rights reserved

Friedman ANOVA, trend test	
Analysed variables	Quarter I
	Quarter II
	Quarter III
	Quarter IV
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Accept missing data (Durbin/Skillings-Mack)	No
C	350
A	420
T1 statistic Friedman	23.9143
Degrees of freedom	3
p-value	<0.0001
T2 statistic Iman-Davenport	17.1896
Degrees of freedom	3 / 39
p-value	<0.0001
Skillings-Mack statistic	23.9143
Degrees of freedom	3
p-value	<0.0001

$p < 0.05$ for the Friedman test and for the Iman-Davenport correction, statistically significant, $\alpha = 0.05$

Sales of the bar are not the same in every quarter.

The Dunn POST-HOC analysis with the Bonferroni correction indicates differences in sales between quarters I and III, and between I and IV, while the equivalent analysis using the stronger Conover-Iman test indicates differences between all quarters except III and IV.

Copyright ©2010-2026 U-smile Stats – All rights reserved

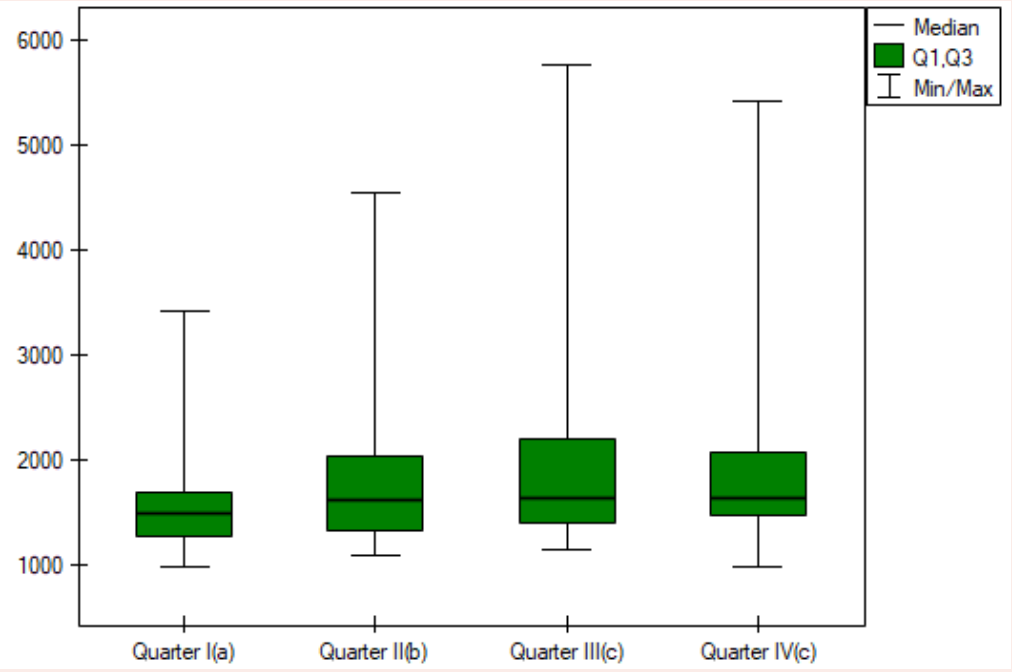
POST-HOC (Dunn Bonferroni)				
	Quarter I	Quarter II	Quarter III	Quarter IV
The mean ran				
Quarter I		1.0714	2.1429	1.9286
Quarter II	1.0714		1.0714	0.8571
Quarter III	2.1429	1.0714		0.2143
Quarter IV	1.9286	0.8571	0.2143	
CD	Quarter I	Quarter II	Quarter III	Quarter IV
Quarter I		1.2873	1.2873	1.2873
Quarter II	1.2873		1.2873	1.2873
Quarter III	1.2873	1.2873		1.2873
Quarter IV	1.2873	1.2873	1.2873	
Statistic	Quarter I	Quarter II	Quarter III	Quarter IV
Quarter I		2.1958	4.3916	3.9524
Quarter II	2.1958		2.1958	1.7566
Quarter III	4.3916	2.1958		0.4392
Quarter IV	3.9524	1.7566	0.4392	
p-value	Quarter I	Quarter II	Quarter III	Quarter IV
Quarter I		0.1686	0.0001	0.0005
Quarter II	0.1686		0.1686	0.4739
Quarter III	0.0001	0.1686		1
Quarter IV	0.0005	0.4739	1	
Homogeneous	Quarter I(a)	Quarter II(a	Quarter III(l	Quarter IV(b
A	*	*		
B		*	*	*

Copyright ©2010-2026 U-smile Stats – All rights reserved

POST-HOC (Conover-Iman)

	Quarter I	Quarter II	Quarter III	Quarter IV
The sum rank				
Quarter I		15	30	27
Quarter II	15		15	12
Quarter III	30	15		3
Quarter IV	27	12	3	
CD	Quarter I	Quarter II	Quarter III	Quarter IV
Quarter I		9.4095	9.4095	9.4095
Quarter II	9.4095		9.4095	9.4095
Quarter III	9.4095	9.4095		9.4095
Quarter IV	9.4095	9.4095	9.4095	
Statistic	Quarter I	Quarter II	Quarter III	Quarter IV
Quarter I		3.2244	6.4489	5.804
Quarter II	3.2244		3.2244	2.5795
Quarter III	6.4489	3.2244		0.6449
Quarter IV	5.804	2.5795	0.6449	
p-value	Quarter I	Quarter II	Quarter III	Quarter IV
Quarter I		0.0026	<0.0001	<0.0001
Quarter II	0.0026		0.0026	0.0138
Quarter III	<0.0001	0.0026		0.5228
Quarter IV	<0.0001	0.0138	0.5228	
Homogeneous	Quarter I(a)	Quarter II(b	Quarter III(c	Quarter IV(c
A	*			
B		*		
C			*	*

The graph shows the homogeneous groups determined by the Conover-Iman test.



Copyright ©2010-2026 U-smile Stats – All rights reserved

We can provide a detailed description of the data by selecting Descriptive statistics in the analysis window Σμ.

Summary				
Group	Quarter I	Quarter II	Quarter III	Quarter IV
Sample size	14	14	14	14
Median	1481	1607.5	1634	1628
Minimum	983	1086	1135	977
Maximum	3415	4556	5772	5432
Lower quartile	1272.25	1326.75	1398.25	1473.5
Upper quartile	1683.75	2026.25	2193.5	2068.25

If the data had been described with an ordinal scale with few categories, it would be worth presenting them also as counts and percentages – in this example, however, that would not be a good method of description.

FOR THOSE INTERESTED

Hypotheses: they relate to the equality of the sum of ranks for successive measurements or are simplified to medians:

$$\begin{aligned} \mathcal{H}_0: &\quad \theta_1 = \theta_2 = ... = \theta_k, \\ \mathcal{H}_1: &\quad \text{not all } \theta_j \text{ are equal } (j = 1, 2, ..., k), \end{aligned}$$

where $\theta_1, \theta_2, ...\theta_k$ – medians of the analysed feature at successive measurements in the population studied.

Formulas: two test statistics are determined: the Friedman statistic

$$T_1 = \frac{1}{C}\left(\frac{12}{nk(k+1)}\left(\sum_{j=1}^{k}\left(\sum_{i=1}^{n} R_{ij}\right)^2\right) - 3n(k+1)\right),$$

and its Iman-Davenport modification

$$T_2 = \frac{(n_j - 1)T_1}{n_j(k-1) - T_1},$$

where:
n – sample size,
R_{ij} – ranks assigned to successive measurements $(j = 1, 2, ...k)$, separately for each object studied $(i = 1, 2, ...n)$,
$C = 1 - \frac{\sum(t^3 - t)}{n(k^3 - k)}$ – correction for ties,
t – number of cases included in a tie. The correction C is used when ties occur (otherwise $C = 1$).

The T_1 statistic asymptotically (for large sample sizes) has the χ^2 distribution with $df = k - 1$ degrees of freedom. The T_2 statistic has the F Snedecor distribution with $df_1 = k - 1$ and $df_2 = (n_j - 1)(k - 1)$ degrees of freedom.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The p value, determined on the basis of the test statistic, is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

POST-HOC tests

An introduction to contrasts and POST-HOC tests is provided in section 17.1.2, which concerns the one-way analysis of variance.

The Dunn test

Used for simple comparisons (the counts in each measurement are always the same). The Dunn test (Dunn, 1964[53]) is corrected for multiple testing – most often with the **Bonferroni** or **Sidak** correction (other, newer corrections are described in more detail in the Multiple comparisons section).

(i) The critical difference is calculated using the formula:

$$CD = Z_{\alpha(corrected)}\sqrt{\frac{k(k+1)}{6n}},$$

where $Z_{\alpha(corrected)}$ is the critical value (statistic) of the normal distribution for the significance level α adjusted for the number of possible simple comparisons c according to the selected correction.

(ii) The test statistic is defined by:

$$Z = \frac{\sum_{j=1}^{k} c_j \overline{R}_j}{\sqrt{\frac{k(k+1)}{6n}}},$$

where $\overline{R}_j$ – mean rank of measurement j, for $(j = 1, 2, ...k)$. The test statistic asymptotically (for large sample sizes) has the normal distribution, and the p value is adjusted for the number of possible simple comparisons according to the selected correction.

The Conover-Inman test

The non-parametric equivalent of LSD Fisher[48], used for simple comparisons (the counts in each measurement are always the same).

(i) The critical difference is calculated using the formula:

$$CD = \sqrt{F_{\alpha,1,df_2}} \cdot \sqrt{\frac{2\left(n_j A - \sum_{j=1}^{t} R_j^k\right)}{(n_j - 1)(k-1)}},$$

where $A = \sum_{i=1}^{n_j}\sum_{j=1}^{k} R_{ij}^2$ – sum of squared ranks, and $F_{\alpha,1,df_2}$ is the critical value (statistic) of the F Snedecor distribution for a given significance level α and for degrees of freedom 1 and df_2, respectively.

(ii) The test statistic is defined by:

$$t = \frac{\sum_{j=1}^{k} c_j R_j}{\sqrt{\frac{2\left(n_j A - \sum_{j=1}^{t} R_j^k\right)}{(n_j - 1)(k-1)}}},$$

where R_j – sum of ranks of measurement j, for $(j = 1, 2, ...k)$. The test statistic has the t-Student distribution with df_2 degrees of freedom.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Page test for trend

WHEN TO USE THIS TEST:

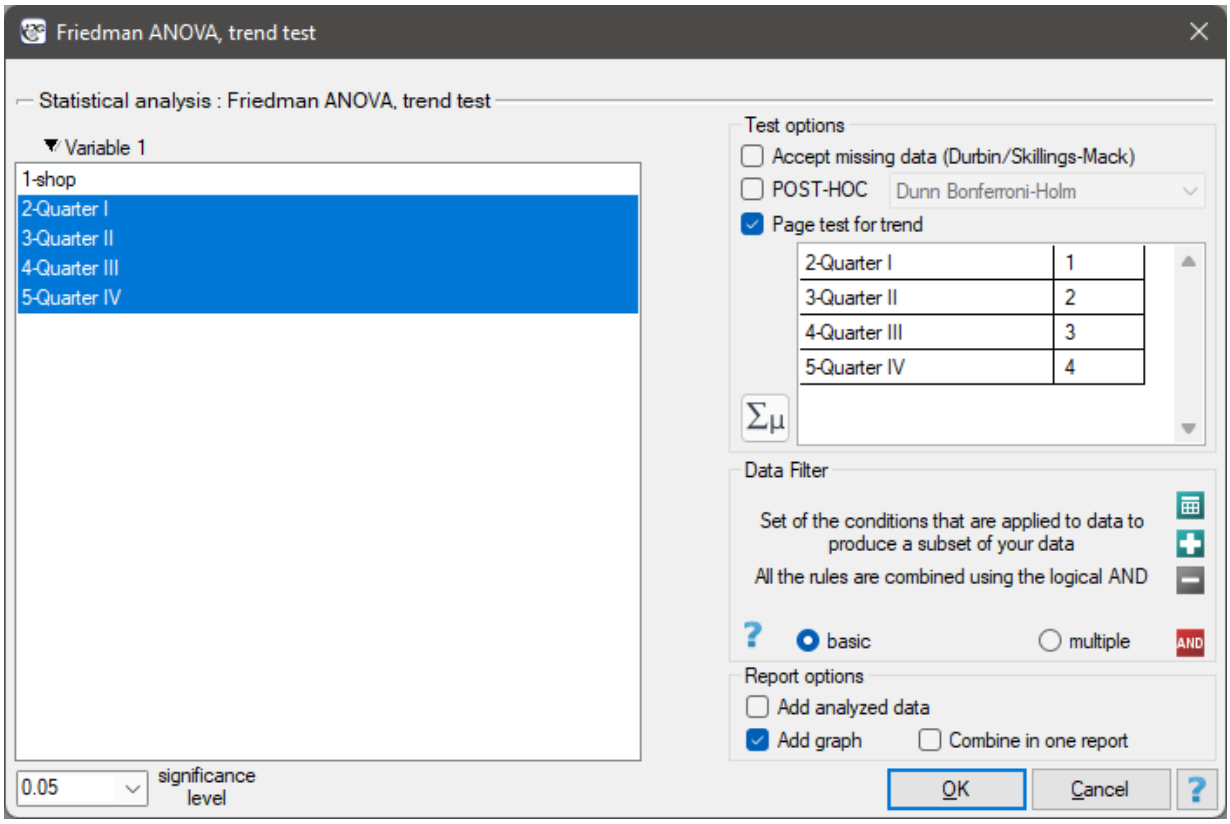
The Page test for ordered alternatives, described in 1963 by Page[134], can be performed in the same situation as the Friedman ANOVA, since it is based on the same assumptions. Unlike the Friedman ANOVA, however, it does not check for arbitrary differences between measurements, but whether the medians of successive measurements increase or decrease in the expected direction.

ASSUMPTIONS

- measurement on an ordinal scale or an interval scale,
- a dependent model,
- a known or assumed order of the measurements.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Friedman ANOVA (missing data allowed) (or via the Wizard).



Copyright ©2010-2026 U-smile Stats – All rights reserved

Step 1. Select the variables corresponding to successive measurements.

Step 2. Select the Page test for trend option.

Step 3. Indicate the expected order of the measurements by assigning them consecutive natural numbers.

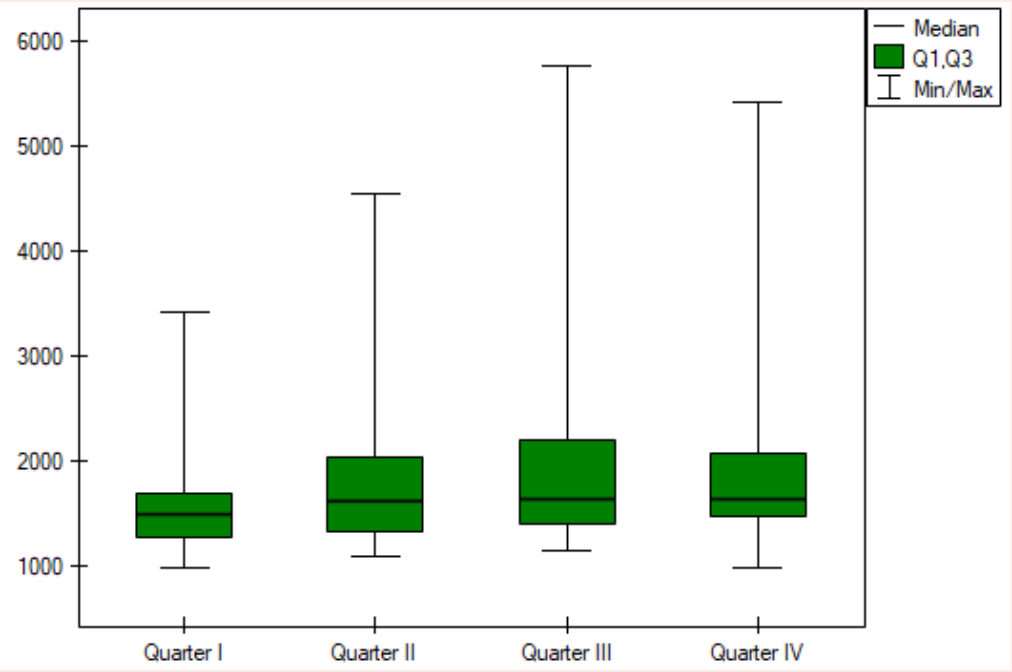
Step 4. Run the analysis.

Note! The phrase "with at least one strict inequality", used in the alternative hypothesis of this test, means that at least one median should be greater than the median of another group of measurements, in the expected order.

PERFORMING AND INTERPRETING THE ANALYSIS

Continuation of Example 17.6 (chocolate bar.pqs file): the expected result of the intensive advertising campaign conducted by the company is a steady increase in sales of the advertised bar. We will therefore check whether sales of the bar change in the expected direction across successive quarters.

Page test for trend	
The ordinal number for [Quarter I]	1
The ordinal number for [Quarter II]	2
The ordinal number for [Quarter III]	3
The ordinal number for [Quarter IV]	4
L statistic	398
Z statistic	4.4439
One sided p-value	<0.0001
Two sided p-value	<0.0001



Copyright ©2010-2026 U-smile Stats – All rights reserved

$p = 0.000004$ **(one-sided, statistically significant, $\alpha = 0.05$)**

The advertising campaign produced the expected trend of increasing product sales.

FOR THOSE INTERESTED

Hypotheses: relate to the equality of the sum of ranks for successive measurements or are simplified to medians:

$$\begin{aligned} \mathcal{H}_0: &\quad \theta_1 = \theta_2 = ... = \theta_k, \\ \mathcal{H}_1: &\quad \theta_1 \geq \theta_2 \geq ... \geq \theta_k, \text{ with at least one strict inequality} \end{aligned}$$

Formulas: the test statistic is defined by:

$$Z = \frac{L - \left[\frac{nk(k+1)^2}{4}\right]}{\sqrt{\frac{n(k^3-k)^2}{144(k-1)}}}$$

where:
$L = \sum_{j=1}^{k} R_j c_j$,
R_j – sum of ranks of the j-th measurement,
c_j – weight for the j-th measurement, indicating the natural order of this measurement among the others (the weights are consecutive natural numbers).

Comment: the formula for the test statistic Z does not include a correction for tied ranks, which makes it somewhat more conservative when tied ranks are present. Using a correction for tied ranks is, however, not recommended for this test. The Z statistic asymptotically (for large sample sizes) has the normal distribution.

Interpretation: with the expected direction of the trend known, the alternative hypothesis is one-sided and the one-sided p value is interpreted. Interpreting the two-sided p value means that the researcher does not know (does not assume) the direction of the possible trend.

The p value, determined on the basis of the test statistic, is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The Durbin's ANOVA (missing data)

WHEN TO USE THIS TEST:

Durbin's analysis of variance of repeated measurements for ranks (Durbin, 1951[54]) is used in the same situation as the Friedman ANOVA – when measurements of the analysed variable are taken several times on the same objects. The Durbin test has an advantage, though: it can also be calculated for an incomplete data set, provided the missing values form a so-called balanced and incomplete block (each object has the same number of measurements, and each pair of measurements occurs together equally often).

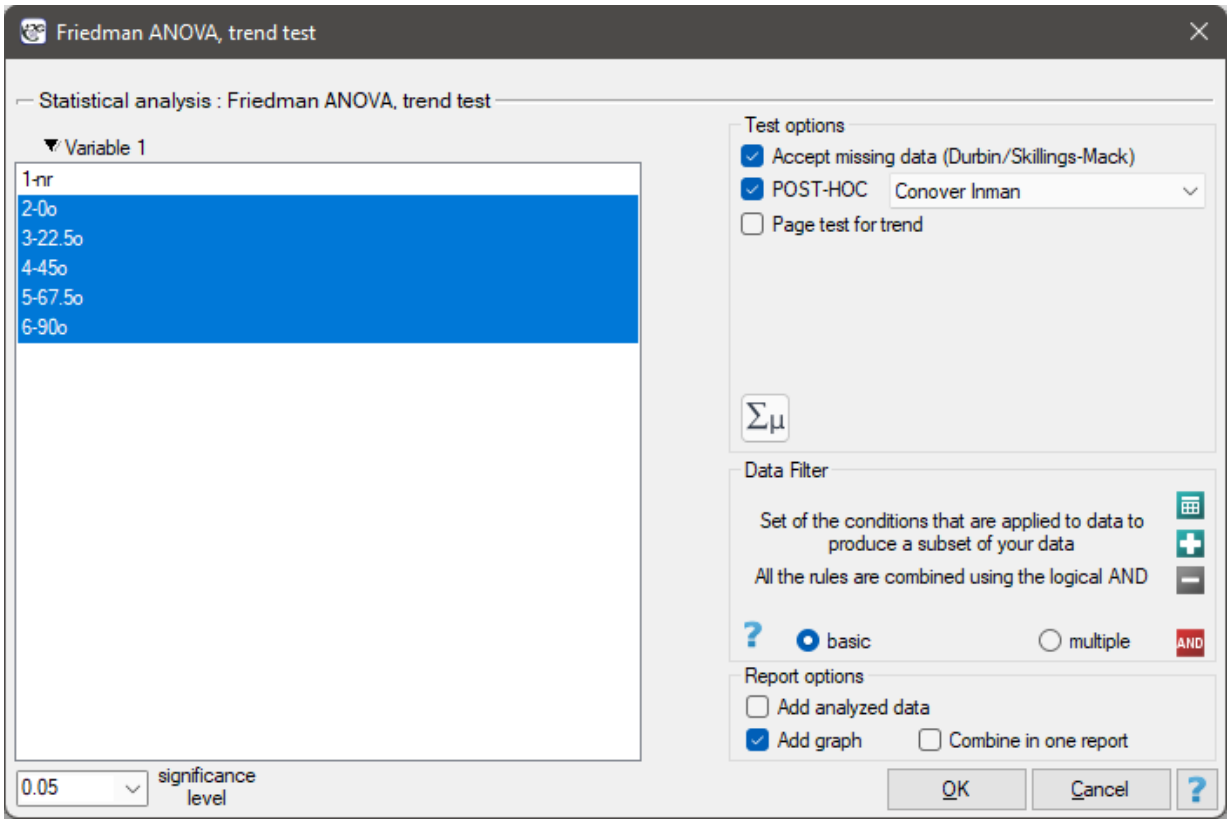
Copyright ©2010-2026 U-smile Stats – All rights reserved

ASSUMPTIONS

- measurement on an ordinal scale or an interval scale,
- a dependent model,
- if present, missing data form a balanced and incomplete block: each object has the same number k of measurements, each measurement concerns the same number r of objects, and the number of objects for which the same pair of measurements was taken simultaneously is constant.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Friedman ANOVA (missing data allowed).



Step 1. Select the variables corresponding to successive measurements.

Step 2. Select the Accept missing data option if the data contain gaps.

Step 3. If the test shows a significant difference, select the Conover-Inman POST-HOC test.

Step 4. Run the analysis.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! For records with missing data to be taken into account, the Accept missing data option must be selected. Empty cells and cells with non-numeric values are treated as missing data. Only records containing more than one numeric value are included in the analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

Example 17.7. (mirror.pqs file)
An experiment was conducted among 20 patients in a psychiatric hospital (Ogilvie, 1965[131]), involving drawing straight lines according to a presented pattern – 5 lines drawn at different angles (0°, 22.5^o, 45°, 67.5^o, 90°) relative to the indicated center. The patients' task was to reproduce the lines with their hand covered, and the outcome was the time it took to draw each line. Because of fatigue and limited patient cooperation, the project was planned in a balanced and incomplete block: each of the 20 patients drew a line at two of the five possible angles, so that each angle was drawn eight times in total. We want to check whether the time spent drawing each line is entirely random, or whether some lines took more or less time to draw. The time (in seconds) each patient took to draw the line at a given angle was recorded in the table.

patient number	0°	22.5^o	45°	67.5^o	90°
1	7	15			
2	20		72		
3	8			26	
4	33				36
5	7	16			
6		68	67		
7		33		64	
8		34			12
9	10		96		
10		29	59		
11			17	9	
12			100		15
13	16			32	
14		19		32	
15			36	39	
16				44	54
17	16				38
18		17			12
19			37		11
20				56	6

Copyright ©2010-2026 U-smile Stats – All rights reserved

Friedman ANOVA, trend test	
Analysed variables	0o
	22.5o
	45o
	67.5o
	90o
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Accept missing data (Durbin/Skillings-Mack)	Yes
C	90
A	100
T1 statistic Durbin (F)	10.4
Degrees of freedom	4
p-value	0.0342
T2 statistic Durbin (F)	4.3333
Degrees of freedom	4 / 16
p-value	0.0145
Skillings-Mack statistic	10.4
Degrees of freedom	4
p-value	0.0342

$p = 0.014541$ for the T_2 statistic (or $p = 0.034203$ for T_1), statistically significant, $\alpha = 0.05$

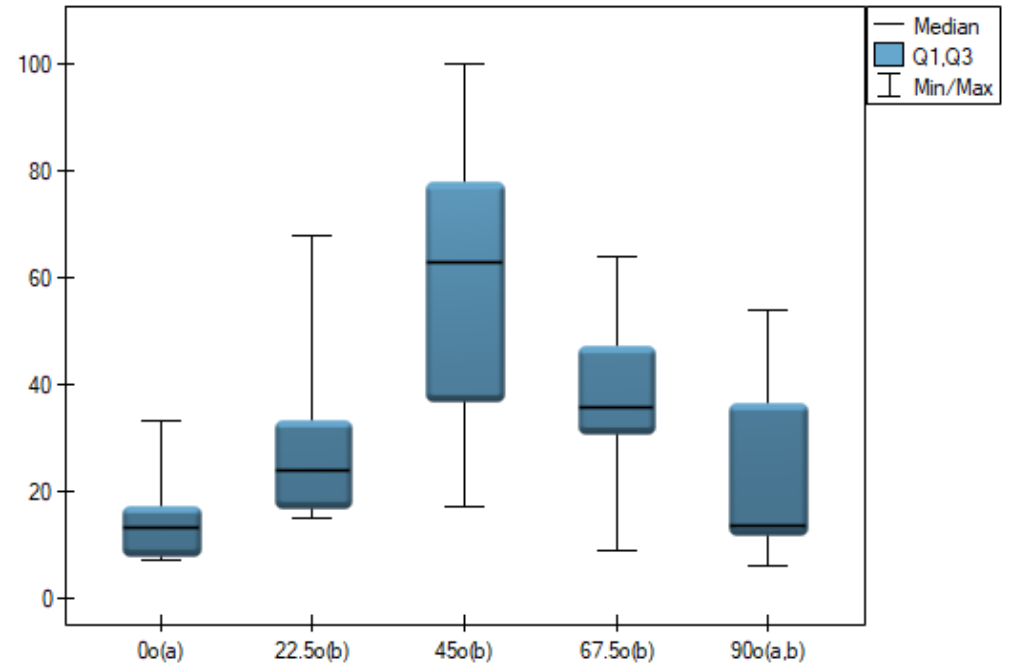
The time spent drawing each line is not the same.

The POST-HOC analysis performed indicates a difference in the time spent drawing the line at angle 0° – it is drawn faster than the lines at 22.5^o, 45° and 67.5^o.

Copyright ©2010-2026 U-smile Stats – All rights reserved

POST-HOC (Conover-Iman)					
	0o	22.5o	45o	67.5o	90o
The sum rank					
0o		5	6	6	3
22.5o	5		1	1	2
45o	6	1		0	3
67.5o	6	1	0		3
90o	3	2	3	3	
CD	0o	22.5o	45o	67.5o	90o
0o		4.6445	4.6445	4.6445	4.6445
22.5o	4.6445		4.6445	4.6445	4.6445
45o	4.6445	4.6445		4.6445	4.6445
67.5o	4.6445	4.6445	4.6445		4.6445
90o	4.6445	4.6445	4.6445	4.6445	
Statistic	0o	22.5o	45o	67.5o	90o
0o		2.2822	2.7386	2.7386	1.3693
22.5o	2.2822		0.4564	0.4564	0.9129
45o	2.7386	0.4564		0	1.3693
67.5o	2.7386	0.4564	0		1.3693
90o	1.3693	0.9129	1.3693	1.3693	
p-value	0o	22.5o	45o	67.5o	90o
0o		0.0365	0.0146	0.0146	0.1898
22.5o	0.0365		0.6542	0.6542	0.3749
45o	0.0146	0.6542		1	0.1898
67.5o	0.0146	0.6542	1		0.1898
90o	0.1898	0.3749	0.1898	0.1898	
Homogeneous	0o(a)	22.5o(b)	45o(b)	67.5o(b)	90o(a,b)
A	*				*
B		*	*	*	*

The graph shows the homogeneous groups indicated by the POST-HOC test.



Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Hypotheses: they relate to the equality of the sum of ranks for successive measurements (R_j) or are simplified to medians (θ_j):

$$\begin{aligned} \mathcal{H}_0: &\quad \theta_1 = \theta_2 = ... = \theta_t, \\ \mathcal{H}_1: &\quad \text{not all } \theta_j \text{ are equal } (j = 1, 2, ..., t), \end{aligned}$$

where t – total number of measurements considered, b – total number of objects studied.

Formulas: two test statistics are determined:

$$T_1 = \frac{(t-1)\left[\sum_{j=1}^{t} R_j^2 - tC\right]}{A - C},$$

$$T_2 = \frac{T_1/(t-1)}{(b(k-1) - T_1)/(bk - b - t + 1)},$$

where:
R_j – sum of ranks for successive measurements $(j = 1, 2, ...t)$,
R_{ij} – ranks assigned to successive measurements, separately for each object studied $(i = 1, 2, ...b)$,
$A = \sum_{i=1}^{b}\sum_{j=1}^{t} R_{ij}^2$ – sum of squared ranks,
$C = \frac{bk(k+1)^2}{4}$ – correction coefficient. The formula for T_1 and T_2 includes a correction for tied ranks.
For complete data, the T_1 statistic is identical to the Friedman test and asymptotically (for large sample sizes) has the χ^2 distribution with $df = t - 1$ degrees of freedom. The T_2 statistic is the equivalent of the Iman-Davenport correction of the Friedman ANOVA, and follows the F Snedecor distribution with $df_1 = t - 1$ and $df_2 = bk - b - t + 1$ degrees of freedom – it is now considered more precise than the T_1 statistic and is recommended for use[48].

The p value, determined on the basis of the test statistic, is compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

POST-HOC tests

An introduction to contrasts and POST-HOC tests is provided in section 17.1.2, which concerns the one-way analysis of variance.

The Conover-Inman test
Used for simple comparisons (the counts in each measurement are always the same).

(i) The critical difference is calculated using the formula:

$$CD = t_{1-\alpha/2, bk-b-t+1}\sqrt{\frac{(A-C)2r}{bk-b-t+1}\left(1 - \frac{T_1}{b(k-1)}\right)},$$

where $t_{1-\alpha/2,bk-b-t+1}$ is the critical value (statistic) of the t-Student distribution for the significance level α and $df = bk - b - t + 1$ degrees of freedom.

Copyright ©2010-2026 U-smile Stats – All rights reserved

(ii) The test statistic is defined by:

$$t = \frac{\sum_{j=1}^{k} c_j R_j}{\sqrt{\frac{(A-C)2r}{bk-b-t+1}\left(1 - \frac{T_1}{b(k-1)}\right)}}.$$

The test statistic has the t-Student distribution with $df = bk - b - t + 1$ degrees of freedom.

The Skillings-Mack ANOVA (missing data)

WHEN TO USE THIS TEST:

The analysis of variance of repeated measurements for Skillings-Mack ranks (Skillings and Mack, 1981[158]) is used in the same situation as the Friedman ANOVA, but when the data contain gaps that do not follow any particular pattern (each object only needs at least two observations). When there are no gaps and no tied ranks, the test is identical to the Friedman ANOVA; when gaps occur in a balanced arrangement, it gives the same results as the Durbin ANOVA.

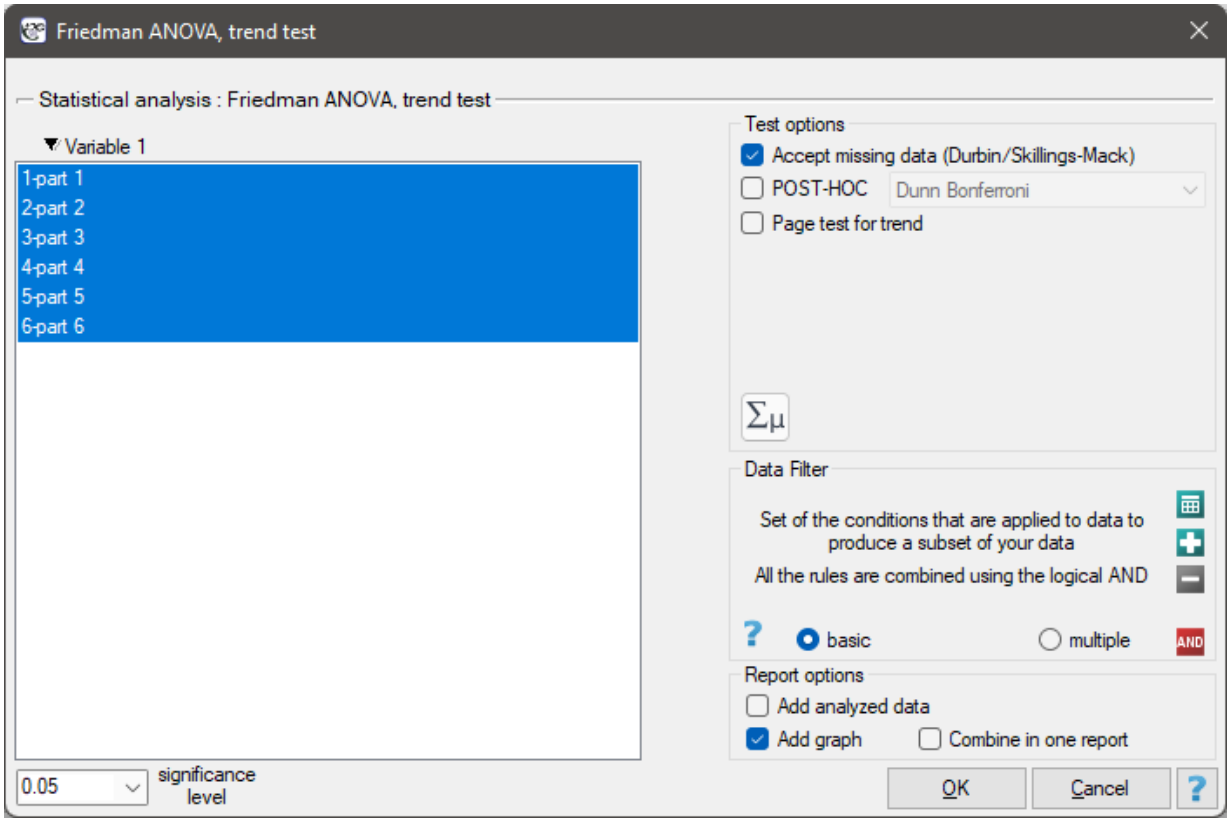
ASSUMPTIONS

- measurement on an ordinal scale or an interval scale,
- a dependent model,
- each object has at least two observations.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests (ordered categories) → Friedman ANOVA (missing data allowed).

Copyright ©2010-2026 U-smile Stats – All rights reserved



Step 1. Select the variables corresponding to successive measurements.

Step 2. Select the Accept missing data option.

Step 3. Run the analysis.

Note! For records with missing data to be taken into account, the Accept missing data option must be selected. Empty cells and cells with non-numeric values are treated as missing data. Only records containing more than one numeric value are included in the analysis. There is no POST-HOC analysis available for this test.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 17.8. (polling.pqs file)
A university teacher, wanting to improve the way he conducted his classes, decided to verify his teaching skills. In several randomly selected student groups, during the last class, he asked students to fill in a short, anonymous questionnaire consisting of 6 questions about how six specified parts of the course material were illustrated (five-point scale, where 1 – a completely incomprehensible way of presenting the material, 5 – a very clear and interesting way of illustrating the material). The data obtained turned out to be incomplete: in the 30-person group completing the survey, only 15 people provided a complete set of answers. An analysis that does not take the gaps into account (Friedman ANOVA) would have limited power due to such a drastic reduction in group size. The gaps were also not planned and do not form a balanced block, so the Durbin ANOVA cannot be used either. We will check whether the ratings of the individual parts of the material differ from one another.

Copyright ©2010-2026 U-smile Stats – All rights reserved

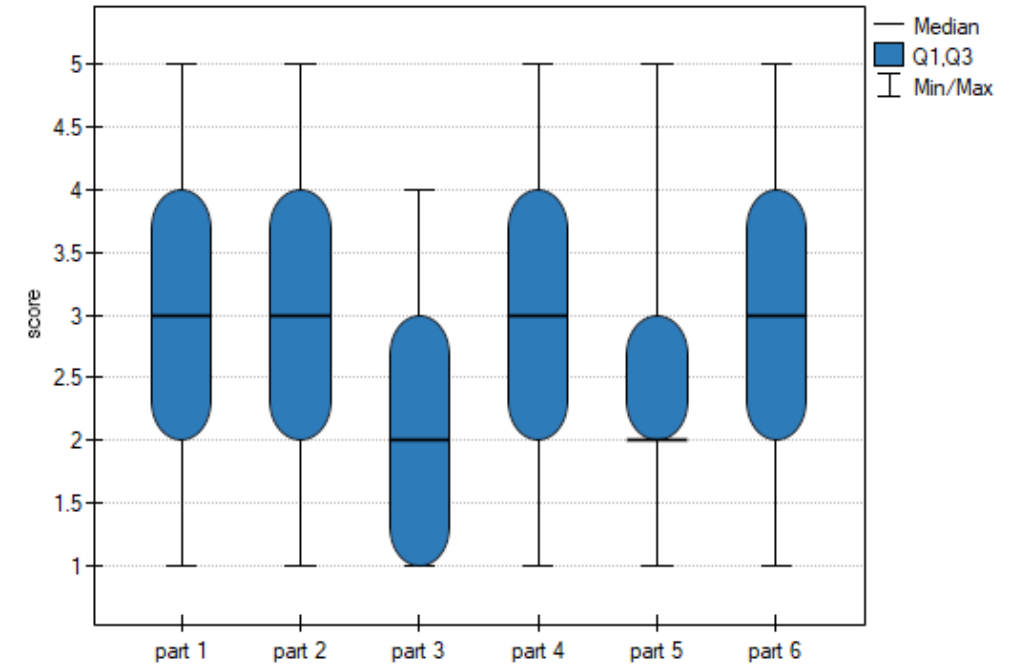
Friedman ANOVA, trend test	
Analysed variables	part 1
	part 2
	part 3
	part 4
	part 5
	part 6
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Accept missing data (Durbin/Skillings-Mack)	Yes
C	2205
A	2086.5
T1 statistic Friedman	NA
Degrees of freedom	5
p-value	NA
T2 statistic Iman-Davenport	NA
Degrees of freedom	5 / 145
p-value	NA
Skillings-Mack statistic	16.037
Degrees of freedom	5
p-value	0.0067

$p = 0.006739$ **(statistically significant,** $\alpha = 0.05$**)**

The ratings of the individual parts of the material differ statistically significantly.

The obtained p value should be treated with caution due to possible tied ranks – in this case, however, it is well below the assumed significance level. The differences in responses can be observed in the graph; there is no POST-HOC analysis available for this test.

Copyright ©2010-2026 U-smile Stats – All rights reserved



FOR THOSE INTERESTED

Hypotheses: they relate to the equality of the sum of ranks for successive measurements (R_j) or are simplified to medians (θ_j):

$$\begin{aligned} \mathcal{H}_0: &\quad \theta_1 = \theta_2 = ... = \theta_k, \\ \mathcal{H}_1: &\quad \text{not all } \theta_j \text{ are equal } (j = 1, 2, ..., k), \end{aligned}$$

Formulas: the test statistic is defined by:

$$\chi^2 = A\Sigma_0^{-1}A^T,$$

where:
$A = (A_1, A_2, ..., A_{k-1})$,
$A_j = \sum_{i=1}^{n} \sqrt{\frac{12}{s_i+1}} \left(R_{ij} - \frac{s_i+1}{2}\right)$,
s_i – number of observations for object i,
R_{ij} – ranks assigned to successive measurements ($j = 1, 2, ...k$), separately for each object studied ($i = 1, 2, ...n$), with ranks for missing data equal to the mean rank for that object,
Σ_0 – matrix determining the covariances of A under $\mathcal{H}_0$[158].

When each pair of measurements occurs simultaneously for at least one observation, this statistic asymptotically (for large sample sizes) has the χ^2 distribution with $k-1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Chi-square test for multidimensional contingency tables

WHEN TO USE THIS TEST:

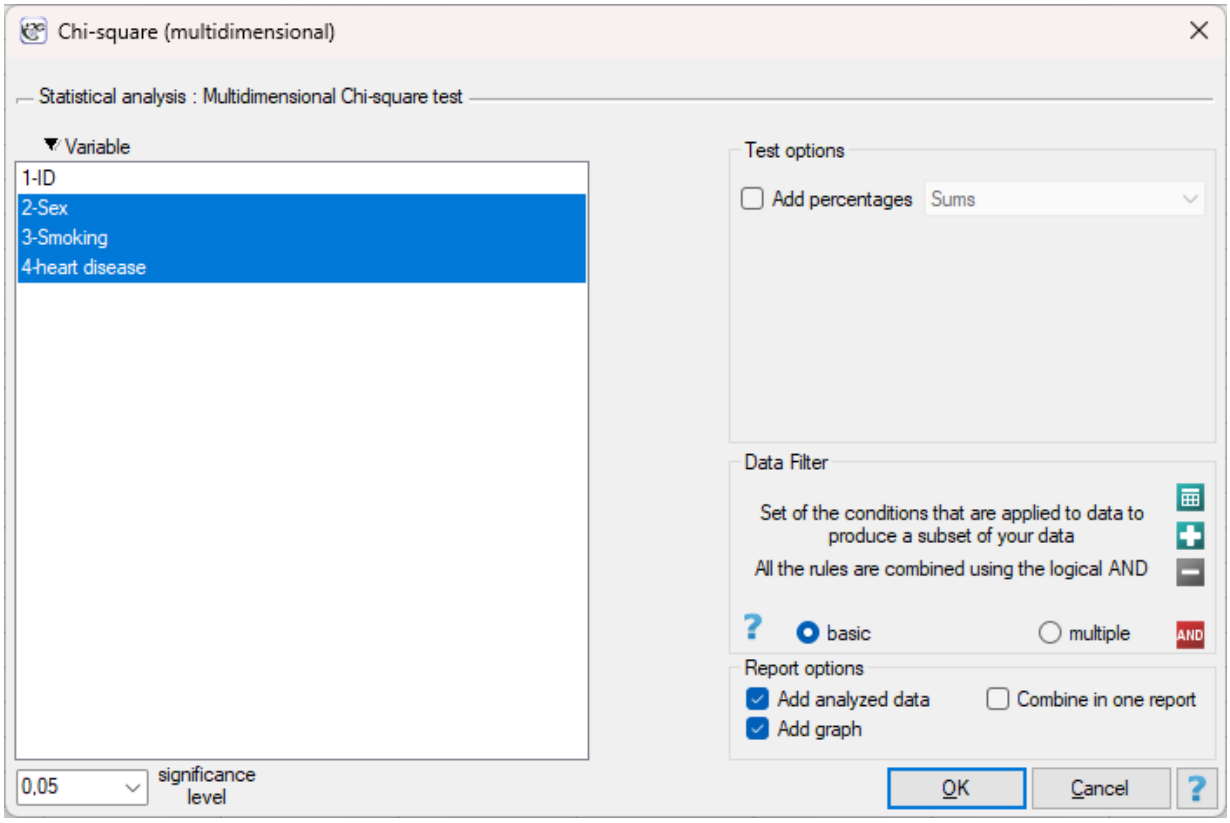
The χ^2 test for multidimensional contingency tables is used when we want to check the dependence between more than two features at once – it is an extension of the χ^2 test for $(R \times C)$ tables to three or more features.

ASSUMPTIONS

- measurement on a nominal scale – any ordering of categories is not taken into account,
- an independent model,
- large expected frequencies according to the Cochran interpretation (1952)[42].

HOW TO SET UP THE ANALYSIS

Statistics → NonParametric tests (unordered categories) → Chi-square (multidimensional) (or via the Wizard).



Step 1. Indicate all the analysed features (three or more).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Step 2. Run the analysis.

Note! This test can only be calculated on the basis of raw data.

PERFORMANCE AND INTERPRETATION

EXAMPLE 17.9. (heart_disease.pqs file)
A sample of 200 people was examined for three features: sex (X_1: f, m), cigarette smoking (X_2: yes, no), and the presence of heart disease (X_3: yes, no). The aim of the study was to check whether there is a dependence between these three features in the analysed population.

Data :

	Sex	Smoking	heart diseas	obs.	exp.
1	F	yes	yes	11	9,12135
2	F	yes	no	23	28,10865
3	F	no	yes	9	15,86865
4	F	no	no	59	48,90135
5	M	yes	yes	25	8,76365
6	M	yes	no	14	27,00635
7	M	no	yes	4	15,24635
8	M	no	no	55	46,98365

Multidimensional Chi-square test	
Analysed variables	Sex
	Smoking
	heart disease
Number of unspecified	0
Number of missing data	0
Significance level	0,05
Size	200
Chi-square statistic	52,382342
Degrees of freedom	4
p-value	<0,000001

$\chi^2 = 52.382342$, $df = 4$, $p < 0.000001$ **(statistically significant,** $\alpha = 0.05$**)**

There is a statistically significant dependence between sex, cigarette smoking, and the presence of heart disease in the analysed population.

The expected frequencies shown in the data window are large enough (Cochran's condition is satisfied), so the test result is reliable. Comparing the observed and expected frequencies shows where the

Copyright ©2010-2026 U-smile Stats – All rights reserved

dependence is strongest: among male smokers, heart disease occurs much more often than would be expected under no dependence (25 observed cases versus 8.76 expected) – almost three times more often than expected. The opposite pattern is seen among male non-smokers: heart disease occurs much less often than expected (4 observed cases versus 15.25 expected). Among women, the deviations between observed and expected frequencies are clearly smaller. The dependence therefore mainly concerns men – for them, cigarette smoking is associated with a markedly increased risk of heart disease.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad O_{ij\ldots} = E_{ij\ldots} \text{ for all categories,} \\ \mathcal{H}_1: &\quad O_{ij\ldots} \neq E_{ij\ldots} \text{ for at least one category,} \end{aligned}$$

where $O_{ij\ldots}$ and $E_{ij\ldots}$ are observed frequencies in a contingency table and the corresponding expected frequencies.

Formulas:

The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum\cdots\sum\frac{(O_{ij\ldots}-E_{ij\ldots})^2}{E_{ij\ldots}}.$$

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with a number of degrees of freedom calculated, for a 3-dimensional table, using the formula:

$$df = (r-l)(c-1)(l-1)+(r-l)(c-1)+(r-1)(l-1)+(c-1)(l-1).$$

The Q-Cochran ANOVA

WHEN TO USE THIS TEST:

The Q-Cochran analysis of variance (Cochran, 1950[41]) extends the McNemar test to $k \geq 2$ dependent groups. It lets you check whether the results of several measurements of a two-valued feature (e.g. correct/wrong answer) are symmetric, i.e. whether the proportion of one of the values is the same across all measurements.

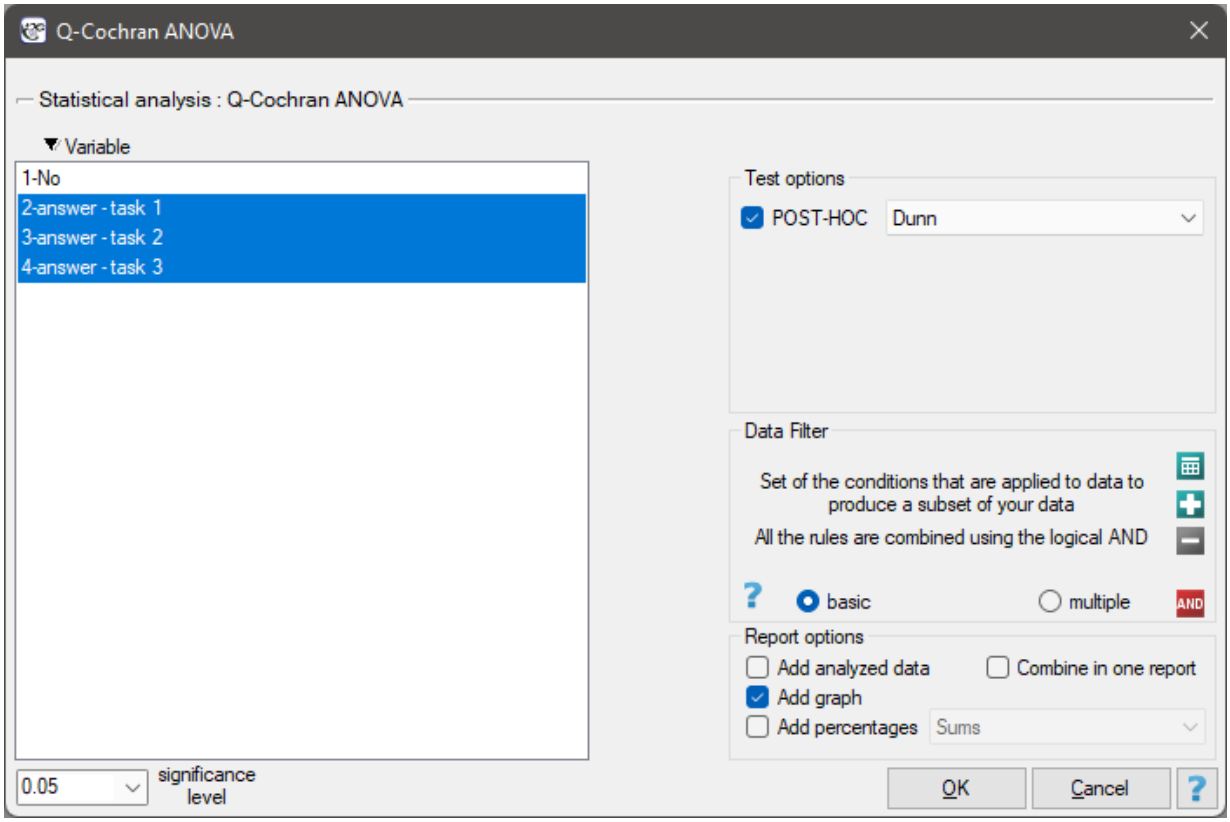
ASSUMPTIONS

- measurement on a nominal scale (dichotomous variables, i.e. with 2 categories, coded as 0 and 1),
- a dependent model.

Copyright ©2010-2026 U-smile Stats – All rights reserved

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests (unordered categories) → Q-Cochran ANOVA (or via the Wizard).



Step 1. Select the variables corresponding to successive measurements (coded as 0/1).

Step 2. If the test shows a significant difference, select the Dunn POST-HOC test.

Step 3. Run the analysis.

Note! This test can only be calculated from raw data.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 17.10. (test.pqs file)
We want to compare the difficulty of 3 test questions. To do this, we select a sample of 20 people from the population the test is aimed at. Each person in the sample answers 3 questions from the test, after which we check the correctness of the answers (an answer can be correct or wrong). The results are shown in the table below.

Copyright ©2010-2026 U-smile Stats – All rights reserved

No.	Question 1 answer	Question 2 answer	Question 3 answer
1	correct	correct	wrong
2	wrong	correct	wrong
3	correct	correct	correct
4	wrong	correct	wrong
5	wrong	correct	wrong
6	wrong	correct	correct
7	wrong	wrong	wrong
8	wrong	correct	wrong
9	correct	correct	wrong
10	wrong	correct	wrong
11	wrong	wrong	wrong
12	wrong	wrong	correct
13	wrong	correct	wrong
14	wrong	wrong	correct
15	correct	wrong	wrong
16	wrong	wrong	wrong
17	wrong	correct	wrong
18	wrong	correct	wrong
19	wrong	wrong	wrong
20	correct	correct	wrong

Q-Cochran ANOVA	
Analysed variables	answer - task 1
	answer - task 2
	answer - task 3
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	20
Degrees of freedom	2
Statistic Q	9.7333
p-value	0.0077

$p = 0.007699$ **(statistically significant, $\alpha = 0.05$)**

The test questions have different levels of difficulty.

We resume the analysis with the button to perform a POST-HOC test, and in the test options window we select the POST-HOC Dunn.

Copyright ©2010-2026 U-smile Stats – All rights reserved

POST-HOC (Dunn)			
	answer - tas	answer - tas	answer - tas
Difference of			
answer - task 1		0.4	0.05
answer - task 2	0.4		0.45
answer - task 3	0.05	0.45	
CD	answer - tas	answer - tas	answer - tas
answer - task 1		0.3785	0.3785
answer - task 2	0.3785		0.3785
answer - task 3	0.3785	0.3785	
Statistic Z	answer - tas	answer - tas	answer - tas
answer - task 1		2.5298	0.3162
answer - task 2	2.5298		2.846
answer - task 3	0.3162	2.846	
p-value	answer - tas	answer - tas	answer - tas
answer - task 1		0.0342	1
answer - task 2	0.0342		0.0133
answer - task 3	1	0.0133	
Homogeneous	answer - tas	answer - tas	answer - tas
A	*		*
B		*	

The POST-HOC analysis performed shows that the differences concern question 2 and 1, and question 2 and 3. The difference is that question 2 is easier than questions 1 and 3 (the number of correct answers is higher there).

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\ "incompatible" observed frequencies are all equal, \\ \mathcal{H}_1: &\ "incompatible" observed frequencies are not all equal, \end{aligned}$$

where ”incompatible” observed frequencies are the observed frequencies calculated when the value of the analysed feature differs across successive measurements.

Formulas:

The test statistic is defined by:

$$Q = \frac{(k-1)\left(kC - T^2\right)}{kT - R}$$

where:

$T = \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}$,

$R = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} x_{ij}\right)^2$,

Copyright ©2010-2026 U-smile Stats – All rights reserved

$C = \sum_{j=1}^{k} \left(\sum_{i=1}^{n} x_{ij}\right)^2$,
x_{ij} – the value of the j-th measurement for the i-th object (i.e. 0 or 1).

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with $df = k - 1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

POST-HOC tests

An introduction to contrasts and POST-HOC tests is provided in section 17.1.2, which concerns the one-way analysis of variance.

The Dunn test
Used for simple comparisons (the counts in each measurement are always the same).

Example – **simple comparisons** (for the difference in proportion for one selected pair of measurements).
Hypotheses:

$\mathcal{H}_0$: the selected "incompatible" observed frequencies are equal,
$\mathcal{H}_1$: the selected "incompatible" observed frequencies are different,

(i) The critical difference is calculated using the formula:

$$CD = Z_{\frac{\alpha}{c}} \sqrt{2\frac{kT - R}{n^2 k(k-1)}},$$

where $Z_{\frac{\alpha}{c}}$ is the critical value (statistic) of the normal distribution for the significance level α adjusted for the number of possible simple comparisons c.

(ii) The test statistic is defined by:

$$Z = \frac{\sum_{j=1}^{k} c_j p_j}{\sqrt{2\frac{kT-R}{n^2 k(k-1)}}},$$

where p_j – proportion of the j-th measurement $(j = 1, 2, ...k)$. This statistic asymptotically (for large sample sizes) has the normal distribution, and the p value is adjusted for the number of possible simple comparisons c.

Copyright ©2010-2026 U-smile Stats – All rights reserved

18 MULTIPLE COMPARISONS

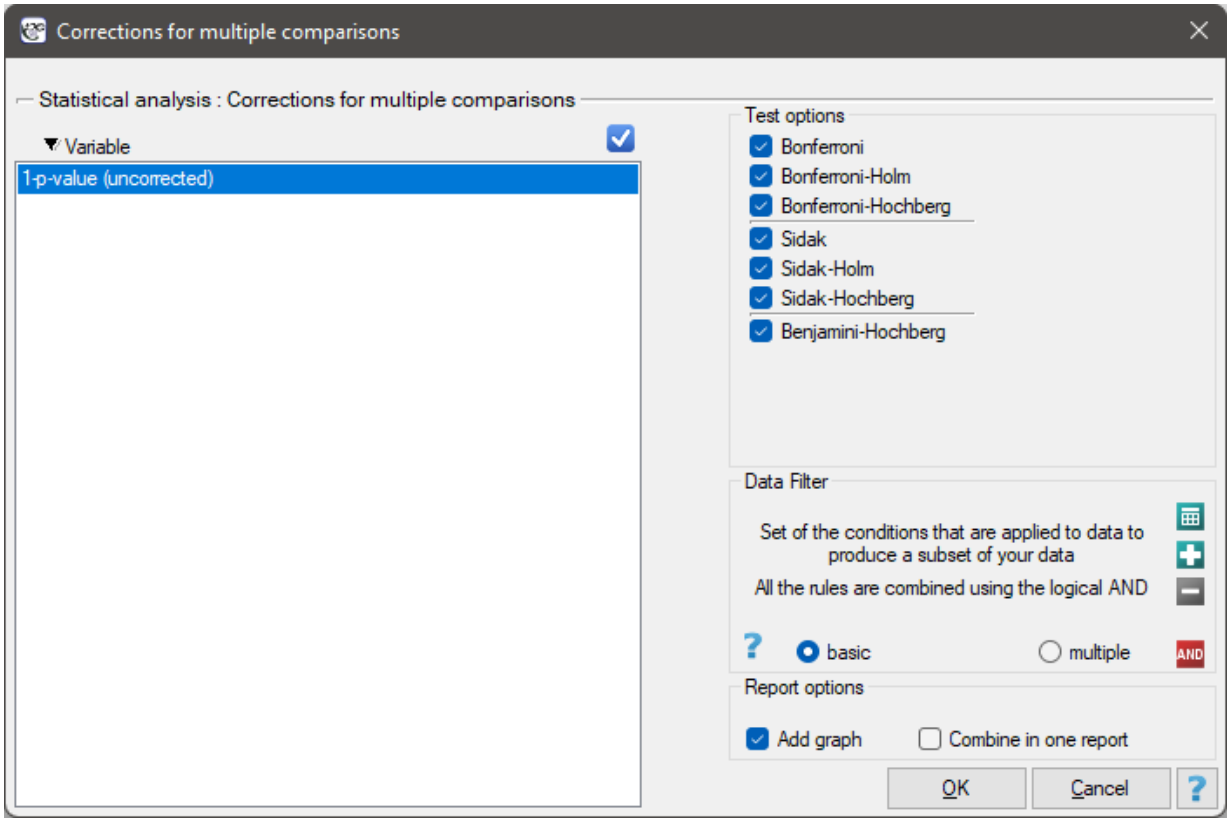
WHEN TO USE:

Simultaneously testing many hypotheses (the so-called **family of hypotheses**) carries the risk of an increase in the error α, which is the main problem in the field of multiple comparisons. When the error α increases, it means that we reject the null hypothesis too often when it is in fact true – that is, we too often indicate that differences exist when in reality they do not. To guard against an increase in α, one strategy is to adjust (reduce) the level of α, or to adjust (increase) the p values of the tests accordingly.

Note! A family of hypotheses can be defined in many ways. Most often these are hypotheses arising within a POST-HOC procedure, i.e. performing many tests as part of a simultaneous comparison of several groups studied. Such a family can also be formed by tests performed as part of a Hotelling-type analysis. Families of hypotheses also occur in many geographical analyses. Wherever many smaller hypotheses are analysed within an overall hypothesis, a multiple-comparisons correction may apply.

HOW TO SET UP THE ANALYSIS

To perform a multiple-comparisons correction, enter successive p values into a single column of the data sheet. The settings window for the multiple comparisons options is opened via the Statistics→Correction for multiple comparisons menu.



PERFORMING AND INTERPRETING THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

Continuation of Example 17.4 (jobSatisfaction.pqs file): this study examined whether the level of job satisfaction was the same across four education categories. Here, the family of hypotheses consisted of the hypotheses resulting from comparing all groups pairwise. To compare all 4 groups, 6 pairs of comparisons were formed – in each case the null hypothesis stated that there was no difference in satisfaction for the pair analysed. To make use of several of the proposed multiple-comparisons corrections, the analysis was performed using the uncorrected Dunn POST-HOC test.

POST-HOC (Dunn uncorrected)*

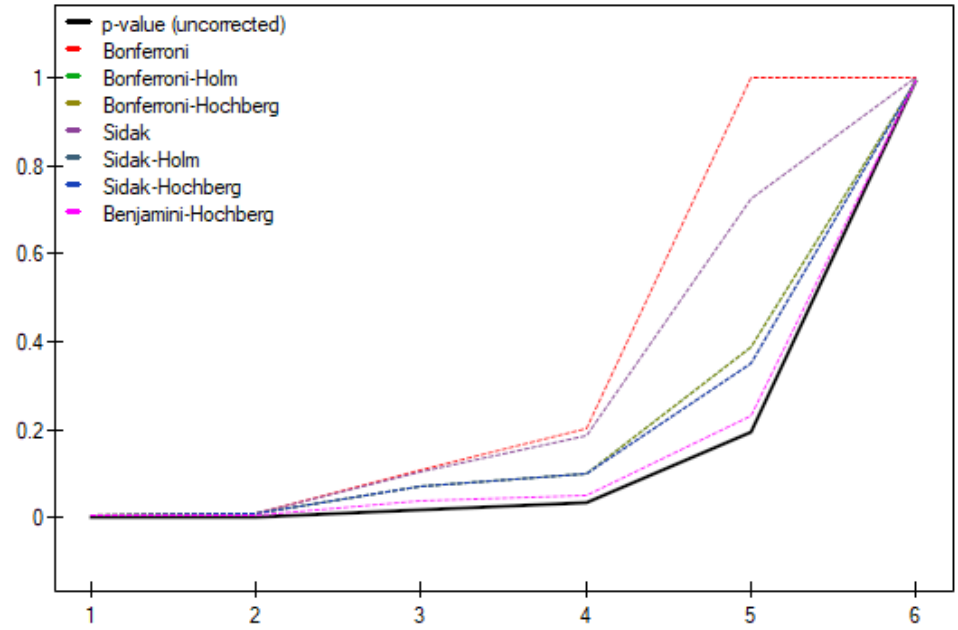
p-value	primary	vocational	secondary	university
primary		0.1937	0.0006	0.0018
vocational	0.1937		0.0181	0.0336
secondary	0.0006	0.0181		0.9919
university	0.0018	0.0336	0.9919	

The p values obtained in this way were entered as the data for the multiple-comparisons correction, yielding the following results:

Corrections for multiple comparisons

Number of missing data	0
Frequency	6
Significance level	0.05

Variable: p-value (uncorrected)	Bonferroni	Bonferroni - Holm	Bonferroni - Hochberg	Sidak	Sidak - Holm	Sidak - Hochberg	Benjamini - Hochberg
0.1937	1	0.3874	0.3874	0.7252	0.3499	0.3499	0.2324
0.0006	0.0036	0.0036	0.0036	0.0036	0.0036	0.0036	0.0036
0.0018	0.0108	0.009	0.009	0.0108	0.009	0.009	0.0054
0.0181	0.1086	0.0724	0.0724	0.1038	0.0705	0.0705	0.0362
0.0336	0.2016	0.1008	0.1008	0.1854	0.0975	0.0975	0.0504
0.9919	1	0.9919	0.9919	1	0.9919	0.9919	0.9919



statistically significant differences for 2 (or 3, with the Benjamini-Hochberg correction) of the 6 pairs of comparisons, $\alpha = 0.05$

The differences concern job satisfaction for primary versus secondary education, and primary versus higher education.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Only when the Benjamini-Hochberg correction is applied can differences be located in three pairs of comparisons.

Other examples of applying the multiple-comparisons correction: Example 19.1 and Example 19.2 (Hotelling-type analysis).

FOR THOSE INTERESTED

If we denote the number of hypotheses tested by c, the corrections for multiple comparisons can be described as follows:

The Bonferroni correction
The best known and, at the same time, the most conservative of the corrections[2]. It consists in multiplying each test probability by the total number of tests performed (or dividing the significance level by that number).

Formula – p-value adjustment:

$$p_{(Bonferroni,i)} = p_i \cdot c$$

Formula – significance-level α adjustment:

$$\alpha_{(Bonferroni,i)} = \frac{\alpha_i}{c}$$

The Sidak correction
The Sidak correction (1967)[156] has greater power than the Bonferroni correction, which is why it is increasingly used.

Formula – p-value adjustment:

$$p_{(Sidak,i)} = 1 - (1 - p_i)^c$$

Formula – significance-level α adjustment:

$$\alpha_{(Sidak,i)} = 1 - (1 - \alpha_i)^{1/c}$$

The Bonferroni-Holm correction
Consists in applying Holm's (1979)[85] multi-step procedure to the Bonferroni correction. The procedure begins by sorting the successive p_i values $(p_1, p_2, ...p_c)$ in ascending order. The Bonferroni correction is then applied to each successive p_i value (with a corresponding reduction in the number of hypotheses remaining to be tested). As a result, all hypotheses tested after the first statistically non-significant p_i value are also non-significant.

Formula – p-value adjustment:

$$p_{(Bonferroni,i)} = p_i \cdot (c - i + 1)$$

The Sidak-Holm correction
Consists in applying Holm's multi-step procedure to the Sidak correction, analogously to the Bonferroni-Holm correction above.

Formula – p-value adjustment:

$$p_{(Sidak,i)} = 1 - (1 - p_i)^{c-i+1}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Bonferroni-Hochberg correction
Consists in applying Hochberg's (1988)[84] multi-step procedure to the Bonferroni correction. The procedure begins by sorting the successive p_i values $(p_c, p_{c-1}, ..., p_1)$ in descending order. The Bonferroni correction is then applied to each successive p_i value (with a corresponding reduction in the number of hypotheses remaining to be tested). As a result, all hypotheses tested after the first statistically non-significant p_i value are also non-significant.

Formula – p-value adjustment:

$$p_{(Bonferroni,i)} = p_i \cdot (c - i + 1)$$

The Sidak-Hochberg correction
Consists in applying Hochberg's multi-step procedure to the Sidak correction, analogously to the Bonferroni-Hochberg correction above.

Formula – p-value adjustment:

$$p_{(Sidak,i)} = 1 - (1 - p_i)^{c-i+1}$$

The Benjamini-Hochberg correction
Consists in applying Hochberg's multi-step procedure to Benjamini's (1995)[19] correction, which is a modified version of the Bonferroni correction. The procedure begins by sorting the successive p_i values $(p_c, p_{c-1}, ..., p_1)$ in descending order. Benjamini's correction is then applied to each successive p_i value (with a corresponding reduction in the number of hypotheses remaining to be tested). As a result, all hypotheses tested after the first statistically non-significant p_i value are also non-significant. Of the corrections proposed, it has the greatest power – it does not directly control the error α, but minimises the expected proportion of false differences among the differences detected.

Formula – p-value adjustment:

$$p_{(BH,i)} = p_i \frac{c}{i}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

19 HOTELLING T-SQUARE

Hotelling's analysis, proposed by Hotelling (1931)[86] (1947)[87] and Mahalanobis (1930[112], 1936[113]), extends the t-Student tests to the domain of multiple variables. As a result, we can subject many different features of the objects studied to simultaneous analysis.

Hotelling's T-square test for a single sample

WHEN TO USE THIS TEST:

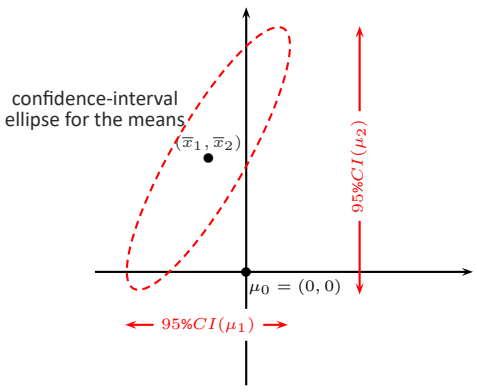
This test lets you check whether k variables in the population studied, $X_1, X_2, ..., X_k$, are characterised by the means specified by the researcher – it is a multivariate extension of the single-sample t-test.

ASSUMPTIONS

- measurement on an interval scale,
- a multivariate normal distribution, or normality of distribution of each variable studied.

Note! The principle behind Hotelling's test is identical to constructing a multivariate confidence-interval ellipse around the centre determined by the means. Because of this, a univariate analysis (which does not take the interrelationships between the variables into account) will often fail to give an identical result.

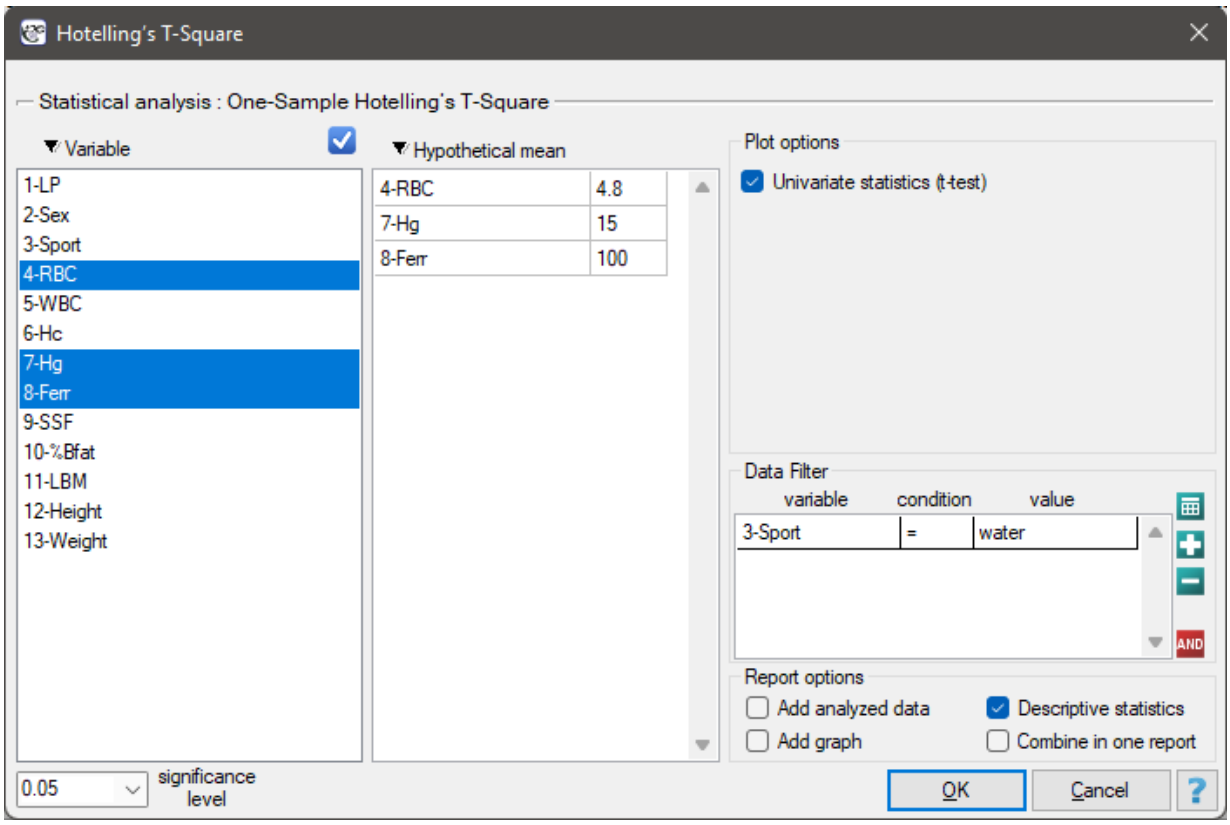
Example – interpreting the Hotelling test ellipse for two variables
The given point described by the mean values ($\mu_0 = (0, 0)$) lies outside the ellipse, which means that Hotelling's test rejects $\mathcal{H}_0$, but with the univariate approach (confidence intervals for each variable separately) we are unable to reject $\mathcal{H}_0$ and thus cannot identify the variable the difference concerns.



Copyright ©2010-2026 U-smile Stats – All rights reserved

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → T-square Hotelling.



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 19.1. (sport.pqs file)
A group of athletes was studied to obtain, among other things, information on health parameters such as:
RBC – red blood cell count,
Hg – haemoglobin [g/dl],
Ferr – ferritin [µg/l].
We want to know how close to the values expected by the researchers are the average levels of RBC, haemoglobin, and ferritin for athletes practising so-called "water"sports. The expected means are:
RBC = 4.8,
Hg = 15 [g/dl],
Ferr = 100 [µg/l].

Since the data sheet contains information on the parameters studied for a larger group of athletes, in the analysis window we select the people practising water sports using a data filter.

Copyright ©2010-2026 U-smile Stats – All rights reserved

One-Sample Hotelling's T-Square	
Analysed variables	RBC
	Hg
	Ferr
Data Filter	Sport=water
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Frequency	59
T-Square	20.3618
F statistic	6.5532
Degrees of freedom 1	3
Degrees of freedom 2	56
p-value	0.0007

$p = 0.000708$ **(statistically significant,** $\alpha = 0.05$**)**

The mean levels of the studied parameters differ from the specified values.

We can look for the differences in the calculated simultaneous confidence intervals or in the intervals with the Bonferroni correction. For ferritin, the simultaneous 95% confidence interval lies below the specified value, indicating a lower population value of ferritin than that specified by the researchers. The intervals for RBC and haemoglobin contain the specified values, which indicates no statistically significant differences.

Confidence intervals [Simultaneous]			
	mean	-95% CI for	+95% CI for
RBC	4.6915	4.5389	4.8442
Hg	14.6305	14.2322	15.0288
Ferr	76.4068	60.4229	92.3907

Somewhat narrower intervals are obtained using the Bonferroni correction – in that case not only the interval for ferritin but also the interval for haemoglobin lies below the specified value.

Confidence intervals [Benferroni]			
	mean	-95% CI for	+95% CI for
RBC	4.6915	4.5632	4.8198
Hg	14.6305	14.2958	14.9653
Ferr	76.4068	62.9731	89.8404

Because of its simplicity, the univariate approach is the one most often used. Here we can choose less conservative multiple-comparison corrections than the Bonferroni or Sidak correction, obtaining in this way differences for all the parameters studied. To adjust the p values of the t-Student test, copy those values into a single column of a new data sheet and, from the Statistics menu, choose Correction for multiple comparisons.

Copyright ©2010-2026 U-smile Stats – All rights reserved

t-test	Frequency	-95% CI for	+95% CI for	t-statistic	Degrees of f	p-value
RBC	59	4.5874	4.7957	-2.0845	58	0.0415
Hg	59	14.3587	14.9023	-2.7213	58	0.0086
Ferr	59	65.4997	87.3139	-4.3299	58	0.0001

Corrections for multiple comparisons	
Number of missing data	0
Frequency	3
Significance level	0.05

Variable: p-value (t test)	Bonferroni	Bonferroni - Holm	Bonferroni - Hochberg	Sidak	Sidak - Holm	Sidak - Hochberg	Benjamini - Hochberg
0.0415	0.1245	0.0415	0.0415	0.1194	0.0415	0.0415	0.0415
0.0086	0.0258	0.0172	0.0172	0.0256	0.0171	0.0171	0.0129
0.0001	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu = \mu_0, \\ \mathcal{H}_1: &\quad \text{not all } \mu_j \text{ equal the specified values,} \end{aligned}$$

where:
$\mu = (\mu_1, \mu_2, ..., \mu_k)$ – means of the variables in the population represented by the sample,
$\mu_0 = (\mu_{01}, \mu_{02}, ..., \mu_{0k})$ – mean values specified by the researcher.

Formulas: The test statistic is defined by:

$$F = \frac{n-k}{k(n-1)} T^2$$

where:
$n = n_1 = n_2 = ... = n_k$ – the sample sizes of the individual variables,
$T^2 = n(\overline{x} - \mu_0)^T S^{-1} (\overline{x} - \mu_0)$ – the original Hotelling test statistic, which has the χ^2 distribution (recommended for samples with large sizes),
$\overline{x} = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_k)$ – means of the variables in the sample,
S – the covariance matrix.

The test statistic has the F Snedecor distribution with k and $n-k$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Simultaneous confidence intervals

When, after performing the analysis, we look for the variables the differences concern, we calculate simultaneous confidence intervals for the means:

$$\overline{x}_j \pm \sqrt{\frac{k(n-1)}{n-k} F_{\alpha, df_1, df_2}} \cdot SE_j$$

or the intervals with the Bonferroni correction, to check whether the specified value falls within them. If the specified value falls within the calculated interval, this means the true mean of that variable may

Copyright ©2010-2026 U-smile Stats – All rights reserved

actually equal the specified value. When using this method, however, keep in mind that the resulting intervals do not take into account the interrelationships between the individual variables (which Hotelling's test does take into account), only the effect of multiple testing.

When looking for the variables the differences concern, we can also use a univariate approach, performing separate comparisons for each variable with the single-sample t-test. Unfortunately, this does not account for the interrelationships between variables, but the p values obtained from the t-Student test can be adjusted in the Multiple comparisons section.

Hotelling's T-square test for dependent groups

WHEN TO USE THIS TEST:

Used when k variables under study are measured twice, under different conditions (assuming that the variances of the variables are close to each other in both measurements). If we denote the first measurement by $X_1, X_2, ..., X_k$ and the second by $Y_1, Y_2, ..., Y_k$, we test the hypothesis that the population means of the variables from the first measurement are the same as from the second measurement.

ASSUMPTIONS

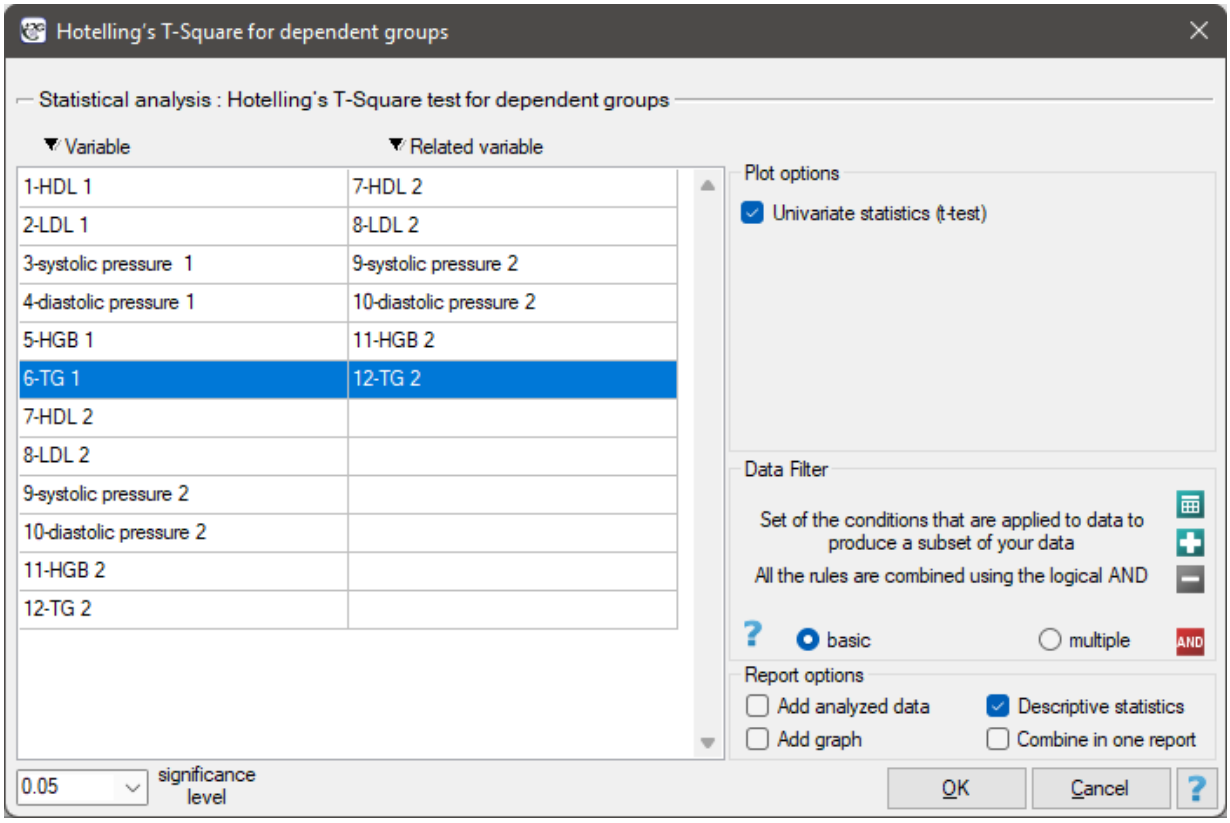
- measurement on an interval scale,
- a multivariate normal distribution, or normality of distribution of each variable studied,
- a dependent model.

Note! The principle behind Hotelling's test is identical to constructing a multivariate confidence-interval ellipse around the centre determined by the mean differences (see the example interpreting the Hotelling test ellipse for a single sample). Because of this, a univariate analysis (which does not take the interrelationships between the variables into account) will often fail to give an identical result.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → T-square Hotelling for dependent groups.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 19.2. (hypertension.pqs file)
In a group of people suffering from hypertension, the effect of treatment on changes in indicators such as: HDL and LDL cholesterol, haemoglobin (HGB), triglycerides (TG), and systolic and diastolic blood pressure was studied. Measurements from 44 patients were taken twice (before treatment and after 3 months of treatment). The resulting values were then compared.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Hotelling's T-Square test for dependent groups	
Analysed variables	HDL 1
	LDL 1
	systolic pressure 1
	diastolic pressure 1
	HGB 1
	TG 1
	HDL 2
	LDL 2
	systolic pressure 2
	diastolic pressure 2
	HGB 2
	TG 2
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Frequency	44
T-Square	39.1998
F statistic	5.7736
Degrees of freedom 1	6
Degrees of freedom 2	38
p-value	0.0002

$p = 0.00024$ **(statistically significant, $\alpha = 0.05$)**

The mean values of the parameters studied change as a result of treatment (the difference is significantly different from 0).

The confidence intervals for systolic and diastolic blood pressure lie above the value 0, indicating a significant decrease in these parameters as a result of treatment. The intervals for the remaining parameters contain the value 0, so we have no evidence of a change resulting from treatment.

Confidence intervals [Simultaneous]			
	difference of	-95% CI	+95% CI
HDL 1/HDL 2	0.0191	-0.1995	0.2377
LDL 1/LDL 2	0.0957	-0.3477	0.5391
systolic pressur	8.0227	1.1112	14.9343
diastolic pressur	5.7955	0.2053	11.3856
HGB 1/HGB 2	0.2323	-0.1747	0.6393
TG 1/TG 2	-0.1102	-0.5188	0.2983

Confidence intervals [Benferroni]			
	difference of	-95% CI	+95% CI
HDL 1/HDL 2	0.0191	-0.1323	0.1705
LDL 1/LDL 2	0.0957	-0.2114	0.4027
systolic pressur	8.0227	3.236	12.8095
diastolic pressur	5.7955	1.9239	9.667
HGB 1/HGB 2	0.2323	-0.0496	0.5141
TG 1/TG 2	-0.1102	-0.3932	0.1727

Copyright ©2010-2026 U-smile Stats – All rights reserved

Because of its simplicity, the univariate approach is the one most often used. Using this approach together with a multiple-comparison correction, we likewise conclude that the differences concern only the blood-pressure values. To adjust the p values of the t-Student test, copy those values into a single column of a new data sheet and, from the Statistics menu, choose Correction for multiple comparisons.

t-test						
	Frequency	-95% CI for	+95% CI for	t-statistic	Degrees of f	p-value
HDL 1/HDL 2	44	-0.0913	0.1295	0.3488	43	0.729
LDL 1/LDL 2	44	-0.1282	0.3196	0.8618	43	0.3936
systolic pressur	44	4.5325	11.513	4.6356	43	<0.0001
diastolic pressur	44	2.9725	8.6184	4.1403	43	0.0002
HGB 1/HGB 2	44	0.0267	0.4378	2.2792	43	0.0277
TG 1/TG 2	44	-0.3165	0.0961	-1.0775	43	0.2872

Corrections for multiple comparisons	
Number of missing data	0
Frequency	6
Significance level	0.05

Variable: p-value (t test)	Bonferroni	Bonferroni - Holm	Bonferroni - Hochberg	Sidak	Sidak - Holm	Sidak - Hochberg	Benjamini - Hochberg
0.729	1	0.8616	0.729	0.9996	0.729	0.729	0.729
0.3936	1	0.8616	0.729	0.9503	0.6378	0.6323	0.4723
0.0001	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
0.0002	0.0012	0.001	0.001	0.0012	0.001	0.001	0.0006
0.0277	0.1662	0.1108	0.1108	0.1551	0.1063	0.1063	0.0554
0.2872	1	0.8616	0.729	0.8688	0.6378	0.6323	0.4308

FOR THOSE INTERESTED

Definition: equivalently, if we calculate the differences between the pairs of measurements $d_1, d_2, ..., d_k$, the hypothesis states that the means of the differences in the population studied equal 0.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_0 = 0, \\ \mathcal{H}_1: &\quad \text{not all } \mu_{j0} \text{ equal zero,} \end{aligned}$$

where $\mu_0 = (\mu_{01}, \mu_{02}, ..., \mu_{0k})$ – the population means of the differences between the first and second measurement.

Formulas: The test statistic is defined by:

$$F = \frac{n-k}{k(n-1)} T^2$$

where:
$n = n_1 = n_2 = ... = n_k$ – the sample sizes of the individual differences,
$T^2 = n(\overline{x} - \overline{y})^T S^{-1} (\overline{x} - \overline{y})$ – the original Hotelling test statistic, which has the χ^2 distribution (recommended for samples with large sizes),
$\overline{x} = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_k)$ – means of the variables in the sample for the first measurement,
$\overline{y} = (\overline{y}_1, \overline{y}_2, ..., \overline{y}_k)$ – means of the variables in the sample for the second measurement,
S – the covariance matrix of the differences between the first and second measurement.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The test statistic has the F Snedecor distribution with k and $n-k$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Simultaneous confidence intervals

When, after performing the analysis, we look for the variables the differences concern, we calculate simultaneous confidence intervals for the mean differences:

$$\overline{x}_j - \overline{y}_j \pm \sqrt{\frac{k(n-1)}{n-k} F_{\alpha,df_1,df_2}} \cdot SE_{(diff)_j}$$

or the intervals with the Bonferroni correction, to check whether the value 0 falls within them. If the difference can equal 0, this means the difference between the values studied may not actually exist. When using this method, keep in mind that the resulting intervals do not take into account the interrelationships between the variables (which Hotelling's test does take into account), only the effect of multiple testing.

When looking for the variables the differences concern, we can also use a univariate approach, performing separate comparisons for each variable with the t-test for dependent groups. Unfortunately, this does not account for the interrelationships between variables, but the p values obtained from the t-Student test can be adjusted in the Multiple comparisons section.

Hotelling's T-square test for independent groups

WHEN TO USE THIS TEST:

This test lets you check the equality of the means of k variables studied, $X_1, X_2, ..., X_k$, from the first population, and the means of the same k variables, $Y_1, Y_2, ..., Y_k$, from the second population.

ASSUMPTIONS

- measurement on an interval scale,
- a multivariate normal distribution in both populations, or normality of distribution of each variable studied in both populations,
- an independent model,
- equality of variances of the variables studied in both populations, or equality of the covariance matrices – a condition that is especially important for groups of different sizes; when it is not met, a correction should be calculated.

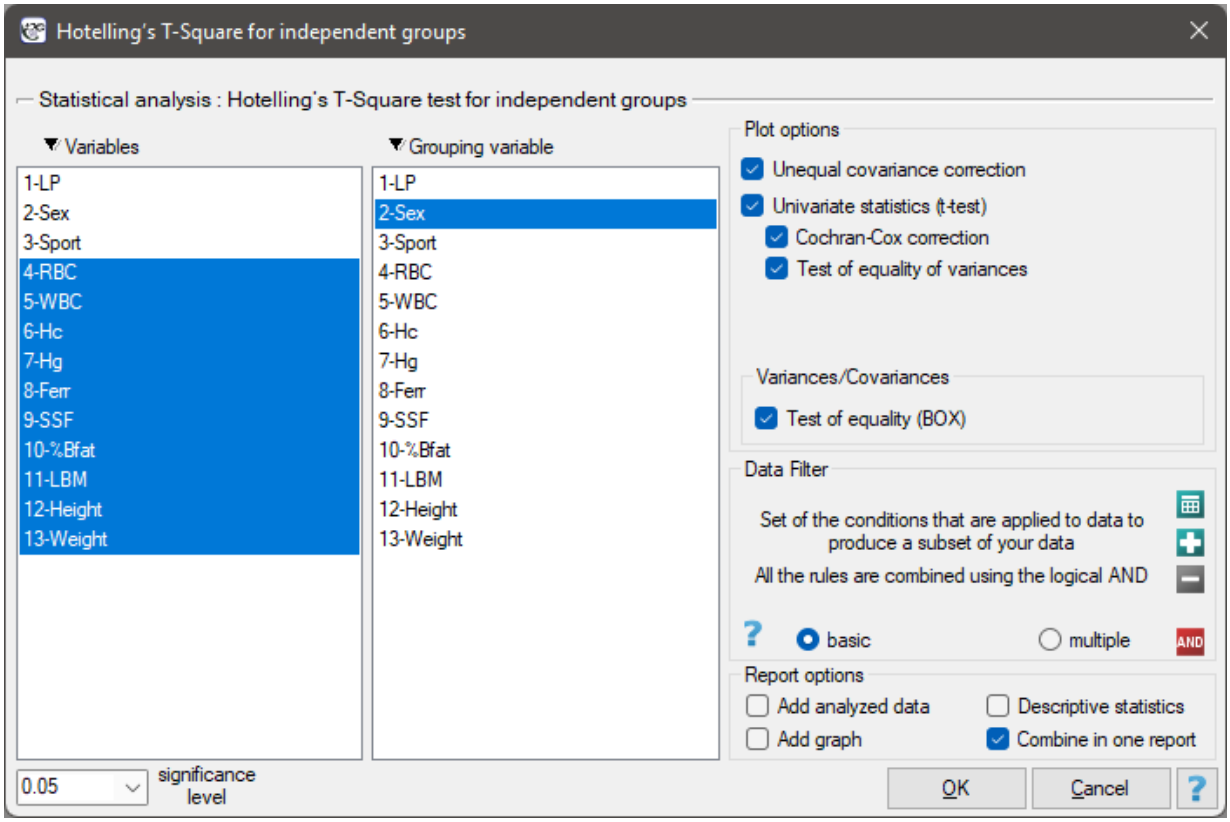
Note! The principle behind Hotelling's test is identical to constructing multivariate confidence-interval ellipses around the centres determined by the means (see the example interpreting the Hotelling test

Copyright ©2010-2026 U-smile Stats – All rights reserved

ellipse for a single sample). Because of this, a univariate analysis (which does not take the interrelationships between the variables into account) will often fail to give an identical result.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → T-square Hotelling for independent groups.



PERFORMING AND INTERPRETING THE ANALYSIS

Example 19.3. (sport.pqs file)
A group of athletes was studied to obtain information on health parameters such as:

RBC – red blood cell count,	SSF – sum of skinfolds [mm],
WBC – white blood cell count,	% Bfat – body fat percentage,
Hc – haematocrit [%],	LBM – lean body mass [kg],
Hg – haemoglobin [g/dl],	Height [cm],
Ferr – ferritin [µg/l],	Body weight [kg].

We want to know whether women and men who practise the sports studied professionally differ in the levels of these parameters.

Because of differences in the covariance of the parameters analysed (Box's test $p < 0.000001$), the analysis is performed with the correction for unequal covariance matrices enabled.

Copyright ©2010-2026 U-smile Stats – All rights reserved

BOX	
Degrees of freedom 1	55
Degrees of freedom 2	67477.1082
Box's M statistic	260.303
F statistic	4.3949
p-value	<0.0001

The resulting corrected Hotelling statistic is shown below.

Hotelling's T-Square test for independent groups	
Analysed variables	RBC
	WBC
	Hc
	Hg
	Ferr
	SSF
	%Bfat
	LBM
	Height
	Weight
	Sex
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	Sex
Frequency	151
Unequal covariance correction	Yes
T-Square	1048.2697
F statistic	98.4951
Degrees of freedom 1	10
Degrees of freedom 2	146.8964
p-value	<0.0001

$p < 0.000001$ **(statistically significant,** $\alpha = 0.05$**)**

Men and women in the population studied have different mean values of the parameters analysed.

According to the simultaneous confidence intervals (and according to the intervals with the Bonferroni correction), the differences concern: RBC, Hc, Hg, ferritin, and LBM (women score lower on average), as well as SSF and % Bfat (men score lower on average). The remaining parameters – WBC, height, and body weight – do not differ significantly.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Confidence intervals [Simultaneous]

	difference of	-95% CI	+95% CI
RBC	-0.614	-0.8679	-0.3601
WBC	0.3517	-0.8314	1.5349
Hc	-4.9612	-6.9153	-3.007
Hg	-2.0044	-2.6855	-1.3233
Ferr	-35.9445	-64.7919	-7.0971
SSF	43.9535	25.7809	62.1261
%Bfat	10.1645	7.2735	13.0554
LBM	-17.6416	-23.5534	-11.7298
Height	-4.1451	-11.0836	2.7935
Weight	-7.0924	-16.7059	2.5211

Confidence intervals [Benferroni]

	difference of	-95% CI	+95% CI
RBC	-0.614	-0.8734	-0.3546
WBC	0.3517	-0.8563	1.5597
Hc	-4.9612	-6.9569	-2.9655
Hg	-2.0044	-2.7	-1.3088
Ferr	-35.9445	-65.4503	-6.4387
SSF	43.9535	25.3664	62.5407
%Bfat	10.1645	7.2072	13.1218
LBM	-17.6416	-23.6889	-11.5943
Height	-4.1451	-11.2341	2.9439
Weight	-7.0924	-16.9173	2.7325

Because of its simplicity, the univariate approach – although it does not account for multiple testing or the interrelationships between variables – is the one most often used. Its results indicate that the differences concern all the parameters studied except WBC, and the results of the t-Student test are interpreted with the Cochran-Cox correction or without it, depending on whether the assumption of equal variances is met (the result of the Fisher-Snedecor test). To account for multiple testing here, one of the p-value corrections described in the Multiple comparisons section should be applied.

t-test

	Frequency	-95% CI for	+95% CI for	t-statistic	Degrees of f	p-value
RBC	151	-0.7236	-0.5045	-11.0762	149	<0.0001
WBC	151	-0.1737	0.8772	1.3227	149	0.188
Hc	151	-5.8143	-4.108	-11.4908	149	<0.0001
Hg	151	-2.3019	-1.7069	-13.3146	149	<0.0001
Ferr	151	-48.0804	-23.8086	-5.8526	149	<0.0001
SSF	151	35.3963	52.5108	10.1496	149	<0.0001
%Bfat	151	8.8008	11.5281	14.7287	149	<0.0001
LBM	151	-20.1245	-15.1587	-14.0401	149	<0.0001
Height	151	-7.1333	-1.1569	-2.7411	149	0.0069
Weight	151	-11.2021	-2.9828	-3.4102	149	0.0008

Copyright ©2010-2026 U-smile Stats – All rights reserved

t-test CC

	Frequency	-95% CI for	+95% CI for	t-statistic	Degrees of f	p-value
RBC	151	-0.7259	-0.5022	-10.8625	149	<0.0001
WBC	151	-0.1688	0.8723	1.3353	149	0.1838
Hc	151	-5.8213	-4.101	-11.4033	149	<0.0001
Hg	151	-2.3042	-1.7046	-13.2185	149	<0.0001
Ferr	151	-48.6814	-23.2076	-5.5968	149	<0.0001
SSF	151	35.93	51.977	10.864	149	<0.0001
%Bfat	151	8.8877	11.4412	15.7926	149	<0.0001
LBM	151	-20.2523	-15.0309	-13.4039	149	<0.0001
Height	151	-7.2018	-1.0884	-2.6834	149	0.0083
Weight	151	-11.3301	-2.8547	-3.3138	149	0.0012

Fisher-Snedecor test

	F statistic	p-value
RBC	0.6331	0.0489
WBC	1.246	0.3514
Hc	0.838	0.444
Hg	0.8456	0.4675
Ferr	0.3199	<0.0001
SSF	8.2552	<0.0001
%Bfat	9.189	<0.0001
LBM	0.3034	<0.0001
Height	0.6059	0.0309
Weight	0.5029	0.0032

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned}\mathcal{H}_0: &\quad \mu_x = \mu_y,\\ \mathcal{H}_1: &\quad \text{not all } \mu_{xj} = \mu_{yj},\end{aligned}$$

where:
$\mu_x = (\mu_{x1}, \mu_{x2}, ..., \mu_{xk})$ – means of the variables in the first population,
$\mu_y = (\mu_{y1}, \mu_{y2}, ..., \mu_{yk})$ – means of the variables in the second population.

Formulas: The test statistic is defined by:

$$F = \frac{n-k}{k(n-1)} T^2$$

where:
$n = n_x + n_y - 1$,
n_x, n_y – the sizes of the first and second sample (the sizes of the individual variables are the same),
$T^2 = (\overline{x} - \overline{y})^T \left(S \left(\frac{1}{n_x} + \frac{1}{n_y} \right) \right)^{-1} (\overline{x} - \overline{y})$ – the original Hotelling test statistic, which has the χ^2 distribution (recommended for samples with large sizes),
$\overline{x} = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_k)$ – means of the variables in the first sample,
$\overline{y} = (\overline{y}_1, \overline{y}_2, ..., \overline{y}_k)$ – means of the variables in the second sample,
S – the pooled covariance matrix for both samples.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The test statistic has the F Snedecor distribution with k and $n-k$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Simultaneous confidence intervals

When, after performing the analysis, we look for the variables the differences concern, we calculate simultaneous confidence intervals for the mean differences:

$$\overline{x}_j - \overline{y}_j \pm \sqrt{\frac{k(n-1)}{n-k} F_{\alpha,df_1,df_2}} \cdot SE_{(diff)_j}$$

or the intervals with the Bonferroni correction, to check whether the value 0 falls within them. If the difference can equal 0, this means the difference between the values studied may not actually exist. When using this method, keep in mind that the resulting intervals do not take into account the interrelationships between the variables (which Hotelling's test does take into account), only the effect of multiple testing.

When looking for the variables the differences concern, we can also use a univariate approach, performing separate comparisons for each variable with the t-test for independent groups. Unfortunately, this does not account for the interrelationships between variables, but the p values obtained from the t-Student test can be adjusted in the Multiple comparisons section.

Hotelling's T-square test for independent groups with correction for unequal variances

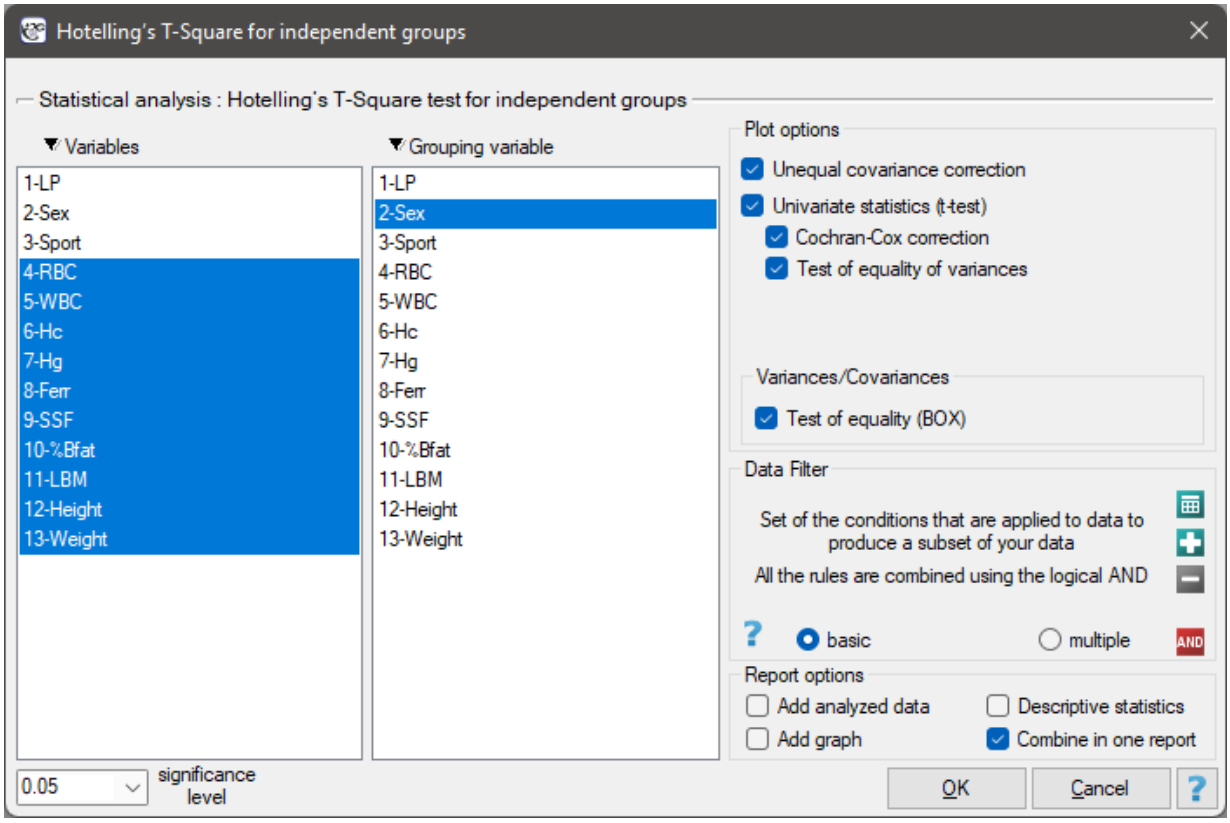
WHEN TO USE THIS TEST:

This correction applies to the Hotelling's T-square test for independent groups and is calculated when the variances of the variables studied differ between the two populations.

HOW TO SET UP THE ANALYSIS

The correction is an option available in the same window as the Hotelling's T-square test for independent groups.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

See Example 19.3 above – because of the significant result of Box's test, the analysis there was in fact performed using this correction.

FOR THOSE INTERESTED

Formulas: The test statistic is defined by:

$$F = \frac{n-k}{k(n-1)} T^2$$

where:
$n = n_x + n_y - 1$,
n_x, n_y – the sizes of the first and second sample (the sizes of the individual variables are equal),
$T^2 = (\overline{x} - \overline{y})^T \left(\frac{S_x}{n_x} + \frac{S_y}{n_y} \right)^{-1} (\overline{x} - \overline{y})$ – the original Hotelling test statistic, which has the χ^2 distribution (recommended for samples with large sizes),
$\overline{x} = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_k)$ – means of the variables in the first sample,
$\overline{y} = (\overline{y}_1, \overline{y}_2, ..., \overline{y}_k)$ – means of the variables in the second sample,
S_x, S_y – the covariance matrices for the first and second sample.

The test statistic has the F Snedecor distribution with k and m degrees of freedom (where m is the number of degrees of freedom adjusted for the unequal variances).

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Simultaneous confidence intervals

When, after performing the analysis, we look for the variables the differences concern, we calculate simultaneous confidence intervals for the mean differences:

$$\overline{x}_j - \overline{y}_j \pm \sqrt{\frac{k(n-1)}{n-k} F_{\alpha,df_1,df_2}} \cdot SE_{(diff)_j}$$

or the intervals with the Bonferroni correction, to check whether the value 0 falls within them. If the difference can equal 0, this means the difference between the values studied may not actually exist. When using this method, keep in mind that the resulting intervals do not take into account the interrelationships between the variables (which Hotelling's test does take into account), only the effect of multiple testing.

When looking for the variables the differences concern, we can also use a univariate approach, performing separate comparisons for each variable with the t-test with the Cochran-Cox correction. Unfortunately, this does not account for the interrelationships between variables, but the p values obtained from the t-Student test can be adjusted in the Multiple comparisons section.

Box's test for equality of covariance matrices

WHEN TO USE THIS TEST:

This test is used to compare two or more ($m \geq 2$) covariance matrices describing independent populations.

ASSUMPTIONS

- measurement on an interval scale,
- a multivariate normal distribution in the populations studied, or normality of distribution of each variable studied in each population,
- an independent model.

Note! Box's test is especially sensitive to a violation of the normality assumption.

HOW TO SET UP THE ANALYSIS

Box's test is calculated as an option in Hotelling's test for independent groups or in the MANOVA analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

See Example 19.3.

Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \Sigma_1 = \Sigma_2 = ... = \Sigma_m, \\ \mathcal{H}_1: &\quad \text{not all } \Sigma_j \text{ are equal } (j = 1, 2, ..., m), \end{aligned}$$

where $\Sigma_1, \Sigma_2, ..., \Sigma_m$ – the population covariance matrices.

Formulas: The test statistic is defined by:

$$F = \frac{M}{b}$$

where:
$M = (n - m) \ln |S| - \sum_{j=1}^{m} (n_j - 1) \ln |S_j|$,
S – the pooled covariance matrix,
S_j – the covariance matrix for the j-th sample,
$b = \frac{df_1}{1 - c_1 - \frac{df_1}{df - 2}}$,
$df_1 = \frac{k(k+1)(m-1)}{2}$,
$df_2 = \frac{df_1 + 2}{|c_2 - c_1^2|}$,
k – the number of variables analysed,
$n = n_1 = n_2 = ... = n_k$ – the sample sizes of the individual variables.

The test statistic has the F Snedecor distribution with df_1 and df_2 degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

20 UNIVARIATE MANOVA

One-way MANOVA for independent groups

WHEN TO USE THIS TEST:

One-way MANOVA (multivariate analysis of variance) is an extension of the one-way ANOVA for independent groups. It lets you check the equality of the means of k variables studied across several ($m \geq 2$) populations. Unlike using ANOVA repeatedly (separately for each variable), MANOVA examines the differences between populations simultaneously for multiple variables, while taking their mutual correlation into account. This approach is also used as an alternative to the ANOVA for dependent groups, since it does not require the sphericity assumption to be met. We recommend reading about Hotelling's T-square beforehand.

ASSUMPTIONS

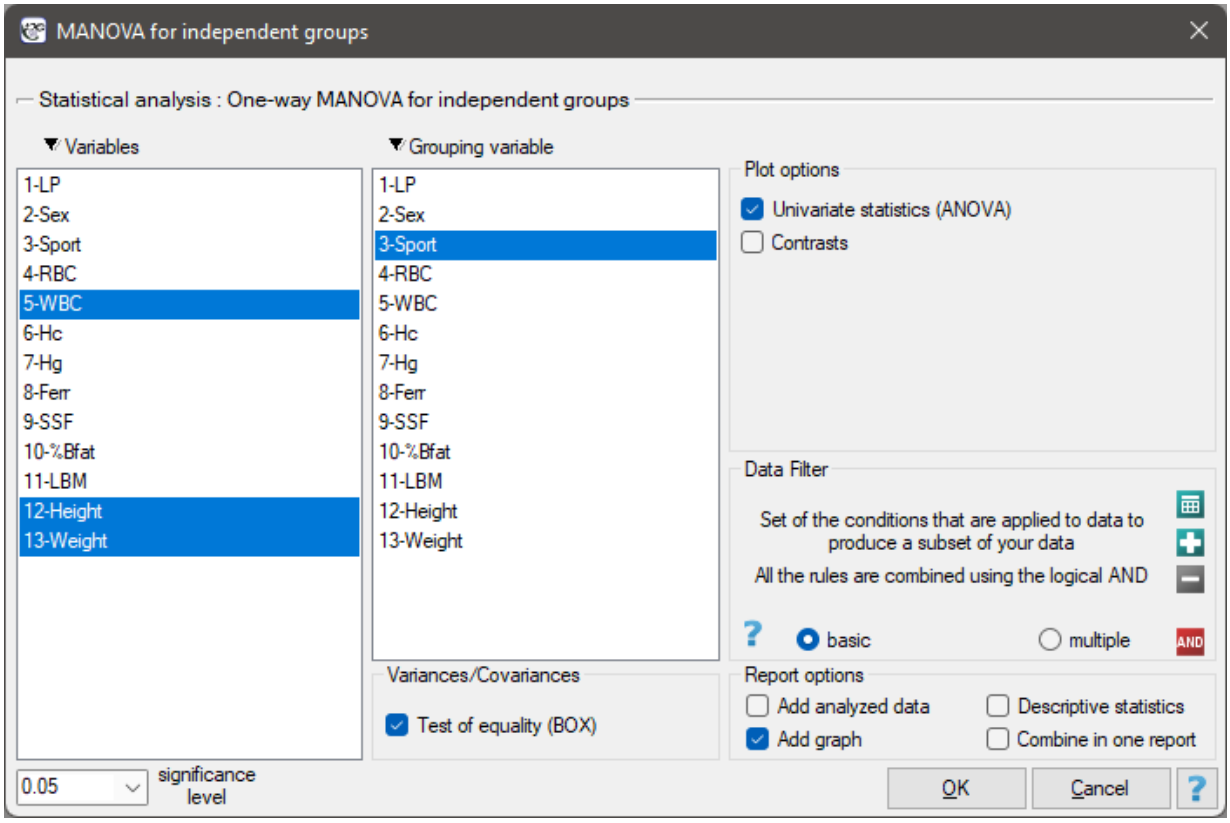
- measurement on an interval scale,
- a multivariate normal distribution in each population, or normality of distribution of each variable studied in each population,
- an independent model,
- equality of the covariance matrices, or equality of variances of the variables studied for the populations compared – a condition that is especially important for groups of different sizes.

Note! The principle behind MANOVA (as with Hotelling's tests) is identical to constructing multivariate confidence-interval ellipses around the centres determined by the means (see the example interpreting the Hotelling test ellipse for a single sample). Because of this, a univariate analysis (which does not take the interrelationships between the variables into account) will often fail to give an identical result.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → One-way MANOVA for independent groups.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 20.1. (sport.pqs file)
A group of athletes was studied to obtain information on health parameters, among others such as:
WBC – white blood cell count,
Height [cm],
Body weight [kg].

We want to know:

1. whether athletes who professionally practise three types of sport – "team sports"(such as basketball, volleyball, etc.), "track"(such as 100m, 400m, etc.), and "water sports"(such as swimming, rowing, etc.) – differ in the levels of these parameters;

2. whether athletes practising sports that require great effort, such as "track"and "water sports", differ in the levels of these parameters from those practising "team sports".

Re. 1)
Hypotheses:

$\mathcal{H}_0$: the means of the parameters analysed are the same for athletes practising the individual types of sport,
$\mathcal{H}_1$: at least one parameter has a different mean value for the populations compared.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The result of Box's test ($p = 0.63025$) allows a MANOVA-type analysis to be calculated.

One-way MANOVA for independent groups	
Analysed variables	WBC
	Height
	Weight
	Sport
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	Sport
Size	151
Number of groups	3
Number of variables in the model	3
Wilk's Lambda	0.8929
Eta-square	0.0551
df1	6
df2	292
F statistic	2.8353
p-value	0.0107
Hotelling-Lawley Trace	0.1174
Eta-square	0.0554
df1	6
df2	290
F statistic	2.8368
p-value	0.0106
Pillai-Bartlett Trace	0.1093
Eta-square	0.0547
df1	6
df2	294
F statistic	2.8335
p-value	0.0107
BOX	
Degrees of freedom 1	12
Degrees of freedom 2	93631.9269
Box's M statistic	10.1406
F statistic	0.8198
p-value	0.6302

significant Wilks' Lambda, Hotelling-Lawley trace, and Pillai-Bartlett trace (statistically significant, $\alpha = 0.05$)

The populations of athletes studied differ with regard to the parameters analysed.

To determine which variables the differences concern, we perform a univariate ANOVA analysis.

One-way ANOVA			
	WBC	Height	Weight
SS[BG]	15.9682	554.0601	1262.241
SS[WG]	383.4849	12857.2188	24771.6939
df[BG]	2	2	2
df[WG]	148	148	148
F statistic	3.0813	3.1889	3.7707
p-value	0.0489	0.0441	0.0253

Copyright ©2010-2026 U-smile Stats – All rights reserved

ANOVA-POST-HOC (Fisher LSD) [WBC]			
	team	track	water
p-value	team	track	water
team		0.0585	0.0204
track	0.0585		0.773
water	0.0204	0.773	
Homogeneous	team(b)	track(a,b)	water(a)
A		*	*
B	*	*	

ANOVA-POST-HOC (Fisher LSD) [Height]			
	team	track	water
p-value	team	track	water
team		0.0175	0.5718
track	0.0175		0.0513
water	0.5718	0.0513	
Homogeneous	team(a)	track(b)	water(a,b)
A	*		*
B		*	*

ANOVA-POST-HOC (Fisher LSD) [Weight]			
	team	track	water
p-value	team	track	water
team		0.0076	0.0747
track	0.0076		0.2808
water	0.0747	0.2808	
Homogeneous	team(a)	track(b)	water(a,b)
A	*		*
B		*	*

The results obtained should be treated with caution. Although they indicate significant differences in all the parameters compared, the p values returned are close to the boundary of statistical significance (for WBC $p = 0.04885$, for height $p = 0.04406$, for weight $p = 0.02528$). In addition, when interpreting them, keep in mind that they take into account neither the mutual correlation of the parameters nor multiple testing. Accounting for multiple testing here would require applying one of the p-value corrections described in the Multiple comparisons section.

Re. 2)
Hypotheses:

$\mathcal{H}_0$: the means of the parameters analysed for “team sports”
do not differ from the corresponding means of athletes from the other two groups,
$\mathcal{H}_1$: at least one parameter has a different mean value
for the populations compared.

To check whether the above hypotheses hold, we set an appropriate contrast in the analysis window. We enter 2 as the contrast value for team sports, and -1 for track and for the disciplines defined as water sports.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Multivariate analysis contrasts:	
Contrast for group [team]	2
Contrast for group [track]	-1
Contrast for group [water]	-1
Wilk's Lambda	0.9184
Eta-square	0.0816
df1	3
df2	146
F statistic	4.3255
p-value	0.0059
Hotelling-Lawley Trace	0.0889
Eta-square	0.0816
df1	3
df2	146
F statistic	4.3255
p-value	0.0059
Pillai-Bartlett Trace	0.0816
Eta-square	0.0816
df1	3
df2	146
F statistic	4.3255
p-value	0.0059

Confidence intervals for effect [Simultaneous]			
Group	Effect	-95% CI	+95% CI
WBC	1.3737	-0.234	2.9814
Height	-5.7006	-15.0096	3.6085
Weight	-11.8187	-24.7401	1.1027

Confidence intervals for effect [Benferroni]			
Group	Effect	-95% CI	+95% CI
WBC	1.3737	0.0066	2.7408
Height	-5.7006	-13.6164	2.2153
Weight	-11.8187	-22.8063	-0.8312

$p = 0.00592$ for Wilks' Lambda, the Hotelling-Lawley trace, and the Pillai-Bartlett trace (statistically significant, $\alpha = 0.05$)

Athletes practising sports that require great effort differ in the parameters analysed from those practising team sports.

We do not observe these differences in the simultaneous intervals, but on the basis of the Bonferroni intervals we can conclude that the difference concerns weight and WBC. WBC values are higher in the team-sports group, and body weight is significantly lower in this group.

DEFINITIONS OF MEASURES

Partial η^2 – a measure of effect size expressing the proportion of explained variance to total variance associated with a given factor. In the one-way MANOVA model for independent groups, it indicates what

Copyright ©2010-2026 U-smile Stats – All rights reserved

proportion of the within-subject variability in outcomes can be attributed to the factor under study that defines the independent groups.

Contrasts (univariate analysis) – when the analysis performed aims to compare selected populations or a selected pattern of populations, we perform a contrast analysis, analogous to contrasts in univariate analysis, but taking into account the interrelationships between the variables.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2 = ... = \mu_m, \\ \mathcal{H}_1: &\quad \text{not all } \mu_i \text{ are equal,} \end{aligned}$$

where:
$\mu_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{ik})$ – means of the variables in the i-th population,
$(i = 1, 2, ..., m)$,
$(j = 1, 2, ..., k)$.

Explanation:
Several coefficients are used in MANOVA analyses. The best known is Wilks' Lambda. The Pillai-Bartlett trace is the most conservative, but relatively robust to violations of the MANOVA assumptions, and is preferred for small samples. The Hotelling-Lawley trace, in turn, is the least conservative of the three tests proposed. Pioneering work on the development of these techniques was begun by Wilks (1932)[180], Pillai (1955)[137], Lawley (1938)[102], Hotelling (1951)[88], and Roy (1939)[145].

The test statistics are based on matrices of sums of squares and cross products ($SSCP$). The total matrix $T = SSCP$ is split into two matrices: the first is associated with the hypothesis analysed and is denoted by H (the between-groups sum-of-squares and cross-products matrix), and the second concerns the residuals (errors) and is denoted by E (the within-groups sum-of-squares and cross-products matrix).

Wilks' Lambda – formulas:
The value of Lambda is defined as follows:

$$\Lambda = \frac{|E|}{|H + E|}$$

The test statistic is defined by:

$$F = \frac{1 - \Lambda^{\frac{1}{b}}}{\Lambda^{\frac{1}{b}}} \frac{df_2}{df_1}$$

where $df_1 = k(m-1)$, $df_2 = ab - c$, and a, b, c are coefficients that depend on the number of variables analysed and the number of populations compared.

The Hotelling-Lawley trace – formulas:
The Hotelling-Lawley trace is defined as follows:

$$T_0 = trace(HE^{-1})$$

The test statistic is defined by:

$$F = \frac{T_0^2}{s} \frac{df_2}{df_1}$$

where $df_1 = s(2t + s + 1)$, $df_2 = 2(su + 1)$, and s, t, u are coefficients that depend on the number of variables analysed and the number of populations compared.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Pillai-Bartlett trace – formulas:
The Pillai-Bartlett trace is defined as follows:

$$V = trace(H(H+E)^{-1})$$

The test statistic is defined by:

$$F = \frac{V}{s-V}\frac{df_2}{df_1}$$

where $df_1 = s(2t+s+1)$, $df_2 = s(2u+s+1)$, and s, t, u are coefficients that depend on the number of variables analysed and the number of populations compared.

Each of the above test statistics has the F Snedecor distribution with df_1 and df_2 degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Partial η^2 – formula:

$$\eta^2 = \frac{F \cdot df_1}{F \cdot df_1 + df_2}$$

Contrasts (univariate analysis)
For effect sizes, **simultaneous confidence intervals** or confidence intervals with the Bonferroni correction can also be calculated. When using these intervals, however, keep in mind that they do not take into account the interrelationships between the individual variables (which MANOVA does take into account), only the effect of multiple testing.

When looking for the variables the differences concern, we can also use a univariate approach, performing separate comparisons for each variable with the ANOVA for independent groups. Unfortunately, this does not account for the interrelationships between variables, but the p values obtained from ANOVA can be adjusted in the Multiple comparisons section.

Copyright ©2010-2026 U-smile Stats – All rights reserved

21 ANALYSIS FOR STRATAS

21.1 The Mantel-Haenszel method for several 2x2 tables

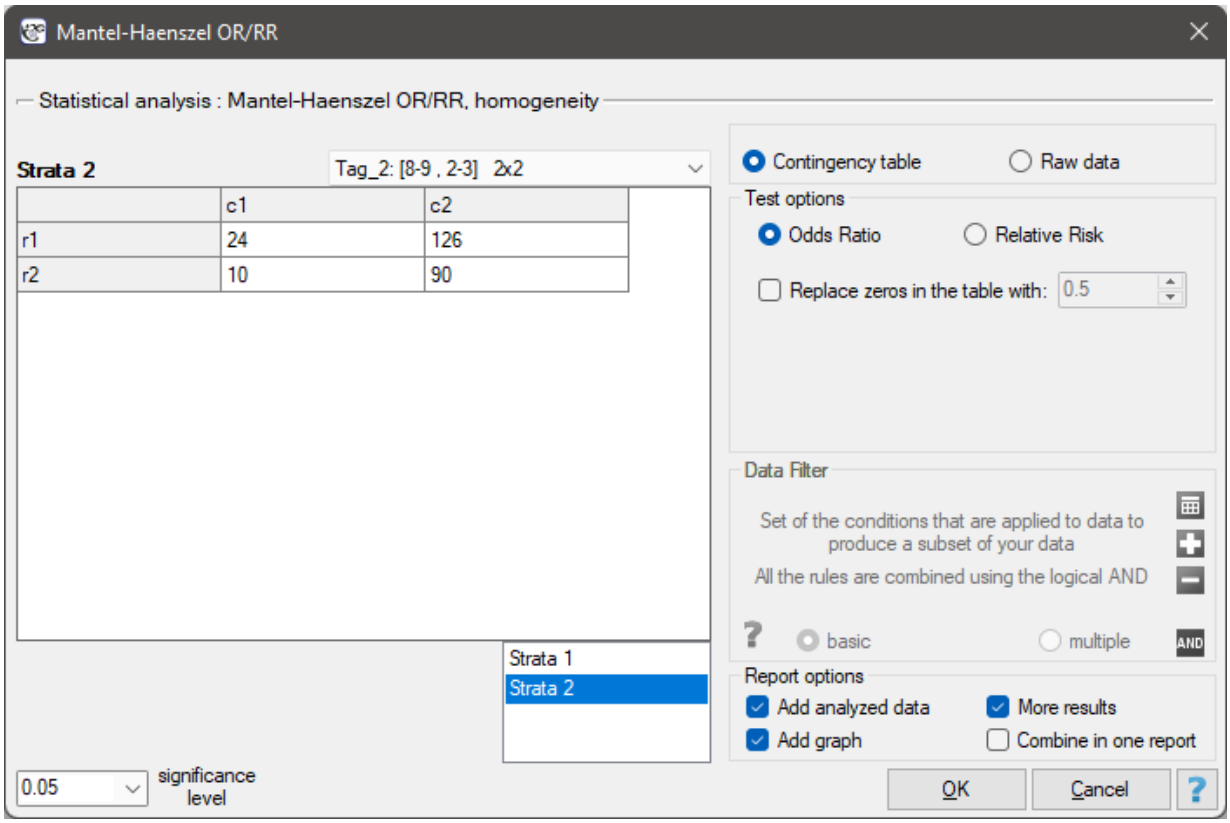
WHEN TO USE:

The Mantel-Haenszel method for 2×2 tables, proposed by Mantel and Haenszel (1959)[115] and later extended by Mantel (1963)[116] (a broader review of the development of these methods was carried out by, among others, Newman (2001)[129]), can be used, among other things, to analyse 2×2 tables occurring in several ($w \geq 2$) strata formed by a so-called confounding variable. For successive strata ($s = 1, ..., w$), 2×2 contingency tables of observed frequencies are built:

Observed frequencies stratum s $\left(O_{ij}^{(s)}\right)$		phenomenon studied		
		occurs	does not occur	Total
risk factor	present	$O_{11}^{(s)}$	$O_{12}^{(s)}$	$O_{11}^{(s)} + O_{12}^{(s)}$
	absent	$O_{21}^{(s)}$	$O_{22}^{(s)}$	$O_{21}^{(s)} + O_{22}^{(s)}$
	Total	$O_{11}^{(s)} + O_{21}^{(s)}$	$O_{12}^{(s)} + O_{22}^{(s)}$	$n^{(s)} = O_{11}^{(s)} + O_{12}^{(s)} + O_{21}^{(s)} + O_{22}^{(s)}$

HOW TO SET UP THE ANALYSIS

The settings window for the Mantel-Haenszel OR/RR options is opened via the Statistics→Stratified analysis→Mantel-Haenszel OR/RR menu.



Copyright ©2010-2026 U-smile Stats – All rights reserved

The Mantel-Haenszel odds ratio

WHEN TO USE:

If all the tables (formed by the individual strata) are homogeneous, a pooled odds ratio, together with confidence intervals, can be determined on the basis of these tables. This odds ratio is a weighted mean of the odds ratios determined for the individual strata. Using the weighted method proposed by Mantel and Haenszel makes it possible to account for the contribution (weight) each stratum makes to the pooled odds ratio (the larger the stratum, the greater its influence on the resulting odds ratio).

Two significance tests accompany the odds ratio determined this way. The **Mantel-Haenszel χ^2 test for** OR_{MH} checks whether the pooled odds ratio differs significantly from 1, i.e. whether a relationship between the features studied exists at all once strata are accounted for. The χ^2 **test of homogeneity for** OR, in turn, checks the assumption underlying the whole method – whether the odds ratio is indeed similar across all the strata, or whether the variable forming the strata significantly modifies the strength of the relationship (in which case pooling the strata into a single common odds ratio would not be justified).

DEFINITIONS OF MEASURES

The Mantel-Haenszel odds ratio – a weighted mean of the odds ratios determined for the individual strata, taking into account the size of each stratum.

PERFORMING AND INTERPRETING THE ANALYSIS

Example 21.1. (leptospirosis.pqs file)
The table shown presents hypothetical results of a survey conducted to detect risk factors for the occurrence of leptospirosis among rural and urban residents of the West Indies (where a rural area is treated as the exposure factor)[20]. The occurrence of leptospirosis antibodies is indirect evidence of infection.

Observed frequencies O_{ij}		leptospirosis antibodies	
		yes	no
place of residence	rural	60	140
	urban	60	140

The odds of the occurrence of leptospirosis antibodies is the same for urban and rural residents (OR=1). We want to know what the odds of the occurrence of leptospirosis antibodies will be once we also take sex into account in the analysis. We will therefore split the sample into 2 strata by sex (these are stored in the file as a saved selection):

Observed frequencies for men		leptospirosis antibodies	
		yes	no
place of residence	rural	36	14
	urban	50	50

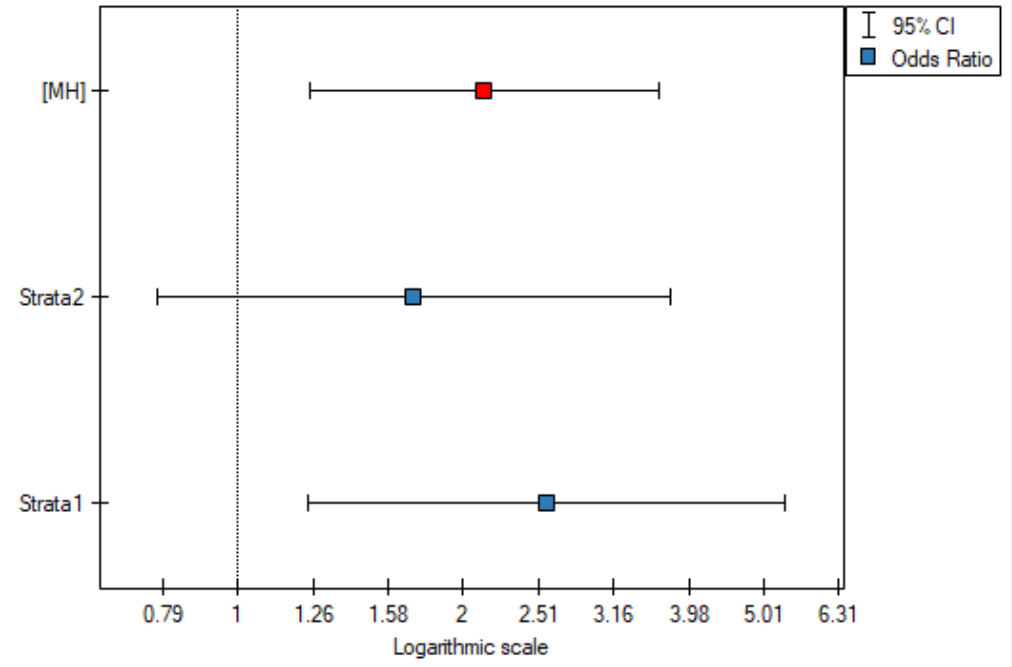
Observed frequencies for women		leptospirosis antibodies	
		yes	no
place of residence	rural	24	126
	urban	10	90

Copyright ©2010-2026 U-smile Stats – All rights reserved

Sex is associated with both factors (the occurrence of leptospirosis antibodies and place of residence in the West Indies). It is therefore a confounding factor, and ignoring it can lead to incorrect results.

Mantel–Haenszel OR/RR, homogeneity	
Analysed variables	Contingency table
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	400
Strata 1	
Odds Ratio	2.5714
-95% CI for the Odds Ratio	1.2376
+95% CI for the Odds Ratio	5.3427
Statistic for the Odds Ratio	2.5314
p-value	0.0114
Strata 2	
Odds Ratio	1.7143
-95% CI for the Odds Ratio	0.7813
+95% CI for the Odds Ratio	3.7612
Statistic for the Odds Ratio	1.3445
p-value	0.1788
Odds Ratio [MH]	2.1264
-95% CI for the Odds Ratio [MH]	1.2443
+95% CI for the Odds Ratio [MH]	3.6336
Degrees of freedom	1
Statistic for the Odds Ratio [MH]	7.8194
p-value	0.0052
Homogeneity of the Odds Ratio	
Degrees of freedom	1
Statistic	0.5486
p-value	0.4589

Copyright ©2010-2026 U-smile Stats – All rights reserved



OR_{MH}[95%CI]=2.13[1.24, 3.65], Mantel-Haenszel χ^2 test: $p = 0.005169$ (statistically significant, $\alpha = 0.05$)

The odds of the occurrence of leptospirosis antibodies is significantly higher for people living in rural areas.

Before we read off the pooled odds ratio, we check the χ^2 test of homogeneity for OR – the tables are homogeneous ($p = 0.465049$), so pooling the strata into a single common odds ratio is justified. The odds of the occurrence of leptospirosis antibodies is higher for rural residents, both for women (OR[95%CI]=2.57[1.24, 5.34]) and for men (OR[95%CI]=1.71[0.78, 3.76]). The Mantel-Haenszel χ^2 test confirms that the pooled odds ratio determined this way is statistically significant.

FOR THOSE INTERESTED

Formulas:

Weights for each stratum are determined according to the formula:

$$g^{(s)} = \frac{O_{21}^{(s)} \cdot O_{12}^{(s)}}{n^{(s)}},$$

and the **Mantel-Haenszel odds ratio**:

$$OR_{MH} = \frac{R}{S},$$

where:

$$R = \sum_{s=1}^{w} \frac{O_{11}^{(s)} \cdot O_{22}^{(s)}}{n^{(s)}},$$

$$S = \sum_{s=1}^{w} g^{(s)}.$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

The confidence interval for $\log OR_{MH}$ is determined on the basis of the standard error (RBG – Robins-Breslow-Greenland[140][141]) calculated using the formula:

$$SE_{MH} = \sqrt{\frac{T}{2R^2} + \frac{U+Y}{2RS} + \frac{W}{2S^2}},$$

where:

$$T = \sum_{s=1}^{w} T^{(s)}, \qquad T^{(s)} = \frac{O_{11}^{(s)} \cdot O_{22}^{(s)} \cdot \left(O_{11}^{(s)} + O_{22}^{(s)}\right)}{\left(n^{(s)}\right)^2},$$

$$U = \sum_{s=1}^{w} U^{(s)}, \qquad U^{(s)} = \frac{O_{21}^{(s)} \cdot O_{12}^{(s)} \cdot \left(O_{11}^{(s)} + O_{22}^{(s)}\right)}{\left(n^{(s)}\right)^2},$$

$$Y = \sum_{s=1}^{w} Y^{(s)}, \qquad Y^{(s)} = \frac{O_{11}^{(s)} \cdot O_{22}^{(s)} \cdot \left(O_{21}^{(s)} + O_{12}^{(s)}\right)}{\left(n^{(s)}\right)^2},$$

$$W = \sum_{s=1}^{w} W^{(s)}, \qquad W^{(s)} = \frac{O_{21}^{(s)} \cdot O_{12}^{(s)} \cdot \left(O_{21}^{(s)} + O_{12}^{(s)}\right)}{\left(n^{(s)}\right)^2}.$$

The Mantel-Haenszel χ^2 test for OR_{MH}

The Mantel-Haenszel χ^2 test is used to verify the hypothesis of the significance of the calculated odds ratio (OR_{MH}), and should be calculated with large sample sizes, i.e. when both conditions of the so-called "rule of 5"are met:

- $\min(O_{11}^{(s)} + O_{12}^{(s)}, O_{11}^{(s)} + O_{21}^{(s)}) - \sum_{s=1}^{w} E_{11}^{(s)} \geq 5$ for all strata $s = 1, 2, ..., w$,
- $\max(0, O_{11}^{(s)} - O_{22}^{(s)}) \geq 5$ for all strata $s = 1, 2, ..., w$.

When zero values are present in the table, a continuity correction (increasing the frequencies by 0.5) is applied to both the observed and the expected frequencies.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \quad OR_{MH} = 1, \\ \mathcal{H}_1 &: \quad OR_{MH} \neq 1. \end{aligned}$$

Formulas:

The test statistic is defined by:

$$\chi^2_{MH} = \frac{\left(\sum_{s=1}^{w} O_{11}^{(s)} - \sum_{s=1}^{w} E_{11}^{(s)}\right)^2}{V},$$

where:

$E_{11}^{(s)} = \dfrac{\left(O_{11}^{(s)} + O_{21}^{(s)}\right)\left(O_{11}^{(s)} + O_{12}^{(s)}\right)}{n^{(s)}}$ are the expected frequencies in the first cell of the contingency table, for the individual strata $s = 1, 2, ..., w$,

$$V = \sum_{s=1}^{w} V^{(s)},$$

$$V^{(s)} = \frac{\left(O_{11}^{(s)} + O_{12}^{(s)}\right)\left(O_{21}^{(s)} + O_{22}^{(s)}\right)\left(O_{11}^{(s)} + O_{21}^{(s)}\right)\left(O_{12}^{(s)} + O_{22}^{(s)}\right)}{\left(n^{(s)}\right)^2\left(n^{(s)} - 1\right)}.$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with one degree of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The χ^2 test of homogeneity for OR
The χ^2 test of homogeneity for OR is used to verify the hypothesis that the variable forming the strata is an effect modifier, i.e. that it affects the calculated odds ratio in such a way that it is significantly different across the individual strata.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad OR_{MH} = OR^{(s)}, \text{ for all strata } s = 1, 2, ..., w, \\ \mathcal{H}_1: &\quad OR_{MH} \neq OR^{(s)}, \text{ for at least one stratum.} \end{aligned}$$

Formulas:
The test statistic (Breslow-Day (1980)[26], Tarone (1985)[27][163]) is defined by:

$$\chi^2 = \sum_{s=1}^{w} \frac{\left(O_{11}^{(s)} - E^{(s)}\right)^2}{Var^{(s)}} - \frac{\left(\sum_{s=1}^{w} O_{11}^{(s)} - \sum_{s=1}^{w} E^{(s)}\right)^2}{\sum_{s=1}^{w} Var^{(s)}}$$

where $E^{(s)}$ is the solution of the quadratic equation:

$$\frac{E^{(s)}\left(O_{22}^{(s)} - O_{11}^{(s)} + E^{(s)}\right)}{\left(O_{11}^{(s)} + O_{21}^{(s)} - E^{(s)}\right)\left(O_{11}^{(s)} + O_{12}^{(s)} - E^{(s)}\right)} = OR_{MH},$$

$$Var^{(s)} = \left(\frac{1}{E^{(s)}} + \frac{1}{O_{22}^{(s)} - O_{11}^{(s)} + E^{(s)}} + \frac{1}{O_{11}^{(s)} + O_{21}^{(s)} - E^{(s)}} + \frac{1}{O_{11}^{(s)} + O_{12}^{(s)} - E^{(s)}}\right)^{-1}.$$

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with $df = w - 1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The Mantel-Haenszel relative risk

WHEN TO USE:

If all the tables (formed by the individual strata) are homogeneous, a pooled relative risk, together with confidence intervals, can be determined on the basis of these tables. This relative risk is a weighted mean of the values of relative risk determined for the individual strata, taking into account the contribution (weight) each stratum makes to the pooled relative risk.

As with the odds ratio, two significance tests accompany the relative risk determined this way. The **Mantel-Haenszel χ^2 test for RR_{MH}** checks whether the pooled relative risk differs significantly from 1. The **χ^2 test of homogeneity for RR**, in turn, checks whether the relative risk is similar across all the strata, or whether the variable forming the strata significantly modifies its value (in which case pooling the strata into a single common relative risk would not be justified).

Copyright ©2010-2026 U-smile Stats – All rights reserved

DEFINITIONS OF MEASURES

The Mantel-Haenszel relative risk – a weighted mean of the values of relative risk determined for the individual strata, taking into account the size of each stratum.

PERFORMING AND INTERPRETING THE ANALYSIS

The performance and interpretation are analogous to the Mantel-Haenszel odds ratio – see Example 21.1 above: before reading off the pooled relative risk, we check the χ^2 test of homogeneity for RR, and we confirm its statistical significance with the Mantel-Haenszel χ^2 test for RR_{MH}.

FOR THOSE INTERESTED

Formulas:

Weights for each stratum are determined according to the formula:

$$g^{(s)} = \frac{O_{21}^{(s)}\left(O_{11}^{(s)} + O_{12}^{(s)}\right)}{n^{(s)}},$$

and the **Mantel-Haenszel relative risk**:

$$RR_{MH} = \frac{R}{S},$$

where:

$$R = \sum_{s=1}^{w} \frac{O_{11}^{(s)}\left(O_{21}^{(s)} + O_{22}^{(s)}\right)}{n^{(s)}},$$

$$S = \sum_{s=1}^{w} g^{(s)}.$$

The confidence interval for log RR_{MH} is determined on the basis of the standard error calculated using the formula:

$$SE_{MH} = \sqrt{\frac{V}{RS}},$$

where:

$$V = \sum_{s=1}^{w} V^{(s)},$$

$$V^{(s)} = \frac{\left(O_{11}^{(s)} + O_{12}^{(s)}\right)\left(O_{21}^{(s)} + O_{22}^{(s)}\right)\left(O_{11}^{(s)} + O_{21}^{(s)}\right) - \left(O_{11}^{(s)} * O_{21}^{(s)} * n^{(s)}\right)}{\left(n^{(s)}\right)^2}.$$

The Mantel-Haenszel χ^2 test for RR_{MH}

The Mantel-Haenszel χ^2 test is used to verify the hypothesis of the significance of the calculated relative risk (RR_{MH}), and should be calculated with large sample sizes in the contingency table.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad RR_{MH} = 1, \\ \mathcal{H}_1: &\quad RR_{MH} \neq 1. \end{aligned}$$

Formulas:

The test statistic is defined by:

$$\chi^2_{MH} = \frac{\left(\sum_{s=1}^{w} O_{11}^{(s)} - \sum_{s=1}^{w} E_{11}^{(s)}\right)^2}{V},$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where $E_{11}^{(s)} = \frac{\left(O_{11}^{(s)}+O_{21}^{(s)}\right)\left(O_{11}^{(s)}+O_{12}^{(s)}\right)}{n^{(s)}}$ are the expected frequencies in the first cell of the contingency table, for the individual strata $s = 1, 2, ..., w$.

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with one degree of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The χ^2 test of homogeneity for RR

The χ^2 test of homogeneity for RR is used to verify the hypothesis that the variable forming the strata is an effect modifier, i.e. that it affects the calculated relative risk in such a way that it is significantly different across the individual strata.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad RR_{MH} = RR^{(s)}, \text{ for all strata } s = 1, 2, ..., w, \\ \mathcal{H}_1: &\quad RR_{MH} \neq RR^{(s)}, \text{ for at least one stratum.} \end{aligned}$$

Formulas:

The test statistic, based on the weighted least-squares method, is defined by:

$$\chi^2 = \sum_{s=1}^{w} v^{(s)} \left(\ln(RR^{(s)}) - \ln(RR_{MH})\right)^2$$

where:

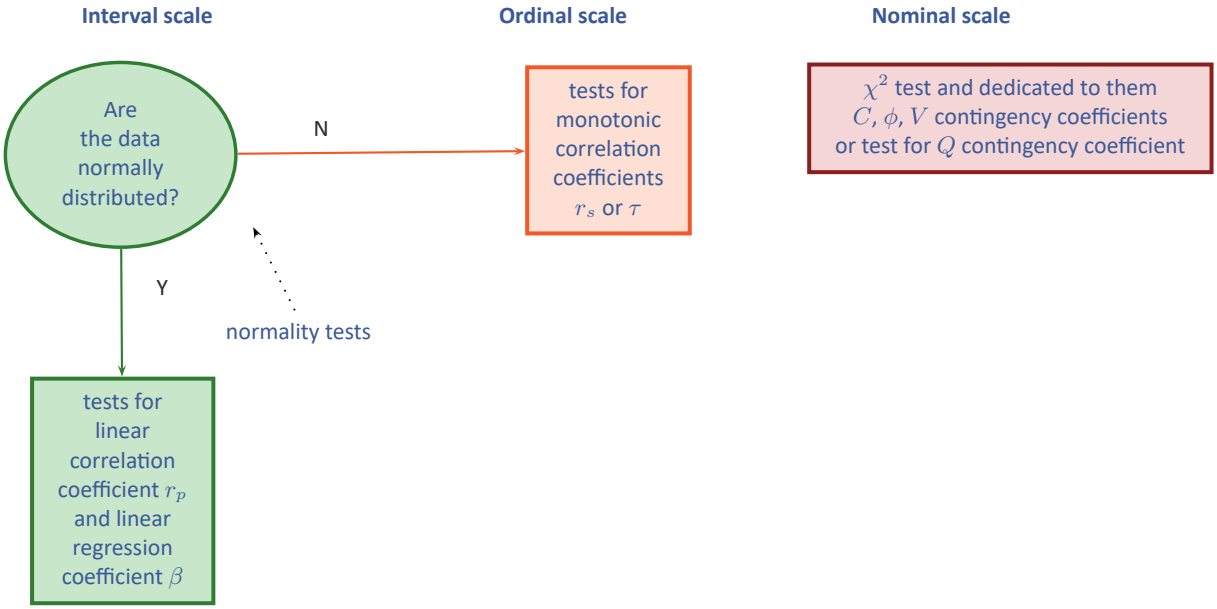
$$v^{(s)} = \left(\frac{O_{12}^{(s)}}{O_{11}^{(s)}\left(O_{11}^{(s)}+O_{12}^{(s)}\right)} + \frac{O_{22}^{(s)}}{O_{21}^{(s)}\left(O_{21}^{(s)}+O_{22}^{(s)}\right)} \right)^{-1}.$$

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with $df = w - 1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

22 CORRELATION



Correlation coefficients are one of the measures of descriptive statistics that represent the degree of correlation (dependence) between 2 or more features (variables). The choice of a particular coefficient depends mainly on the scale on which the measurements were made. Determining it is one of the first steps in correlation analysis. The statistical significance of the obtained coefficients can then be analysed using statistical tests.

Note! A dependence between variables does not always reflect a cause-and-effect relationship.

Copyright ©2010-2026 U-smile Stats – All rights reserved

22.1 PARAMETRIC TESTS

The linear correlation coefficients

WHEN TO USE:

The Pearson product-moment correlation coefficient r_p *(Pearson (1896, 1900))* is used to examine the strength of the linear relationship between two features.

ASSUMPTIONS

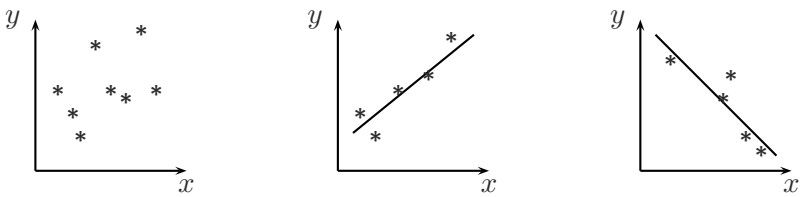
- measurement on an interval scale,
- no outliers,
- normality of distribution of the analysed features or normality of the model residuals.

DEFINITIONS OF MEASURES

The Pearson product-moment correlation coefficient r_p – describes the strength and direction of the linear relationship between two interval features. The value $r_p \in \langle -1; 1 \rangle$ is interpreted as follows:

- $r_p \approx 1$ means a strong positive linear relationship, i.e. the measurement points lie close to a straight line, and an increase in the independent variable corresponds to an increase in the dependent variable;
- $r_p \approx -1$ means a strong negative linear relationship, i.e. the measurement points lie close to a straight line, but an increase in the independent variable corresponds to a decrease in the dependent variable;
- when the coefficient is equal or very close to zero, there is no linear relationship between the analysed features (a non-linear relationship may still exist).

Wykres 22.1. Graphical interpretation of the r_p coefficient.



$r_p \approx 0$ $r_p \approx 1$ $r_p \approx -1$

When one of the two analysed features is constant (regardless of changes in the other feature), the two features are not related, and the r_p coefficient cannot be calculated.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! It is an error to calculate the correlation coefficient when there are outliers in the sample, which can completely distort the value and sign of the Pearson correlation coefficient, when the sample is clearly heterogeneous, or when the analysed relationship clearly takes a shape other than linear.

The coefficient of determination r_p^2 – expresses the percentage of variability in the dependent variable explained by variability in the independent variable.

The resulting correlation model represents a linear relationship of the form:

$$Y = \beta X + \alpha,$$

where the **coefficients** β **and** α **of the linear regression equation** are the slope and the intercept of the regression line.

FOR THOSE INTERESTED

Formulas:

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

where:
x_i, y_i – successive values of feature X and Y,
$\overline{x}, \overline{y}$ – means of the values of feature X and feature Y,
n – sample size.

R_p denotes the Pearson correlation coefficient in the population, while r_p denotes it in the sample.

The coefficients of the linear regression equation are calculated using the formulas:

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \alpha = \overline{y} - \beta\overline{x}.$$

The Pearson correlation coefficient significance

WHEN TO USE THIS TEST:

The test of significance for the Pearson product-moment correlation coefficient lets you check whether a linear relationship exists between the analysed features of a population. It is based on the Pearson product-moment correlation coefficient r_p calculated for the sample – the closer the value of r_p is to 0, the weaker the relationship joining the analysed features.

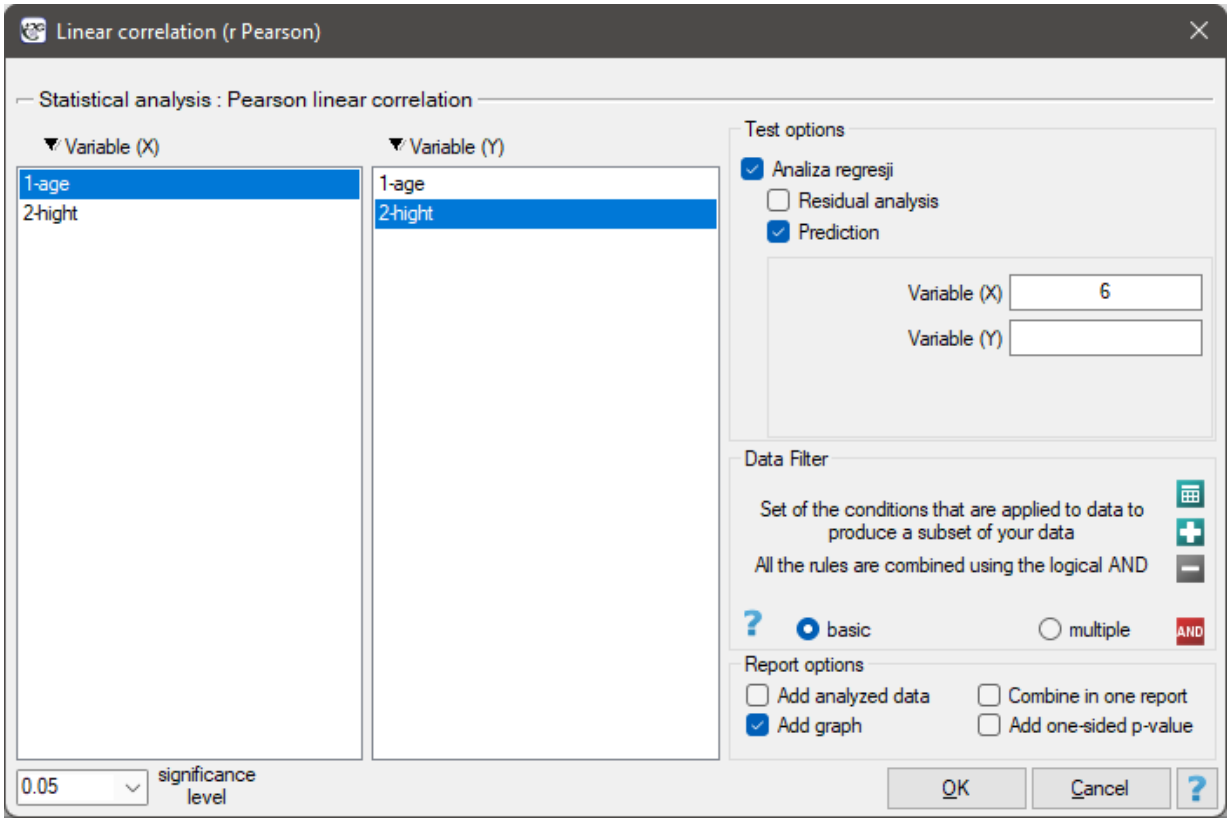
ASSUMPTIONS

- measurement on an interval scale,
- normality of distribution of the analysed features in the population, or normality of the model residuals.

Copyright ©2010-2026 U-smile Stats – All rights reserved

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → linear correlation (r-Pearson) (or via the Wizard).



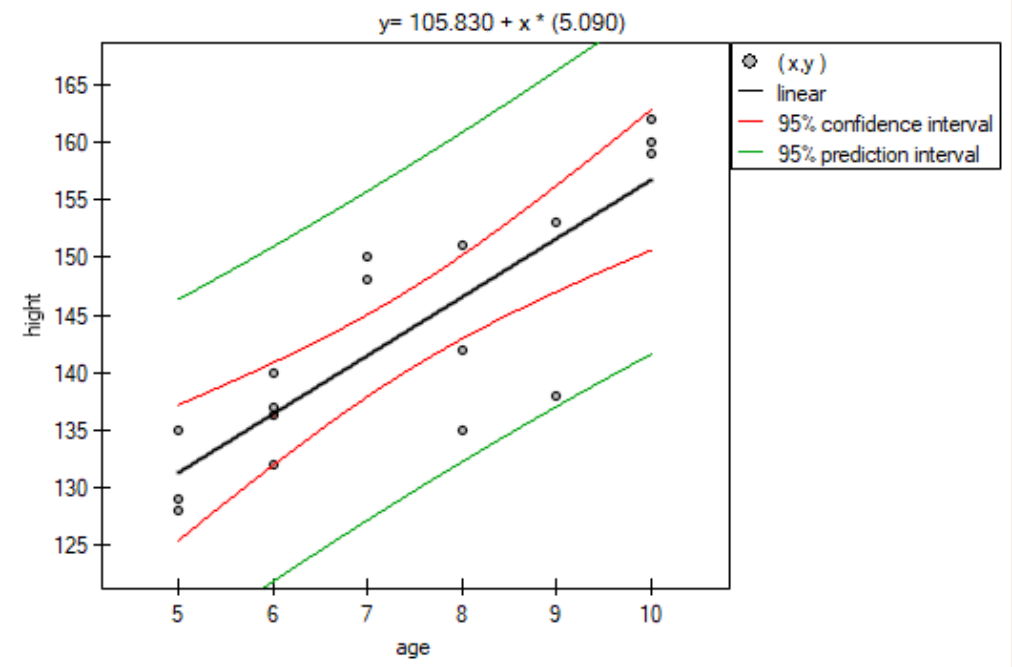
PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 22.1. (age-height.pqs file)
Among the students of a certain ballet school, the relationship between age and height was studied. A sample of sixteen children was drawn, and the following measurements of these features were recorded:
(age, height): (5, 128) (5, 129) (5, 135) (6, 132) (6, 137) (6, 140) (7, 148) (7, 150) (8, 135) (8, 142) (8, 151) (9, 138) (9, 153) (10, 159) (10, 160) (10, 162).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Pearson linear correlation	
Analysed variables	age
	hight
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of pairs	16
Name	age
Mean	7.4375
Standard deviation	1.8246
Name	hight
Mean	143.6875
Standard deviation	11.1876
Standard deviation of the residuals	6.4564
Linear correlation (r Pearson)	
r	0.8302
Std. err. of r	0.149
-95% CI for r coefficient	0.5683
+95% CI for r coefficient	0.9393
t-statistic for r	5.5712
Degrees of freedom	14
Two sided p-value	0.0001
Linear regression	
r2	0.6892
a - slope	5.0901
Std. err. of a	0.9136
-95% CI for a coefficient	3.1305
+95% CI for a coefficient	7.0497
t-test statistic for a	5.5712
Degrees of freedom	14
Two sided p-value	0.0001
b - Y intercept	105.8298
Std. err. of b	6.9843
-95% CI for b coefficient	90.8499
+95% CI for b coefficient	120.8097
t-test statistic for b	15.1525
Degrees of freedom	14
Two sided p-value	<0.0001
prediction of Y value for X= 6	136.3705
-95% CI for the prediction of Y	121.8213
+95% CI for the prediction of Y	150.9196

Copyright ©2010-2026 U-smile Stats – All rights reserved



$p = 0.000069$ **(statistically significant,** $\alpha = 0.05$**)**

There is a linear relationship between age and height for the population of children at the school studied.

This relationship is directly proportional, meaning that as the children's age increases, so does their height. The Pearson product-moment correlation coefficient, and thus the strength of the linear relationship between age and height, is $r_p = 0.8302$. The coefficient of determination $r_p^2 = 0.6892$ means that about 69% of the variability in height is explained by the variability in age.

From the regression equation:

$$height = 5.09 \cdot age + 105.83$$

we can calculate a predicted value for a child aged, for example, 6 years. The predicted height of such a child is 136.37 cm.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad R_p = 0, \\ \mathcal{H}_1: &\quad R_p \neq 0. \end{aligned}$$

Formulas:

The test statistic is defined by:

$$t = \frac{r_p}{SE},$$

where $SE = \sqrt{\dfrac{1 - r_p^2}{n - 2}}$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The value of the test statistic cannot be calculated when $r_p = 1$ or $r_p = -1$, or when $n < 3$. The test statistic has the t-Student distribution with $n - 2$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The slope coefficient significance

WHEN TO USE THIS TEST:

The t-test for the significance of the coefficients of the linear regression equation lets you check whether a linear relationship exists between the analysed features of a population, based on the slope coefficient β calculated for the sample – the closer the value of β is to 0, the weaker a relationship the fitted line represents. In a regression with a single independent variable, this test is equivalent to the test of significance for the Pearson correlation coefficient – it gives exactly the same p value.

ASSUMPTIONS

- measurement on an interval scale,
- normality of distribution of the analysed features in the population, or normality of the model residuals.

PERFORMING AND INTERPRETING THE ANALYSIS

Because of its equivalence to the test of significance for the Pearson correlation coefficient with a single independent variable, the performance and interpretation of this test coincide with Example 22.1 above.

Analysis of model residuals – explained in the Multiple Linear Regression module.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \beta = 0, \\ \mathcal{H}_1: &\quad \beta \neq 0. \end{aligned}$$

Formulas:
The test statistic is defined by:

$$t = \frac{\beta}{SE}$$

where:
$SE = \frac{s_{yx}}{sd_x\sqrt{n-1}}$,

Copyright ©2010-2026 U-smile Stats – All rights reserved

$s_{yx} = sd_y\sqrt{\frac{n-1}{n-2}(1-r^2)}$,

sd_x, sd_y – standard deviation of the values of feature X and feature Y.

The value of the test statistic cannot be calculated when $r_p = 1$ or $r_p = -1$, or when $n < 3$. The test statistic has the t-Student distribution with $n - 2$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Prediction

WHEN TO USE:

Prediction consists in predicting the value of one variable (most often the dependent variable y_0) on the basis of the value of another variable (most often the independent variable x_0), using the calculated linear regression equation. The accuracy of the predicted value is expressed by the prediction intervals calculated for it.

DEFINITIONS OF MEASURES

Interpolation – consists in predicting the value of a given variable that lies inside the range for which the regression model was built. Interpolation is generally a safe procedure – it only assumes continuity of the function expressing the relationship between the two variables.

Extrapolation – consists in predicting the value of a given variable that lies outside the range for which the regression model was built. Unlike interpolation, extrapolation is often a risky procedure and is only performed a short distance from the range for which the regression model was built.

HOW TO SET UP THE ANALYSIS

In the Pearson linear correlation analysis window, select the Prediction button and indicate the values of the independent variable for which you want to carry out the prediction.

PERFORMING AND INTERPRETING THE ANALYSIS

See Example 22.1 above – the predicted height of a 6-year-old child, calculated from the regression equation, is exactly the result of such a prediction (interpolation, since age 6 falls within the range of the data on which the model was built).

Comparison of correlation coefficients

WHEN TO USE THIS TEST:

The t-test for checking the equality of Pearson product-moment correlation coefficients coming from 2 independent populations lets you check whether two Pearson product-moment correlation coefficients (R_{p_1}, R_{p_2}), determined in two different populations, are equal.

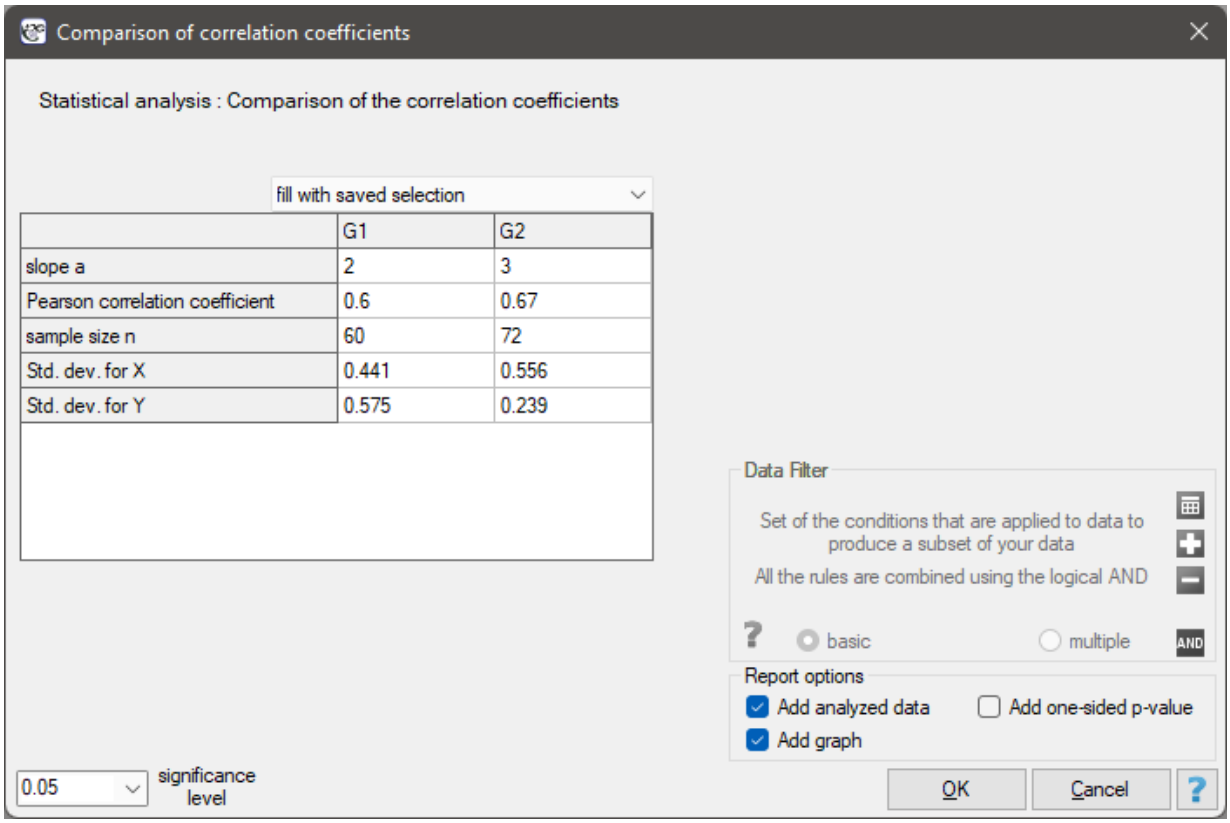
Copyright ©2010-2026 U-smile Stats – All rights reserved

ASSUMPTIONS

- the coefficients r_{p_1} and r_{p_2} come from 2 samples drawn from independent populations,
- the coefficients r_{p_1} and r_{p_2} examine the relationship between the same features X and Y,
- the sizes of both samples n_1 and n_2 are known.

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → comparison of dependence coefficients.



FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad R_{p_1} = R_{p_2}, \\ \mathcal{H}_1: &\quad R_{p_1} \neq R_{p_2}. \end{aligned}$$

Formulas: the test statistic is defined by:

$$t = \frac{z_{r_{p_1}} - z_{r_{p_2}}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}},$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where:
$z_{r_{p_1}} = \frac{1}{2} \ln\left(\frac{1+r_{p_1}}{1-r_{p_1}}\right)$,
$z_{r_{p_2}} = \frac{1}{2} \ln\left(\frac{1+r_{p_2}}{1-r_{p_2}}\right)$.

The test statistic has the t-Student distribution with $n_1 + n_2 - 4$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Comparison of the slope of regression lines

WHEN TO USE THIS TEST:

The t-test for checking the equality of linear regression coefficients coming from 2 independent populations lets you check whether two linear regression coefficients β_1 and β_2, determined in the two populations studied, are equal.

ASSUMPTIONS

- the coefficients β_1 and β_2 come from 2 samples drawn from independent populations,
- the coefficients β_1 and β_2 examine the relationship between the same features X and Y,
- the sizes of both samples n_1 and n_2 are known,
- the standard deviations of the values of both features in both samples are known: sd_{x_1}, sd_{y_1} and sd_{x_2}, sd_{y_2},
- the Pearson product-moment correlation coefficients of both samples are known: r_{p_1} and r_{p_2}.

HOW TO SET UP THE ANALYSIS

The settings window is shared with the test for comparing correlation coefficients (see the screenshot above) – in the comparison of dependence coefficients window, indicate the slope coefficients of the regression lines instead.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \quad \beta_1 = \beta_2, \\ \mathcal{H}_1 &: \quad \beta_1 \neq \beta_2. \end{aligned}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

Formulas: the test statistic is defined by:

$$t = \frac{\beta_1 - \beta_2}{\sqrt{\frac{s^2_{yx_1}}{sd^2_{x_1}(n_1-1)} + \frac{s^2_{yx_2}}{sd^2_{x_1}(n_2-1)}}},$$

where:

$s_{yx_1} = sd_{y_1}\sqrt{\frac{n_1-1}{n_1-2}(1-r^2_{p_1})},$

$s_{yx_2} = sd_{y_2}\sqrt{\frac{n_2-1}{n_2-2}(1-r^2_{p_2})}.$

The test statistic has the t-Student distribution with $n_1 + n_2 - 4$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

22.2 NON-PARAMETRIC TESTS

The monotonic correlation coefficients

WHEN TO USE:

A monotonic relationship can be described as a monotonic increase or a monotonic decrease. The relationship between 2 features shows a monotonic increase if an increase in one feature is accompanied by an increase in the other, and a monotonic decrease if an increase in one feature is accompanied by a decrease in the other. The **Spearman's rank-order correlation coefficient** r_s and the **Kendall's tau correlation coefficient** $\tilde{\tau}$ *(Kendall (1938)[95])* are used to examine the strength of such a relationship between two features X and Y.

ASSUMPTIONS

- measurement on an ordinal scale or an interval scale.

DEFINITIONS OF MEASURES

The Spearman's rank-order correlation coefficient r_s – the value $r_s \in \langle -1; 1 \rangle$ is interpreted as follows:

- $r_s \approx 1$ means a strong positive (increasing) monotonic relationship, i.e. an increase in the independent variable corresponds to an increase in the dependent variable;
- $r_s \approx -1$ means a strong negative (decreasing) monotonic relationship, i.e. an increase in the independent variable corresponds to a decrease in the dependent variable;
- when the coefficient is equal or very close to zero, there is no monotonic relationship between the analysed features (a non-monotonic relationship, e.g. a sinusoidal one, may still exist).

R_s denotes the Spearman rank-order correlation coefficient in the population, while r_s denotes it in the sample.

The Kendall's tau correlation coefficient $\tilde{\tau}$ – the value $\tilde{\tau} \in \langle -1; 1 \rangle$ is interpreted as follows:

- $\tilde{\tau} \approx 1$ means a strong agreement in the ordering of ranks (an increasing monotonic relationship), i.e. an increase in the independent variable corresponds to an increase in the dependent variable;
- $\tilde{\tau} \approx -1$ means a strong disagreement in the ordering of ranks (a decreasing monotonic relationship), i.e. an increase in the independent variable corresponds to a decrease in the dependent variable;
- when the coefficient is equal or very close to zero, there is no monotonic relationship between the analysed features (a non-monotonic relationship, e.g. a sinusoidal one, may still exist).

τ denotes the Kendall correlation coefficient in the population, while $\tilde{\tau}$ denotes it in the sample.

The Spearman r_s coefficient versus the Kendall $\tilde{\tau}$ coefficient

Copyright ©2010-2026 U-smile Stats – All rights reserved

- for an interval scale with normality of distribution of both features, the r_s value gives results closer to r_p, while the $\tilde{\tau}$ value may differ considerably from r_p,
- the value of $\tilde{\tau}$ is less than or equal to the value of r_p,
- $\tilde{\tau}$ is an unbiased estimator of the population parameter τ, while the value of r_s is not an unbiased estimator of the parameter R_s.

FOR THOSE INTERESTED

Formulas:
The value of the Spearman rank-order correlation coefficient is calculated using the formula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where:
$d_i = R_{x_i} - R_{y_i}$ – the difference of ranks for feature X and feature Y,
n – the number of d_i.

This formula is modified when tied ranks are present:

$$r_s = \frac{\Sigma_X + \Sigma_Y - \sum_{i=1}^{n} d_i^2}{2\sqrt{\Sigma_X \Sigma_Y}},$$

where $\Sigma_X = \frac{n^3-n-T_X}{12}$, $\Sigma_Y = \frac{n^3-n-T_Y}{12}$, $T_X = \sum_{i=1}^{s}(t_{i_{(X)}}^3 - t_{i_{(X)}})$, $T_Y = \sum_{i=1}^{s}(t_{i_{(Y)}}^3 - t_{i_{(Y)}})$, and t is the number of cases included in a tie. The correction for ties should be used when ties are present – otherwise the formula reduces to the form described by the previous equation.

The value of the Kendall's tau correlation coefficient is calculated using the formula:

$$\tilde{\tau} = \frac{2(n_C - n_D)}{\sqrt{n(n-1)-T_X}\sqrt{n(n-1)-T_Y}},$$

where:
n_C – the number of pairs of observations for which the values of the ranks of feature X and feature Y change in the same direction (the number of concordant pairs),
n_D – the number of pairs of observations for which the values of the ranks of feature X change in a different direction than those of feature Y (the number of discordant pairs),
$T_X = \sum_{i=1}^{s}(t_{i_{(X)}}^2 - t_{i_{(X)}})$, $T_Y = \sum_{i=1}^{s}(t_{i_{(Y)}}^2 - t_{i_{(Y)}})$, and t is the number of cases included in a tie.

The formula for $\tilde{\tau}$ includes a correction for tied ranks. This correction should be used when ties are present (when there are no ties, the correction is not calculated, since then $T_X = 0$ and $T_Y = 0$).

The t-test for the significance of the Spearman rank-order correlation coefficient

WHEN TO USE THIS TEST:

The test of significance for the Spearman rank-order correlation coefficient lets you check whether a monotonic relationship exists between the analysed features of a population. It is based on the Spear-

Copyright ©2010-2026 U-smile Stats – All rights reserved

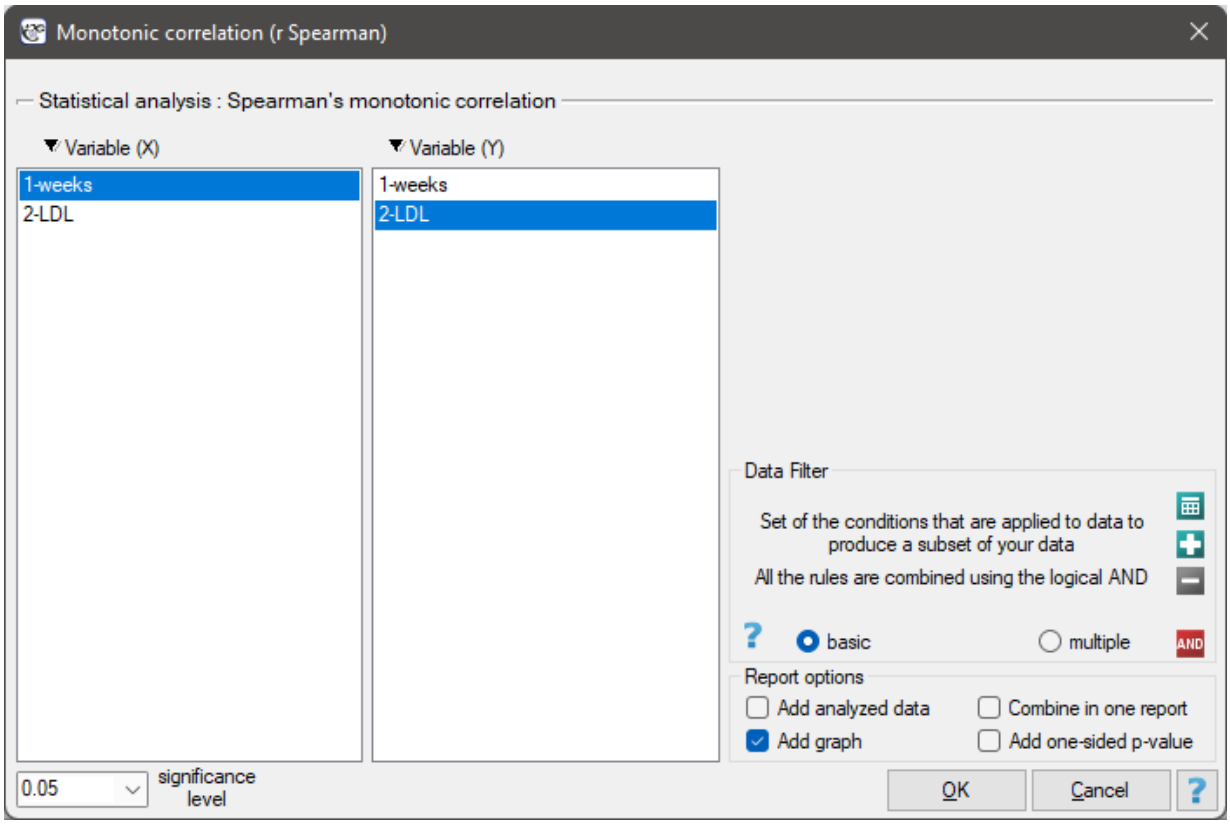
man rank-order correlation coefficient r_s calculated for the sample – the closer the value of r_s is to 0, the weaker the monotonic relationship joining the analysed features.

ASSUMPTIONS

– measurement on an ordinal scale or an interval scale.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → monotonic correlation (r-Spearman) (or via the Wizard).



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 22.2. (LDL weeks.pqs file)
The effectiveness of a new therapy designed to lower LDL cholesterol levels was studied. 88 people at different stages of the treatment were examined. We will check whether LDL cholesterol levels fall and stabilize with the duration of treatment (time in weeks).

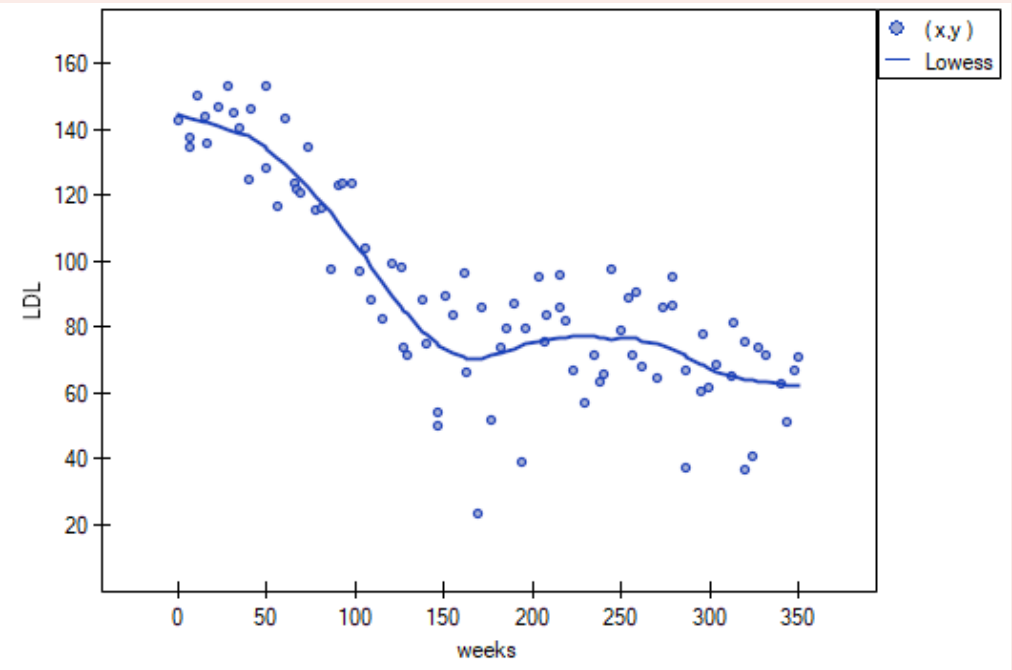
Copyright ©2010-2026 U-smile Stats – All rights reserved

Spearman's monotonic correlation	
Analysed variables	weeks
	LDL
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of pairs	88
r	-0.782
Std. err. of r	0.0672
-95% CI for r coefficient	-0.8536
+95% CI for r coefficient	-0.6813
t-statistic for r	-11.6347
Degrees of freedom	86
Two sided p-value	<0.0001

$p < 0.0001$ **(statistically significant,** $\alpha = 0.05$**)**

There is a statistically significant monotonic relationship between treatment time and LDL level.

This relationship is initially decreasing, and after 150 weeks it begins to stabilize. The Spearman monotonic correlation coefficient, and thus the strength of the monotonic relationship, is fairly high at $r_s = -0.7806$. The graph was drawn by fitting a curve using local LOWESS linear smoothing techniques.



Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad R_s = 0, \\ \mathcal{H}_1: &\quad R_s \neq 0. \end{aligned}$$

Formulas:
The test statistic is defined by:

$$t = \frac{r_s}{SE},$$

where $SE = \sqrt{\frac{1 - r_s^2}{n - 2}}$.

The value of the test statistic cannot be calculated when $r_s = 1$ or $r_s = -1$, or when $n < 3$. The test statistic has the t-Student distribution with $n - 2$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The test for the significance of the Kendall's tau correlation coefficient

WHEN TO USE THIS TEST:

The test of significance for the Kendall's tau correlation coefficient lets you check whether a monotonic relationship exists between the analysed features of a population. It is based on the Kendall correlation coefficient calculated for the sample – the closer the value of $\tilde{\tau}$ is to 0, the weaker the monotonic relationship joining the analysed features.

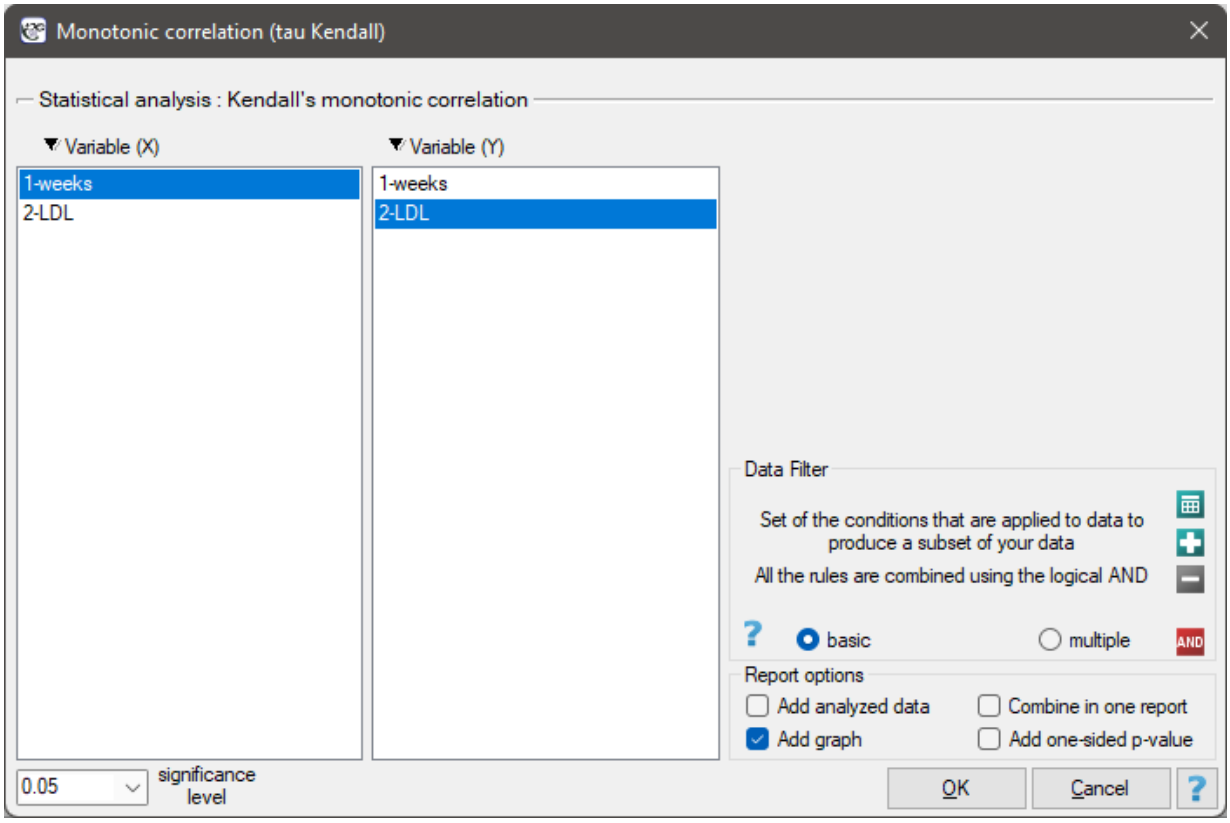
ASSUMPTIONS

- measurement on an ordinal scale or an interval scale.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → monotonic correlation (tau-Kendall) (or via the Wizard).

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

Continuation of Example 22.2 (LDL weeks.pqs file): we check the same relationship between treatment time and LDL level, this time using the Kendall tau coefficient.

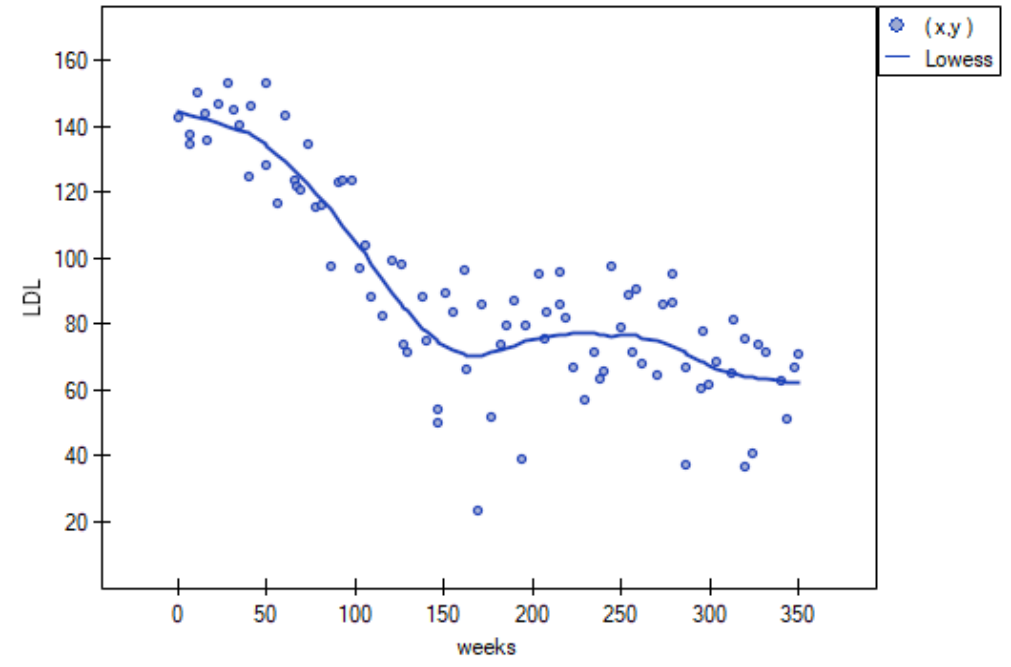
Kendall's monotonic correlation	
Analysed variables	weeks
	LDL
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of pairs	88
tau	-0.5959
Z statistic for tau	-8.2216
Two sided p-value	<0.0001

$p < 0.0001$ **(statistically significant,** $\alpha = 0.05$**)**

There is a statistically significant monotonic relationship between treatment time and LDL level.

This relationship is initially decreasing, and after 150 weeks it begins to stabilize. The Kendall monotonic correlation coefficient, and thus the strength of the monotonic relationship, is fairly high at $\tilde{\tau} = -0.5975$. The graph was drawn by fitting a curve using local LOWESS linear smoothing techniques.

Copyright ©2010-2026 U-smile Stats – All rights reserved



FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \tau = 0, \\ \mathcal{H}_1: &\quad \tau \neq 0. \end{aligned}$$

Formulas:
The test statistic is defined by:

$$Z = \frac{3\tilde{\tau}\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}.$$

The test statistic asymptotically (for large sample sizes) has the normal distribution.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Contingency table coefficients and their statistical significance

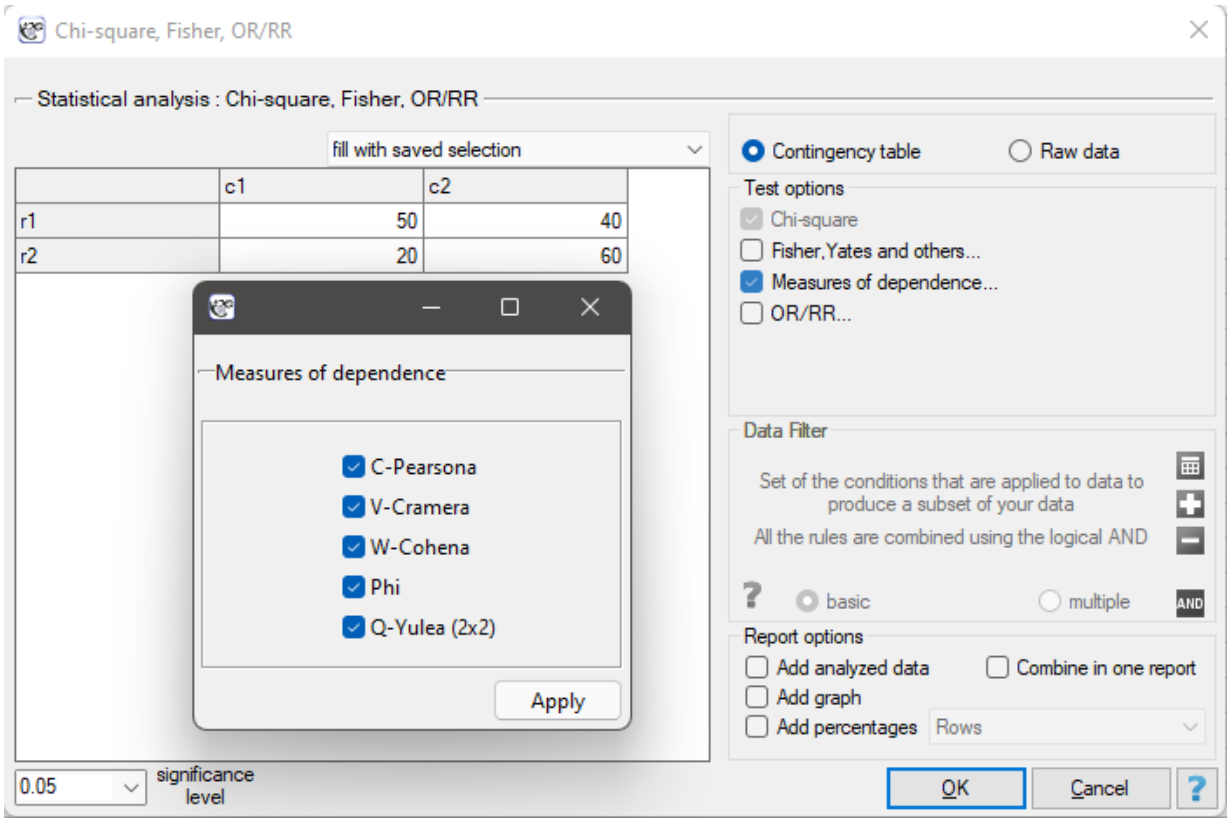
WHEN TO USE:

Contingency coefficients let you assess the strength of the relationship between two features recorded on a nominal scale. They are calculated for raw data or for data gathered in a contingency table.

HOW TO SET UP THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

Statistics → Nonparametric tests → Chi-square, Fisher, OR/RR → Measures of dependence... (or via the Wizard).



DEFINITIONS OF MEASURES

The Yule's Q contingency coefficient *(Yule (1900)[186])* – a measure of dependence calculated for 2×2 contingency tables. The value of Q lies in the range $\langle -1; 1 \rangle$: the closer to 0, the weaker the relationship, and the closer to -1 or $+1$, the stronger the relationship (because of possible misinterpretation of a negative value, PQStat also presents the result as an absolute value). The coefficient is not very robust to small observed frequencies – when one of them is 0, it may incorrectly indicate a complete dependence between the features.

The ϕ contingency coefficient – a measure of dependence recommended especially for 2×2 tables, though it can be calculated for any table. The value of ϕ lies in the range $\langle 0; 1 \rangle$: the closer to 0, the weaker the relationship, and the closer to 1, the stronger the relationship.

The Cramer's V contingency coefficient *(Cramer (1946)[50])* – an extension of the ϕ coefficient to $r \times c$ tables. The value of V lies in the range $\langle 0; 1 \rangle$: the closer to 0, the weaker the relationship, and the closer to 1, the stronger the relationship. The value of V depends on the size of the table, so it should not be used to compare tables of different sizes.

The Cohen's W contingency coefficient *(Cohen (1988)[47])* – a modification of the Cramer's V coefficient, calculable for $r \times c$ tables. The value of W lies in the range $\langle 0; \max W \rangle$ (for tables in which at least one variable has only two categories, this range is $\langle 0; 1 \rangle$). Like V, it depends on the size of the table.

The Pearson's C contingency coefficient – a measure of dependence for $r \times c$ tables. The value of C lies in the range $\langle 0; 1 \rangle$, and its upper limit depends on the size of the table, which is why a version adjusted

Copyright ©2010-2026 U-smile Stats – All rights reserved

for table size, C_{adj}, is also calculated, allowing comparisons between tables of different sizes.

Note! The ϕ, Cramer's V, Cohen's W, and Pearson's C coefficients are considered statistically significant if the p value determined on the basis of the χ^2 test statistic and the χ^2 distribution (calculated for the given table) is equal to or less than the significance level α. The significance of the Yule's Q coefficient, on the other hand, is checked with a separate Z test (see *For those interested*).

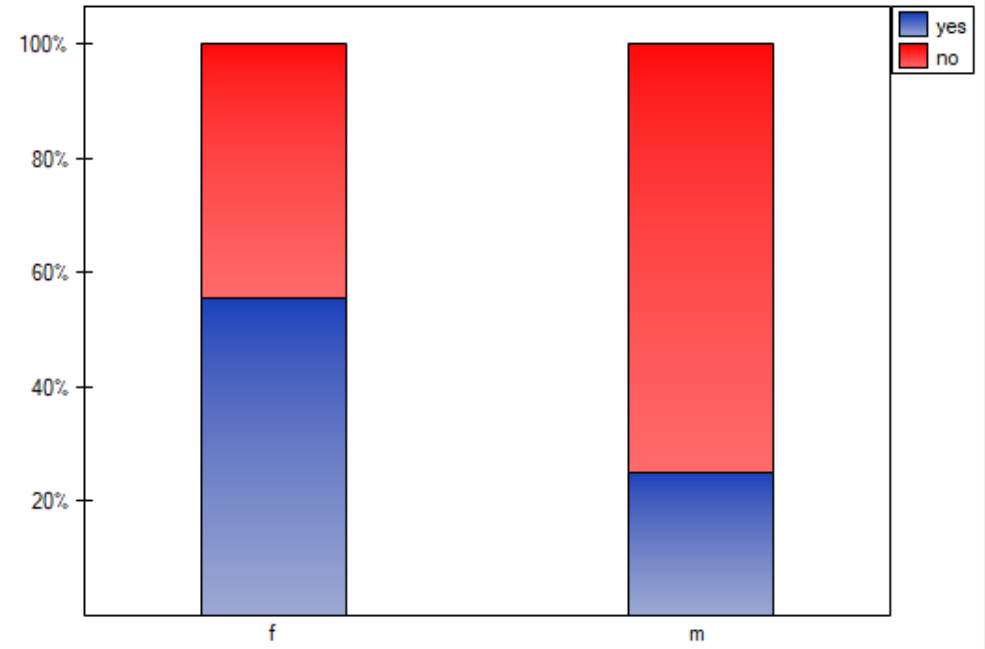
PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 22.3. (sex-exam.pqs file)
Consider a sample of 170 people ($n = 170$) for whom we examine 2 features (X=sex, Y=passing the exam). Each of these features occurs in two categories (X_1=f, X_2=m, Y_1=yes, Y_2=no). Based on this sample, we would like to find out whether there is a relationship between sex and passing the exam in the population studied. The data distribution is shown in the contingency table:

Observed frequencies O_{ij}		passing the exam		
		yes	no	total
sex	f	50	40	90
	m	20	60	80
	total	70	100	170

Chi-square, Fisher, OR/RR	
Analysed variables	Contingency table
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	170
Cochran condition	fulfilled
Pearson's chi-square statistic	16.3254
Degrees of freedom	1
p-value	0.0001
Measures of dependence	
C-Pearson	0.296
C-Pearson (max)	0.7071
C-Pearson (adjusted)	0.4186
V-Cramer	0.3099
W-Cohen	0.3099
W-Cohen (max)	1
Phi	0.3099
Q-Yule	0.5789
Z statistic Yulea	5.212
p-value (asymptotic) Yulea	<0.0001

Copyright ©2010-2026 U-smile Stats – All rights reserved



$\chi^2 = 16.33$, $p = 0.00005$ **(statistically significant,** $\alpha = 0.05$**)**

There is a statistical relationship between sex and passing the exam in the population studied.

The value of the coefficients based on the χ^2 test, and thus the strength of the relationship between the analysed features, is as follows:

The Pearson's C_{adj} contingency coefficient = 0.42.
The Cramer's V = ϕ = the Cohen's W = 0.31.

The Yule's Q contingency coefficient = 0.58, and the p value of the Z test performed, like the significance level of the χ^2 test, indicates the statistical significance of the relationship studied.

FOR THOSE INTERESTED

The Yule's Q contingency coefficient

$$Q = \frac{O_{11}O_{22} - O_{12}O_{21}}{O_{11}O_{22} + O_{12}O_{21}},$$

where $O_{11}, O_{12}, O_{21}, O_{22}$ – observed frequencies in the contingency table.

The statistical significance of the calculated Yule's Q coefficient is determined using a Z test.
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad Q = 0, \\ \mathcal{H}_1: &\quad Q \neq 0. \end{aligned}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

The test statistic is defined by:

$$Z = \frac{Q}{\sqrt{\frac{1}{4}(1-Q^2)^2(\frac{1}{O_{11}}+\frac{1}{O_{12}}+\frac{1}{O_{21}}+\frac{1}{O_{22}})}}.$$

The test statistic asymptotically (for large sample sizes n) has the normal distribution.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The ϕ contingency coefficient

$$\phi = \sqrt{\frac{\chi^2}{n}},$$

where χ^2 – the value of the χ^2 test statistic, n – the total frequency in the contingency table.

The Cramer's V contingency coefficient

$$V = \sqrt{\frac{\chi^2}{n(w'-1)}},$$

where χ^2 – the value of the χ^2 test statistic, n – the total frequency in the table, and w' – the smaller of the two values r and c.

The Cohen's W contingency coefficient

$$W = \sqrt{\frac{\chi^2}{n(w'-1)}}\sqrt{w'-1},$$

where χ^2 – the value of the χ^2 test statistic, n – the total frequency in the table, and w' – the smaller of the two values r and c.

The Pearson's C contingency coefficient

$$C = \sqrt{\frac{\chi^2}{\chi^2+n}},$$

where χ^2 – the value of the χ^2 test statistic, n – the total frequency in the table. The upper limit that the C coefficient can reach for a given table size is given by the formula:

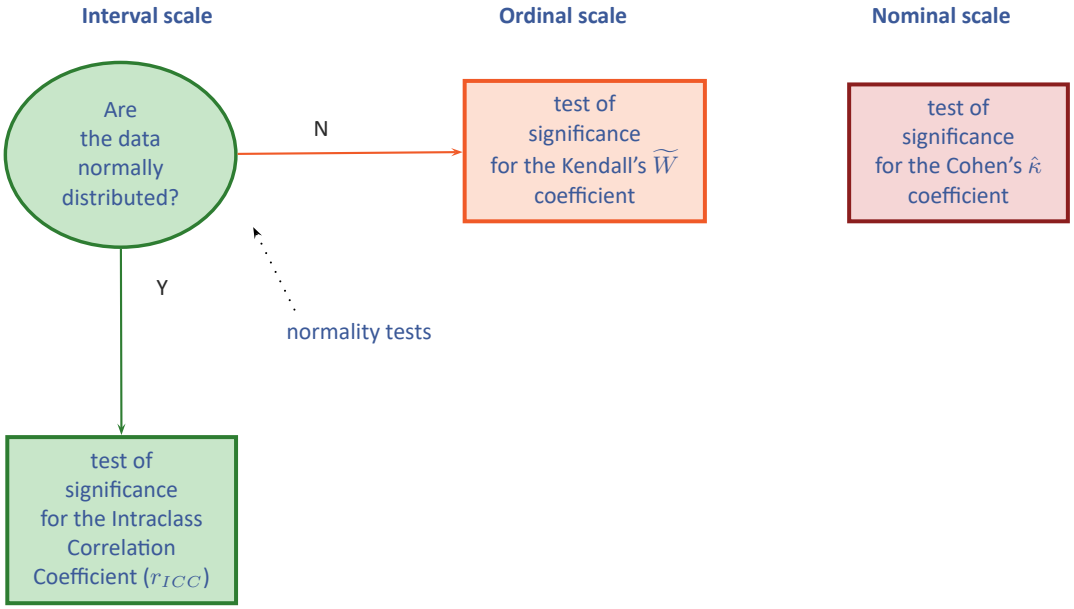
$$C_{max} = \sqrt{\frac{w'-1}{w'}},$$

where w' – the smaller of the two values r and c. The version adjusted for table size:

$$C_{adj} = \frac{C}{C_{max}}.$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

23 AGREEMENT ANALYSIS



Copyright ©2010-2026 U-smile Stats – All rights reserved

23.1 PARAMETRIC TESTS

The intraclass correlation coefficient and a test to examine its significance

WHEN TO USE:

The intraclass correlation coefficient *(ICC)* is used when the measurement of the variable studied is made by several "raters" ($k \geq 2$). It measures the strength of **interrater reliability**, i.e. the degree to which their ratings agree. Since it can be determined in several different situations, there are several variants of it, depending on the model and the type of agreement.

ASSUMPTIONS

- measurement on an interval scale,
- the normal distribution for all variables that are the differences of measurement pairs (or normality of the analysed variable for each measurement),
- for model 1 – an independent model, for model 2 / 3 – a dependent model.

DEFINITIONS OF MEASURES

Depending on the variability present in the data, two main research models and two types of agreement are distinguished.

Model 1 – for each of the n randomly selected objects rated, a set of k raters is randomly selected from the population of raters, and a different set of raters may be drawn for each object. The ICC coefficient is then determined using the ANOVA for independent groups random model. ICC(1,1) answers the question of the reliability of a single rater's ratings, and ICC(1,k) the reliability of a result that is the average of k raters' ratings.

Model 2 – a set of k raters is randomly selected from the population of raters, and each of them rates all n random objects. The ICC coefficient is then determined using the ANOVA for dependent groups random model. Depending on the type of agreement sought, we can estimate:

- **absolute agreement** – raters are in absolute agreement if they give exactly the same ratings, e.g. perfectly agreeing would be ratings given by a pair of raters (2,2), (5,5), (8,8);
- **consistency** – raters may use different ranges of values, but apart from this shift there should be no differences, in order to keep the verdict consistent, e.g. perfectly consistent would be ratings given by a pair of raters (2,5), (5,8), (8,11).

As in model 1, ICC(2,1) answers the question of the reliability of a single rater's ratings, and ICC(2,k) the reliability of a result that is the average of k raters' ratings (the formulas differ depending on the type of agreement chosen).

Note! Sometimes it is necessary to consider **model 3**[155], i.e. a set of k raters is selected and each of them rates all n random objects, but the agreement result applies only to these particular k raters. The

Copyright ©2010-2026 U-smile Stats – All rights reserved

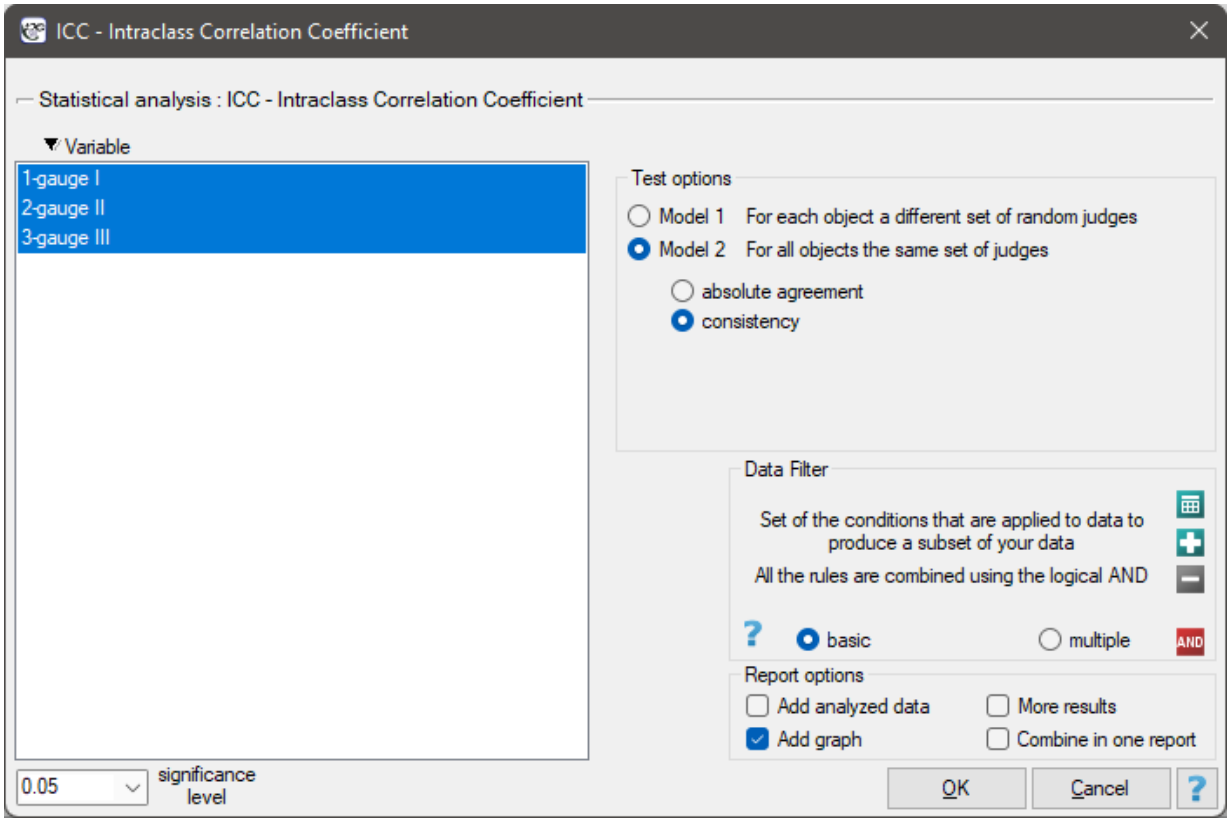
ICC coefficient is then determined in a mixed model (since the randomness concerns only the objects, not the raters). Since we ignore the variability concerning the raters, we examine consistency (rather than absolute agreement), and the coefficients from the second model apply: ICC(2,1) and ICC(2,k), since they are identical to the coefficients ICC(3,1) and ICC(3,k) that would be desired in this case, under the assumption of no interaction between objects and raters.

Interpretation – the value $ICC \in \langle -1; 1 \rangle$ is interpreted as follows:

- $ICC \approx 1$ means strong agreement in the rating of individual objects by the raters, which is reflected in high variance between objects (a significant difference in means among the n objects) and low variance between raters' ratings (a small difference in the mean ratings determined for the k raters);
- $ICC \approx -1$ – a negative intraclass correlation coefficient is treated in the same way as $ICC \approx 0$;
- $ICC \approx 0$ means no agreement in the rating of individual objects by the raters, which is reflected in low variance between objects (a small difference in means among the n objects) and high variance between raters' ratings (a significant difference in the mean ratings determined for the k raters).

HOW TO SET UP THE ANALYSIS

Statistics → Parametric tests → ICC – Intraclass Correlation Coefficient (or via the Wizard).



PERFORMING AND INTERPRETING THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

EXAMPLE 23.1. (sound intensity.pqs file)
In order to effectively protect the hearing of workers at a workplace, it is first necessary to reliably estimate the sound intensity at the various locations where people are present. At one company, it was decided to conduct an experiment before choosing a sound-intensity meter (sonograph). Sound-intensity measurements were made at 42 randomly selected measurement points on the company premises, using 3 randomly drawn analogue sonographs and 3 randomly selected digital sonographs. An excerpt of the measurements collected is shown in the table below.

gauge I	gauge II	gauge III
84.1	84.1	84.7
85.7	85.3	85.3
84.5	83.5	84.3
86.2	86.1	87.1
85.6	84.8	85.5
80.5	80.6	81
81.4	81.7	82.1
85.6	86	87.3
83.2	82.4	80.1
83.5	82.5	85.5
82.4	82.5	81.5

To find out which type of instrument (analogue or digital) better accomplishes the task at hand, the ICC coefficient should be determined in model 2, examining absolute agreement. The type of meter with the higher ICC value will provide more reliable measurements, and will therefore be used at the company site in the future.

The analysis performed for the analogue meters shows significant agreement of the measurements.

Copyright ©2010-2026 U-smile Stats – All rights reserved

ICC - Intraclass Correlation Coefficient	
Analysed variables	gauge I
	gauge II
	gauge III
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Model 2 For all objects the same set of judges	absolute agreement
Between-subjects degrees of freedom (df[BS])	41
Residual degrees of freedom (df[RES])	82
Mean square between-conditions (MS[BC])	105.1517
Mean square between-subjects (MS[BS])	8.9026
Mean square residual (MS[RES])	0.7966
Intraclass correlation coefficient (r[ICC])	0.4516
-95% CI dla ICC(2,1)	0.0308
+95% CI dla ICC(2,1)	0.7304
Average measure r[ICC]	0.7118
-95% CI dla ICC(2,k)	0.087
+95% CI dla ICC(2,k)	0.8904
F statistic	11.1754
p-value	<0.0001

$p < 0.000001$ for the analogue meters (statistically significant, $\alpha = 0.05$)

The reliability of the measurement made with a single analogue meter is $ICC(2,1) = 0.45$, and the reliability of a measurement that is the average of the three analogue meters is $ICC(2,k) = 0.71$.

However, the lower limit of the 95% confidence interval for these coefficients is worryingly low. A similar analysis performed for the digital meters gave better results.

Copyright ©2010-2026 U-smile Stats – All rights reserved

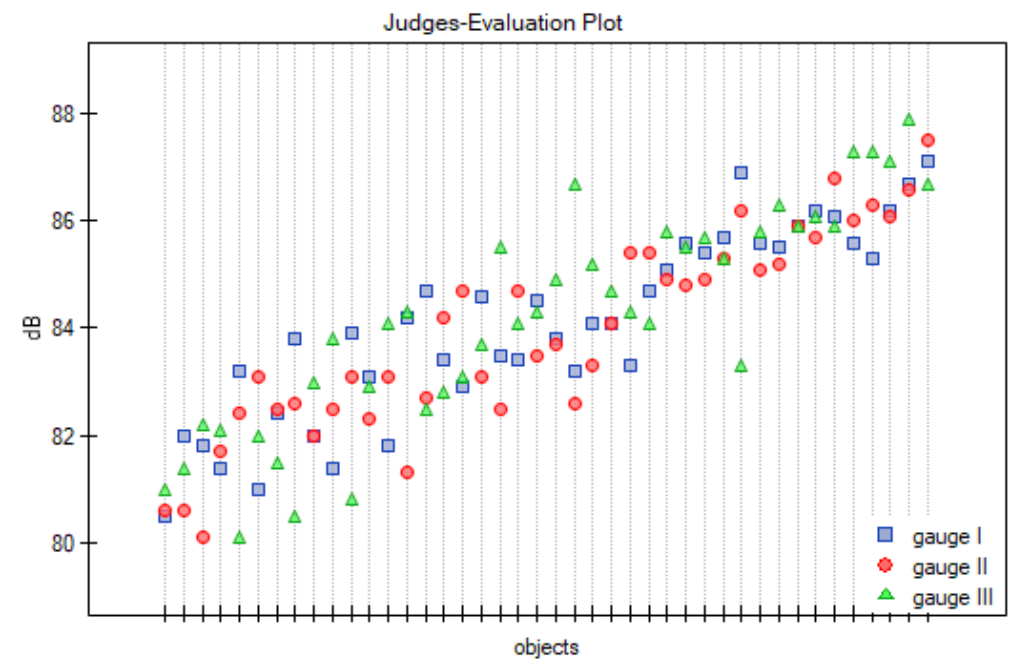
ICC - Intraclass Correlation Coefficient	
Analysed variables	gauge I
	gauge II
	gauge III
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Model 2 For all objects the same set of judges	absolute agreement
Between-subjects degrees of freedom (df[BS])	41
Residual degrees of freedom (df[RES])	82
Mean square between-conditions (MS[BC])	0.916
Mean square between-subjects (MS[BS])	8.7002
Mean square residual (MS[RES])	0.947
Intraclass correlation coefficient (r[ICC])	0.732
-95% CI dla ICC(2,1)	0.6013
+95% CI dla ICC(2,1)	0.834
Average measure r[ICC]	0.8912
-95% CI dla ICC(2,k)	0.819
+95% CI dla ICC(2,k)	0.9378
F statistic	9.187
p-value	<0.0001

$p < 0.000001$ **for the digital meters (statistically significant, $\alpha = 0.05$)**

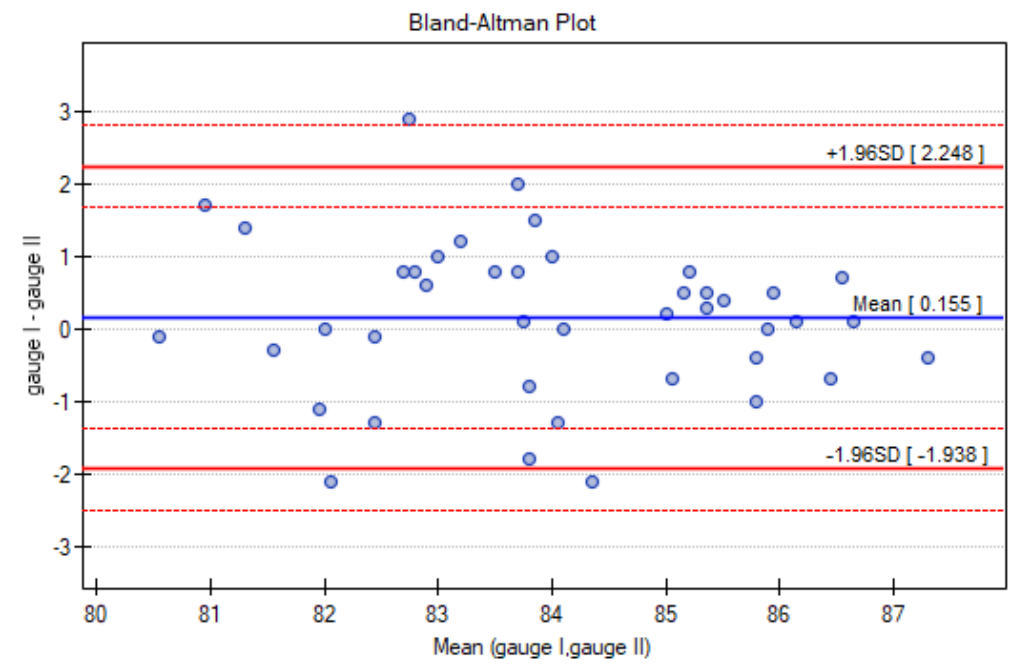
The ICC coefficients and their confidence intervals are much higher than for the analogue meters: $ICC(2,1) = 0.73$, $ICC(2,k) = 0.89$.

Digital meters will therefore be used at the company site from now on. The agreement of the results obtained for the digital meters is shown on a scatter plot, where each measurement point is described by the sound-intensity value obtained for each meter.

Copyright ©2010-2026 U-smile Stats – All rights reserved



By plotting the previously sorted data according to the mean sound-intensity value, we can check whether the degree of agreement increases or decreases as sound intensity increases. For our data, slightly higher agreement (closeness of the points on the graph) is observed at high sound-intensity values. Similarly, the agreement of the results obtained can be observed on Bland-Altman plots[8][23] constructed separately for each pair of meters. The plot for meter I and meter II is shown below.



Here too, we observe higher agreement (points concentrated close to the horizontal axis $y = 0$) for higher sound-intensity values.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! If the researcher were not concerned with estimating the actual sound-intensity value at the company site, but instead wanted to identify locations where this level is higher than elsewhere, or to check whether the noise level changes over time, then model 2 examining consistency would be a sufficient model.

FOR THOSE INTERESTED

Model 1 – formulas:
ICC(1,1):

$$ICC(1,1) = \frac{MS_{WG} - MS_{BG}}{MS_{WG} + (k-1)MS_{BG}}.$$

ICC(1,k):

$$ICC(1,k) = \frac{MS_{WG} - MS_{BG}}{MS_{WG}},$$

where:
MS_{WG} – mean square within-groups,
MS_{BG} – mean square between-objects.

Model 2 – absolute agreement – formulas:
ICC(2,1):

$$ICC(2,1) = \frac{MS_{BS} - MS_{res}}{MS_{BS} + (k-1)MS_{res} + \frac{k}{n}(MS_{BC} - MS_{res})}.$$

ICC(2,k):

$$ICC(2,k) = \frac{MS_{BS} - MS_{res}}{MS_{BS} + (MS_{BC} - MS_{res})/n},$$

where:
MS_{BC} – mean square between-raters,
MS_{BS} – mean square between-objects,
MS_{res} – mean square residual.

Model 2 – consistency – formulas:
ICC(2,1):

$$ICC(2,1) = \frac{MS_{BS} - MS_{res}}{MS_{BS} + (k-1)MS_{res}},$$

ICC(2,k):

$$ICC(2,k) = \frac{MS_{BS} - MS_{res}}{MS_{BS}},$$

where:
MS_{BS} – mean square between-objects,
MS_{res} – mean square residual.

The F-test for the significance of the intraclass correlation coefficient
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad ICC = 0 \\ \mathcal{H}_1: &\quad ICC \neq 0 \quad (ICC = 1) \end{aligned}$$

Formulas:
The test statistic is defined by:

$$F = \frac{MS_{BS}}{MS_{res}}$$ – in the dependent model,

Copyright ©2010-2026 U-smile Stats – All rights reserved

or

$$F = \frac{MS_{WG}}{MS_{BG}} \text{ – in the independent model.}$$

The test statistic has the F Snedecor distribution with the number of degrees of freedom defined in the model.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

23.2 NON-PARAMETRIC TESTS

The Kendall's concordance coefficient and a test to examine its significance

WHEN TO USE:

The Kendall's $\widetilde{W}$ coefficient of concordance *(Kendall's Coefficient of Concordance)*, described in the works of Kendall and Babington-Smith (1939)[96] and Wallis (1939)[171], is used when we have rankings coming from different sources (from different raters) and concerning several ($k \geq 2$) objects, and we want to assess the agreement of these rankings. It is often used to measure the strength of **interrater reliability**, i.e. the degree to which the raters' ratings agree.

ASSUMPTIONS

– measurement on an ordinal scale or an interval scale.

DEFINITIONS OF MEASURES

The Kendall's $\widetilde{W}$ coefficient of concordance – the value $W \in \langle 0; 1 \rangle$ is interpreted as follows:

- $\widetilde{W} \approx 1$ means strong agreement in the rating of individual objects by the raters;
- $\widetilde{W} \approx 0$ means no agreement in the rating of individual objects by the raters.

W denotes the Kendall concordance coefficient in the population, while $\widetilde{W}$ denotes it in the sample.

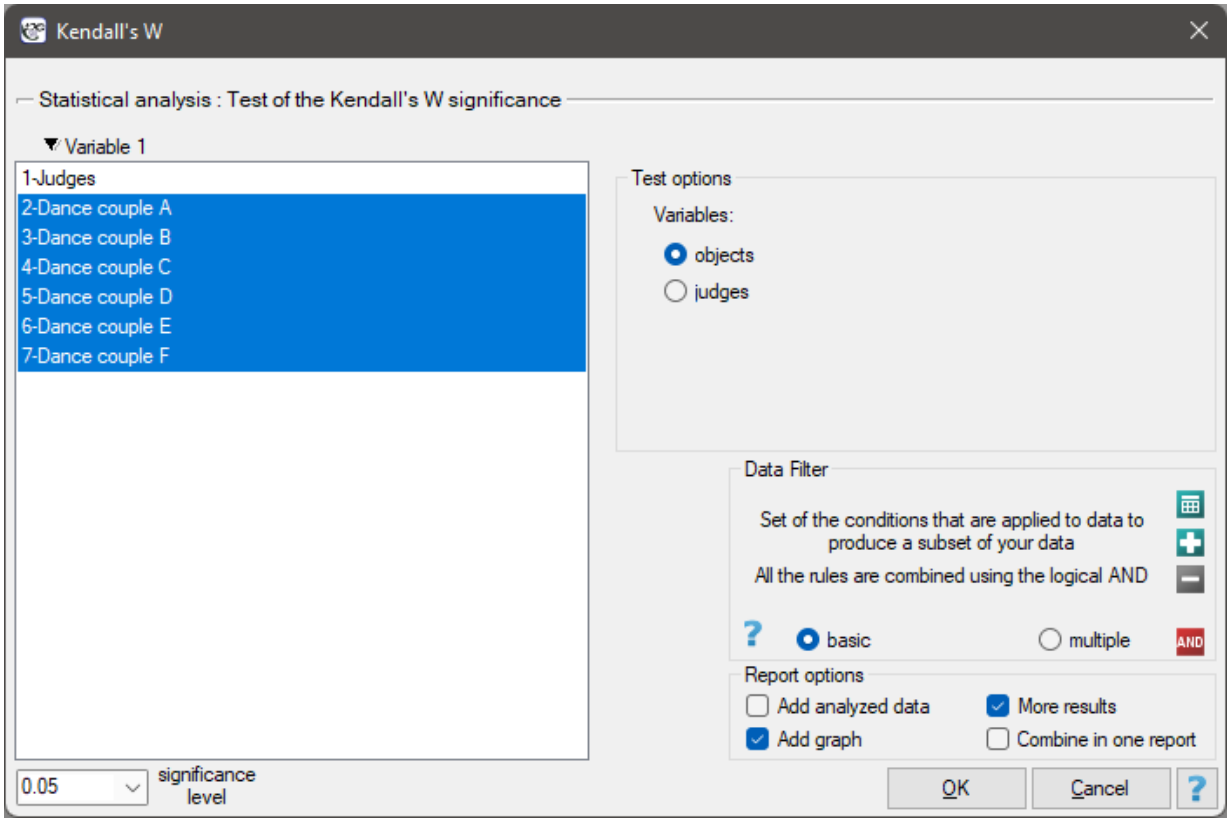
Note! The Kendall's $\widetilde{W}$ coefficient of concordance versus the Spearman r_s coefficient: when we calculate the value of the Spearman correlation coefficient r_s for all possible pairs of rankings, the average r_s coefficient (denoted by $\bar{r}_s$) is a linear function of the $\widetilde{W}$ coefficient calculated from the same data (see *For those interested*).

The Kendall's $\widetilde{W}$ coefficient of concordance versus the Friedman ANOVA: the Kendall's $\widetilde{W}$ coefficient of concordance and the Friedman ANOVA are based on the same mathematical model. As a result, the value of the χ^2 test statistic for testing the significance of the Kendall coefficient of concordance and the value of the Friedman ANOVA test statistic are the same.

Copyright ©2010-2026 U-smile Stats – All rights reserved

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Kendall's W (or via the Wizard).



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 23.2. (judges.pqs file)
In the six-point system, 9 judges score dance couples on, among other things, artistic impression. The judges begin by comparing the competitors against each other and placing them at a given position (thereby creating a ranking). We will check whether the ranks assigned by the judges agree:

Judges	Couple A	Couple B	Couple C	Couple D	Couple E	Couple F
S1	3	6	2	5	4	1
S2	4	6	1	5	3	2
S3	4	6	2	5	3	1
S4	2	6	3	5	4	1
S5	2	6	1	5	4	3
S6	3	5	1	6	4	2
S7	5	4	1	6	3	2
S8	3	6	2	5	4	1
S9	2	6	3	5	4	1

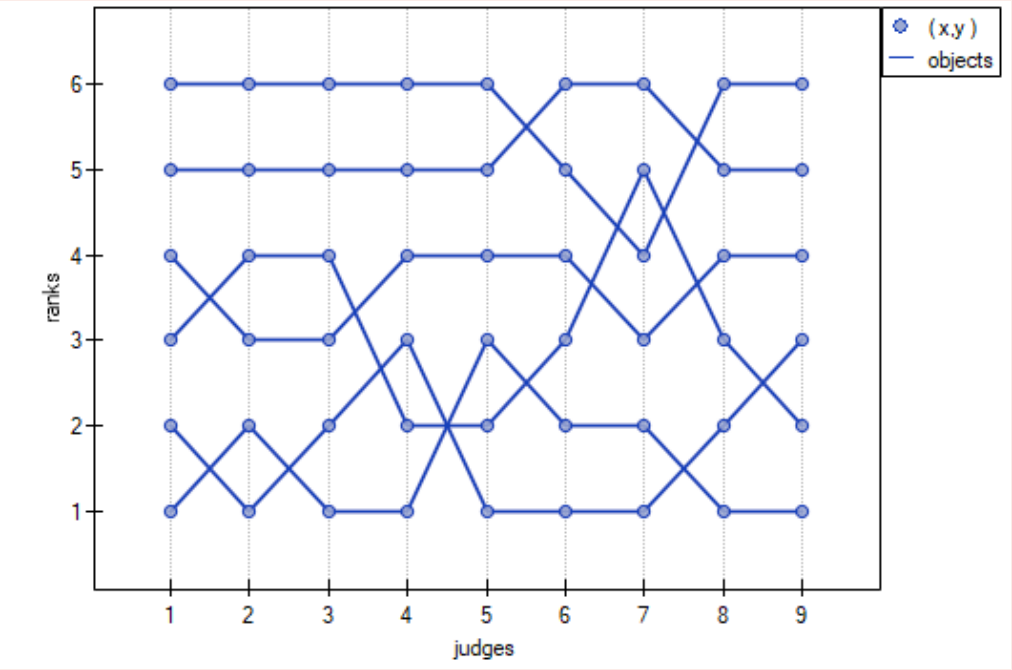
Copyright ©2010-2026 U-smile Stats – All rights reserved

Test of the Kendall's W significance	
Analysed variables	Dance couple A
	Dance couple B
	Dance couple C
	Dance couple D
	Dance couple E
	Dance couple F
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Degrees of freedom	5
Kendall's coefficient of concordance	0.8335
Mean Spearman correlation coefficient	0.8127
Chi-square statistic (adjusted for ties)	37.5079
p-value	<0.0001

$p < 0.000001$ **(statistically significant,** $\alpha = 0.05$**)**

The judges' ratings are statistically concordant.

The strength of this agreement is high, at $\widetilde{W} = 0.83351$, as is the average Spearman monotonic correlation coefficient, $\bar{r}_s = 0.81270$. This result can be shown on a graph in which the X axis represents the successive judges. The more often the lines intersect (they should be parallel to the X axis when agreement is perfect), the weaker the agreement represented by the judges' ratings.



Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Formulas:
The value of the Kendall coefficient of concordance is calculated using the formula:

$$\widetilde{W} = \frac{12U - 3n^2k(k+1)^2}{n^2k(k^2-1) - nC},$$

where:
n – the number of different sets of ratings (the number of raters),
k – the number of ranked objects,
$U = \sum_{j=1}^{k}\left(\sum_{i=1}^{n} R_{ij}\right)^2$,
R_{ij} – ranks assigned to successive objects $(j = 1, 2, ...k)$, separately for each rater $(i = 1, 2, ...n)$,
$C = \sum(t^3 - t)$ – a correction for tied ranks,
t – the number of cases included in a tie.

The formula for the coefficient includes the correction for tied ranks, C. This correction is used when tied ranks are present (when there are no tied ranks, the correction is not calculated, since then $C = 0$).

The relationship between the average Spearman correlation coefficient and the Kendall coefficient of concordance is given by:

$$\bar{r}_s = \frac{n\widetilde{W} - 1}{n - 1}.$$

The χ^2 test for the significance of the Kendall coefficient of concordance
Hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \quad W = 0 \\ \mathcal{H}_1 &: \quad W \neq 0 \end{aligned}$$

Formulas:
The test statistic is defined by:

$$\chi^2 = n(k-1)\widetilde{W}.$$

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with the number of degrees of freedom calculated using the formula $df = k - 1$.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The Cohen's Kappa coefficient and a test to examine its significance

WHEN TO USE:

The Cohen's Kappa coefficient (Cohen J., 1960[45]) determines the degree of agreement between two measurements of the same variable made under different conditions. The same variable may be

Copyright ©2010-2026 U-smile Stats – All rights reserved

measured by two different raters (reproducibility), or by one rater twice (repeatability). The $\hat{\kappa}$ coefficient is calculated for dependent categorical variables, on the basis of raw data or a $c \times c$ contingency table built from it.

ASSUMPTIONS

- measurement on a nominal scale (unweighted Kappa, i.e. Cohen's Kappa) or an ordinal scale (weighted Kappa).

DEFINITIONS OF MEASURES

The Kappa coefficient ($\hat{\kappa}$) – its value lies in the range from -1 to 1. A value of 1 means full agreement, a value of 0 means agreement at the same level that would arise from a random distribution of the data in a contingency table. The range between 0 and -1 is practically never used – a negative $\hat{\kappa}$ value means agreement at a level lower than would arise from a random distribution of the data in the contingency table. $\hat{\kappa}$ denotes the agreement coefficient in the sample, while κ denotes it in the population.

Depending on requirements, unweighted Kappa (i.e. Cohen's Kappa) or weighted Kappa can be determined. The assigned weights (w_{ij}) refer to the individual cells of the contingency table: on the diagonal they equal 1, and off the diagonal they lie in the range $\langle 0; 1)$.

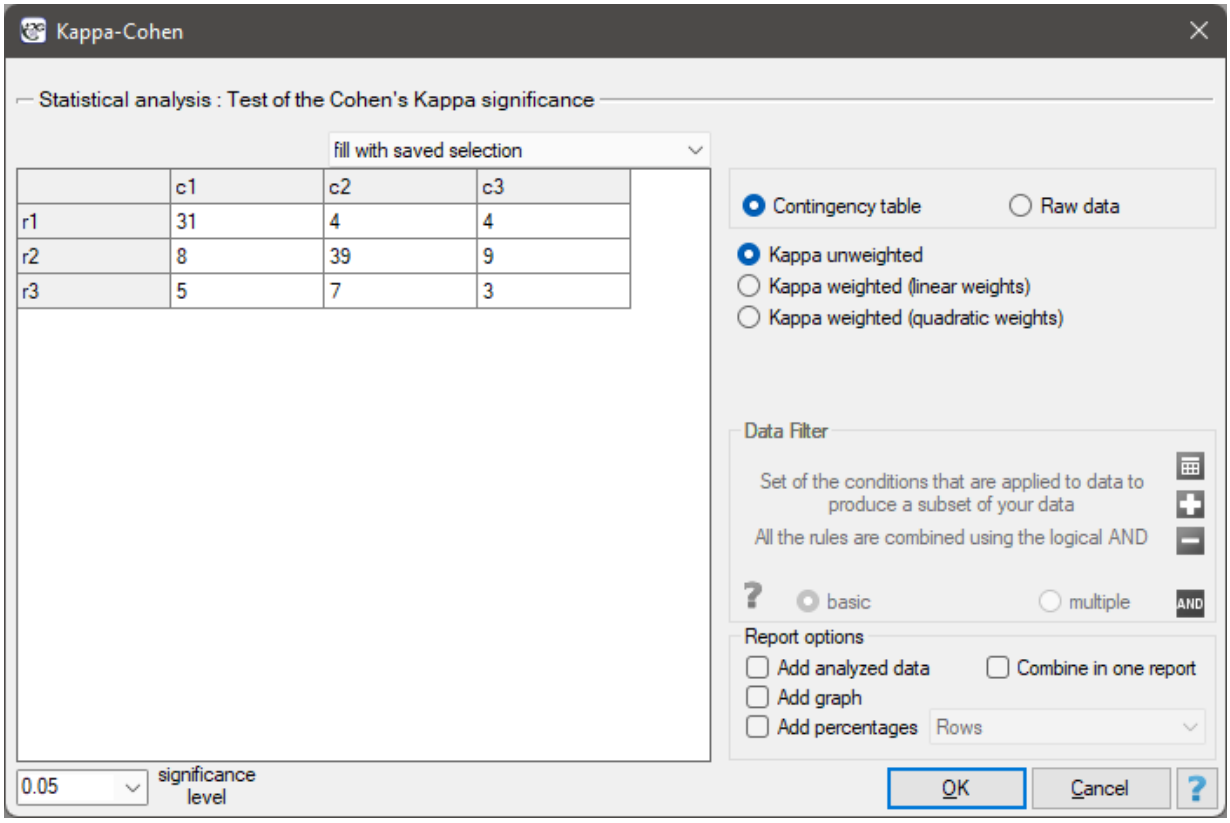
Unweighted Kappa – calculated for data whose categories cannot be ordered, e.g. data comes from patients divided by the type of disease diagnosed, and these diseases cannot be ordered (pneumonia (1), bronchitis (2), other (3)). In such a situation, the agreement of diagnoses made by two doctors can be checked using unweighted Kappa, i.e. Cohen's Kappa. The disagreement of pairs $\{(1), (3)\}$ and $\{(1), (2)\}$ is treated equivalently, so the off-diagonal weights of the weight matrix are set to zero.

Weighted Kappa – in a situation where the data categories can be sorted (e.g. data comes from patients divided by the degree of lesion severity into: no lesion (1), mild lesion (2), suspected cancer (3), cancer (4)), the agreement of ratings given by two radiologists can be built while taking into account the possibility of sorting. The pair of ratings $\{(1), (4)\}$ may then be considered more discordant than $\{(1), (2)\}$. For the order of the categories to affect the agreement result in this way, weighted Kappa should be determined. The assigned weights can take a linear or quadratic form (Cicchetti, 1971[38]; Cohen, 1968[46]) – the further from the diagonal of the matrix, the smaller the weight, with linear weights decreasing proportionally, and quadratic weights decreasing more slowly closer to the diagonal and more rapidly further away. Quadratic weights are of greater interest because of the practical interpretation of the Kappa coefficient, which in that case is identical to the intraclass correlation coefficient[62].

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Kappa-Cohen (or via the Wizard).

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 23.3. (diagnosis.pqs file)
We examine the agreement of the diagnosis made by 2 doctors. To do this, we draw a sample of 110 patients from a children's hospital. The doctors see patients in neighbouring consulting rooms. Each patient is examined first by doctor A and then by doctor B. The diagnoses made by the doctors are shown in the table below.

	pneumonia	bronchitis	others
pneumonia	31	4	4
bronchitis	8	39	9
others	5	7	3

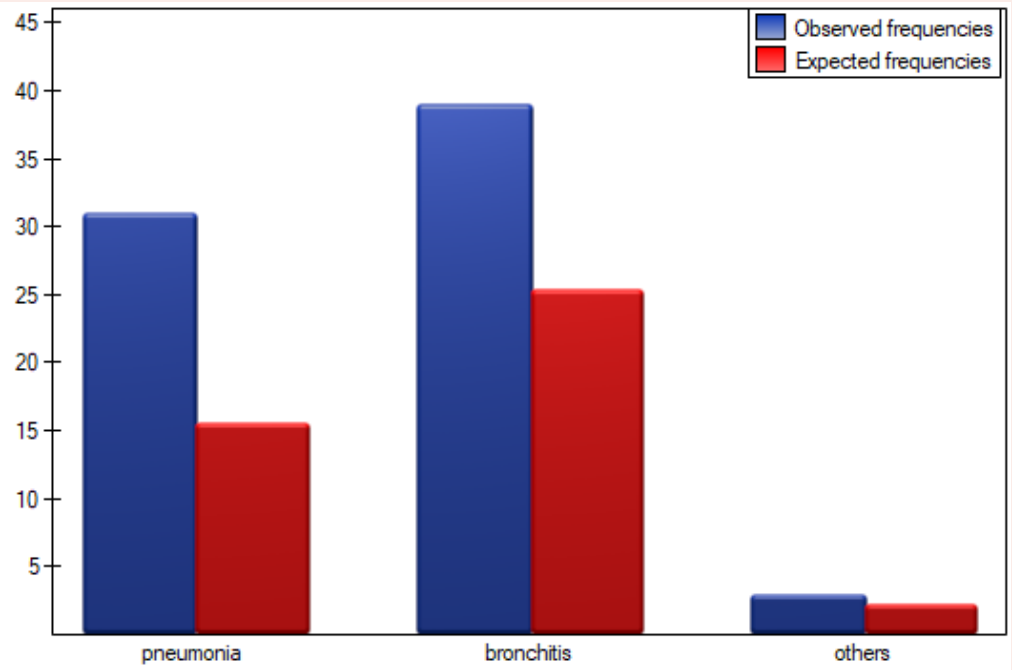
We could examine the agreement of the diagnosis using the plain percentage of concordant values. In our example, the doctors gave a concordant diagnosis for 73 patients (31+39+3=73), which is 66.36% of the group studied. The Kappa coefficient adjusts this value for the chance of agreement (i.e. it corrects for the agreement that arises from a random distribution of the data in the table).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test of the Cohen's Kappa significance	
Analysed variables	Doctor 1
	Doctor 2
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of pairs	110
Type	Kappa unweighted
Kappa coefficient	0.4458
Std. err. of Kappa	0.068
-95% CI for Kappa coefficient	0.3125
+95% CI for Kappa coefficient	0.5791
Std. err. of Kappa distribution	0.0723
Z statistic	6.1639
p-value (asymptotic)	<0.0001

Data :			✓Doctor 2	
↓Doctor 1	pneumonia	bronchitis	others	Summary
pneumonia	31	4	4	39
bronchitis	8	39	9	56
others	5	7	3	15
Summary	44	50	16	110

% of sum :			✓Doctor 2
↓Doctor 1	pneumonia	bronchitis	others
pneumonia	28.182%	3.636%	3.636%
bronchitis	7.273%	35.455%	8.182%
others	4.545%	6.364%	2.727%



$\hat{\kappa} = 0.4458$, $p < 0.000001$ **(statistically significant,** $\alpha = 0.05$**)**

The diagnoses made by the two doctors are statistically concordant.

The agreement expressed by the coefficient $\hat{\kappa} = 44.58\%$ is lower than the agreement not adjusted for the chance of agreement.

EXAMPLE 23.4. (radiology.pqs file)
Liver damage was assessed on radiological images using the following categories: no changes (1), mild changes (2), suspected cancer (3), cancer (4). The assessment was made by two independent radiologists based on a group of 70 patients. We want to check the agreement of the diagnosis made.

Copyright ©2010-2026 U-smile Stats – All rights reserved

	no changes	mild changes	suspicion of cancer	cancer
no changes	23	4	3	0
mild changes	2	12	5	2
suspicion of cancer	5	4	4	0
cancer	1	2	1	2

Because the diagnosis is made on an ordinal scale, the appropriate measure of agreement would be the weighted Kappa coefficient.

Test of the Cohen's Kappa significance	
Analysed variables	Contingency table
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of pairs	70
Type	Kappa weighted (linear weigl
Kappa coefficient	0.3902
Std. err. of Kappa	0.0905
-95% CI for Kappa coefficient	0.2129
+95% CI for Kappa coefficient	0.5675
Std. err. of Kappa distribution	0.0866
Z statistic	4.5037
p-value (asymptotic)	<0.0001

Test of the Cohen's Kappa significance	
Analysed variables	Contingency table
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of pairs	70
Type	Kappa weighted (quadratic w
Kappa coefficient	0.4187
Std. err. of Kappa	0.1109
-95% CI for Kappa coefficient	0.2012
+95% CI for Kappa coefficient	0.6361
Std. err. of Kappa distribution	0.1189
Z statistic	3.5201
p-value (asymptotic)	0.0004

$\hat{\kappa} = 0.390194$ for the linear weights ($p = 0.000007$), $\hat{\kappa} = 0.418658$ for the quadratic weights ($p = 0.000431$) – both statistically significant, $\alpha = 0.05$

The diagnoses made by the two radiologists are statistically concordant, regardless of the weights used.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Because the data are concentrated mainly on the main diagonal of the matrix and in its immediate vicinity, the coefficient weighted by linear weights is lower than the coefficient determined using quadratic weights. If there were substantial disagreement concerning the two extreme cases, and the pair (no changes and cancer), located in the top-right corner of the table, occurred far more often, e.g. 15 times, then such a large lack of agreement would be more apparent when using quadratic weights (the Kappa coefficient would drop drastically) than when using linear weights.

FOR THOSE INTERESTED

Linear weights

$$w_{ij} = 1 - \frac{|i-j|}{c-1}.$$

Example weights for a 5x5 matrix:

1	0.75	0.5	0.25	0
0.75	1	0.75	0.5	0.25
0.5	0.75	1	0.75	0.5
0.25	0.5	0.75	1	0.75
0	0.25	0.5	0.75	1

Quadratic weights

$$w_{ij} = 1 - \frac{(i-j)^2}{(c-1)^2}.$$

Example weights for a 5x5 matrix:

1	0.9375	0.75	0.4375	0
0.9375	1	0.9375	0.75	0.4375
0.75	0.9375	1	0.9375	0.75
0.4375	0.75	0.9375	1	0.9375
0	0.4375	0.75	0.9375	1

Procedure:
To determine agreement using the Kappa coefficient, the data are presented as a table of observed frequencies O_{ij} (**??**), which is then transformed into a probability contingency table $p_{ij} = O_{ij}/n$.

The Kappa coefficient ($\hat{\kappa}$) is then expressed by the formula:

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e},$$

where:
$P_o = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} p_{ij}$,
$P_e = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} p_{i.} p_{.i}$,
$p_{i.}$, $p_{.i}$ – the row and column totals of the probability contingency table.

The standard error for Kappa is expressed by the formula:

$$SE_{\hat{\kappa}} = \frac{1}{(1-P_e)\sqrt{n}} \sqrt{\sum_{i=1}^{c} \sum_{j=1}^{c} p_{i.} p_{.j} [w_{ij} - (\overline{w}_{i.} + (\overline{w}_{.j})]^2 - P_e^2}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where:
$\overline{w}_{i.} = \sum_{j=1}^{c} p_{.j} w_{ij}$,
$\overline{w}_{.j} = \sum_{i=1}^{c} p_{i.} w_{ij}$.

The Z-test for the significance of the Cohen's Kappa coefficient ($\hat{\kappa}$) (Fleiss, 2003[63]) is used to verify the hypothesis about the agreement of the results of two measurements $X^{(1)}$ and $X^{(2)}$ of feature X, and is based on the $\hat{\kappa}$ coefficient calculated for the sample.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \quad \kappa = 0, \\ \mathcal{H}_1 &: \quad \kappa \neq 0. \end{aligned}$$

Formulas:
The test statistic is defined by:

$$Z = \frac{\hat{\kappa}}{SE_{\kappa_{distr}}},$$

where:

$$SE_{\kappa_{distr}} = \frac{1}{(1-P_e)\sqrt{n}} \sqrt{\sum_{i=1}^{c}\sum_{j=1}^{c} p_{ij}[w_{ij} - (\overline{w}_{i.} + \overline{w}_{.j})(1-\hat{\kappa})]^2 - [\hat{\kappa} - P_e(1-\hat{\kappa})]^2}.$$

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

The Fleiss' Kappa coefficient and a test to examine its significance

WHEN TO USE:

The Fleiss' Kappa coefficient determines the agreement of measurements made by several raters (Fleiss, 1971[61]) and is an extension of Cohen's Kappa, which only allows the agreement of two raters to be examined. Each of the n randomly selected objects may be rated by a different random set of k raters. The analysis is based on data transformed into a table with n rows and c columns, where c is the number of possible categories to which raters assign the objects studied.

ASSUMPTIONS

- measurement on a nominal scale – any ordering of the categories is not taken into account.

DEFINITIONS OF MEASURES

The Kappa coefficient ($\hat{\kappa}$) – a value of $\hat{\kappa} = 1$ means full agreement among raters, while $\hat{\kappa} = 0$ means

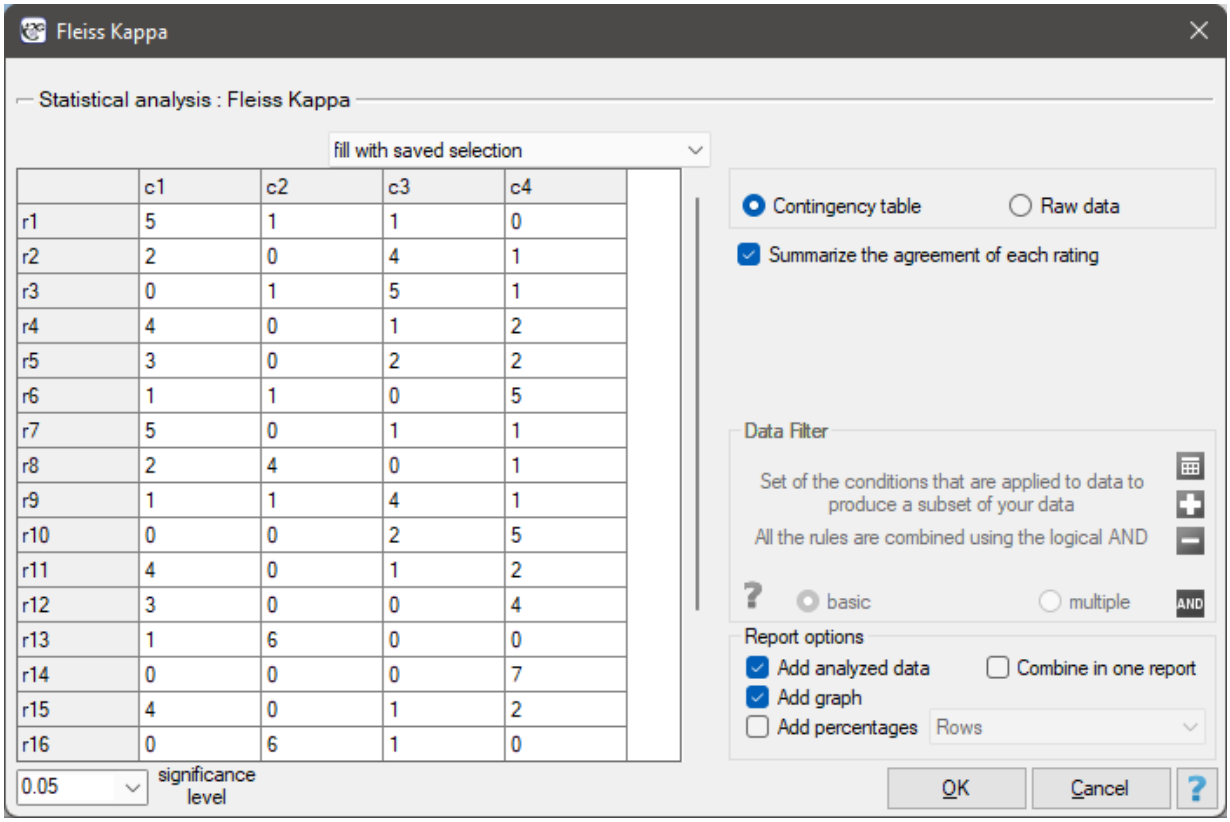
Copyright ©2010-2026 U-smile Stats – All rights reserved

the level of agreement that would arise if the raters' opinions were given at random. Negative values of Kappa, in turn, indicate agreement lower than the level expected by chance.

Note! The determination of the Fleiss' Kappa coefficient is conceptually similar to the Mantel-Haenszel method. The calculated Kappa is a general measure summarising the agreement of all raters' ratings, and can be determined as a Kappa built from individual strata, which are the specific ratings given by the raters (Fleiss, 2003[63]). Therefore, as a summary of each stratum, the interrater agreement (Kappa coefficient) can be determined, summarising each possible rating separately.

HOW TO SET UP THE ANALYSIS

Statistics → Nonparametric tests → Fleiss Kappa.



	c1	c2	c3	c4
r1	5	1	1	0
r2	2	0	4	1
r3	0	1	5	1
r4	4	0	1	2
r5	3	0	2	2
r6	1	1	0	5
r7	5	0	1	1
r8	2	4	0	1
r9	1	1	4	1
r10	0	0	2	5
r11	4	0	1	2
r12	3	0	0	4
r13	1	6	0	0
r14	0	0	0	7
r15	4	0	1	2
r16	0	6	1	0

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 23.5. (temperament.pqs file)
20 volunteers take part in a game aimed at determining their personality type. Each volunteer has ratings given by 7 different observers (most often people from their close circle or family). Each observer was introduced to the basic traits describing temperament for each personality type: choleric, phlegmatic, melancholic, sanguine. We examine the observers' agreement in assigning personality types. An excerpt of the data is shown in the table below.

Copyright ©2010-2026 U-smile Stats – All rights reserved

	spitfire	phlegmatic	melancholic	sanguine
Volunteer 1	5	1	1	0
Volunteer 2	2	0	4	1
Volunteer 3	0	1	5	1
Volunteer 4	4	0	1	2
Volunteer 5	3	0	2	2
Volunteer 6	1	1	0	5

Fleiss Kappa	
Analysed variables	Judge 1
	Judge 2
	Judge 3
	Judge 4
	Judge 5
	Judge 6
	Judge 7
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of raters	7
Number of rating categories	4
Number of objects	20
Total number of issued ratings	140
Kappa coefficient	0.2449
Std. err. of Kappa	0.0286
-95% CI for Kappa coefficient	0.1888
+95% CI for Kappa coefficient	0.301
Z statistic	8.5541
p-value (asymptotic)	<0.0001

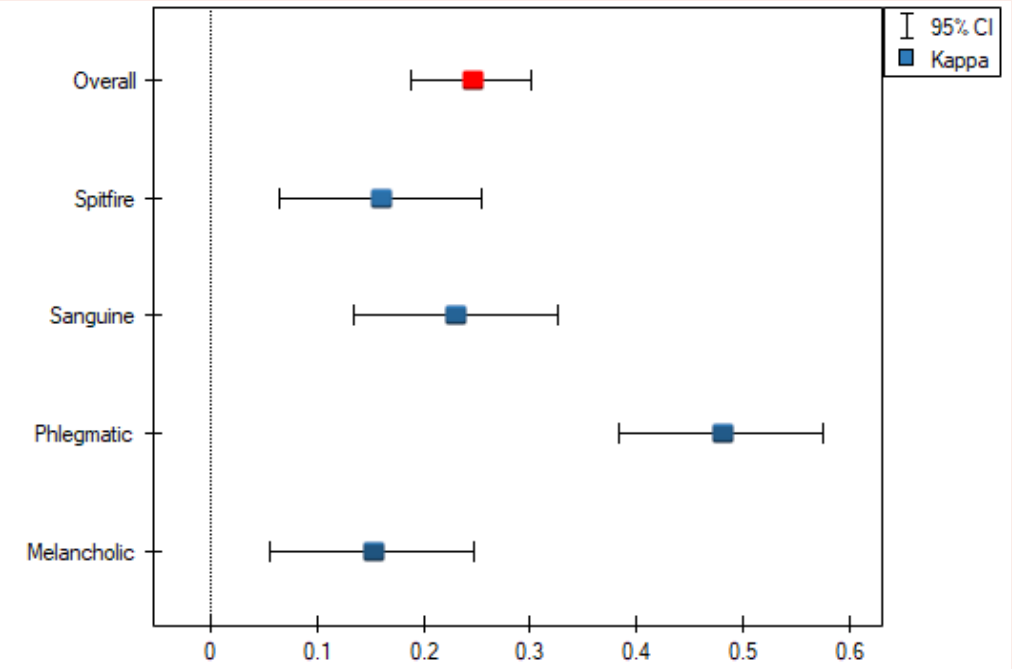
$\hat{\kappa} = 0.244918$, $p < 0.000001$ **(statistically significant,** $\alpha = 0.05$**)**

We observe a small but statistically significant Kappa coefficient, which means non-random agreement among the raters' ratings.

Significant agreement applies to every rating, as confirmed by the report summarising agreement for each stratum (for each rating) and the graph showing the individual Kappa coefficients along with the overall summary Kappa.

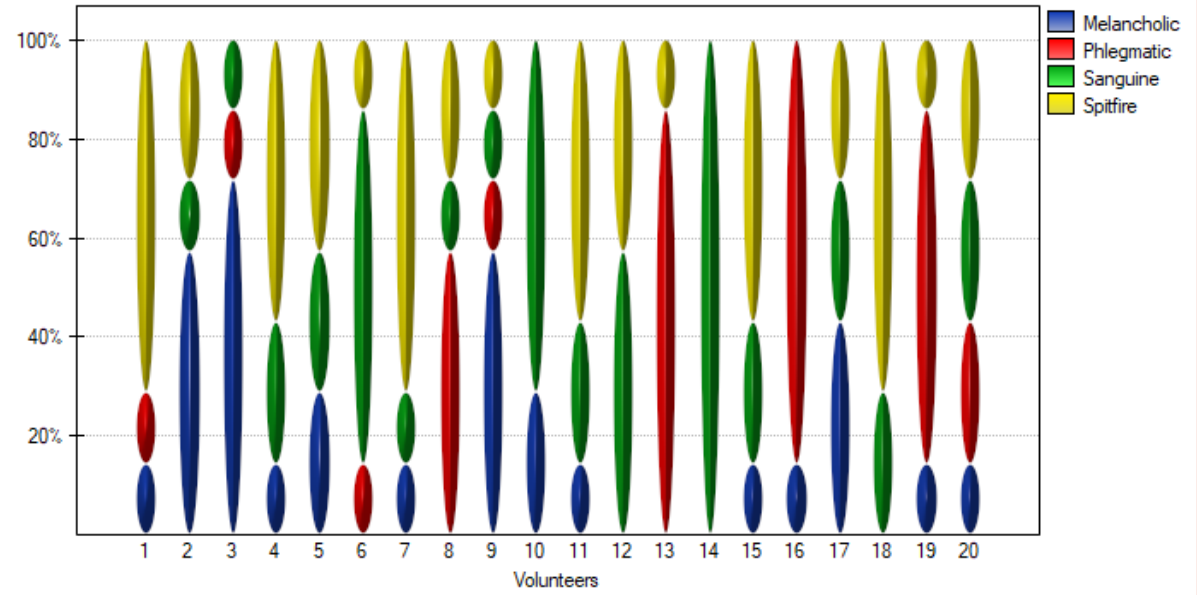
Copyright ©2010-2026 U-smile Stats – All rights reserved

Summary of agreement	Melancholic	Phlegmatic	Sanguine	Spitfire
Kappa coefficient	0.1518	0.48	0.23	0.1595
Std. err. of Kappa	0.0488	0.0488	0.0488	0.0488
-95% CI for Kappa coefficient	0.0561	0.3843	0.1344	0.0638
+95% CI for Kappa coefficient	0.2474	0.5756	0.3256	0.2551
Z statistic	3.1107	9.8361	4.7136	3.2678
p-value (asymptotic)	0.0019	<0.0001	<0.0001	0.0011



It may be interesting to note that the highest agreement concerns the rating of phlegmatics (Kappa=0.479952). With a small number of people observed, it is also worth plotting a graph showing how the observers rated each of them.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In this case, only person no. 14 received an unambiguous personality-type rating – sanguine. Persons no. 13 and 16 were rated as phlegmatic by 6 observers (out of 7 possible). For the remaining persons, agreement among the ratings was somewhat lower. The personality type that seems hardest to define characterises the last person, who received the most varied set of ratings.

FOR THOSE INTERESTED

The Kappa coefficient ($\hat{\kappa}$) is expressed by the formula:

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e},$$

where:
$P_o = \frac{1}{kn(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{c} x_{ij} - kn,$
$P_e = \sum_{i=1}^{c} q_j^2,$
$q_j = \frac{1}{km} \sum_{i=1}^{n} x_{ij},$
x_{ij} – the number of raters assigning object i to category j.

For the coefficient $\hat{\kappa}$, a standard error SE can be determined, which allows its statistical significance to be tested and asymptotic confidence intervals to be calculated.

The Z-test for the significance of the Fleiss' Kappa coefficient ($\hat{\kappa}$) (Fleiss, 2003[63]) is used to verify the hypothesis about the agreement of the ratings of several raters, and is based on the coefficient $\hat{\kappa}$ calculated for the sample.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \quad \kappa = 0, \\ \mathcal{H}_1 &: \quad \kappa \neq 0. \end{aligned}$$

Formulas: The test statistic is defined by:

$$Z = \frac{\hat{\kappa}}{SE},$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

24 DIAGNOSTIC TESTS

Evaluation of a diagnostic test

WHEN TO USE:

Suppose that with the help of a diagnostic test we examine the occurrence of a given feature (most often a disease), and we know the reality (the so-called gold standard), i.e. we know whether this feature actually occurs in the people studied. On the basis of this information, we can build a 2×2 contingency table:

Observed frequencies		Reality (gold standard)		
		disease **(+)**	disease free **(–)**	**Total**
diagnostic test	positive result **(+)**	TP	FP	TP+FP
	negative result **(–)**	FN	TN	FN+TN
	Total	TP+FN	FP+TN	n=TP+FP+FN+TN

where:
TP – true positive results,
FP – false positive results,
FN – false negative results,
TN – true negative results.
For such a table, we can calculate the measures given below.

DEFINITIONS OF MEASURES

Sensitivity – describes the ability to detect people who are genuinely ill (who have a given feature). If we examine a group of ill people, sensitivity tells us what percentage of them have a positive test result.

Specificity – describes the ability to detect people who are genuinely healthy (without a given feature). If we examine a group of healthy people, specificity tells us what percentage of them have a negative test result.

Positive predictive value (PPV) – the probability that a person has the disease, given a positive test result. So if the person examined obtains a positive test result, PPV tells them how confident they can be that they have the disease.

Negative predictive value (NPV) – the probability that a person does not have the disease, given a negative test result. So if the person examined obtains a negative test result, NPV tells them how confident they can be that they do not have the disease. Positive and negative predictive values depend on the prevalence of the disease.

Prevalence – the probability of the disease occurring in the population for which the diagnostic test was carried out.

Likelihood ratio of a positive test (LR_+) – this measure allows the fit of the results of several tests to the gold standard to be compared, and does not depend on the prevalence of the disease. It is the ratio of two odds: the odds that a positive test result is obtained by a person from the group of ill people, to the odds that the same result is observed among healthy people.

Likelihood ratio of a negative test (LR_-) – the ratio of two odds: the odds that a negative test result is obtained by a person from the group of ill people, to the odds that the same result is observed among healthy people.

Copyright ©2010-2026 U-smile Stats – All rights reserved

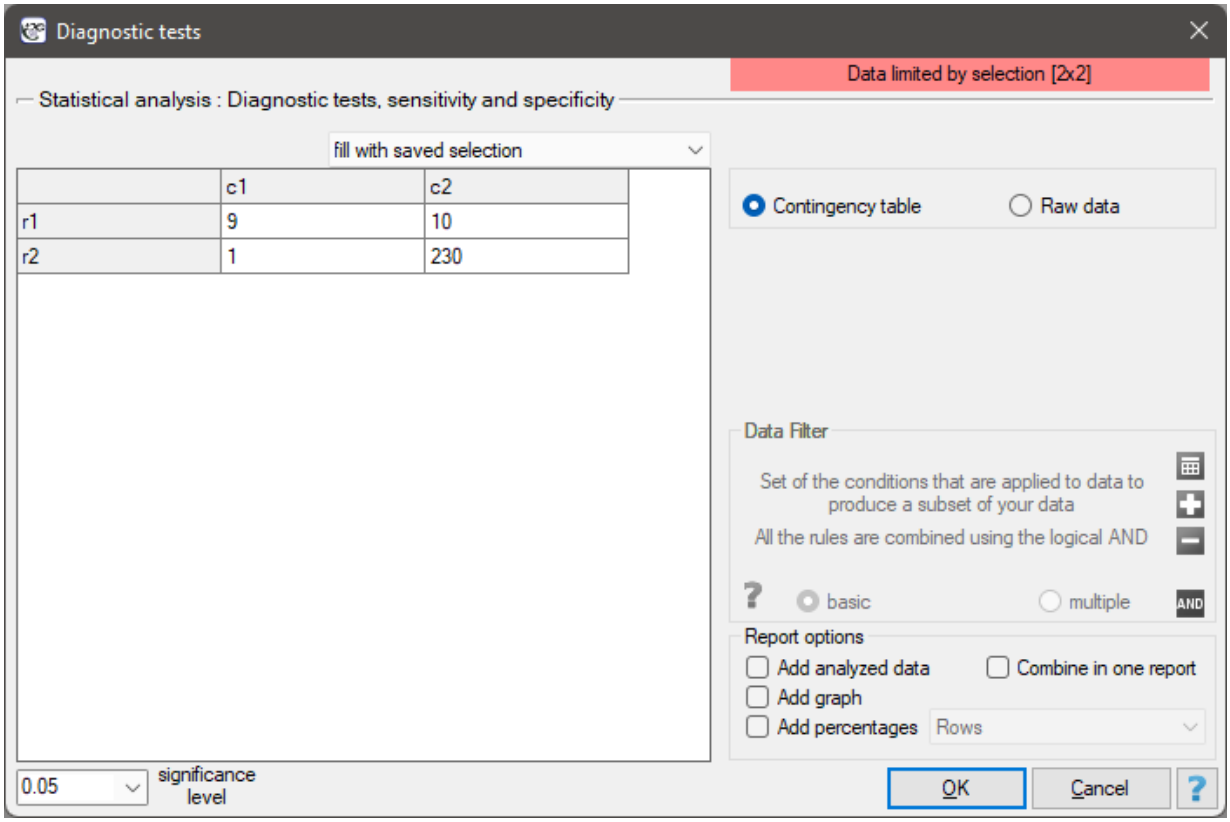
Accuracy (Acc**)** – the probability of a correct diagnosis using the diagnostic test. So if the person examined obtains a positive or negative test result, Acc tells them how confident they can be in the diagnosis made.

Diagnostic Odds Ratio (DOR**)** – the ratio of two odds: the odds of a positive test result for an ill person, to the odds of a positive test result for a healthy person.

Note! Confidence intervals for sensitivity, specificity, PPV, NPV, Acc, and the prevalence are built on the basis of the Clopper-Pearson method for a single proportion.

HOW TO SET UP THE ANALYSIS

Advanced statistics → Diagnostic tests → Diagnostic reliability.



PERFORMING AND INTERPRETING THE ANALYSIS

Example 24.1. (mammography.pqs file)
Mammography is one of the most widely used screening tests for detecting breast cancer. The study below was carried out on a group of 250 so-called "asymptomatic" women aged 40 to 50. Mammography can detect a cancer focus smaller than 5 mm, and also lets changes be identified that are not yet a nodule, but only a change in tissue structure. An example result of a mammography examination is shown below.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Observed frequencies		Reality (histopathology)		
		disease **(+)**	disease free **(–)**	**Total**
diagnostic test	positive result **(+)**	9	10	19
	negative result **(–)**	1	230	231
	Total	10	240	250

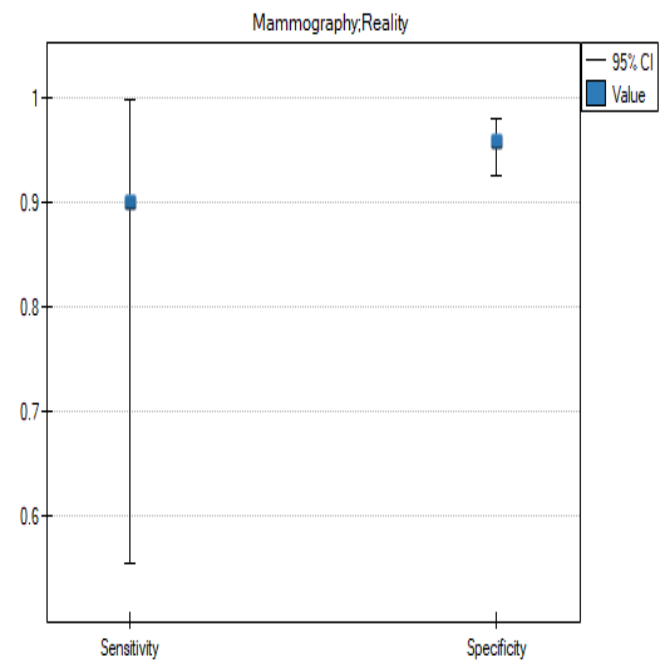
We will determine the values allowing us to evaluate the diagnostic test performed.

Diagnostic tests, sensitivity and specificity	
Analysed variables	Mammography
	Reality
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Sensitivity	0.9
-95% CI	0.555
+95% CI	0.9977
Specificity	0.9583
-95% CI	0.9247
+95% CI	0.9798
Positive predictive value (PPV)	0.4737
-95% CI	0.2445
+95% CI	0.7114
Negative predictive value (NPV)	0.9957
-95% CI	0.9761
+95% CI	0.9999
Positive likelihood ratio (PLR)	21.6
-95% CI	11.3787
+95% CI	41.0031
Negative likelihood ratio (NLR)	0.1043
-95% CI	0.0163
+95% CI	0.67
Accuracy (ACC)	0.956
-95% CI	0.9226
+95% CI	0.9778
Prevalence	0.04
-95% CI	0.0193
+95% CI	0.0723

Data :		↙Reality
↓Mammograph	disease	disease-free
positive result	9	10
negative resul	1	230

% of col :		↙Reality
↓Mammograph	disease	disease-free
positive result	90%	4.167%
negative resul	10%	95.833%

Copyright ©2010-2026 U-smile Stats – All rights reserved



sensitivity=90%, specificity=95.83%, PPV=47.37%, NPV=99.57%

The mammography test shows high sensitivity and specificity, but because of the low prevalence, a positive test result alone does not yet give certainty of diagnosis.

Interpretation of the individual measures:

- 90% of women with breast cancer were correctly diagnosed, i.e. obtained a positive mammography result;
- 95.83% of healthy women (without breast cancer) were correctly diagnosed, i.e. obtained a negative mammography result;
- 4 out of every 100 women examined suffer from breast cancer;
- a woman who obtains a positive mammography result can be 47.37% sure that she has breast cancer;
- a woman who obtains a negative mammography result can be 99.57% sure that she does not have breast cancer;
- the odds that a positive mammography result is obtained by a woman who genuinely has cancer is 21.60 times greater than the odds that a positive mammography result is obtained by a woman who is genuinely healthy;
- the odds that a negative mammography result is obtained by a woman who genuinely has cancer amounts to 10.43% of the odds that a negative mammography result is obtained by a woman who is genuinely healthy;
- a woman undergoing mammography (regardless of the result obtained) can be 96.50% confident in the diagnosis made;

Copyright ©2010-2026 U-smile Stats – All rights reserved

- the odds of a positive test result for a woman who genuinely has breast cancer are 207 times greater than the odds of such a result for a healthy woman.

FOR THOSE INTERESTED

Sensitivity

$$\text{sensitivity} = \frac{TP}{TP+FN}$$

Specificity

$$\text{specificity} = \frac{TN}{FP+TN}$$

Positive predictive value

$$PPV = \frac{TP}{TP+FP}$$

Negative predictive value

$$NPV = \frac{TN}{FN+TN}$$

Prevalence

$$\text{prevalence} = \frac{TP+FN}{n}$$

Likelihood ratio of a positive test

$$LR_+ = \frac{\text{sensitivity}}{1-\text{specificity}} = \frac{TP\,(TP+FN)}{FP\,(FP+TN)}$$

The confidence interval for LR_+ is built on the basis of the standard error:

$$SE = \sqrt{\frac{1-\text{sensitivity}}{TP} + \frac{\text{specificity}}{FP}}.$$

Likelihood ratio of a negative test

$$LR_- = \frac{1-\text{sensitivity}}{\text{specificity}} = \frac{FN\,(TP+FN)}{TN\,(FP+TN)}$$

The confidence interval for LR_- is built on the basis of the standard error:

$$SE = \sqrt{\frac{\text{sensitivity}}{FN} + \frac{1-\text{specificity}}{TN}}.$$

Accuracy

$$Acc = \frac{TP+TN}{n}$$

Diagnostic Odds Ratio

$$DOR = \frac{TP/FN}{FP/TN}$$

The confidence interval for DOR is built on the basis of the standard error:

$$SE = \sqrt{\frac{1}{TP} + \frac{1}{FN} + \frac{1}{FP} + \frac{1}{TN}}.$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

The ROC curve

WHEN TO USE:

A diagnostic test is used to distinguish objects with a given feature (marked as **(+)**, e.g. ill people) from objects without the feature (marked as **(–)**, e.g. healthy people). For a diagnostic test to be considered valuable, it should yield a relatively small number of misclassifications. If the test is based on a dichotomous variable, the proper tool for evaluating its quality is the analysis of a 2×2 contingency table. Most often, however, diagnostic tests are based on continuous variables or variables with ordered categories. In such a situation, the proper means of evaluating a test's ability to distinguish **(+)** from **(–)** are ROC curves *(Receiver Operating Characteristic)*.

It is often observed that as the value of the diagnostic variable increases, the odds of the phenomenon studied occurring increase too, or the other way round: as the value of the diagnostic variable increases, the odds of the phenomenon studied occurring decrease. In such cases, ROC curves are used to choose the optimal cut-off point, i.e. a value of the diagnostic variable that best divides the population studied into two groups: **(+)**, in which the phenomenon occurs, and **(–)**, in which it does not. When two or more ROC curves are built on the basis of studies carried out on the same objects, these curves can be compared in terms of classification quality.

ASSUMPTIONS

- measurement of the diagnostic variable on an ordinal or interval scale,
- the true group membership of each object, **(+)** or **(–)**, is known (the so-called gold standard).

DEFINITIONS OF MEASURES

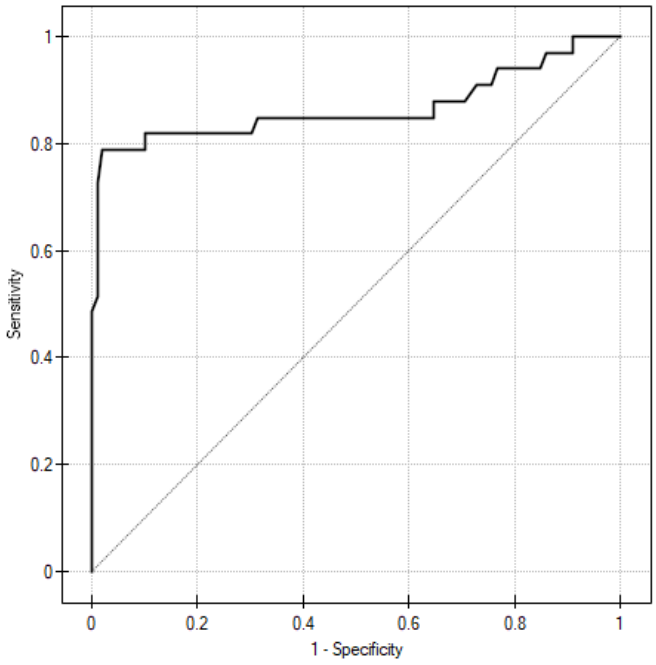
Let us assume we have at our disposal a sample of n elements, in which each object obtains one of k values of the diagnostic variable. Each of the values obtained for the diagnostic variable, $x_1, x_2, ...x_k$, becomes a potential cut-off x_{cat}. If the diagnostic variable is a:

- **stimulant** (as its value increases, the odds of the phenomenon studied occurring increase), then values greater than or equal to the cut-off ($x_i \geq x_{cat}$) are classified into group **(+)**;
- **destimulant** (as its value increases, the odds of the phenomenon studied occurring decrease), then values less than or equal to the cut-off ($x_i \leq x_{cat}$) are classified into group **(+)**.

For each of the k cut-offs, we determine the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values, and on their basis sensitivity, specificity, and the remaining measures described in the Evaluation of a diagnostic test section.

AUC *(area under curve)* – the size of the area under the ROC curve, which lies in the range $\langle 0; 1 \rangle$. The ROC curve is built from the sensitivity and specificity values determined for successive cut-offs – the abscissa axis represents $x = 1-$specificity, and the ordinate axis $y =$sensitivity, and the resulting points are joined together.

Copyright ©2010-2026 U-smile Stats – All rights reserved



The greater the AUC, the more accurately we can classify objects into group **(+)** and group **(–)** on the basis of the diagnostic variable analysed, and so the more effectively that diagnostic variable can be used as a classifier. When the ROC curve coincides with the diagonal $y = x$ (i.e. $AUC = 0.5$), the decision made on the basis of the diagnostic variable is exactly as good as a random split of the objects studied into groups **(+)** and **(–)**.

Note! The area AUC, the error SE_{AUC}, and the confidence interval for AUC are calculated on the basis of:

- the nonparametric **DeLong** method (DeLong E.R. et al., 1988[52]; Hanley J.A. and Hajian-Tilaki K.O., 1997[76]) – **recommended**,
- the nonparametric **Hanley-McNeil** method (Hanley J.A. and McNeil M.D., 1982[77]),
- the **Hanley-McNeil** method assuming a bi-negative exponential distribution (Hanley J.A. and McNeil M.D., 1982[77]) – calculated only when groups **(+)** and **(–)** are equal in size.

PERFORMING AND INTERPRETING THE ANALYSIS

The AUC significance test is always performed together with the selection of the optimal cut-off – see Example 24.2 in the next section.

FOR THOSE INTERESTED

For the classification to be better than a random split of the objects into two classes, the area under the ROC curve should be significantly greater than the area under the line $y = x$, i.e. than 0.5.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \quad AUC = 0.5, \\ \mathcal{H}_1 &: \quad AUC \neq 0.5. \end{aligned}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

Formulas:
The test statistic is defined by:

$$Z = \frac{AUC - 0.5}{SE_{0.5}},$$

where:
$SE_{0.5} = \sqrt{\frac{n_{(+)}+n_{(-)}+1}{12 n_{(+)} n_{(-)}}}$,
$n_{(+)}$ – the size of group **(+)**, in which the phenomenon studied genuinely occurs,
$n_{(-)}$ – the size of group **(–)**, in which the phenomenon studied genuinely does not occur.

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Additionally, once we assume that the diagnostic parameter yields a high AUC, we can select the optimal cut-off.

Revised predictive values
The PQStat program calculates the prevalence on the basis of the sample available. The calculated prevalence will reflect the occurrence of the phenomenon studied (disease) in the population when the study is a screening study covering a large sample representative of the population. When only people suspected of the disease are referred for testing, the prevalence calculated for them may be considerably higher than the prevalence in the population. Since both the positive and negative predictive values depend on the prevalence, if we know this coefficient a priori for the population, we can use it to calculate, for each cut-off x_{cat}, revised predictive values according to Bayes's formulas:

$$PPV_{revised} = \frac{\text{Sensitivity} \cdot P_{apriori}}{\text{Sensitivity} \cdot P_{apriori} + (1 - \text{Specificity}) \cdot (1 - P_{apriori})}$$

$$NPV_{revised} = \frac{\text{Specificity} \cdot (1 - P_{apriori})}{\text{Specificity} \cdot (1 - P_{apriori}) + (1 - \text{Sensitivity}) \cdot P_{apriori}}$$

where $P_{apriori}$ – the prevalence specified by the user, the so-called pre-test probability of disease.

Summary:
The measures determined for each cut-off are gathered in a table:

x_{cat}	**sensitivity**	**specificity**	**PPV**	**NPV**	**LR$_+$**	**LR$_-$**	**Acc**	**PPV$_{rev}$**	**NPV$_{rev}$**
x_1	sensitivity$_1$	specificity$_1$	PPV_1	NPV_1	LR_{+1}	LR_{-1}	Acc_1	PPV_{rev1}	NPV_{rev1}
x_2	sensitivity$_2$	specificity$_2$	PPV_2	NPV_2	LR_{+2}	LR_{-2}	Acc_2	PPV_{rev2}	NPV_{rev2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	sensitivity$_k$	specificity$_k$	PPV_k	NPV_k	LR_{+k}	LR_{-k}	Acc_k	PPV_{revk}	NPV_{revk}

Selection of the optimal cut-off

WHEN TO USE:

Copyright ©2010-2026 U-smile Stats – All rights reserved

The point sought is a certain value of the diagnostic variable that optimally divides the population studied into two groups: **(+)**, in which the phenomenon occurs, and **(–)**, in which it does not. Choosing the optimal cut-off is not easy, since it requires specialist knowledge of the subject of the study. For example, a different cut-off will be needed for a test used in screening studies covering a large group of people, e.g. in a mammography study, than in invasive studies conducted to confirm an earlier suspicion, e.g. in histopathology. Using advanced mathematical tools, we can find such a point, the so-called cut-off, that is most favourable from a mathematical point of view.

DEFINITIONS OF MEASURES

The PQStat program allows the optimal cut-off to be selected using:

The tangent method (cost index) – calculated on the basis of sensitivity, specificity, the costs of wrong decisions, and the prevalence. The errors that can be made when assigning the objects studied to group **(+)** and group **(–)** are false positive results (FP) and false negative results (FN). If making these errors is equally costly (ethical, financial, etc. costs), then in the cost FP field and the cost FN field we enter the same positive value – usually 1. If, on the other hand, we judge that one type of error carries a greater cost than the other, we assign it a correspondingly greater weight. The optimal cut-off value is calculated using m – the slope of the tangent to the ROC curve, defined with reference to the costs of wrong decisions and the prevalence.

Youden's index – conceptually, this is the maximum distance between the line forming the diagonal of a unit square and a point on the ROC curve[182].

Distance to the top-left corner – conceptually, this is the minimum distance between the top-left corner of a unit square (i.e. the point at which sensitivity and specificity can be highest) and a point on the ROC curve.

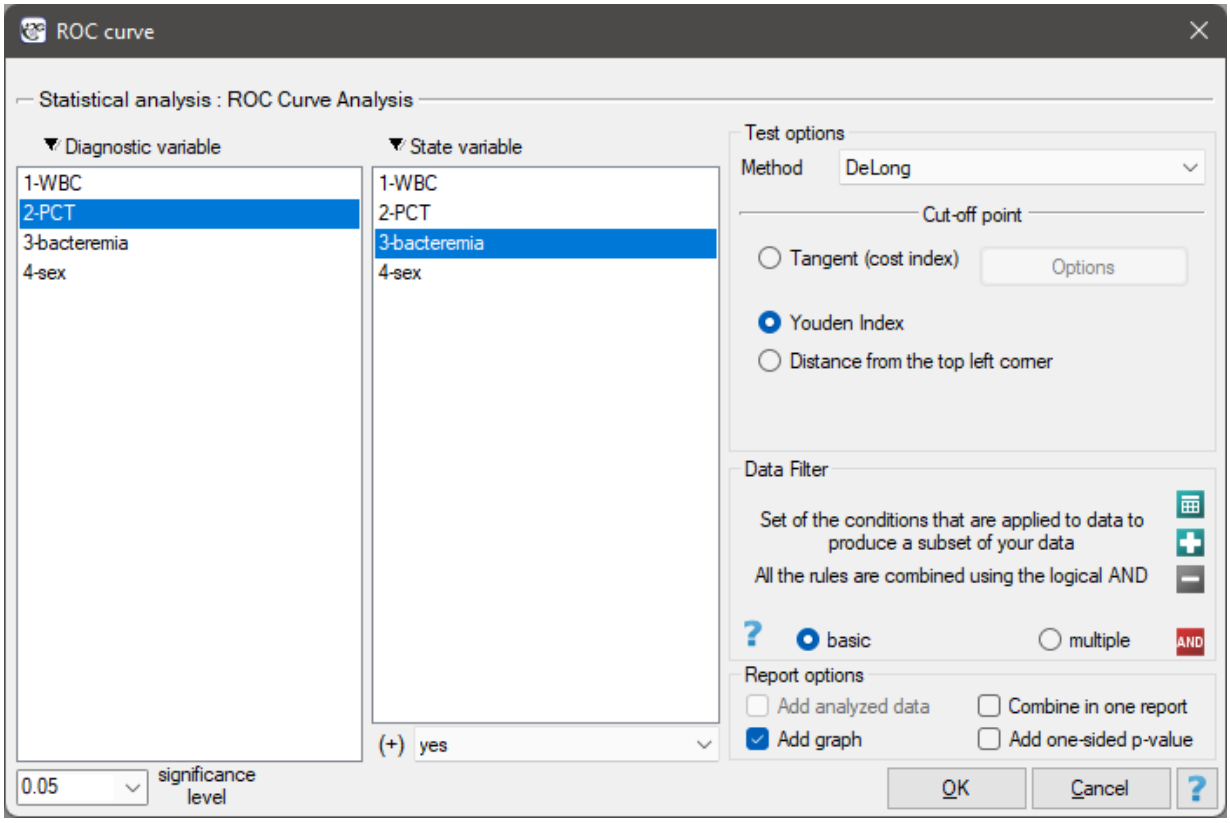
Costs graph – shows the calculated values of a wrong diagnosis together with their costs. The point marked on the graph is the minimum of the cost function.

Sensitivity and specificity intersection graph – lets you locate the point at which the values of sensitivity and specificity are simultaneously greatest.

HOW TO SET UP THE ANALYSIS

Advanced statistics → Diagnostic tests → ROC curve.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 24.2. (bacteremia.pqs file)
Persistent high fever in an infant or young child without a clear, established cause is an indication for testing for bacteremia. The parameters considered most useful and reliable for screening and monitoring bacterial infections are:

WBC – white blood cell count,

PCT – procalcitonin.

It is assumed that in a healthy infant or young child, WBC should not exceed 15 thousand/µl, and PCT should be below 0.5 ng/ml. Example values of these indicators for 136 children up to 3 years of age with persistent fever $> 39^0C$ are shown in the table excerpt below:

WBC	PCT	bacteremia	sex
11.3	0.023	no	f
11	0.022	no	f
6.7	0.009	no	f
5.9	0.004	no	f
6.1	0.006	no	f
12.5	0.031	no	f
4.9	0.002	no	f
6.9	0.011	no	f
11.6	0.025	no	f
20.9	5.919	yes	f
20.8	6.405	yes	f

Copyright ©2010-2026 U-smile Stats – All rights reserved

One possible way of analysing the PCT indicator is to convert it into a dichotomous variable by choosing a cut-off (e.g. $x_{cat} = 0.5$ ng/ml), above which the test is considered "positive". How good this split is will be shown by the sensitivity and specificity values. We want, however, to use a more comprehensive approach – that is, to calculate sensitivity and specificity not just for a single value, but for every PCT value obtained in the sample – in other words, to build a ROC curve. On the basis of the information obtained this way, we want to check whether the PCT indicator is genuinely useful for diagnosing bacteremia, and if so, what the optimal cut-off is, above which we can consider the test "positive"– detecting bacteremia.

Since bacteremia is accompanied by an elevated PCT level, we treat this indicator as a stimulant. In the state variable, we must specify which value in the bacteremia column indicates its presence – here we select the value "yes". In the report, besides the result of the statistical test, we can find a detailed description of every possible cut-off.

ROC Curve Analysis	
Analysed variables	PCT
	bacteremia
Number of unspecified	0
Number of missing data	3
Significance level	0.05
Size	133
Size STATE + (yes)	33
Size STATE - (no)	100
Direction of diagnostic variable	stimulant
Prevalence	0.2481
-95% CI	0.1774
+95% CI	0.3304
DeLong's method	
AUC	0.8892
SE(AUC)	0.0481
-95% CI	0.7949
+95% CI	0.9836
Z statistic	6.6914
Two sided p-value	<0.0001

PCT	STATE -	STATE +	FP	TP	TN	FN	Sensitivity	Specificity	PPV	NPV	LR+	LR-	ACC
0.001	3	0	100	33	0	0	1	0	0.24812030	NA	1	NA	0.24812030
0.002	3	0	97	33	3	0	1	0.03	0.25384615	1	1.030927835	0	0.27067669
0.003	3	1	94	33	6	0	1	0.06	0.25984251	1	1.06382978	0	0.29323308
0.004	2	0	91	32	9	1	0.96969696	0.09	0.26016260	0.9	1.06560106	0.33670033	0.30827067
0.005	1	0	89	32	11	1	0.96969696	0.11	0.26446280	0.91666666	1.08954715	0.27548209	0.32330827
0.006	1	0	88	32	12	1	0.96969696	0.12	0.26666666	0.92307692	1.10192837	0.25252525	0.33082706
0.007	3	1	87	32	13	1	0.96969696	0.13	0.26890756	0.92857142	1.11459421	0.23310023	0.33834586
0.008	1	0	84	31	16	2	0.93939393	0.16	0.26956521	0.88888888	1.11832611	0.37878787	0.35338345
0.009	2	0	83	31	17	2	0.93939393	0.17	0.27192982	0.89473684	1.13179992	0.35650623	0.36090225
0.011	2	1	81	31	19	2	0.93939393	0.19	0.27678571	0.90476190	1.15974560	0.31897926	0.37593984
0.012	3	0	79	30	21	3	0.90909090	0.21	0.27522935	0.875	1.15074798	0.43290043	0.38345864
0.013	2	0	76	30	24	3	0.90909090	0.24	0.28301886	0.88888888	1.19617224	0.37878787	0.40601503
0.014	2	0	74	30	26	3	0.90909090	0.26	0.28846153	0.89655172	1.22850122	0.34965034	0.42105263
0.015	2	1	72	30	28	3	0.90909090	0.28	0.29411764	0.90322580	1.26262626	0.32467532	0.43609022
0.016	3	0	70	29	30	4	0.87878787	0.3	0.29292929	0.88235294	1.25541125	0.40404040	0.44360902
0.017	2	0	67	29	33	4	0.87878787	0.33	0.30208333	0.89189189	1.31162369	0.36730945	0.46616541
0.018	2	0	65	29	35	4	0.87878787	0.35	0.30851063	0.89743589	1.35198135	0.34632034	0.48120300
0.019	2	0	63	29	37	4	0.87878787	0.37	0.31521739	0.90243902	1.39490139	0.32760032	0.49624060

Copyright ©2010-2026 U-smile Stats – All rights reserved

$AUC = 0.889$, $p < 0.000001$ **(statistically significant,** $\alpha = 0.05$**)**

Diagnosing bacteremia using the PCT indicator is significantly more effective than a random split of patients into 2 groups.

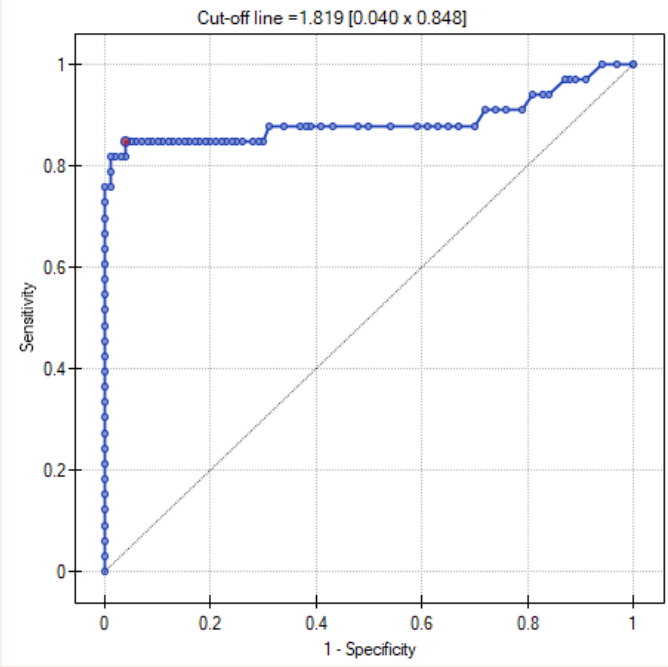
We therefore return to the analysis (the button) to determine the optimal cut-off. The algorithm for finding the optimal cut-off takes into account the costs of wrong decisions and the prevalence:

(1) FN cost – wrong diagnosis is the cost of concluding that the patient does not have bacteremia when they genuinely do (the cost of a false negative decision);

(2) FP cost – wrong diagnosis is the cost of concluding that the patient has bacteremia when they genuinely do not (the cost of a false positive decision).

Since FN costs are much more serious than FP costs, we enter a greater value in the first field than in the second – we decided this would be 5. The PCT value is to be used in screening studies, so we do not provide the population prevalence (the a priori prevalence), which is very low, but instead keep the prevalence estimated from the sample. We do this so as not to move the PCT cut-off too high and increase the number of false negative results.

For cut-off	
Cut-off point method	Tangent (cost index)
Cut-off point	1.819
Cost FN - wrong diagnosis	5
Cost FP - wrong diagnosis	1

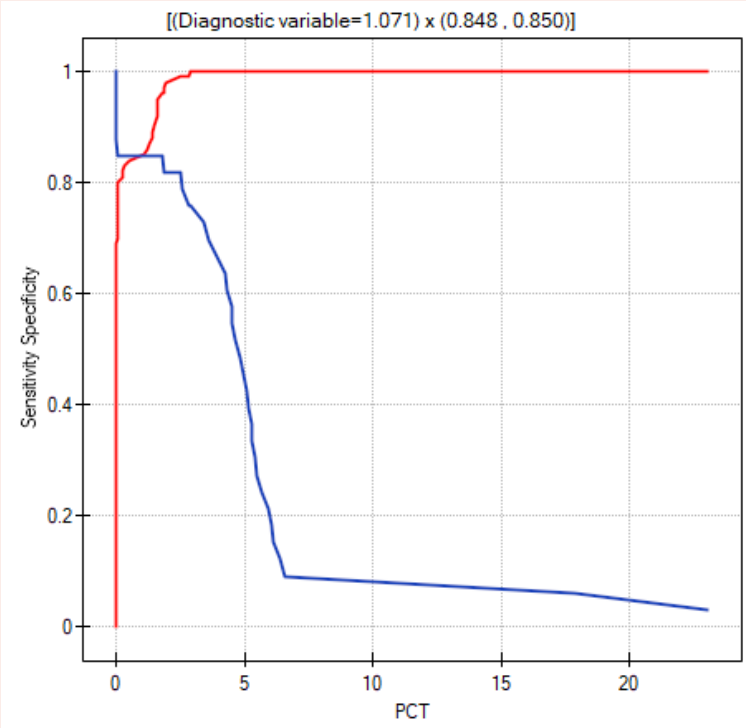


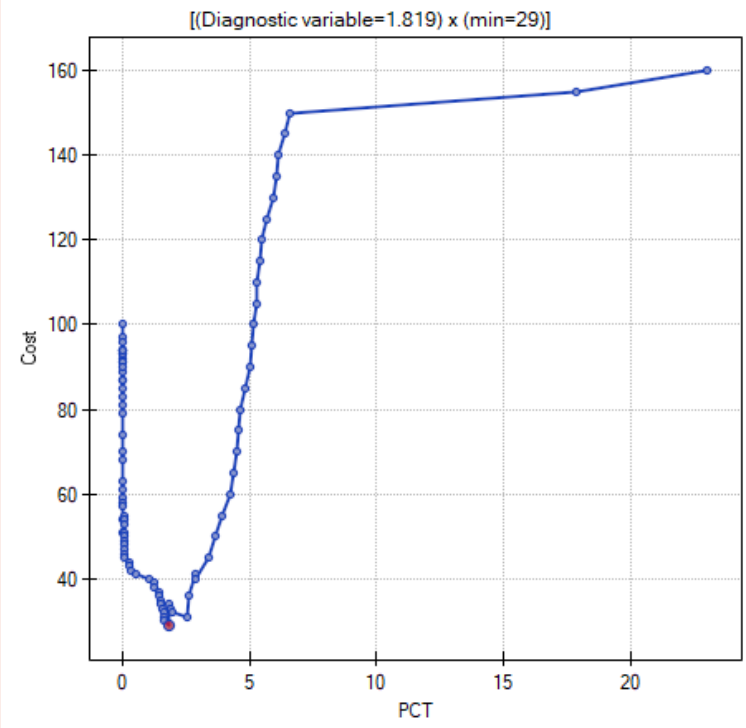
optimal PCT cut-off = 1.819

For this point, sensitivity=0.85 and specificity=0.96.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Another way of choosing the cut-off is to analyse the costs graph and the sensitivity-specificity intersection graph:





The costs graph shows that the minimum cost of wrong decisions falls at PCT=1.819. The sensitivity and specificity values, on the other hand, are similar at PCT=1.071.

Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

The tangent method (cost index)
The optimal cut-off is taken to be the value of the diagnostic variable at which the expression:

$$\text{Sensitivity} - m \cdot (1 - \text{Specificity})$$

reaches its minimum (Zweig M.H., 1993[188]). The slope angle m is defined with reference to the costs of wrong decisions and the prevalence. By default, the costs of wrong decisions equal 1, and the prevalence is estimated from the sample. Knowing the prevalence a priori ($P_{apriori}$) and the costs of wrong decisions, the user can influence the value of m, and thereby the search for the optimal cut-off.

Youden's index

$$d = \text{Sensitivity} + \text{Specificity} - 1$$

Distance to the top-left corner

$$d = \sqrt{(1 - \text{Sensitivity})^2 + (1 - \text{Specificity})^2}$$

Costs graph
The values of a wrong diagnosis are calculated according to the formula:

$$cost = cost_{FP} \cdot FP + cost_{FN} \cdot FN$$

ROC curves comparison

WHEN TO USE:

Very often the aim of a study is to compare the size of the area under one ROC curve (AUC_1) with the area under another ROC curve (AUC_2). A ROC curve with a larger area usually allows more accurate classification of objects. The methods used to compare the areas depend on the study model.

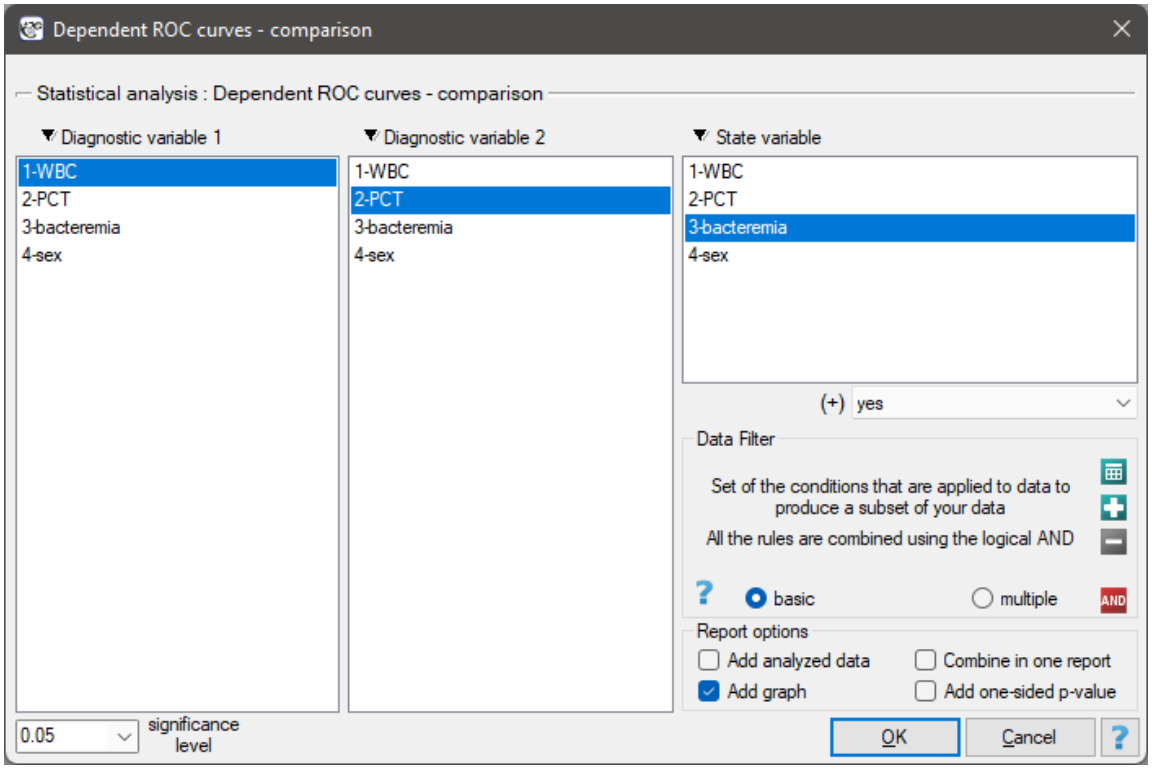
ASSUMPTIONS

- **dependent model** – the ROC curves compared are built from measurements made on the same objects,
- **independent model** – the ROC curves compared are built from measurements made on different objects.

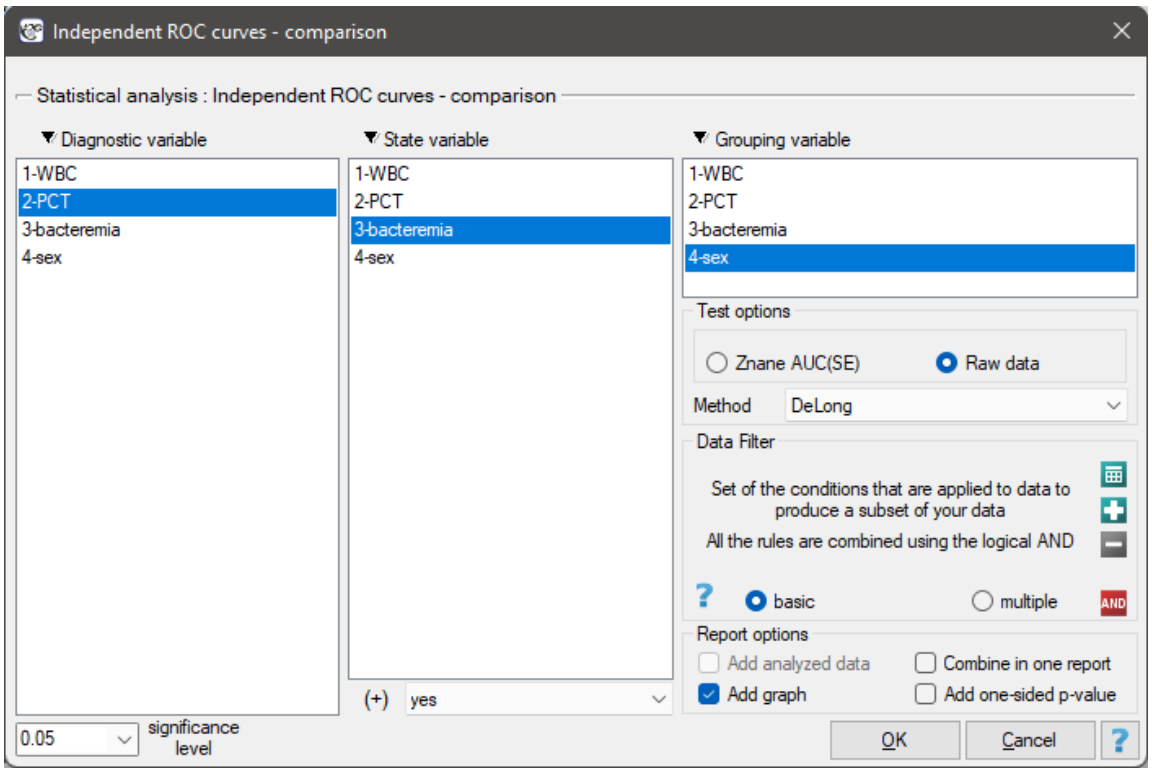
HOW TO SET UP THE ANALYSIS

For the dependent model: Advanced statistics → Diagnostic tests → Dependent ROC Curves – comparison.

Copyright ©2010-2026 U-smile Stats – All rights reserved



For the independent model: Advanced statistics → Diagnostic tests → Independent ROC Curves – comparison.



PERFORMING AND INTERPRETING THE ANALYSIS

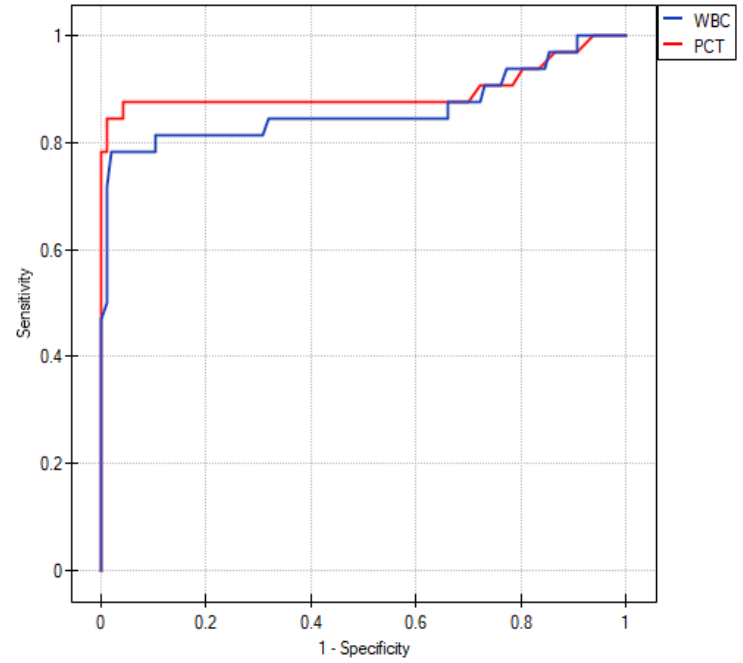
Copyright ©2010-2026 U-smile Stats – All rights reserved

Continuation of Example 24.2 (bacteremia.pqs file): we will perform 2 comparisons – (1) we will build 2 ROC curves to compare the diagnostic value of the WBC and PCT parameters (dependent model – the same children); (2) we will build 2 ROC curves to compare the diagnostic value of the PCT parameter for boys and girls (independent model – different children).

(1) WBC vs PCT (dependent model)
Both WBC and PCT are stimulants (high values of these parameters accompany bacteremia).

Dependent ROC curves - comparison	
Analysed variables	WBC
	PCT
	bacteremia
Number of unspecified	0
Number of missing data	7
Significance level	0.05
Grouping variable	bacteremia
Size	129
Size STATE + (yes)	32
Size STATE - (no)	97
Variable WBC	
Direction of diagnostic variable	stimulant
AUC	0.8613
SE(AUC)	0.0517
-95% CI	0.7599
+95% CI	0.9627
Variable PCT	
Direction of diagnostic variable	stimulant
AUC	0.8956
SE(AUC)	0.049
-95% CI	0.7996
+95% CI	0.9917
DeLong's method	
AUC1-AUC2	0.0343
SE(AUC1-AUC2)	0.0227
-95% CI	0
+95% CI	0.0788
Z statistic	1.5128
Two sided p-value	0.1303

Copyright ©2010-2026 U-smile Stats – All rights reserved



$AUC_{WBC} = 0.8613$, $AUC_{PCT} = 0.8956$, $p = 0.130322$ **(not statistically significant,** $\alpha = 0.05$**)**

We cannot determine which of the parameters – WBC or PCT – is better for diagnosing bacteremia.

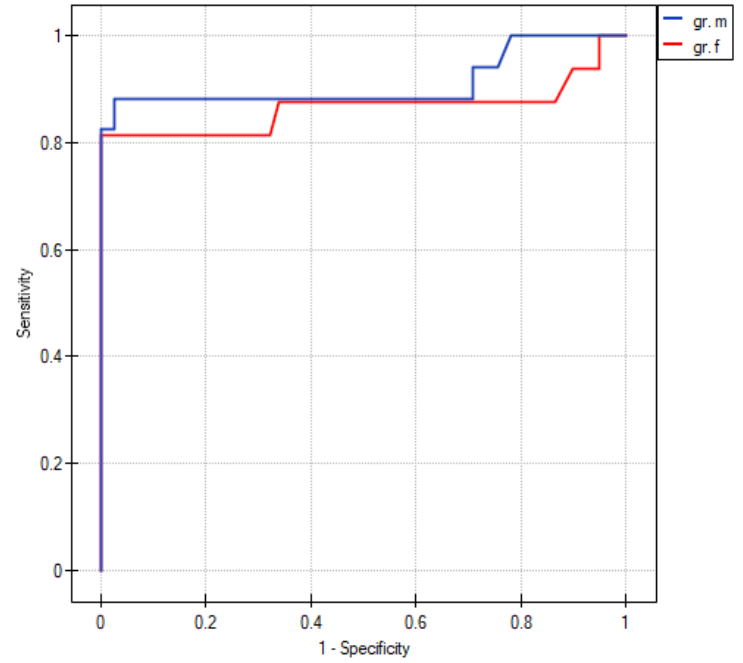
(2) PCT: girls vs boys (independent model)

The PCT parameter is a stimulant (its high values accompany bacteremia); we compare its diagnostic value separately for girls and boys.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Independent ROC curves - comparison	
Analysed variables	PCT
	bacteremia
Number of unspecified	0
Number of missing data	1
Significance level	0.05
Grouping variable	sex
Direction of diagnostic variable	stimulant
Method	DeLong
Group name	m
Size	58
Size STATE + (yes)	17
Size STATE - (no)	41
AUC	0.9118
SE(AUC)	0.0599
-95% CI	0.7943
+95% CI	1
Group name	f
Size	75
Size STATE + (yes)	16
Size STATE - (no)	59
AUC	0.8649
SE(AUC)	0.0792
-95% CI	0.7098
+95% CI	1
AUC1-AUC2	0.0468
SE(AUC1-AUC2)	0.0993
Z statistic	0.4717
p-value	0.6372

Copyright ©2010-2026 U-smile Stats – All rights reserved



$AUC_f = 0.8649$, $AUC_m = 0.9118$, $p = 0.637176$ **(not statistically significant,** $\alpha = 0.05$**)**

We cannot determine the sex for which the PCT parameter is better for diagnosing bacteremia.

FOR THOSE INTERESTED

Hypotheses (common to both models):

$$\begin{aligned} \mathcal{H}_0: &\quad AUC_1 = AUC_2, \\ \mathcal{H}_1: &\quad AUC_1 \neq AUC_2. \end{aligned}$$

Dependent model

The test statistic is defined by:

$$Z = \frac{|AUC_1 - AUC_2|}{SE_{AUC_1 - AUC_2}},$$

where AUC_1, AUC_2, and the standard error of the difference in areas $SE_{AUC_1 - AUC_2}$ are calculated on the basis of the nonparametric method proposed by **DeLong** (DeLong E.R. et al., 1988[52]; Hanley J.A. and Hajian-Tilaki K.O., 1997[76]).

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

Independent model

The test statistic (Hanley J.A. and McNeil M.D., 1983[78]) is defined by:

$$Z = \frac{|AUC_1 - AUC_2|}{\sqrt{SE_{AUC_1}^2 - SE_{AUC_2}^2}},$$

where AUC_1, AUC_2, and the standard errors of the areas SE_{AUC_1}, SE_{AUC_2} are calculated on the basis of:

Copyright ©2010-2026 U-smile Stats – All rights reserved

- ⋆ the nonparametric **DeLong** method (DeLong E.R. et al., 1988[52]; Hanley J.A. and Hajian-Tilaki K.O., 1997[76]) – **recommended**,
- ⋆ the nonparametric **Hanley-McNeil** method (Hanley J.A. and McNeil M.D., 1982[77]).

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

For both models: the p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

25 MULTIDIMENSIONAL MODELS

Matching groups

WHEN TO USE:

If we estimate the treatment effect from a **fully randomized experiment**, then by randomly assigning subjects to the treated and untreated groups, we create **groups that are similar in terms of possible confounding factors**. The similarity of the groups is due to the random assignment itself. In such studies, we can examine the pure (independent of confounding factors) effect of the treatment method on the outcome of the experiment. In this case, matching the groups by any means other than random assignment is not necessary.

The possibility of error arises when the difference in treatment outcome between the treated and untreated groups may be due not to the treatment itself, but to a factor that induced people to take part in the treatment. This happens when, for some reason, randomization is not possible, e.g. it is an observational study, or for ethical reasons we cannot assign treatment arbitrarily. In such cases, artificial group matching may be applicable. For example, if the people we assign to the treatment group are healthier, and the people in the control group have a more severe form of the disease, then it is not the treatment method itself but the patient's condition before treatment that may affect the outcome of the experiment. When we see such an imbalance between groups, it is good if we can decide to randomize – this solves the problem, since randomly assigning people to groups makes them similar. However, we can imagine a different situation. This time, the group we are interested in will not be people undergoing treatment but smokers, and the control group will be non-smokers, and the analyses will aim to show the adverse effect of smoking on the occurrence of lung cancer. Then, wanting to check whether smoking genuinely increases the risk of developing the disease, it would be unethical to conduct a fully randomized trial, since it would mean forcing people randomly selected into the risk group to smoke. The solution to this situation is to establish an exposed group, i.e. to draw a certain number of people from among those who already smoke, and then select a control group made up of non-smokers. The control group should be selected, since leaving the selection to chance may result in a non-smoking group that is younger than the smokers merely because smoking is becoming less fashionable in our country, so there are automatically many young people among non-smokers. The control group should be drawn from non-smokers, but in such a way that it is **as similar as possible to the treatment group**. In this way, we come closer to examining the pure (independent of selected confounding factors, such as age) effect of smoking/non-smoking on the outcome of the experiment, i.e. in this case the occurrence of lung cancer. **Such a selection can be made precisely through the matching offered by the program**.

One of the main advantages of researcher-controlled matching is that the control group becomes more similar to the treatment group, but this is also the greatest weakness of this method. It is an advantage because our study increasingly resembles a randomized study – in a randomized trial, the study and control groups are similar with respect to almost all features, including those we are not studying (the random assignment provides this similarity). As a result of applying researcher-controlled matching, the treatment and control groups become similar only with respect to selected features.

ASSUMPTIONS

Copyright ©2010-2026 U-smile Stats – All rights reserved

- the features with respect to which matching is performed should be measured for all objects in both groups,
- the group from which the match is drawn (usually the control group) should be sufficiently large and contain objects genuinely similar to the treatment group.

DEFINITIONS OF MEASURES

The first two methods listed below are based on matching groups through **Propensity Score Matching (PSM)**. This type of matching was proposed by Rosenbaum and Rubin[142]. In practice, it is a technique for matching a control group (untreated or minimally/standardly treated subjects) to a treatment group, based on a probability describing the subjects' propensity to be assigned to treatment depending on the observed associated variables. The probability score describing this propensity, referred to as the *Propensity Score*, is a balancing score, so that as a result of matching the control group to the treatment group, the distribution of the measured associated variables becomes more similar between treated and untreated subjects. The third method does not determine the probability for each individual, but determines a distance/dissimilarity matrix that indicates the objects that are closest/most similar with respect to multiple selected features.

Given probability – i.e. a value between 0 and 1 given for each person studied, describing the probability of belonging to the treatment group, i.e. the Propensity Score. Such a probability can be determined beforehand using various methods, e.g. a logistic regression model, neural networks, or many other methods. If a person from the group from which we draw the control obtains a Propensity Score similar to that obtained by a person from the treatment group, they can enter the analysis, since the two individuals are similar with respect to the features that were considered when determining the Propensity Score.

Determined from a logistic regression model – since logistic regression is the most commonly used matching method, PQStat provides the ability to automatically determine the Propensity Score value in the analysis window based on this method. The matching then proceeds using the Propensity Score obtained in this way.

Similarity/distance matrix – with this option, no Propensity Score value is determined; instead, a matrix is built indicating the distance of each person in the treatment group to a person in the control group. The user can set boundary conditions, e.g. specify that the person matched to a given person from the study group cannot differ from them by more than 3 years of age and must be of the same sex. Distances in the constructed matrix are determined based on any metric or method describing dissimilarity. This way of matching the control group to the treated group is very flexible – besides the free choice of how distances/dissimilarity are determined, many metrics allow the indication of weights specifying how important each variable is to the researcher. However, great caution is advised when choosing a distance/dissimilarity matrix – many features and many ways of determining distances require prior standardization or normalization of the data, and choosing the inverse of distance or similarity (instead of dissimilarity) may result in finding the most distant and dissimilar objects, whereas we normally use these methods to find similar objects. If the researcher has no specific reason to change the metric, it is standard practice to use statistical distance, i.e. the Mahalanobis metric – it is the most universal, does not require prior standardization of the data, and is robust to correlation between variables. A more detailed description of the distances and dissimilarity/similarity measures available in the program, and how to interpret the results obtained, can be found in the Similarity matrix section.

In practice, there are many methods for indicating how close the compared objects are. Two are pro-

Copyright ©2010-2026 U-smile Stats – All rights reserved

posed in the program:

Nearest neighbour method – a standard way of selecting objects, not only with a similar Propensity Score, but also those whose distance/dissimilarity in the matrix is smallest.

Nearest neighbour method, closer than... – works the same way as the nearest neighbour method, with the difference that only objects that are close enough can be matched. The limit of this closeness is set by giving a value describing the threshold beyond which objects are so dissimilar to the objects studied that we do not want to give them a chance to join the newly built control group. When the analysis is based on the Propensity Score or on a matrix defined by dissimilarity, the most dissimilar objects are those at a distance of 1, and the most similar are those at a distance of 0 – when choosing this method, we should give a value closer to 0 when matching more restrictively, or closer to 1 when the threshold is placed further away. When distances rather than dissimilarities are determined in the matrix, the minimum value is still 0, but the maximum value is not predetermined.

We can match without replacing objects already drawn, or with replacement of these objects back into the group from which we are drawing.

Matching without replacement – when matching without replacement is used, once an untreated person has been selected to be matched to a given treated person, that untreated person is no longer available for consideration as a potential match for subsequent treated persons. As a result, each untreated person is included in at most one matched set.

Matching with replacement – matching with replacement allows a given untreated person to be included more than once in a single matched set. When matching with replacement is used, subsequent analyses, in particular the estimation of variance, must take into account the fact that the same untreated person may appear in multiple matched sets.

Note! When it is not possible to uniquely match an untreated person to a treated one, because in the group from which we are choosing there is more than one person matching the treated person equally well, one of these persons is joined, chosen at random. For a repeated analysis, a fixed *seed* is set by default, so the results of a repeated draw will be the same; however, when the analysis is performed anew, the seed is changed and the result of the draw may be different.

When it is not possible to match an untreated person to a treated one, because there are no more people left to join in the group from which we are choosing (e.g. matching persons have already been joined to other treated persons, or the set from which we are choosing has no similar persons), that person remains without a pair.

Most often, **1:1** matching is performed, i.e. for one treated person, one untreated person is matched. However, if the original control group from which we draw is large enough and we need to draw more individuals, **1:k** matching can be chosen, where k indicates the number of individuals that should be matched to each treated individual.

Evaluation of the matching

After matching the control group to the treatment group, the results of such matching can be returned to the worksheet, i.e. a new control group can be obtained. However, it should not be assumed that applying matching will always yield satisfactory results – in many situations, the group from which we draw does not have enough objects that are sufficiently similar to the treatment group. Therefore, the matching performed should always be evaluated. There are many methods for evaluating the matching of groups. The program uses methods based on the standardized difference between groups, described more extensively in, among others, the work of P.C. Austin[13][14]. This approach allows comparison of the relative balance of variables measured in different units, and its result is not affected by sample size. The estimation of agreement using statistical tests was abandoned, since the matched control group is

Copyright ©2010-2026 U-smile Stats – All rights reserved

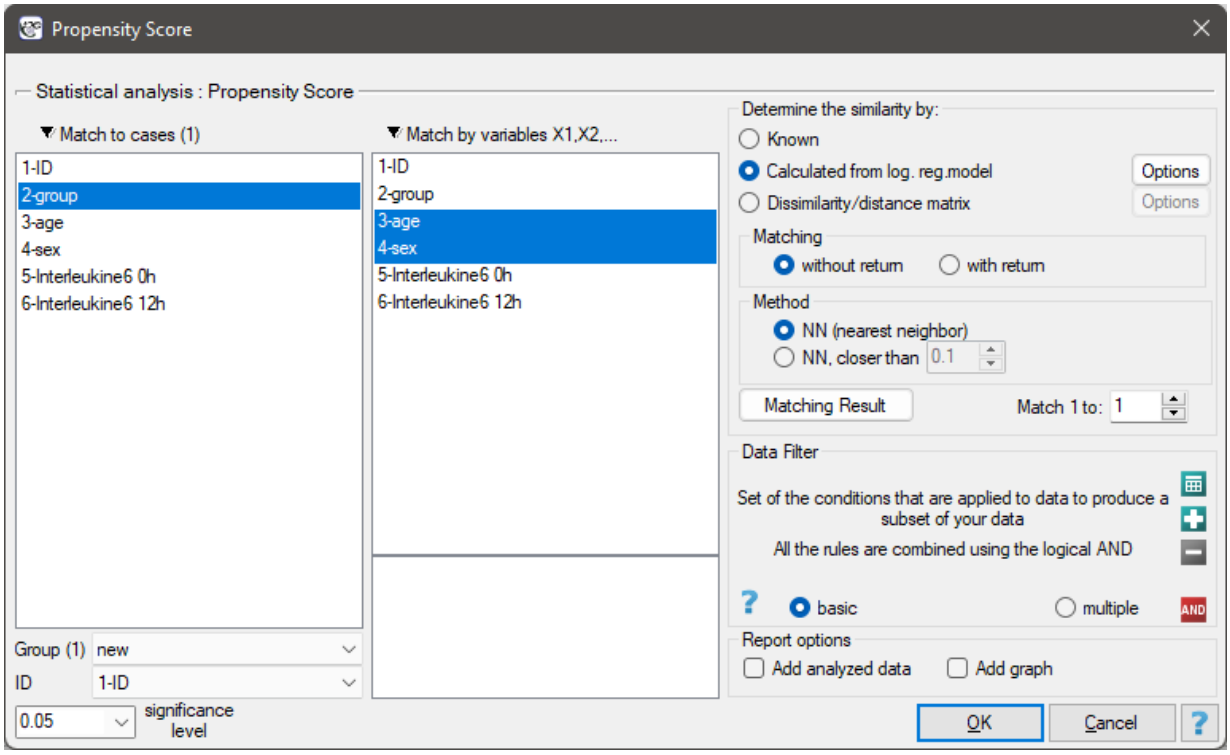
usually much smaller than the original control group, so the resulting p values of tests comparing the study group to the smaller control group more often remain consistent with the null hypothesis, and thus do not show significant differences due to the reduced sample size.

Note! Although there is no universally agreed criterion for what threshold of standardized difference should be used to indicate a significant imbalance, a standardized difference of less than 0.1 (for both means and frequencies) can be used as guidance[130]. Therefore, to conclude that the groups are well matched, we should observe standardized differences close to 0, and preferably not falling outside the range from -0.1 to 0.1. Graphically, these results are presented on a scatter plot. Negative differences indicate lower means/frequencies in the treatment group, positive ones in the control group.

The 1:1 matching obtained in reports means the summary for the study group and the corresponding control group obtained in the first matching; the 1:2 matching means the summary for the study group and the corresponding control group obtained in the first + second matching (i.e. it does not refer to the study group and the corresponding control group obtained only in the second matching), etc.

HOW TO SET UP THE ANALYSIS

Advanced statistics → Multivariate models → Matching groups.

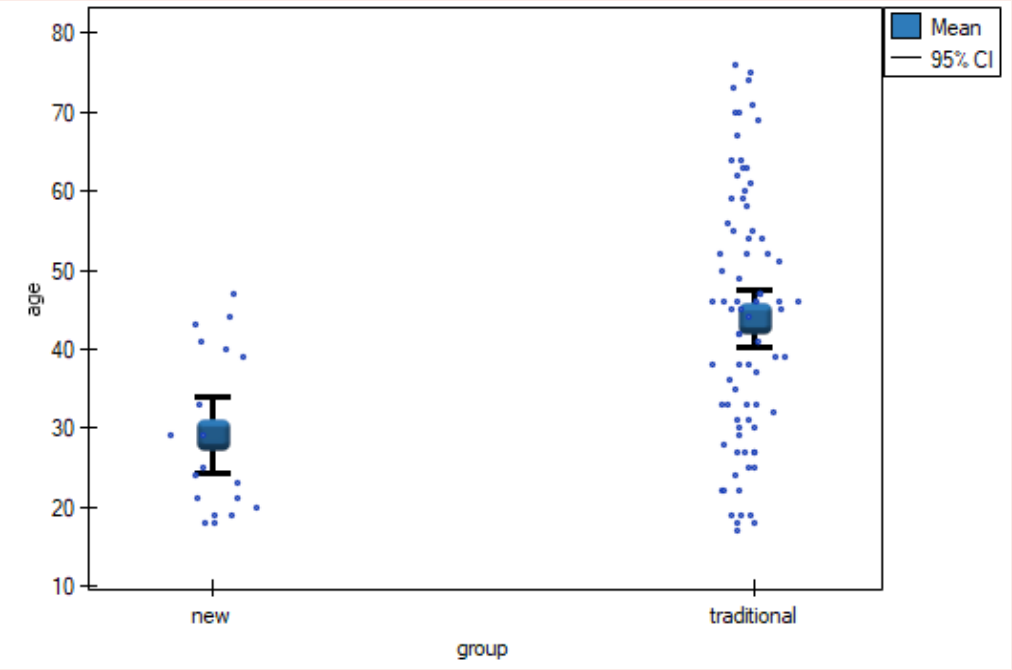


PERFORMING AND INTERPRETING THE ANALYSIS

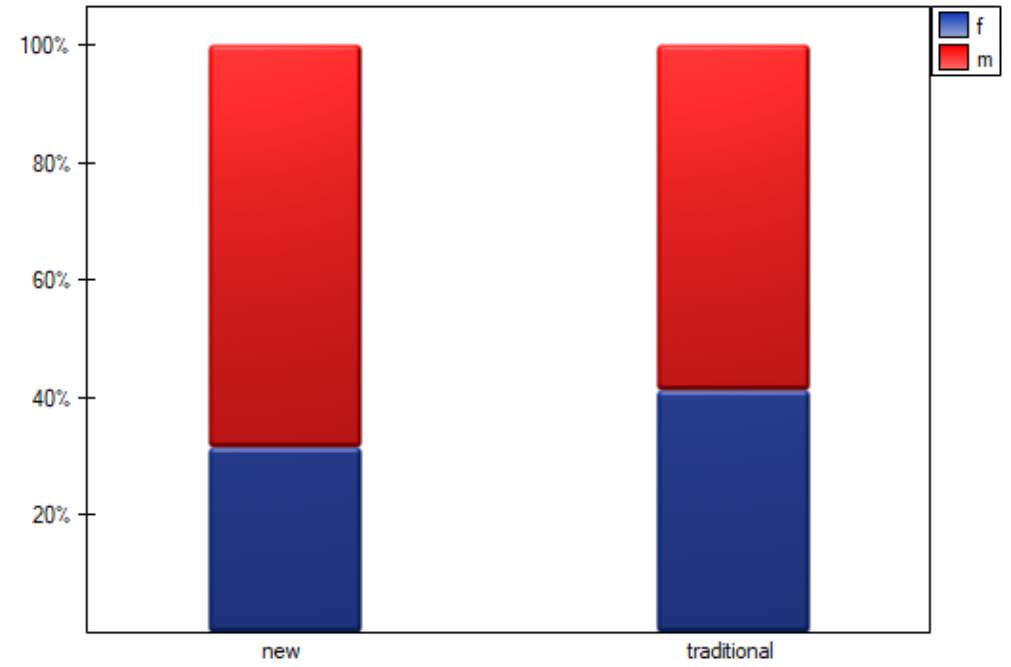
EXAMPLE 25.1. (matching.pqs file)
We want to compare two ways of treating patients after accidents: the traditional way and a new way.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The correct effect of both treatments should be observed as a decreasing level of selected cytokines. To compare the effectiveness of these two treatments, both should be carried out on patients who are quite similar – then we can be sure that any differences in the effectiveness of these methods will result from the treatment effect itself and not from other differences between patients assigned to different groups. The study is conducted a posteriori, that is, it is based on data collected from patients' treatment histories. Therefore, the researchers had no influence on the assignment of patients to the group treated with the new drug and the group treated traditionally. It was noted that the traditional treatment was mainly prescribed to older patients, while the new treatment was prescribed to younger patients, in whom it is easier to lower cytokine levels. The groups were fairly similar in terms of sex structure, but not identical.



Copyright ©2010-2026 U-smile Stats – All rights reserved



If the planned study had been carried out on such selected groups of patients, the new treatment would have had an easier task, since younger organisms might respond better to treatment. The conditions of the experiment would not be equal for both treatments, which could distort the results of the analyses and the conclusions drawn. It was therefore decided to select the traditionally treated group so that it would be similar to the study group treated with the new method.

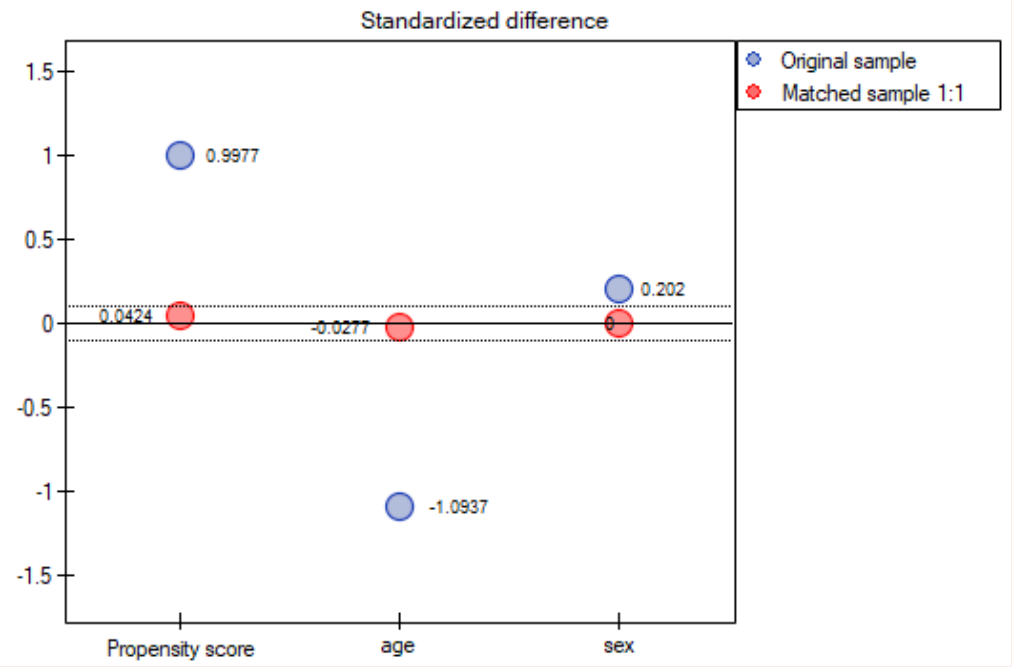
We plan to perform the matching with respect to two features, i.e. age and sex. The traditionally treated group is larger (80 patients) than the group treated with the new drug (19 patients), so there is a good chance that we will be able to select groups that are similar. The random selection is performed using the logistic regression model algorithm built into PSM. We remember that sex must be coded numerically, since only numerical values are used in logistic regression analysis. We choose the nearest neighbour as the method. We want the same person to be unable to be selected twice, so we choose matching without replacement. We will try 1:1 matching, i.e. for each person treated with the new drug we will match one person treated traditionally. Remember that the selection is random, and so depends on the random *seed* value set by our computer, so the draw performed by the reader may differ from the values presented here.

We can view the summary of the selection in tables and graphs.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Original sample					
	cases	control	difference	SE of differe	std. differen
Propensity	0.3073	0.1645	0.1428	0.1431	0.9977
age	29.1053	43.9125	-14.8072	13.5386	-1.0937
sex	0.6842	0.5875	0.0967	0.4788	0.202
Matching 1:1					
Propensity	0.3073	0.3015	0.0057	0.1353	0.0424
age	29.1053	29.3684	-0.2632	9.493	-0.0277
sex	0.6842	0.6842	0	0.4648	0

In the original sample, the mean age is more than 14 years higher in traditionally treated patients (the difference between the means is 14.8072), while the sex structure differs by less than 10% (0.0967). Much smaller differences are observed between patients treated with the new method and the traditionally treated patients matched to them. We obtain the most information about the quality of the matching from the standardized differences (the last column of the table and the graph).



standardized difference in Propensity Score = 0.0424, standardized difference in age = -0.0277, standardized difference in the proportion of sexes = 0

After matching, the groups become well balanced – all standardized differences fall within the range from -0.1 to 0.1.

The line at 0 indicates group equilibrium (a difference between groups equal to 0). When the groups are in equilibrium with respect to the given features, all points on the graph lie close to this line, i.e. around the interval from -0.1 to 0.1. In the case of the original sample (blue colour), we see a considerable

Copyright ©2010-2026 U-smile Stats – All rights reserved

departure of the Propensity Score – as we know, this mismatch results mainly from the age mismatch (its standardized difference lies at a large distance from 0), and to a lesser extent from the sex mismatch. By performing the matching, we obtained groups more similar to each other (red colour on the graph). The age of both groups is now similar – the traditionally treated group differs from the group treated with the new method by less than a year on average (the difference between the means shown in the table is 0.2632). In the case of sex, the match is perfect, i.e. the proportion of women and men is the same in both groups. We can return the data prepared in this way to the worksheet and subject it to the analyses we have planned.

Looking at the summary we have just obtained, we can see that despite the good balancing of the groups and the perfect matching of many individuals, there are individual people who are not as similar to each other as we might expect.

cases				control 1			
ID	Propensity	age	sex	ID	Propensity	age	sex
1	0.3671	20	0	80	0.3566	27	1
2	0.1745	33	0	5	0.1745	33	0
3	0.1575	41	1	68	0.1575	41	1
4	0.1681	40	1	7	0.1745	33	0
5	0.4689	21	1	56	0.4496	22	1
6	0.3853	19	0	3	0.3853	19	0
7	0.4689	21	1	11	0.4496	22	1
8	0.4116	24	1	66	0.4227	17	0
9	0.527	18	1	24	0.527	18	1
10	0.3929	25	1	46	0.3929	25	1
11	0.2239	29	0	59	0.2376	28	0
12	0.1792	39	1	27	0.1909	38	1
13	0.3853	19	0	44	0.3566	27	1
14	0.105	47	1	65	0.105	47	1
15	0.129	44	1	79	0.129	44	1
16	0.138	43	1	10	0.1435	36	0
17	0.4305	23	1	16	0.4496	22	1
18	0.3218	29	1	33	0.3218	29	1
19	0.4038	18	0	71	0.3051	30	1

Sometimes, besides obtaining well-balanced groups, researchers want to determine precisely how individual people are selected, i.e. to gain greater control over the similarity of objects with respect to the Propensity Score value, or the similarity of objects with respect to the values of specific features. In such a case, if the group from which we draw is large enough, the analysis may yield results more favourable from the researcher's point of view; but if the group from which we draw lacks objects meeting our criteria, then for some individuals it will not be possible to find a match meeting our conditions.

Suppose we would like to obtain such groups whose Propensity Score (i.e. propensity to take part in the study) differs by no more than a given value. How do we determine this value? We can look at the report from the earlier analysis, where the smallest and largest distance between the drawn objects is given.

Copyright ©2010-2026 U-smile Stats – All rights reserved

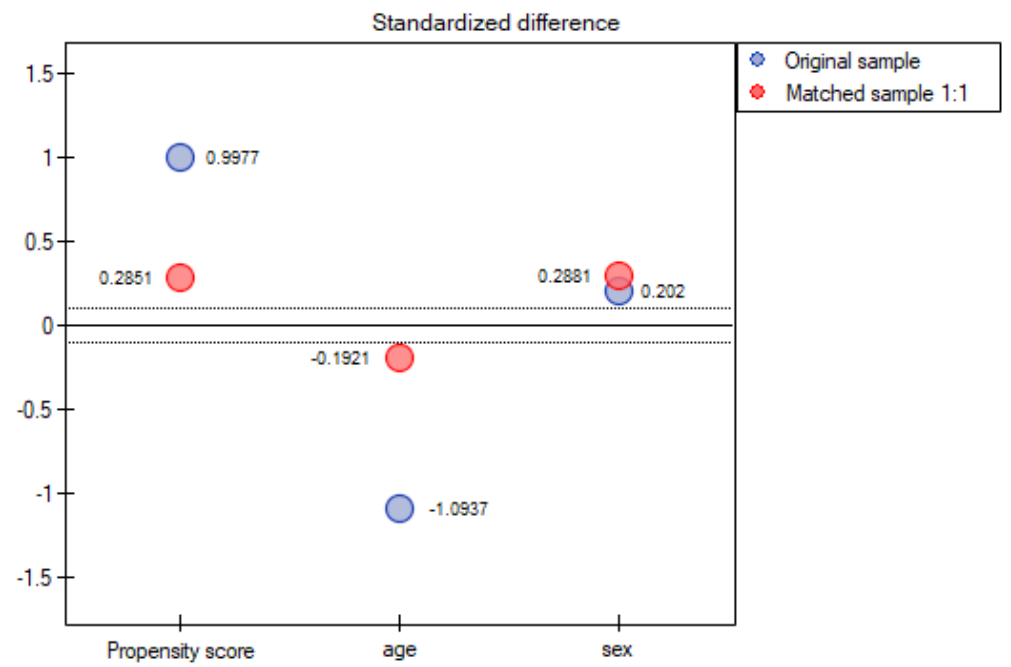
Minimum distance	0
Maksimum distance	0.5183

In our case, the closest objects differ by min=0, and the furthest by max=0.5183. Let us therefore try to check what kind of selection we obtain when we match, to people treated with the new method, people treated traditionally whose Propensity Score is very close, e.g. less than 0.01.

cases				control 1			
ID	Propensity	age	sex	ID	Propensity	age	sex
1	0.3671	20	0	no match			
2	0.1745	33	0	7	0.1745	33	0
3	0.1575	41	1	68	0.1575	41	1
4	0.1681	40	1	5	0.1745	33	0
5	0.4689	21	1	no match			
6	0.3853	19	0	3	0.3853	19	0
7	0.4689	21	1	no match			
8	0.4116	24	1	no match			
9	0.527	18	1	14	0.527	18	1
10	0.3929	25	1	46	0.3929	25	1
11	0.2239	29	0	no match			
12	0.1792	39	1	no match			
13	0.3853	19	0	no match			
14	0.105	47	1	65	0.105	47	1
15	0.129	44	1	79	0.129	44	1
16	0.138	43	1	10	0.1435	36	0
17	0.4305	23	1	66	0.4227	17	0
18	0.3218	29	1	33	0.3218	29	1
19	0.4038	18	0	no match			

We can see that this time we failed to select the whole group. Comparing the Propensity Score for each pair (treated with the new drug and treated traditionally), we can see that the differences are indeed small. However, since the matched group is now much smaller, summarising the whole selection process, we must note that the Propensity Score, age, and sex are not close enough to the line at 0. Our attempt to improve the situation did not lead to the desired effect, and the resulting groups are not sufficiently well balanced.

Copyright ©2010-2026 U-smile Stats – All rights reserved



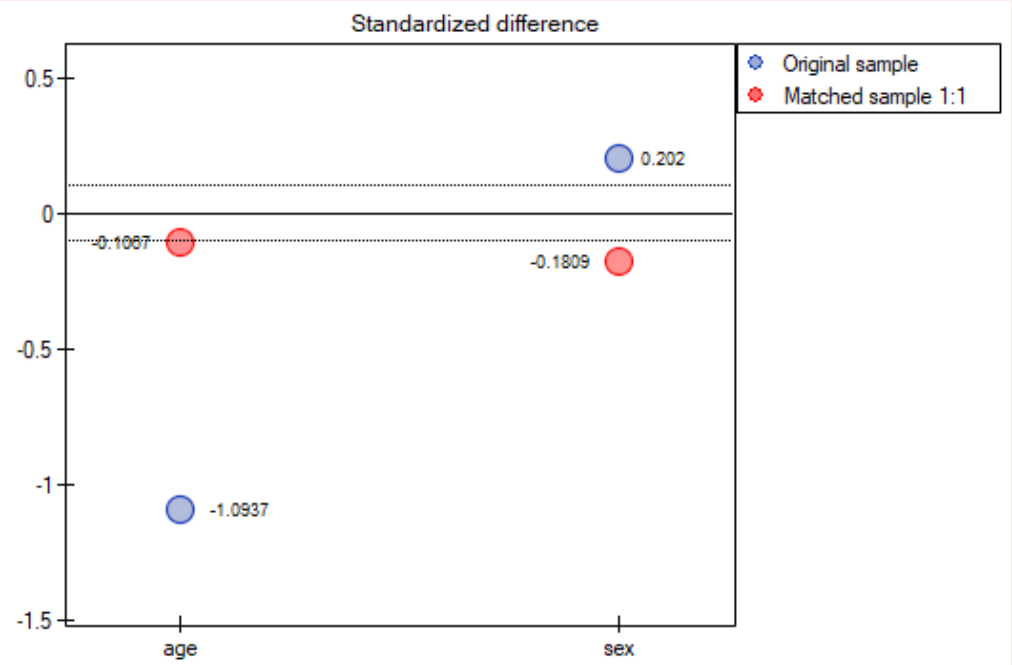
Variant 2: Mahalanobis distance matrix

Suppose we wanted to obtain pairs (people treated with the new method and people treated traditionally) who are of the same sex and whose ages do not differ by more than 3 years. In the Propensity Score-based matching, we did not have this kind of ability to control the extent of agreement of each variable. For this purpose, we will use a different method, not based on the Propensity Score but on distance/dissimilarity matrices. After selecting the Options button, we select the proposed Mahalanobis statistical distance matrix and set the neighbourhood matching to a maximum distance of 3 for age and 0 for sex. As a result, for two people no match could be found, but the remaining matches meet the set criteria.

Copyright ©2010-2026 U-smile Stats – All rights reserved

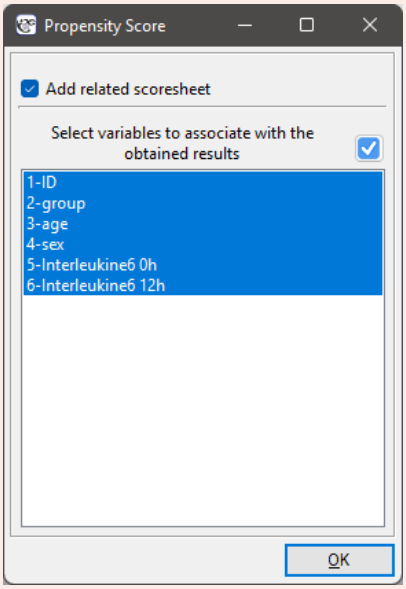
cases				control 1			
ID	Propensity	age	sex	ID	Propensity	age	sex
1		20	0	3		19	0
2		33	0	7		33	0
3		41	1	68		41	1
4		40	1	53		38	1
5		21	1	11		22	1
6		19	0	66		17	0
7		21	1	16		22	1
8		24	1	46		25	1
9		18	1	24		18	1
10		25	1	44		27	1
11		29	0	59		28	0
12		39	1	38		38	1
13		19	0	no match			
14		47	1	65		47	1
15		44	1	43		45	1
16		43	1	77		42	1
17		23	1	56		22	1
18		29	1	33		29	1
19		18	0	no match			

Summarising the whole draw, we must note that although it satisfies our assumptions, the resulting groups are not as well balanced as in our first draw, based on the Propensity Score. The points marked in red, representing the quality of the matching in terms of age and the quality of the matching in terms of sex, deviate somewhat from the line of identity set at level 0, which means the average difference in age and sex structure is now greater than in the first matching.



Copyright ©2010-2026 U-smile Stats – All rights reserved

It is up to the researcher to decide which way of preparing the data will be more advantageous for them. Ultimately, once a decision has been made, the data can be returned to a new worksheet. To do this, we go back to the report we have chosen and, in the project tree, right-click and select the Redo analysis menu. In the same analysis window, we select the Matching result button and specify which additional variables should be returned to the new worksheet.



As a result, we obtain a new data sheet in which, side by side, we find the data of people treated with the new method and the traditionally treated people matched to them.

FOR THOSE INTERESTED

Formulas

For comparing continuous variables, we determine the standardized mean difference:

$$d = \frac{(\bar{x}_{treated} - \bar{x}_{control})}{\sqrt{\frac{sd^2_{treated} + sd^2_{control}}{2}}}$$

where:

$\bar{x}_{treated}$, $\bar{x}_{control}$ – the mean value of the variable in the treatment group and the mean value of the variable in the control group,

$sd^2_{treated}$, $sd^2_{control}$ – the variance in the treatment group and the variance in the control group.

For comparing binary variables (with two categories, usually 0 and 1), we determine the standardized frequency difference:

$$d = \frac{(\hat{p}_{treated} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treated}(1-\hat{p}_{treated}) + \hat{p}_{control}(1-\hat{p}_{control})}{2}}}$$

where:

$\hat{p}_{treated}$, $\hat{p}_{control}$ – the frequency of the value described as 1 in the treatment group and the frequency of the value described as 1 in the control group.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Procedure: variables with multiple categories should be broken down, in logistic regression analysis, into dummy variables with two categories, and, by checking the match of both groups, the standardized frequency difference should be determined for them.

Multivariate regression models provide an opportunity to study the effects of multiple independent variables (multiple factors) and their interactions on a single dependent variable. Through multivariate models, it is also possible to build many simplified models at the same time – univariate (single-factor) models. Information about which model we want to build (multivariate or univariate) is visible in the window of the selected analysis. When multiple independent variables are selected simultaneously in the analysis window, it is possible to choose the model.

Preparation of variables for analysis in multivariate models

Variable coding in multivariate models

WHEN TO USE:

A challenge in preparing data for multivariate analysis is the appropriate coding of nominal and ordinal variables. This is an important element of data preparation, since it has a fundamental impact on the interpretation of the model coefficients. Nominal or ordinal variables divide the analysed objects into two or more categories, with dichotomous variables (with two categories, $k = 2$) only needing to be coded appropriately, while variables with many categories ($k > 2$) need to be broken down into dummy variables with two categories and coded.

HOW TO SET UP THE ANALYSIS

$k = 2$ If a variable is dichotomous, the researcher decides how to enter the data representing it, so any numerical codes can be entered, e.g. 0 and 1. In the program, one's own coding can be changed into effect coding by selecting this option in the window of the selected multivariate analysis. This coding replaces the smaller value with -1, and the larger value with 1.

$k > 2$ If a variable has many categories, in the window of the selected multivariate analysis we select the Dummy variables button and set the reference/base category for those variables that we want to break down into dummy variables. These variables will be dummy-coded, unless the effect coding option is selected in the analysis window – in which case they will be coded as -1, 0, and 1.

DEFINITIONS OF MEASURES

Dummy coding is used to answer, with the help of multivariate models, the question: how do the results (Y) in each analysed category differ from the results of the reference category? This coding consists in assigning a value of 0 or 1 to each category of a given variable. The category coded as 0 is then the **reference category**.

Copyright ©2010-2026 U-smile Stats – All rights reserved

$k = 2$ When the coded variable is dichotomous, placing it in a regression model gives us a calculated coefficient (b_i). This coefficient is the reference of the value of the dependent variable Y for category 1 to the reference category (adjusted for the other variables in the model).

$k > 2$ When the analysed variable has more than two categories, k categories are represented by $k - 1$ dummy variables with dummy coding. When creating dummy-coded variables, a category is selected for which no dummy variable is created. This category is treated in the model as the reference category (since in every dummy-coded variable it corresponds to a value of 0).

When the variables $X_1, X_2, ..., X_{k-1}$ obtained this way, with dummy coding, are placed in a regression model, coefficients $b_1, b_2, ..., b_{k-1}$ will be calculated for them.

b_1 is the reference of the Y results (for codes 1 in X_1) to the reference category (adjusted for the other variables in the model);

b_2 is the reference of the Y results (for codes 1 in X_2) to the reference category (adjusted for the other variables in the model);

...

b_{k-1} is the reference of the Y results (for codes 1 in X_{k-1}) to the reference category (adjusted for the other variables in the model).

Example

We will code the sex variable, with two categories (male sex will be selected as the reference category), and the education variable, with 4 categories (primary education will be selected as the reference category), using dummy coding.

Sex	Coded sex
f	1
f	1
f	1
m	0
m	0
f	1
f	1
m	0
m	0
f	1
m	0
f	1
m	0
f	1
m	0
m	0
...	...

Education	Coded education		
	vocational	secondary	tertiary
primary	0	0	0
primary	0	0	0
primary	0	0	0
vocational	1	0	0
vocational	1	0	0
vocational	1	0	0
vocational	1	0	0
secondary	0	1	0
secondary	0	1	0
secondary	0	1	0
secondary	0	1	0
tertiary	0	0	1
tertiary	0	0	1
tertiary	0	0	1
tertiary	0	0	1
tertiary	0	0	1
...	...	...	...

Building on the basis of dummy variables, in a multiple regression model, we might want to check how these variables affect a certain dependent variable, e.g. Y = the amount of earnings (expressed in thousands of PLN). As a result of such an analysis, we obtain, for each dummy variable, coefficients such as:

– for sex, a statistically significant coefficient $b_i = -0.5$ – meaning that women's average earnings are half a thousand PLN lower than men's, assuming that the other variables in the model remain unchanged;

Copyright ©2010-2026 U-smile Stats – All rights reserved

– for vocational education, a statistically significant coefficient $b_i = 0.6$ – meaning that the average earnings of people with vocational education are 0.6 thousand PLN higher than for people with primary education, assuming that the other variables in the model remain unchanged;
– for secondary education, a statistically significant coefficient $b_i = 1$ – meaning that the average earnings of people with secondary education are a thousand PLN higher than for people with primary education, assuming that the other variables in the model remain unchanged;
– for tertiary education, a statistically significant coefficient $b_i = 1.5$ – meaning that the average earnings of people with tertiary education are 1.5 thousand higher than for people with primary education, assuming that the other variables in the model remain unchanged.

Effect coding is used to answer, with the help of multivariate models, the question: how do the results (Y) in each analysed category differ from the (unweighted) sample mean? This coding consists in assigning a value of -1 or 1 to each category of a given variable. The category coded as -1 is then the **base category**.

$k = 2$ When the coded variable is dichotomous, placing it in a regression model gives us a calculated coefficient (b_i). This coefficient is the reference of Y for category 1 to the unweighted overall mean (adjusted for the other variables in the model).

When the analysed variable has more than two categories, k categories are represented by $k-1$ dummy variables with effect coding. When creating variables with effect coding, a category is selected for which no separate variable is created. This category is treated in the model as the base category (since in every variable recorded with effect coding it corresponds to a value of -1).

When the variables $X_1, X_2, ..., X_{k-1}$ obtained this way, with effect coding, are placed in a regression model, coefficients $b_1, b_2, ..., b_{k-1}$ will be calculated for them.

b_1 is the reference of the Y results (for codes 1 in X_1) to the unweighted overall mean (adjusted for the other variables in the model);

b_2 is the reference of the Y results (for codes 1 in X_2) to the unweighted overall mean (adjusted for the other variables in the model);

...

b_{k-1} is the reference of the Y results (for codes 1 in X_{k-1}) to the unweighted overall mean (adjusted for the other variables in the model).

Example
We will code, using effect coding, the sex variable with two categories (male sex will be selected as the base category), and a variable indicating the region of residence within the country analysed. Five regions are distinguished: north, south, east, west, and central – the central region will be selected as the base category.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Sex	Coded sex
f	1
f	1
f	1
m	-1
m	-1
f	1
f	1
m	-1
m	-1
f	1
m	-1
f	1
m	-1
f	1
m	-1
m	-1
...	...

Region of residence	Coded region			
	west	east	north	south
central	-1	-1	-1	-1
central	-1	-1	-1	-1
central	-1	-1	-1	-1
west	1	0	0	0
west	1	0	0	0
west	1	0	0	0
west	1	0	0	0
east	0	1	0	0
east	0	1	0	0
east	0	1	0	0
east	0	1	0	0
north	0	0	1	0
north	0	0	1	0
south	0	0	0	1
south	0	0	0	1
south	0	0	0	1
...	...	...	...	...

Building on the basis of dummy variables, in a multiple regression model, we might want to check how these variables affect a certain dependent variable, e.g. Y = the amount of earnings (expressed in thousands of PLN). As a result of such an analysis, we obtain, for each dummy variable, coefficients such as:
– for sex, a statistically significant coefficient $b_i = -0.5$ – meaning that women's average earnings are half a thousand PLN lower than the average earnings in the country, assuming that the other variables in the model remain unchanged;
– for the western region, a statistically significant coefficient $b_i = 0.6$ – meaning that the average earnings of people living in the western region of the country are 0.6 thousand PLN higher than the average earnings in the country, assuming that the other variables in the model remain unchanged;
– for the eastern region, a statistically significant coefficient $b_i = -1$ – meaning that the average earnings of people living in the eastern region of the country are a thousand PLN lower than the average earnings in the country, assuming that the other variables in the model remain unchanged;
– for the northern region, a statistically significant coefficient $b_i = 0.4$ – meaning that the average earnings of people living in the north are 0.4 thousand higher than the average earnings in the country, assuming that the other variables in the model remain unchanged;
– for the southern region, a non-statistically-significant coefficient $b_i = 0.1$ – meaning that the average earnings of people living in the south do not differ significantly from the average earnings in the country, assuming that the other variables in the model remain unchanged.

Interactions

Interactions are considered in multivariate models, and their presence means that the effect of an independent variable (X_1) on the dependent variable (Y) differs depending on the level of another independent variable (X_2) or a series of further independent variables. To be able to consider interactions in multivariate models, it is necessary to indicate the variables representing the probable interactions, i.e. the products of the relevant variables. For this purpose, we select the Interactions button in the window of the selected multivariate analysis. In the interaction settings window, with the CTRL button held, we select the variables that are to form interactions and move them to the neighbouring list using

Copyright ©2010-2026 U-smile Stats – All rights reserved

the arrow. Pressing the OK button gives us the appropriate columns in the data sheet.

Note! In the analysis of interactions, choosing the appropriate coding for dichotomous variables allows us to avoid over-parameterisation associated with interactions. Over-parameterisation causes the lower-order effects for dichotomous variables to be redundant with respect to the confounded higher-order interactions. As a result, including a higher-order interaction in the model cancels out the effect of the lower-order interactions, preventing their proper evaluation. To avoid over-parameterisation in a model in which interactions of dichotomous variables occur, it is recommended to select the effect coding option.

In models with interactions, remember to trim them appropriately, so that when removing main effects, the higher-order effects that depend on them are also removed. That is: if a model contains the following variables (main effects): X_1, X_2, X_3, and interactions: $X_1 * X_2$, $X_1 * X_3$, $X_2 * X_3$, $X_1 * X_2 * X_3$, then removing variable X_1 from the model also requires removing the interactions in which it appears, i.e. $X_1 * X_2$, $X_1 * X_3$, and $X_1 * X_2 * X_3$.

25.1 MULTIPLE LINEAR REGRESSION

Multiple linear regression

WHEN TO USE:

The constructed multiple regression model allows the study of the influence of many independent variables (X_1, X_2, ..., X_k) on one dependent variable (Y). The most frequently used variety of multiple regression is Multiple Linear Regression. It is an extension of linear regression models based on the Pearson product-moment correlation coefficient. It assumes the existence of a linear relationship between the variables studied. The linear multiple regression model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon,$$

where:
Y – the dependent variable, explained by the model,
$X_1, X_2, \ldots X_k$ – independent, explanatory variables,
$\beta_0, \beta_1, \beta_2, \ldots \beta_k$ – parameters,
ϵ – the random component (model residual).

ASSUMPTIONS

- measurement of the dependent variable Y on an interval scale,
- a linear relationship between the dependent and independent variables,
- a sufficiently large number of observations.

Note! When constructing the model, remember that the number of observations must be large, i.e. it should satisfy the assumption $n \geq 50 + 8k$, where k is the number of explanatory variables in the model[74].

Copyright ©2010-2026 U-smile Stats – All rights reserved

DEFINITIONS OF MEASURES

Dummy variables and interactions in the model – a discussion of the preparation of dummy variables and interactions is presented in section 25.0.2 Preparation of variables for analysis in multivariate models.

Regression coefficients – if the model was built using a sample of size n, the above equation can be represented in matrix form:

$$Y = X\beta + \epsilon,$$

where: $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$

The solution of the equation is then the vector of parameter estimates $\beta_0, \beta_1, \dots, \beta_k$, called the **regression coefficients**:

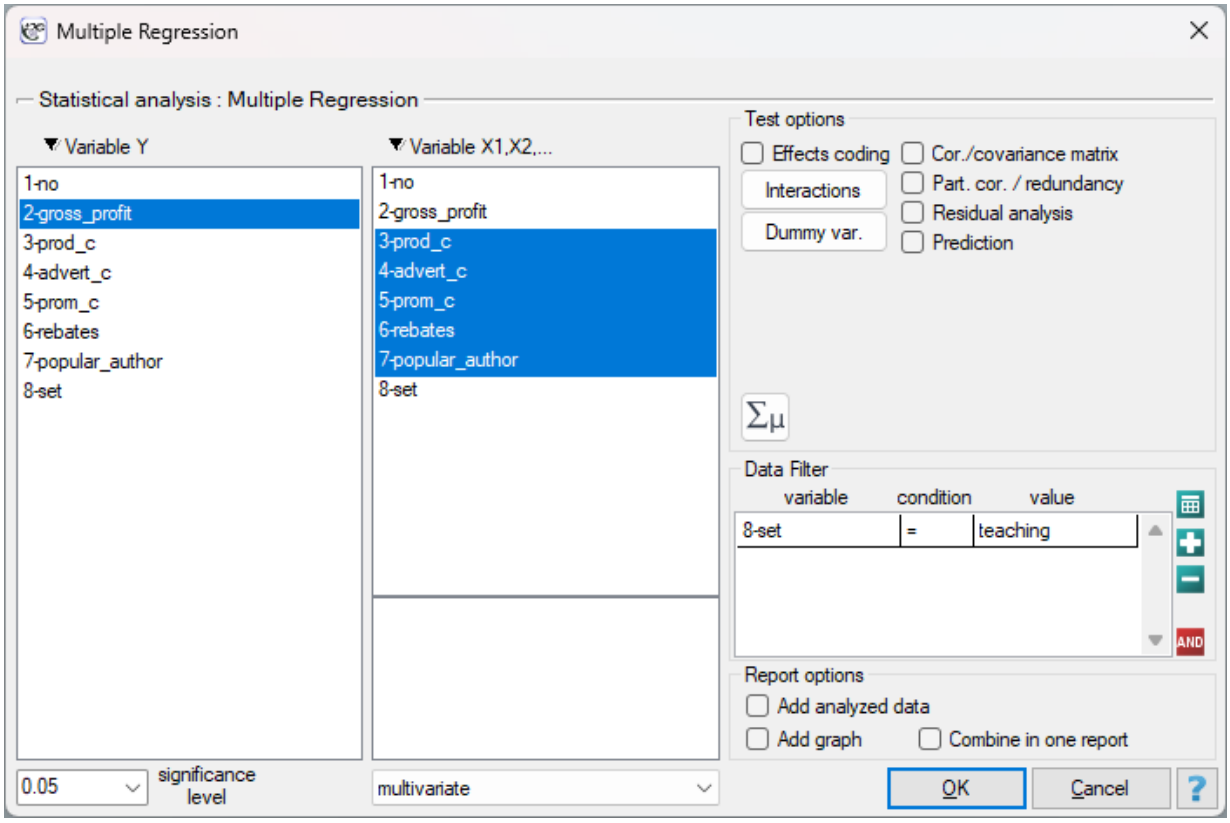
$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}.$$

These coefficients are estimated using the classical **least squares method**. On the basis of these values, we can draw conclusions about the magnitude of the effect of the independent variable (for which the coefficient was estimated) on the dependent variable – they tell us by how many units the dependent variable will change when the independent variable changes by 1 unit. Each coefficient carries a certain estimation error, calculated from the formula $SE_b = \sqrt{\frac{1}{n-(k+1)} e^T e (X^T X)^{-1}}$, where $e = Y - \widehat{Y}$ is the vector of **model residuals** (the difference between the actual values of the dependent variable Y and the values $\widehat{Y}$ predicted from the model).

HOW TO SET UP THE ANALYSIS

Advanced statistics → Multivariate models → Multiple regression.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Model verification

Statistical significance of individual variables in the model – on the basis of the coefficient and its estimation error, we can infer whether the independent variable for which this coefficient was estimated has a significant effect on the dependent variable. For this purpose, we use the t-test.

Quality of the constructed model – the quality of a multiple linear regression model can be assessed using several measures:

- **Standard error of estimation** – a measure of model fit, based on the model residuals $e_i = y_i - \widehat{y}_i$, i.e. the discrepancy between the actual values of the dependent variable y_i in the sample and the values of the dependent variable $\widehat{y}_i$ calculated from the constructed model. For the model to be well fitted, the standard error of estimation should be as small as possible.
- **Multiple correlation coefficient** $R = \sqrt{R^2} \in \langle 0; 1 \rangle$ – indicates the strength of the effect of the set of variables $X_1, X_2, \ldots X_k$ on the dependent variable Y.
- **Multiple coefficient of determination** R^2 – a measure of model fit. The value of this coefficient falls within the range $\langle 0; 1 \rangle$, where 1 means a perfect fit and 0 means a complete lack of fit. It expresses the percentage of variability in the dependent variable explained by the model. Since the value of R^2 depends on the model fit but is also sensitive to the number of variables in the model and the sample size, an adjusted value R^2_{adj} is also calculated.
- **Information criteria** (AIC, $AICc$, BIC) – based on the entropy of the information carried by the model (the model's uncertainty), i.e. they estimate the information lost when a given model is used to describe the phenomenon studied. We should choose the model with the minimum value of the given information criterion. An information criterion is not, however, an absolute

Copyright ©2010-2026 U-smile Stats – All rights reserved

measure – if all the models compared describe reality poorly, looking for a warning sign in the information criterion makes no sense.

- **Error analysis for ex post forecasts** (MAE, MPE, $MAPE$) – describe the accuracy of the forecast, indicating on average how much the actual realisations of the dependent variable will deviate from the forecasts.
- **Statistical significance of all variables in the model** – the basic tool for estimating the significance of all variables in the model is the analysis of variance test (F-test).

More information about the variables in the model:

- **Standardized** $b_1, b_2, \ldots, b_k$ – unlike raw parameters (which, depending on the variable described, are expressed in different units of measurement and cannot be directly compared), standardized parameter estimates allow the contribution of individual variables to explaining the variability of the dependent variable Y to be compared.
- **Correlation matrix** – contains information about the strength of the relationship between individual variables, i.e. the Pearson correlation coefficient $r_p \in \langle -1; 1 \rangle$, calculated for each pair of variables, without taking into account the influence of the other variables in the model.
- **Covariance matrix** – like the correlation matrix, contains information about the linear relationship between individual variables, except that this value is not standardized.
- **Partial correlation coefficient** – falls within the range $\langle -1; 1 \rangle$ and is a measure of the correlation between a specific independent variable X_i (taking into account its correlation with the other variables in the model) and the dependent variable Y (taking into account its correlation with the other variables in the model). The square of this coefficient is the **partial coefficient of determination** – falls within the range $\langle 0; 1 \rangle$ and represents the ratio of the exclusive variability of a given independent variable X_i to the variability of the dependent variable Y that has not been explained by the other variables in the model. The closer these coefficients are to 0, the more useless the information carried by the variable studied, i.e. the more redundant it is.
- **Semi-partial correlation coefficient** – falls within the range $\langle -1; 1 \rangle$ and is a measure of the correlation between a specific independent variable X_i (taking into account its correlation with the other variables in the model) and the dependent variable Y (NOT taking into account its correlation with the other variables in the model). The square of this coefficient is the **semi-partial coefficient of determination** – falls within the range $\langle 0; 1 \rangle$ and represents the ratio of the exclusive variability of a given independent variable X_i to the total variability of the dependent variable Y.
- **R-squared (**$R^2 \in \langle 0; 1 \rangle$**)** – expresses the percentage of variability in a given independent variable X_i explained by the other independent variables. The closer to 1, the more strongly the variable studied is linearly related to the other independent variables, which may mean it is a redundant variable.
- **Variance inflation factor (**$VIF \in \langle 1; \infty \rangle$**)** – indicates how much the variance of the estimated regression coefficient is increased due to collinearity. The closer to 1, the lower the collinearity. Strong collinearity is assumed to occur when $VIF > 5$[154].
- **Tolerance** $= 1 - R^2 \in \langle 0; 1 \rangle$ – expresses the percentage of variability in a given independent variable X_i NOT explained by the other independent variables. The closer the tolerance value is to 0, the more strongly the variable studied is linearly related to the other independent variables, which may mean it is a redundant variable.

Copyright ©2010-2026 U-smile Stats – All rights reserved

- **Comparison of the full model with the model after removing a given variable** – performed using the F-test (when more than one variable is removed from the model) or the t-test (when only one variable is removed – this is the same test used to examine the significance of individual variables in the model; the results of both tests are then identical). If the difference between the compared models is statistically significant, the full model is significantly better than the reduced model, which means the variable studied is not redundant.
- **Scatter plots** – allow a subjective assessment of the linearity of the relationship between variables and the identification of outliers, and can also be used in the analysis of model residuals.

Analysis of model residuals

To obtain a correct regression model, we should check the basic assumptions concerning the model residuals:

- **Outliers** – examining the model residuals quickly gives us information about outlying values. Such observations can greatly disturb the regression equation, since they have a large influence on the values of its coefficients. If a given residual e_i is more than 3 standard deviations from the mean value, it may be suspected that removing this observation will improve the model. **Cook's distance** describes the magnitude of changes in the regression coefficients resulting from omitting a given case (in the program, Cook's distances are marked in bold for cases exceeding the 50th percentile of the Fisher-Snedecor distribution statistic $F(0.5, k+1, n-k-1)$). **Mahalanobis distance** is dedicated to detecting outliers – high values indicate a considerable distance of a given case from the centre of the independent variables.
- **Normality of the distribution of model residuals** – we check this assumption visually using a Q-Q plot of the normal distribution. A large difference between the distribution of the residuals and the normal distribution may disturb the assessment of the significance of the coefficients of the individual variables in the model.
- **Homoscedasticity (constancy of variance)** – to check whether there are areas where the variance of the model residuals is increased or decreased, we use plots of the residuals (or the square of the residuals) against the predicted and observed values.
- **Autocorrelation of model residuals** – for the constructed model to be considered correct, the residual values should not be correlated with each other. This assumption can be checked using the Durbin-Watson test statistic.

PERFORMING AND INTERPRETING THE ANALYSIS

Example 25.2. (publisher.pqs file)
A certain book publisher wanted to find out what effect variables such as: production costs, advertising costs, direct promotion costs, the sum of discounts offered, and the author's popularity have on gross profit from sales. To do this, they analysed 40 titles published over the previous year (training set). An excerpt of the data is shown below:

Copyright ©2010-2026 U-smile Stats – All rights reserved

no	gross_profit	prod_c	advert_c	prom_c	rebates	popular_author
1	58	7.9	9	0.38	1.8	1
2	63	10.1	10	0.59	2.4	0
3	27	3	7	0.7	1.7	0
4	35	6	3	0.21	2.6	1
5	34	6.6	2.1	0.13	2.2	0
6	48	10.7	1	0.08	2.1	1
7	14	2.7	0.7	0.06	0.3	0
8	63.5	12	5	0.56	1.7	0

The first five variables are expressed in thousands of dollars – these are therefore variables collected on an interval scale. The last variable, author's popularity, is a dichotomous variable, where 1 means a known author and 0 means an unknown author. Based on the knowledge gained, the publisher plans to predict the gross profit from the next book published by a known author. The costs they intend to incur are: production costs ≈ 11, advertising costs ≈ 13, direct promotion costs ≈ 0.5, sum of discounts offered ≈ 0.5.

We build a multiple linear regression model for the training set, choosing: gross profit – as the dependent variable Y, production costs, advertising costs, direct promotion costs, sum of discounts offered, author's popularity – as the independent variables X_1, X_2, X_3, X_4, X_5. As a result, the coefficients of the regression equation will be calculated, along with measures allowing the quality of the model to be assessed.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Multiple Regression	
Analysed variables	gross_profit
	prod_c
	advert_c
	prom_c
	rebates
	popular_author
Data Filter	set=teaching
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Sample size assumption n ≥ 50+8k	No
Size	40
Number of variables in the model	5
R	0.9225
R2	0.851
Adjusted R2	0.8291
Standard error of estimation	8.0865
Residual sum of squares	2223.3112
Total sum of squares	14918.9978
Explained sum of squares	12695.6865
F	38.8298
p-value	<0.0001
AIC - Akaike criterion	172.7149
AICc - corrected Akaike criterion	177.0786
BIC - Bayesian criterion	190.226
MAE (mean absolute error)	4.621
MPE (mean percentage error)	-21.483%
MAPE (mean absolute percentage error)	31.575%

Model

	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. erro
intercept	4.1752	4.7728	-5.5243	13.8746	0.8748	0.3878		
prod_c	2.5607	0.5015	1.5415	3.5799	5.106	<0.0001	0.4228	0.0828
advert_c	1.9982	0.3591	1.2685	2.7279	5.5651	<0.0001	0.4613	0.0829
prom_c	4.6682	4.7916	-5.0696	14.406	0.9742	0.3368	0.0662	0.068
rebates	1.4232	1.4048	-1.4317	4.2781	1.0131	0.3182	0.0675	0.0666
popular_author	10.1537	2.7826	4.4989	15.8086	3.649	0.0009	0.2616	0.0717

$SE_e = 8.086501$, $R^2 = 0.850974$, $R^2_{adj} = 0.829059$, **F-test:** $p < 0.000001$

The model is well fitted and statistically significant.

Based on the estimated value of the coefficient b, the relationship between gross profit and all the

Copyright ©2010-2026 U-smile Stats – All rights reserved

independent variables can be described by the equation:

The coefficients obtained are interpreted as follows:

-
-
-
-
-

The result of the t-test obtained for each variable shows that only production costs, advertising costs, and author's popularity have a significant effect on the profit obtained. At the same time, the standardized coefficients b are largest for these variables.

Based on the interpretation of the results so far, we can suspect that some of the variables have no significant effect on profit and may be redundant. For the model to be well formulated, the interval-scaled independent variables should be strongly correlated with the dependent variable and relatively weakly correlated with each other. We can check this by calculating the correlation matrix and the covariance matrix:

Correlation matrix

	gross_profit	prod_c	advert_c	prom_c	rebates	popular_autl
gross_profit	1	0.7707	0.7948	0.0711	0.1319	0.5538
prod_c	0.7707	1	0.5588	-0.0798	0.093	0.3406
advert_c	0.7948	0.5588	1	0.1199	0.0567	0.3269
prom_c	0.0711	-0.0798	0.1199	1	-0.0565	-0.0493
rebates	0.1319	0.093	0.0567	-0.0565	1	0.0104
popular_author	0.5538	0.3406	0.3269	-0.0493	0.0104	1

Copyright ©2010-2026 U-smile Stats – All rights reserved

Covariance matrix

	gross_profit	prod_c	advert_c	prom_c	rebates	popular_autl
gross_profit	382.5384	48.6794	70.1909	0.3859	2.3928	5.4573
prod_c	48.6794	10.4294	8.1486	-0.0715	0.2784	0.5542
advert_c	70.1909	8.1486	20.3856	0.1502	0.2374	0.7436
prom_c	0.3859	-0.0715	0.1502	0.077	-0.0145	-0.0069
rebates	2.3928	0.2784	0.2374	-0.0145	0.86	0.0049
popular_author	5.4573	0.5542	0.7436	-0.0069	0.0049	0.2538

The most coherent information, allowing us to find those variables in the model that are redundant, is given by the analysis of partial and semi-partial correlation and redundancy:

Part. Semipart. Cor.

	partial	semipartial	tolerance	R^2	VIF	t stat.	p-value
prod_c	0.6588	-0.338	0.6392	0.3608	1.5644	5.106	<0.0001
advert_c	0.6904	-0.3684	0.6379	0.3621	1.5675	5.5651	<0.0001
prom_c	0.1648	-0.0645	0.9481	0.0519	1.0547	0.9742	0.3368
rebates	0.1712	-0.0671	0.988	0.012	1.0122	1.0131	0.3182
popular_author	0.5305	-0.2416	0.8531	0.1469	1.1722	3.649	0.0009

The values of the partial and semi-partial correlation coefficients indicate that direct promotion costs and the sum of discounts offered make the smallest contribution to the constructed model. However, these are the variables least correlated with the others in the model, as indicated by the low value of R^2 and the high tolerance value. Ultimately, from a statistical point of view, models without these variables would not be worse than the current model (see the result of the t-test for comparing models). It is up to the researcher's decision whether to keep this model or build a new one without direct promotion costs and the sum of discounts offered. We will keep the current model.

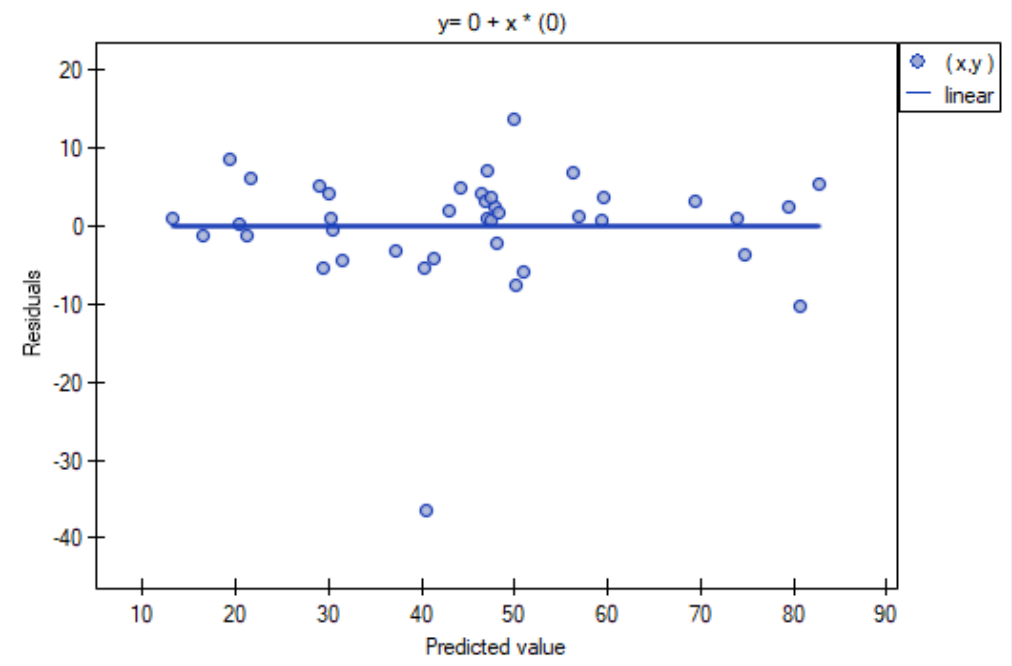
Finally, we will perform an analysis of the residuals. An excerpt of this analysis is shown below:

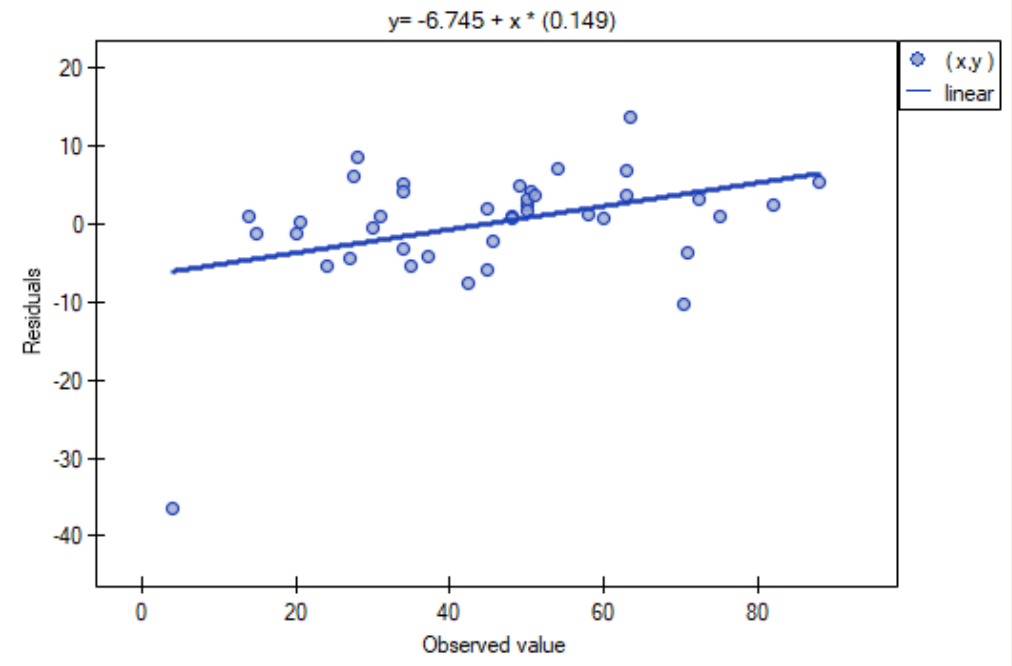
Residual analysis

	predicted va	residual	standard res	<=-3sd	(-3sd;2sd]	(-2sd;sd]	(-sd;sd)	[sd;2sd)	[2sd;3sd)	>=3sd	Cook di	Mahalanobis
1	56.8783	1.1217	0.1387				*				0.0002	1.5764
2	56.1906	6.8094	0.8421				*				0.0201	4.0536
3	31.5321	-4.5321	-0.5605				*				0.0096	4.3613
4	40.3684	-5.3684	-0.6639				*				0.0126	4.0889
5	29.01	4.99	0.6171				*				0.0099	3.7332
6	47.0888	0.9112	0.1127				*				0.0007	7.0912
7	13.1949	0.8051	0.0996				*				0.0006	7.6478
8	49.9285	13.5715	1.6783					*			0.118	5.744
9	46.5015	3.9985	0.4945				*				0.0119	6.4611
10	47.7767	2.2233	0.2749				*				0.0025	4.7472
11	47.9896	-2.2896	-0.2831				*				0.0026	4.6303
12	79.5361	2.4639	0.3047				*				0.0039	5.7187
13	29.9215	4.0785	0.5044				*				0.0034	1.7025
14	30.206	0.794	0.0982				*				0.0005	6.907
15	46.8409	3.1591	0.3907				*				0.0037	3.5041
16	40.4679	-36.4679	-4.5097	*							0.3122	2.0769
17	47.11	6.89	0.852				*				0.0213	4.194
18	50.1979	-7.6979	-0.9519			*					0.0173	2.6974
19	43.0358	1.9642	0.2429				*				0.0039	8.0621

Copyright ©2010-2026 U-smile Stats – All rights reserved

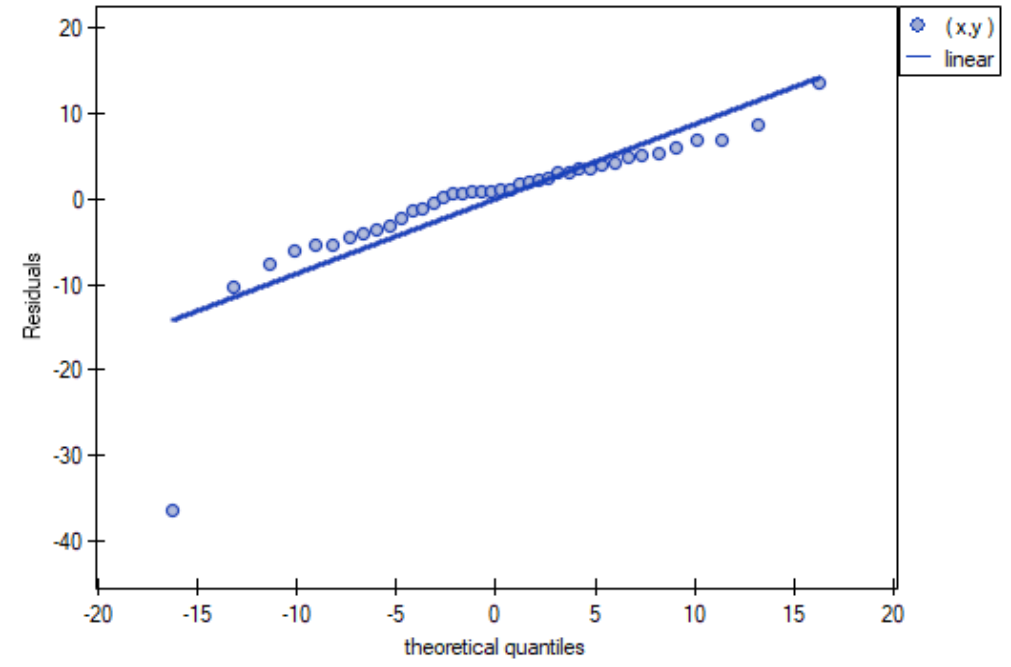
We can see that one of the model residuals is an outlier – it is more than 3 standard deviations from the mean value. This is observation number 16. We can easily find this observation by plotting a graph of residuals against the observed or predicted values of the variable Y.

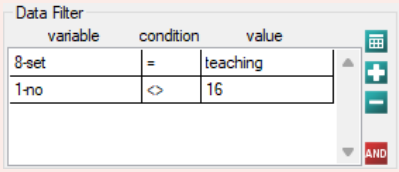




This outlying point disturbs the assumption of homoscedasticity – it would be satisfied (i.e. the variance of the residuals shown on the Y axis would be similar as we move along the X axis) if we rejected this point. Additionally, the distribution of the residuals deviates slightly from the normal distribution (the p value of the Lilliefors test is $p = 0.016415$):

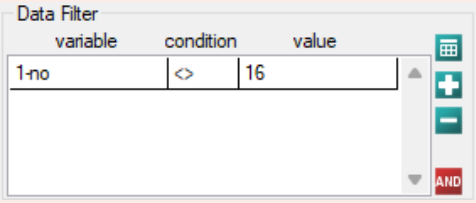
Copyright ©2010-2026 U-smile Stats – All rights reserved





Looking more closely at the outlying point (position 16 in the task data), we see that this is the only book for which the costs are higher than the gross profit (gross profit = 4 thousand dollars, sum of costs = (8+6+0.33+1.6) = 15.93 thousand dollars).

The resulting model can be improved by removing the outlier from it. This requires running the analysis again with a filter enabled that excludes the outlier.



As a result, we obtain a very similar model, but with a smaller error and a better fit:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Multiple Regression	
Analysed variables	gross_profit
	prod_c
	advert_c
	prom_c
	rebates
	popular_author
Data Filter	set=teaching
	and no<>16
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Sample size assumption n ≥ 50+8k	No
Size	39
Number of variables in the model	5
R	0.9699
R2	0.9408
Adjusted R2	0.9318
Standard error of estimation	4.8633
Residual sum of squares	780.4997
Total sum of squares	13173.1708
Explained sum of squares	12392.6711
F	104.7939
p-value	<0.0001
AIC - Akaike criterion	128.8585
AICc - corrected Akaike criterion	133.3585
BIC - Bayesian criterion	146.167
MAE (mean absolute error)	3.5963
MPE (mean percentage error)	-1.478%
MAPE (mean absolute percentage error)	9.137%

Model

	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. erro
intercept	6.8926	2.8914	1.01	12.7752	2.3838	0.023		
prod_c	2.6789	0.302	2.0645	3.2933	8.871	<0.0001	0.4706	0.053
advert_c	2.0809	0.2162	1.641	2.5208	9.6248	<0.0001	0.5112	0.0531
prom_c	1.9202	2.9031	-3.9862	7.8267	0.6614	0.5129	0.0288	0.0436
rebates	1.3254	0.845	-0.3937	3.0445	1.5686	0.1263	0.0669	0.0426
popular_author	7.3826	1.7107	3.9023	10.863	4.3157	0.0001	0.1992	0.0462

$$profit_{gross} = 6.89 + 2.68(c_{prod}) + 2.08(c_{adv}) + 1.92(c_{prom}) + 1.33(discounts) + 7.38(popul_{author}) + [4.86]$$

We will finally use the constructed model for prediction. Based on the predicted expenses in the amount of: production costs ≈ 11 thousand dollars, advertising costs ≈ 13 thousand dollars, direct promotion

Copyright ©2010-2026 U-smile Stats – All rights reserved

costs ≈ 0.5 thousand dollars, sum of discounts offered ≈ 0.5 thousand dollars, and the fact that the author is known (author's popularity ≈ 1), we calculate the predicted gross profit together with the confidence interval:

Prediction for prod_c	11
Prediction for advert_c	13
Prediction for prom_c	0.5
Prediction for rebates	0.5
Prediction for popular_author	1
Prediction of Y value	72.4182
-95% CI (for point)	61.8563
+95% CI (for point)	82.9801
-95% CI (for expected values)	68.723
+95% CI (for expected values)	76.1134

The predicted profit is 72 thousand dollars.

Finally, it should be noted that this is only a preliminary model. In an actual study, more data would need to be collected, since the number of variables in the model is too small relative to the number of books evaluated, i.e. $n < 50 + 8k$.

FOR THOSE INTERESTED

t-test for the significance of individual variables
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \beta_i = 0, \\ \mathcal{H}_1: &\quad \beta_i \neq 0. \end{aligned}$$

Formulas: the test statistic is defined by:

$$t = \frac{b_i}{SE_{b_i}}.$$

The test statistic has the t-Student distribution with $n - k$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Standard error of estimation

$$SE_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - (k+1)}}.$$

Multiple coefficient of determination
In its determination, we use the following equality:

$$T_{SS} = E_{SS} + R_{SS},$$

where:
T_{SS} – the total sum of squares,

Copyright ©2010-2026 U-smile Stats – All rights reserved

E_{SS} – the sum of squares explained by the model,
R_{SS} – the residual sum of squares.
The coefficient of determination is calculated from the formula:

$$R^2 = \frac{E_{SS}}{T_{SS}}.$$

The adjusted value of this parameter:

$$R^2_{adj} = R^2 - \frac{k(1-R^2)}{n-(k+1)}.$$

Information criteria
AIC, $AICc$, and BIC represent a kind of trade-off between goodness of fit and complexity. The second term in the sum in the formulas for the information criteria (the so-called loss or penalty function) measures the simplicity of the model – it depends on the number of variables in the model (k) and the sample size (n).

Akaike information criterion *(Akaike information criterion)*

$$AIC = n \cdot \ln \frac{R_{SS}}{n} + 2(k+1) + (constant)$$

where the constant can be omitted, since it is the same in each of the compared models. This is an asymptotic criterion – suitable for large samples, i.e. when $\frac{n}{k+2} > 40$. For small samples, it tends to favour models with a large number of variables.

Example of interpreting an AIC comparison
Suppose we determined AIC for three models: AIC_1=100, AIC_2=101.4, AIC_3=110. We can then determine the relative likelihood of each model, calculated relative to another model, usually the one with the smallest AIC value, using the formula: $e^{(AIC_{min}-AIC_i)/2}$. Comparing model 2 to model 1, the probability that it minimizes information loss is about half the probability that model 1 does so (exactly $\exp((100-101.4)/2) = 0.497$). Comparing model 3 to model 1, this probability is only a small fraction of the probability for model 1 (exactly $\exp((100-110)/2) = 0.007$).

Corrected Akaike information criterion

$$AICc = AIC + \frac{2(k+3)(k+4)}{n-k}$$

The Akaike criterion correction relates to sample size, which is why this measure is also recommended for small samples.

Bayesian information criterion (Schwarz criterion)

$$BIC = n \cdot \ln \frac{R_{SS}}{n} + (k+1) \ln n + (constant)$$

Like the corrected Akaike criterion, BIC takes sample size into account.

Error analysis for ex post forecasts
MAE (mean absolute error) – the accuracy of the forecast, as determined by MAE, tells us how much, on average, the actual realisations of the dependent variable deviate (in absolute value) from the forecasts.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

MPE (mean percentage error) – indicates what average percentage of the realisations of the dependent variable the forecast errors represent.

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{y_i}$$

MAPE (mean absolute percentage error) – indicates the average size of forecast errors, expressed as a percentage of the actual values of the dependent variable. MAPE allows the accuracy of forecasts obtained from different models to be compared.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right|$$

F-test – statistical significance of all variables in the model
This test verifies 3 equivalent hypotheses at the same time:

$\mathcal{H}_0$:	all $\beta_i = 0$,	$\mathcal{H}_1$:	there exists $\beta_i \neq 0$;
$\mathcal{H}_0$:	$R^2 = 0$,	$\mathcal{H}_1$:	$R^2 \neq 0$;
$\mathcal{H}_0$:	there is no linear relationship,	$\mathcal{H}_1$:	the relationship is linear.

Formulas: the test statistic is defined by:

$$F = \frac{E_{MS}}{R_{MS}}$$

where:
$E_{MS} = \dfrac{E_{SS}}{df_E}$ – the mean square explained by the model,
$R_{MS} = \dfrac{R_{SS}}{df_R}$ – the residual mean square,
$df_E = k$, $df_R = n - (k+1)$ – the corresponding degrees of freedom.
This statistic has the F Snedecor distribution with df_E and df_R degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Cook's distance and Mahalanobis distance – described above in the practical section.

Autocorrelation of residuals – the Durbin-Watson statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.$$

Interpretation: to test for positive autocorrelation at the significance level α, we check the position of the statistic d relative to the upper ($d_{U,\alpha}$) and lower ($d_{L,\alpha}$) critical value:

- if $d < d_{L,\alpha}$ – the errors are positively correlated;
- if $d > d_{U,\alpha}$ – the errors are not positively correlated;
- if $d_{L,\alpha} < d < d_{U,\alpha}$ – the test result is inconclusive.

To test for negative autocorrelation at the significance level α, we check the position of the value $4 - d$ relative to the upper and lower critical value:

Copyright ©2010-2026 U-smile Stats – All rights reserved

- if $4 - d < d_{L,\alpha}$ – the errors are negatively correlated;
- if $4 - d > d_{U,\alpha}$ – the errors are not negatively correlated;
- if $d_{L,\alpha} < 4 - d < d_{U,\alpha}$ – the test result is inconclusive.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Model-based prediction and test set validation

Validation and prediction for multiple regression

WHEN TO USE:

Validation of a model is a check of its quality. It is first performed on the data on which the model was built (the so-called **training set**), that is, it is returned in the report describing the resulting model. In order to be able to judge with greater confidence how suitable the model is for forecasting new data, an important part of validation is applying the model to data that were not used in estimating the model. If the summary based on the training data is satisfactory, i.e. the calculated errors, R^2 coefficients, and information criteria are at a satisfactory level, and the summary based on the new data (the so-called **test set**) is equally favourable, then it can be concluded with high probability that such a model is suitable for prediction. The test data should come from the same population from which the training data were selected. It is often the case that before building a model we collect data, and then randomly divide it into a training set, i.e. the data that will be used to build the model, and a test set, i.e. the data that will be used for additional validation of the model.

Prediction – most often, the final step in regression analysis is to use the constructed and previously validated model for prediction.

DEFINITIONS OF MEASURES

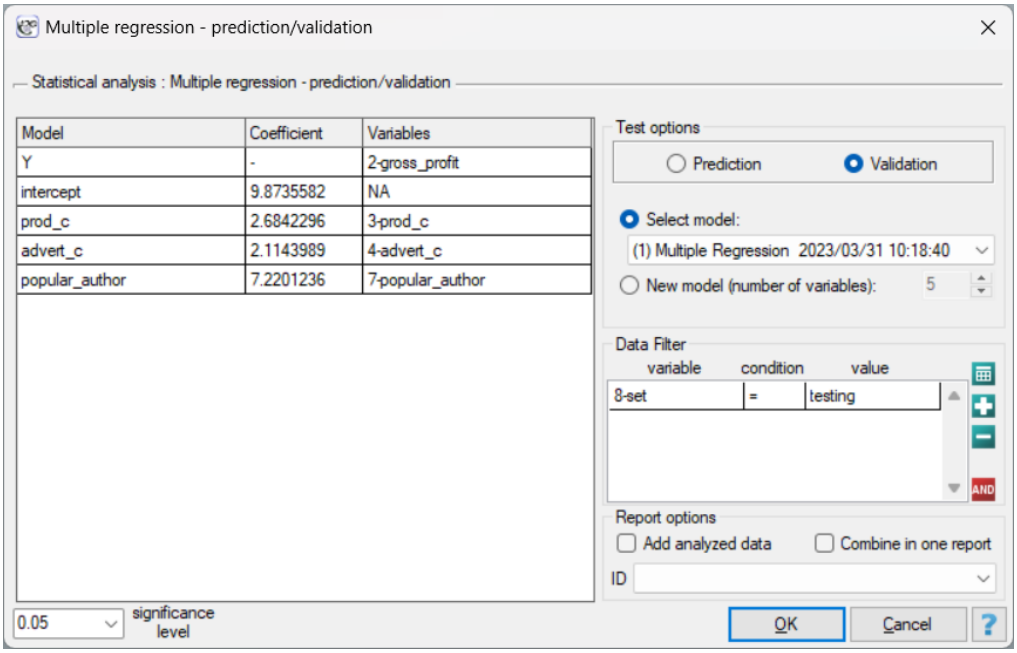
The estimated value is calculated with a certain error. Therefore, in addition, limits resulting from the error are determined for the value predicted by the model:

- for the expected value, **confidence limits** are determined,
- for a single point, **prediction limits** are determined.

HOW TO SET UP THE ANALYSIS

Advanced statistics → Multivariate models → Multiple regression – prediction/validation.

Copyright ©2010-2026 U-smile Stats – All rights reserved



To perform validation or prediction, it is necessary to indicate the model on the basis of which we want to perform it. This can be done on the basis of:

- a multiple regression model built in PQStat – simply select a model from those assigned to the given sheet, and the number of variables and model coefficients will be set automatically; the test set should be in the same sheet as the training set;
- a model not built in PQStat but obtained from another source (e.g. described in a scientific paper we have read) – in the analysis window, enter the number of variables and the coefficients for each of them.

In the analysis window, indicate the new variables that should be used for validation or prediction. Prediction for a single object can be performed together with building the model, that is, in the analysis window Advanced statistics→Multivariate models→Multiple regression, while prediction for a larger group of new data is performed via the menu Advanced statistics→Multivariate models→Multiple regression – prediction/validation.

PERFORMING AND INTERPRETING THE ANALYSIS

Continuation of Example 25.2 (publisher.pqs file): to predict gross profit from book sales, the publisher built a regression model based on a training set with item 16 removed (i.e. 39 books). The model included: production costs, advertising costs, and author's popularity (1=popular author, 0=not). We will build this model once again based on the training set, and then, to make sure the model will work correctly, we will validate it on a test data set. If the model passes this test, we will use it for prediction for new book titles. To use the appropriate sets, we set a data filter each time.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Multiple regression - prediction/validation	
Analysed variables	prod_c
	advert_c
	popular_author
	gross_profit
Data Filter	set=training
	and no<>16
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	39
Validation	
R (correlation Y and pred. Y)	0.9673344
R2 (correlation Y and pred. Y)	0.9357359
R2 Adjusted (correlation Y and pred. Y)	0.9302275
Standard error of estimation	4.9180776
Residual sum of squares	846.5620515
Total sum of squares	13173.1707692
Explained sum of squares	12326.6084199
F	169.8758305
p-value	0
AIC - Akaike criterion	128.0272524
AICc - corrected Akaike criterion	130.4978406
BIC - Bayesian criterion	142.0086223
MAE (mean absolute error)	3.6739289
MPE (mean percentage error)	-0.0171597
MAPE (mean absolute percentage error)	0.0941638

Coefficients of model	
intercept	9.8735582
prod_c	2.6842296
advert_c	2.1143989
popular_author	7.2201236

training set: $R^2_{adj} = 0.93$, $MAE = 3.8$ **thousand dollars**

The model is very well fitted to the data on which it was built.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Multiple regression - prediction/validation	
Analysed variables	prod_c
	advert_c
	popular_author
	gross_profit
Data Filter	set=testing
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	20
Validation	
R (correlation Y and pred. Y)	0.913034
R2 (correlation Y and pred. Y)	0.8336311
R2 Adjusted (correlation Y and pred. Y)	0.802437
Standard error of estimation	8.0620017
Residual sum of squares	1039.9339387
Total sum of squares	5864.992
Explained sum of squares	5751.2377808
F	29.4954006
p-value	0.0000009
AIC - Akaike criterion	87.0236039
AICc - corrected Akaike criterion	92.6236039
BIC - Bayesian criterion	96.9979976
MAE (mean absolute error)	5.8739802
MPE (mean percentage error)	0.0088586
MAPE (mean absolute percentage error)	0.1354043

Coefficients of model	
intercept	9.8735582
prod_c	2.6842296
advert_c	2.1143989
popular_author	7.2201236

test set: $R^2_{adj} = 0.80$, $MAE = 5.9$ thousand dollars

The quality of fit on the test set is somewhat lower than on the training set, but still satisfactory – the model is suitable for prediction.

For this purpose, we will use the data for three new book titles added at the end of the set. We select the Prediction option, set the filter to the new data set, and use our model to predict the gross profit for these books.

Prediction				
Pred. Y	-95% CI (of	+95% CI (ol	-95% CI (of	+95% CI (ol
74.1071	63.6621	84.5521	71.0391	77.1751
24.8392	14.4985	35.1799	22.1477	27.5307
58.7365	48.4804	68.9926	56.3908	61.0822

It turns out that the highest gross profit (between 64 and 85 thousand dollars) is predicted for the first book – the most heavily advertised and most expensively produced book by a popular author.

Copyright ©2010-2026 U-smile Stats – All rights reserved

25.2 COMPARISON OF MULTIPLE LINEAR REGRESSION MODELS

Comparing multiple regression models

WHEN TO USE:

Multiple linear regression provides the ability to analyse many independent variables simultaneously. This raises the problem of choosing the optimal model. In the flood of information carried by a model that is too large, important information can be lost. A model that is too small may omit those features that could reliably describe the phenomenon studied – it is not the number of variables in the model but their quality that determines the quality of the model. Selecting independent variables requires knowledge and experience related to the phenomenon studied. It should be remembered that the model should contain variables strongly correlated with the dependent variable and weakly correlated with each other.

DEFINITIONS OF MEASURES

There is no single simple statistical rule that determines the number of variables needed in a model. We most often use measures of model fit for comparison, such as: R^2_{adj} – the adjusted value of the multiple coefficient of determination (the higher the value, the better the model fit), SE_e – the standard error of estimation (the lower the value, the better the model fit), or the information criteria AIC, AICc, BIC (the lower the value, the better the model). For this purpose, the F-test based on the multiple coefficient of determination R^2 can also be used, which is used to verify the hypothesis that the fit of both compared models is equally good.

If the compared models do not differ significantly, we should select the one with fewer variables – a lack of difference means that the variables present in the full model but absent from the reduced model do not carry significant information. If, on the other hand, the difference in model fit quality is statistically significant, this means that one of them (the one with more variables, a higher R^2, or a lower information criterion value) is significantly better than the other.

Manual comparison of models consists in building 2 models:

- the full model – the model with more variables,
- the reduced model – the model with fewer variables; such a model is created from the full model by removing variables that are unnecessary from the point of view of the phenomenon studied.

The choice of independent variables in the compared models, and then the choice of the better model based on the comparison results, belongs to the researcher.

Automatic model selection – the program builds successive models itself and indicates the best one according to the chosen criterion.

Model terms. The unit that the automatic methods add to or remove from the model is a term: a single variable, a whole dummy variable (all its 0–1 columns at once) or an interaction. The hierarchy principle applies: an interaction can be in the model only when its main effects are in it, and only the interaction can be removed before them.

The selection criterion is chosen in the analysis window:

Copyright ©2010-2026 U-smile Stats – All rights reserved

- AIC, AICc or BIC – information criteria; the model with the lower value of the criterion is better. A move (adding or removing a term) is accepted only when it lowers the criterion;
- p of the nested test – the F-test comparing the model before and after the change, with a given p threshold. A term is removed when p is greater than the threshold, and added when p is smaller than the threshold. The same p value is shown in the report as the comparison of neighbouring models;
- validation error – the prediction error (RMSE) determined on observations that did not take part in fitting the model (described below: Out-of-sample evaluation of models). A move is accepted only when it reduces this error. The criterion requires the Validation option to be switched on – choosing it while validation is off switches cross-validation on.

Selection methods:

- **backward elimination** – starts from the model with all the variables; in each step the term whose removal lowers the criterion the most is removed; it ends when no removal lowers the criterion any further;
- **forward selection** – starts from the null model (intercept only); in each step the term that lowers the criterion the most is added; it ends when no addition lowers the criterion any further;
- **bidirectional stepwise** – starts from the null model; in each step both additions and removals of terms are considered and the best move is chosen; it ends when no move improves the criterion;
- **all subsets** – all models consistent with the hierarchy principle are fitted (a maximum of 15 terms), and the best model of each size is presented in the report; this method uses the information criteria or the validation error (the p criterion is not available in it).

All the models are computed on the same set of observations – rows with missing values in any of the candidate variables are discarded once, before the analysis. Thanks to this, the information criteria and the comparison tests are comparable between models.

The report contains: the selection method and criterion, the number of terms and of fitted models, the information criteria of the null model, the successive models of the selection path (each with the fit measures R^2_{adj}, SE_e, AIC, AICc, BIC and the table of coefficients) and the comparison of each neighbouring pair of models with the F-test (the statistic, degrees of freedom, p value). The best model according to the chosen criterion is marked in the report. If no term improves the null model, the report states that the null model is the best. With the all subsets method, neighbouring models may not be nested – in that case the F-test is not performed.

Note! The p values and confidence intervals in an automatically selected model are too optimistic, because the same data set was used both to select the variables and to estimate their effects. Automatic selection should be treated as an aid in searching for a model, and the final choice should be based on subject-matter knowledge. The selected model should be built again in the Multiple regression window and its assumptions verified.

Out-of-sample evaluation of models

The fit measures (R^2_{adj}, SE_e), the information criteria and the tests are determined on the same observations on which the coefficients of the model were estimated. Such an assessment is too optimistic: the model also adapts to the accidental features of the sample, so it usually predicts new observations

Copyright ©2010-2026 U-smile Stats – All rights reserved

worse than its fit would suggest, and the more variables in the model, the larger this difference. Selection based only on the training set therefore does not always indicate the model that will perform best on new data. For this reason, the analysis window makes it possible to evaluate the models on observations that did not take part in fitting them (the Validation option):

- **k-fold cross-validation** – the observations are randomly divided into k parts (folds) of similar size. Each model is fitted k times, each time leaving one part out, and is used to predict the values of the dependent variable in the part left out. Each observation is thus predicted exactly once – by a model that "has not seen" it. The models presented in the report (coefficients, fit measures, information criteria, tests) are fitted on all the observations, and the validation measures supplement them. The most common choices are $k = 5$ or $k = 10$;
- **random test set** – a given percentage of the observations is randomly set aside as the test set. All the models are fitted only on the remaining part (the training set) – the coefficients, fit measures, information criteria and tests in the report refer to it – and are evaluated on the test set. This approach is simpler and faster, but it requires a larger sample: the models are built on fewer observations, and the evaluation is based on a single, often small, set.

The split is random but fully determined by the Seed: the same seed always gives the same split, so the analysis can be repeated, and all the models of one analysis are evaluated on an identical split. Changing the seed gives a different split and slightly different values of the measures – if the order of the models changes after changing the seed, the differences between them are smaller than the random variability of the split and should not be given much weight.

For each model the report then gives three additional measures:

- **RMSE** – the root mean squared prediction error, expressed in the units of the dependent variable; the counterpart of SE_e determined out of sample (the lower the value, the better the model);
- **MAE** – the mean absolute prediction error; less sensitive to single large errors than RMSE;
- **predictive** R^2 – the part of the variability of the dependent variable explained by the out-of-sample predictions; the counterpart of R^2, usually smaller than it. A negative value means that the model predicts new observations worse than the mean alone.

The point of reference is the **null model**, which assigns to every observation the mean of the dependent variable from the training part – its RMSE and MAE are given in the model selection block of the report. The same block indicates the **best model by validation**, i.e. the model with the smallest out-of-sample RMSE. It is indicated regardless of the selection criterion chosen, which makes it possible to check whether the model selected, for example, by AIC is also the one that predicts new observations best. Out-of-sample evaluation also works in the manual comparison of models – the report then gives the measures of both models.

If validation error is chosen as the Criterion, the successive selection steps and the marking of the best model are also based on the out-of-sample error. Each model considered is then additionally fitted k times, so the computations take longer.

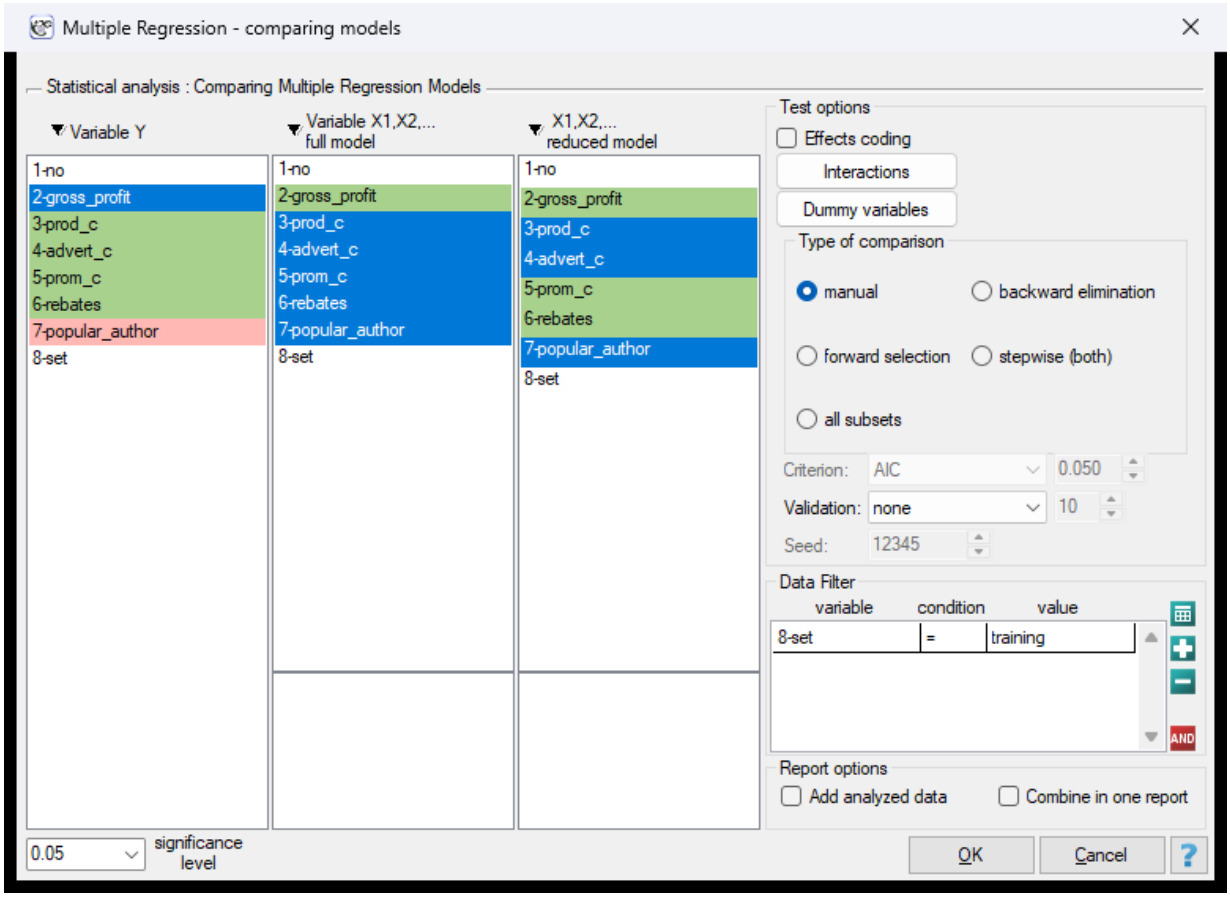
Note! When the validation error is the selection criterion, the validation error of the selected model is slightly too optimistic – out of many models, the one that performed best on the given split was chosen (the report gives a note about this). A reliable assessment of the final model can only be obtained from data that took part neither in fitting nor in selecting the model. With small samples, the differences in

Copyright ©2010-2026 U-smile Stats – All rights reserved

validation error between models are often smaller than the variability resulting from the random split – it is then worth checking the result for several values of the seed.

HOW TO SET UP THE ANALYSIS

Statistics → Multivariate models → Multiple regression – comparing models.



In the window we indicate the dependent variable Y and the independent variables of the full and the reduced model (in automatic selection – the candidate variables). Model selection options:

- Type of comparison – manual or one of the automatic selection methods: backward elimination, forward selection, stepwise (both), all subsets,
- Criterion – AIC, AICc, BIC, p (LR / F test) with the threshold given in the adjacent field, or validation error; in the manual mode the criterion is not used,
- Validation – none (evaluation on the training set only), k-fold CV, k = with the number of folds in the adjacent field (from 2 to 100, 10 by default) or test set, % = with the share of the test set as a percentage of the observations (from 5 to 90, 30 by default),
- Seed – the seed of the random split of the data (12345 by default); the same seed means the same split.

Copyright ©2010-2026 U-smile Stats – All rights reserved

PERFORMING AND INTERPRETING THE ANALYSIS

Continuation of Example 25.2 (publisher.pqs file): to predict gross profit from book sales, the publisher plans to consider variables such as: production costs, advertising costs, direct promotion costs, the sum of discounts offered, and author's popularity. However, not all of these variables need to have a significant effect on profit. Let us try to choose a linear regression model that will contain the optimal (from a statistical point of view) number of variables. We will use the training set data for this analysis.

Manual comparison of models

Based on the full model built earlier, we might suspect that the variables: direct promotion costs and the sum of discounts offered have little influence on the model being built. We will check whether, from a statistical point of view, the full model is better than the model after removing these two variables.

Comparing Multiple Regression Models	
Analysed variables	gross_profit
	prod_c
	advert_c
	prom_c
	rebates
	popular_author
Data Filter	set=teaching
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	40
Number of variables in the model 1	5
Standard error of estimation	8.0865
R	0.9225
R2	0.851
Adjusted R2	0.8291
AIC - Akaike criterion	172.7149
AICc - corrected Akaike criterion	177.0786
BIC - Bayesian criterion	190.226
Number of variables in the model 2	3
Standard error of estimation	8.0725
R	0.918
R2	0.8428
Adjusted R2	0.8296
AIC - Akaike criterion	170.863
AICc - corrected Akaike criterion	173.263
BIC - Bayesian criterion	184.9963
F - comparing models	0.9379
DF1	2
DF2	34
p-value	0.4013

Copyright ©2010-2026 U-smile Stats – All rights reserved

Model 1

	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. erro
intercept	4.1752	4.7728	-5.5243	13.8746	0.8748	0.3878		
prod_c	2.5607	0.5015	1.5415	3.5799	5.106	<0.0001	0.4228	0.0828
advert_c	1.9982	0.3591	1.2685	2.7279	5.5651	<0.0001	0.4613	0.0829
prom_c	4.6682	4.7916	-5.0696	14.406	0.9742	0.3368	0.0662	0.068
rebates	1.4232	1.4048	-1.4317	4.2781	1.0131	0.3182	0.0675	0.0666
popular_author	10.1537	2.7826	4.4989	15.8086	3.649	0.0009	0.2616	0.0717

Model 2

	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. erro
intercept	8.851	3.2776	2.2037	15.4982	2.7004	0.0105		
prod_c	2.5198	0.4928	1.5204	3.5192	5.1134	<0.0001	0.4161	0.0814
advert_c	2.074	0.3507	1.3629	2.7852	5.9148	<0.0001	0.4788	0.0809
popular_author	9.9215	2.7716	4.3003	15.5426	3.5797	0.001	0.2556	0.0714

F-test: $p = 0.4013$ **(not statistically significant,** $\alpha = 0.05$**)**

There is no reason to think that the full model is better than the reduced model.

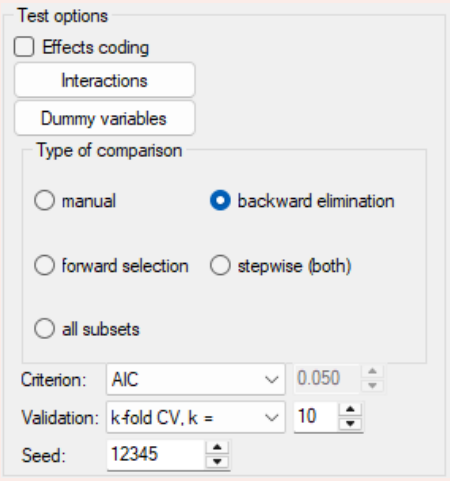
Additionally, the reduced model is slightly better fitted than the full model (for the reduced model $R^2_{adj} = 0.8296$, for the full model $R^2_{adj} = 0.8291$) and has smaller, i.e. more favourable, values of the information criteria AIC, AICc, and BIC.

Automatic model selection

Automatic selection (backward elimination, AIC criterion) gives very similar results. The best model is the one with the highest R^2_{adj} coefficient, the lowest information criteria, and the lowest standard error of estimation SE_e. The suggested best model here is a model containing only 3 independent variables: production costs, advertising costs, author's popularity.

Out-of-sample evaluation

The comparisons so far were based on the same 40 books on which the models were fitted. We will check whether the choice is confirmed when the models are evaluated on observations that did not take part in fitting them. In the analysis window we keep backward elimination and the AIC criterion, and in the Validation field we choose k-fold CV, k = and enter 10 folds (seed 12345).



Copyright ©2010-2026 U-smile Stats – All rights reserved

In the model selection block the report gives the method of evaluation, the error of the null model and the best model by validation, and for each model – three rows with the cross-validation measures.

Out-of-sample evaluation	cross-validation
Number of folds (k)	10
Seed of the random split	12345
Null model: RMSE	19.7867
Null model: MAE	15.9297
Best model by validation (RMSE)	Model 2

Number of variables in the model 3 (best)	3
Standard error of estimation	8.0725
R	0.918
R2	0.8428
Adjusted R2	0.8296
F - comparing models Model 2 vs Model 3	0.928
DF1	1
DF2	35
p-value	0.342
AIC - Akaike criterion	170.863
AICc - corrected Akaike criterion	173.263
BIC - Bayesian criterion	184.9963
Cross-validation: RMSE	8.2794
Cross-validation: MAE	4.9911
Cross-validation: predictive R2	0.8249

		training set			cross-validation ($k = 10$)		
model	variables	SE_e	R^2_{adj}	AIC	RMSE	MAE	predictive R^2
null	0				19.7867	15.9297	
Model 1	5	8.0865	0.8291	172.7149	8.3425	5.2976	0.8222
Model 2	4	8.0806	0.8293	171.8163	8.2309	5.1237	0.8270
Model 3	3	8.0725	0.8296	170.8630	8.2794	4.9911	0.8249

Model 1 is the full model, Model 2 was obtained from it by removing direct promotion costs, and Model 3 – by additionally removing the sum of discounts. All three predict the profit many times better than the null model (RMSE of about 8.3 versus 19.8), and the predictive R^2 (0.82–0.83) is only slightly lower than R^2_{adj} – so the models are not overfitted. At the same time, the out-of-sample error is slightly larger than the standard error of estimation (for Model 3: 8.28 versus 8.07), which was to be expected.

Model 2 has the smallest validation error (RMSE $= 8.2309$), and Model 3, indicated by AIC, is right behind it (RMSE $= 8.2794$) and has the smallest mean absolute error (MAE $= 4.9911$). The differences between these models are negligible compared with the size of the error itself, whereas the full model – although it is the best fitted on the training set (it has the highest R^2) – predicts new observations the worst of the three (RMSE $= 8.3425$). Removing direct promotion costs and the sum of discounts therefore does not worsen the predictions, and validation confirms the choice of the model with 3 variables.

Copyright ©2010-2026 U-smile Stats – All rights reserved

cross-validation ($k = 10$): RMSE $= 8.34$ (Model 1), 8.23 (Model 2), 8.28 (Model 3); null model: 19.79

The smaller models predict new observations no worse than the full model – the out-of-sample evaluation confirms the choice of the model with 3 variables.

Based on the above analyses, from a statistical point of view, the optimal model is the model containing the 3 most important independent variables: production costs, advertising costs, author's popularity. However, the final decision as to which model to choose belongs to the person with specialist knowledge of the field of study – in this case, the publisher. It should be remembered that the chosen model should be built again and its assumptions verified in the Multiple regression window.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad R^2_{FM} = R^2_{RM}, \\ \mathcal{H}_1: &\quad R^2_{FM} \neq R^2_{RM}, \end{aligned}$$

where R^2_{FM}, R^2_{RM} – the multiple coefficients of determination in the compared models (full and reduced).

Formulas: the test statistic is defined by:

$$F = \frac{R^2_{FM} - R^2_{RM}}{k_{FM} - k_{RM}} \cdot \frac{n - k_{FM} - 1}{1 - R^2_{FM}}.$$

This statistic has the F Snedecor distribution with $df_1 = k_{FM} - k_{RM}$ and $df_2 = n - k_{FM} - 1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Out-of-sample evaluation

Let $\hat{y}_{(-i)}$ denote the value predicted for observation i by the model fitted without the part of the data to which observation i belongs (in cross-validation – without its fold, with a test set – on the training set), and $\overline{y}_{(-i)}$ – the mean of the dependent variable in the same training part. The measures are determined jointly for all the n_v evaluated observations (in cross-validation – all the observations, with a test set – the observations of the test set):

$$RMSE = \sqrt{\frac{1}{n_v}\sum_{i=1}^{n_v}\left(y_i - \hat{y}_{(-i)}\right)^2}, \qquad MAE = \frac{1}{n_v}\sum_{i=1}^{n_v}\left|y_i - \hat{y}_{(-i)}\right|,$$

$$R^2_{pred} = 1 - \frac{\sum_{i=1}^{n_v}\left(y_i - \hat{y}_{(-i)}\right)^2}{\sum_{i=1}^{n_v}\left(y_i - \overline{y}_{(-i)}\right)^2}.$$

The denominator of R^2_{pred} is the sum of squared errors of the null model, so R^2_{pred} shows by what part the model reduces the prediction error relative to the null model.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The split: the observations are randomly shuffled (a pseudorandom number generator whose initial state is the seed) and then – in cross-validation – assigned in turn to k parts whose sizes differ by at most 1; with a test set, the given percentage of the observations (rounded to a whole number) goes to the test set. When k is greater than the number of observations, k equal to the number of observations is used. If there are too few observations to set aside a part of the data for validation, the program displays a message.

Copyright ©2010-2026 U-smile Stats – All rights reserved

25.3 LOGISTIC REGRESSION

Logistic regression

WHEN TO USE:

The constructed logistic regression model (similarly to multiple linear regression) allows the study of the effect of many independent variables ($X_1, X_2, ..., X_k$) on one dependent variable (Y). This time, however, the dependent variable only takes two values, e.g. ill/healthy, insolvent/solvent, etc. These two values are coded as (1)/(0), where (1) is the distinguished value – having the feature, and (0) – not having the feature.

The function on which the logistic regression model is based does not calculate the two-level variable Y itself, but the probability of that variable taking the distinguished value:

$$P(Y = 1|X_1, X_2, ..., X_k) = \frac{e^Z}{1 + e^Z},$$

where:
$P(Y = 1|X_1, X_2, ..., X_k)$ – the probability of taking the distinguished value (1), given specific values of the independent variables, the so-called predicted probability for 1,
Z is most often expressed by a linear relationship: $Z = \beta_0 + \sum_{i=1}^{k} \beta_i X_i$,
$X_1, X_2, \ldots X_k$ – independent, explanatory variables,
$\beta_0, \beta_1, \beta_2, \ldots \beta_k$ – parameters.

The logit is the transformation of this model into the form:

$$\ln\left(\frac{P}{1-P}\right) = Z.$$

ASSUMPTIONS

- the dependent variable Y is dichotomous, coded as (1)/(0),
- the independent variables are not collinear,
- a sufficiently large number of observations.

Note! Function Z can also be described by a higher-order relationship, e.g. quadratic – in that case, we introduce into the model a variable containing the square of the given independent variable, X_i^2.

When building the model, remember that the number of observations should be at least ten times the number of estimated model parameters ($n \geq 10(k+1)$). However, a more restrictive criterion, proposed by P. Peduzzi and others in 1996[136], is increasingly used, stating that the number of observations should be at least ten times the ratio of the number of independent variables (v) to the smaller of the proportions of counts (p) described by the dependent variable (i.e. the proportion of ill or healthy subjects), i.e. $n \geq 10v/p$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

When building the model, remember that the independent variables should not be collinear. When collinearity is present, estimation may become unreliable, and the resulting error values very high. Collinear variables should be removed from the model, or a single independent variable should be built from them, e.g. instead of the collinear variables mother's age and father's age, a variable parents' age could be built.

The convergence criterion of the Newton-Raphson iterative algorithm can be controlled using two parameters: the convergence iteration limit (gives the maximum number of iterations in which the algorithm should reach convergence) and the convergence criterion (gives the value below which the improvement in estimation obtained is considered insignificant, and the algorithm terminates).

DEFINITIONS OF MEASURES

Dummy variables and interactions in the model – a discussion of the preparation of dummy variables and interactions is presented in section 25.0.2 Preparation of variables for analysis in multivariate models.

Regression coefficients – the solution of the model equation is the vector of parameter estimates $\beta_0, \beta_1, \ldots, \beta_k$, called the **regression coefficients** ($b_0, b_1, ..., b_k$), estimated using the **maximum likelihood method**, i.e. by searching for the maximum of the likelihood function L (the program uses the iterative Newton-Raphson algorithm). On the basis of these values, we can draw conclusions about the magnitude of the effect of the independent variable (for which the coefficient was estimated) on the dependent variable. Each coefficient carries a certain estimation error, calculated from the formula $SE_b = \sqrt{diag(H^{-1})_b}$, where $diag(H^{-1})$ is the main diagonal of the covariance matrix.

Individual Odds Ratio – on the basis of the coefficients, for each independent variable in the model, an easily interpretable measure is calculated: the individual Odds Ratio:

$$OR_i = e^{\beta_i}.$$

The resulting Odds Ratio expresses the change in the odds of the distinguished value (1) occurring, when the independent variable increases by 1 unit. This result is adjusted for the other independent variables in the model, in the sense that they are assumed to remain at a constant level while the independent variable studied increases by one unit. The OR value is interpreted as follows:

- $OR > 1$ means the independent variable studied has a stimulating effect on obtaining the distinguished value (1), i.e. it tells us by how much the odds of the distinguished value (1) occurring increase when the independent variable increases by one level;
- $OR < 1$ means the independent variable studied has a de-stimulating effect on obtaining the distinguished value (1), i.e. it tells us by how much the odds of the distinguished value (1) occurring decrease when the independent variable increases by one level;
- $OR \approx 1$ means the independent variable studied has no effect on obtaining the distinguished value (1).

U-smile – a plot of four subclasses (non-events: better and worse prediction; events: worse and better prediction) with the selected coefficient (rLR, BA or RB) on the Y axis; the point size corresponds to the share I of the subclass, the class lines are solid when the improvement is significant (for rLR), and the horizontal reference line denotes no change. A smile means an improvement in prediction relative to the null model in both classes (see U-smile).

Odds Ratio – general formula – PQStat calculates the individual Odds Ratio. Its modification, based on the general formula, allows the interpretation of the result obtained to be changed. The odds ratio for

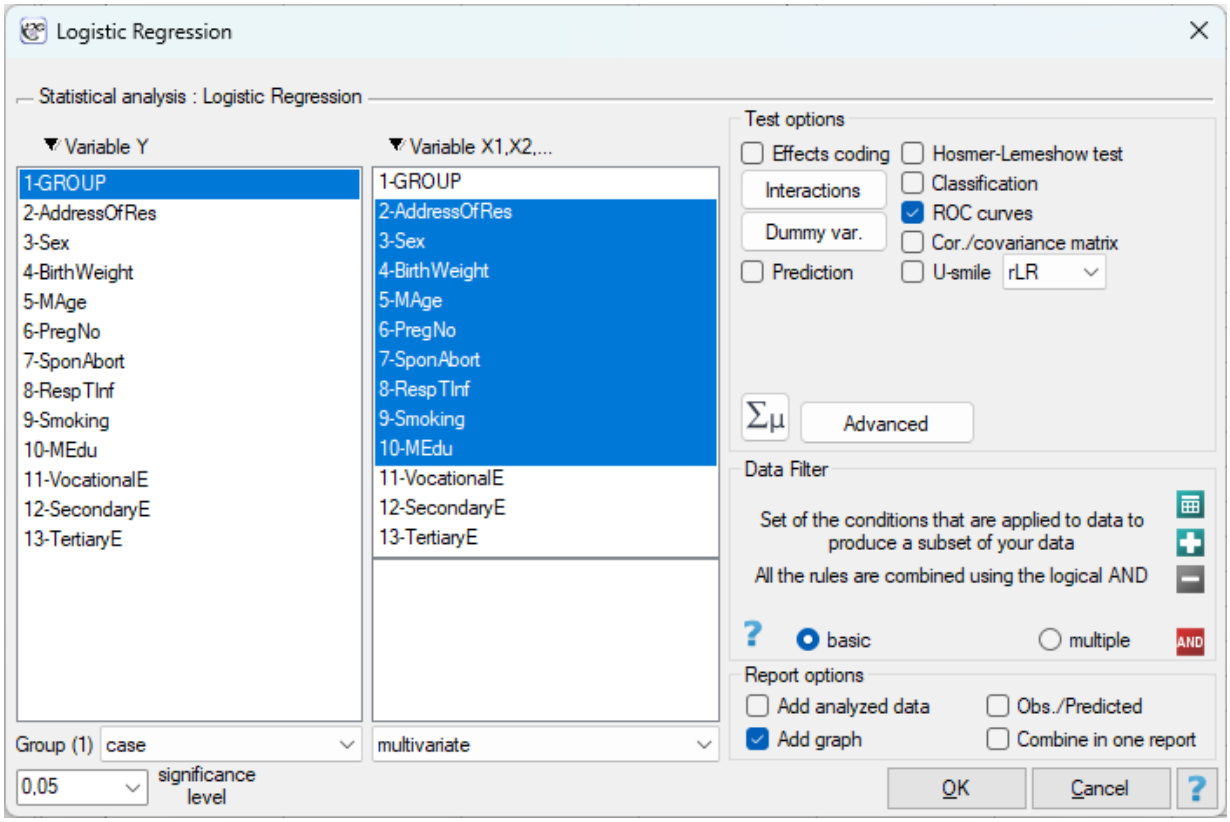
Copyright ©2010-2026 U-smile Stats – All rights reserved

the distinguished state occurring, in the general case, is calculated as the ratio of two odds, computed for two categories of the independent variable being compared.

Note! If the analysis is performed for a model other than linear, or if an interaction is included, then, on the basis of the general formula, we can calculate the corresponding Odds Ratio by changing the formula expressing Z.

HOW TO SET UP THE ANALYSIS

Advanced statistics → Multivariate models → Logistic regression.



Model verification

Statistical significance of individual variables in the model (significance of the odds ratio) – on the basis of the coefficient and its estimation error, we can infer whether the independent variable for which this coefficient was estimated has a significant effect on the dependent variable. For this purpose, we use the Wald test.

Quality of the constructed model – a good model should meet two basic conditions: it should be well fitted and as simple as possible. The quality of a logistic regression model can be assessed using several measures, based on: L_{FM} – the maximum of the likelihood function of the full model (with all variables), L_0 – the maximum of the likelihood function of the model containing only the intercept, n – the sample size:

- **Information criteria** ($AIC, AICc, BIC$) – based on the entropy of the information carried by the model. We should choose the model with the minimum value of the given information criterion.

Copyright ©2010-2026 U-smile Stats – All rights reserved

- **Pseudo R^2** – the so-called McFadden R^2, is a measure of model fit (the equivalent of the multiple coefficient of determination R^2 used in multiple linear regression). The value of this coefficient falls within the range $\langle 0; 1)$, where values close to 1 indicate a very good model fit. Since the R^2_{Pseudo} coefficient never reaches 1 and is sensitive to the number of variables in the model, its adjusted values are also calculated: $R^2_{Nagelkerke}$ and $R^2_{Cox-Snell}$.
- **Statistical significance of all variables in the model** – the basic tool for estimating the significance of all variables in the model is the likelihood ratio test.
- **Hosmer-Lemeshow test** – for different subgroups of the data, compares the observed frequencies of the distinguished value, O_g, with the predicted probability, E_g. If O_g and E_g are sufficiently close, the constructed model can be assumed to be well fitted. For the calculations, the observations are first divided into G subgroups, usually deciles ($G = 10$).
- **AUC – the area under the ROC curve** – the ROC curve, built on the basis of the value of the dependent variable and the predicted probability of the dependent variable P, allows the ability of the constructed logistic regression model to classify cases into two groups, (1) and (0), to be assessed. Additionally, for the ROC curve, a suggested value of the **cut-off point** for the predicted probability is given, together with a table showing sensitivity and specificity for every possible cut-off point.

Note! More options for calculating the cut-off point are provided by the ROC curve module. The analysis is carried out on the basis of the observed values and the predicted probability obtained from the logistic regression analysis.

Classification – based on the chosen cut-off point of the predicted probability, the quality of classification can be checked. The default cut-off point is 0.5. The user can change this value to any value from the range $(0, 1)$, e.g. the value suggested by the ROC curve. As a result, we obtain a classification table and the percentage of correctly classified cases, the percentage of correctly classified (0) – specificity, and the percentage of correctly classified (1) – sensitivity.

U-smile – evaluation of the model against the null model – the U-smile method (*Universal Smile Layout for Explanation*, Kubiak et al. 2024[183], Więckowska et al. 2025[184], 2026[185]) compares two models predicting the probability of an event on the same observations. In the logistic regression window, the reference model is the null model (intercept only, i.e. the event frequency in the sample), and the new model is the model built. The observations are divided into four subclasses:

- non-events (0) with better prediction (the probability decreased),
- non-events (0) with worse prediction,
- events (1) with worse prediction,
- events (1) with better prediction (the probability increased).

For each subclass the following are determined:

- a qualitative coefficient concerning the number of changes, I (the share of the subclass observations in the class),
- and, to choose from, one of the quantitative coefficients concerning the magnitude of the changes:

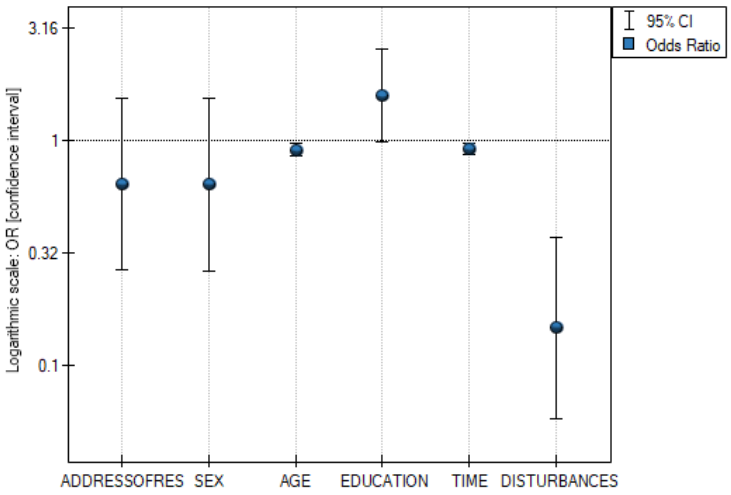
Copyright ©2010-2026 U-smile Stats – All rights reserved

- rLR (relative likelihood ratio) – normalised to the range $\langle 0; 1 \rangle$,
- BA (*Brier Alteration*) – the mean change in the squared residual,
- RB (*Relative Brier*) – the change relative to the Brier score of the reference model.

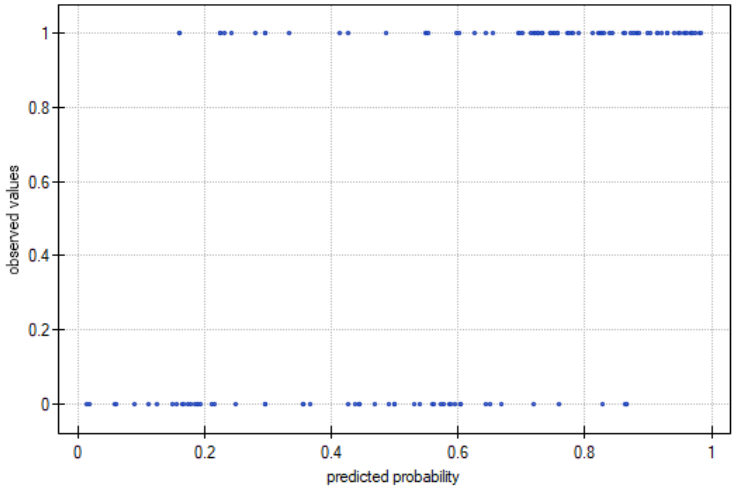
The **net coefficients** for the classes (improvement minus deterioration) and the **overall coefficient** show by how much the model improves prediction, separately for events and non-events.
Significance is checked with the likelihood ratio test (overall), and for rLR additionally with tests for each class based on the gamma distribution.
Formulas: for those interested.

Plots in logistic regression

- **Odds ratio $\pm$ confidence interval** – a plot showing the OR together with the 95% confidence interval for the result of each variable returned in the constructed model. For categorical variables, the line at level 1 indicates the odds ratio value for the reference category.

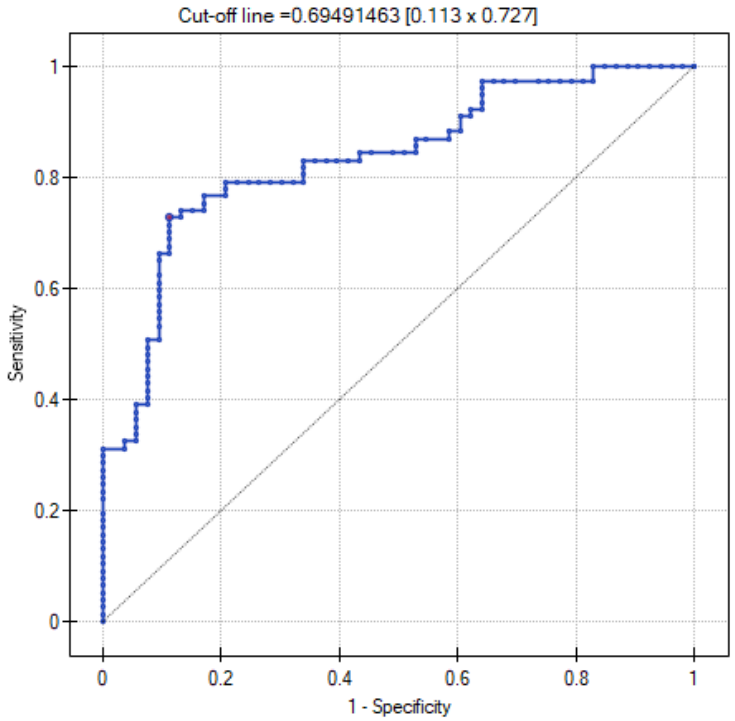


- **Observed values / Expected probability** – a plot showing the predicted probability of an event occurring for each individual (X axis) and the true value, i.e. the occurrence of the event (value 1 on the Y axis) or its absence (value 0 on the Y axis). If the model predicts very well, points will accumulate at the bottom on the left side of the plot and at the top on the right side.

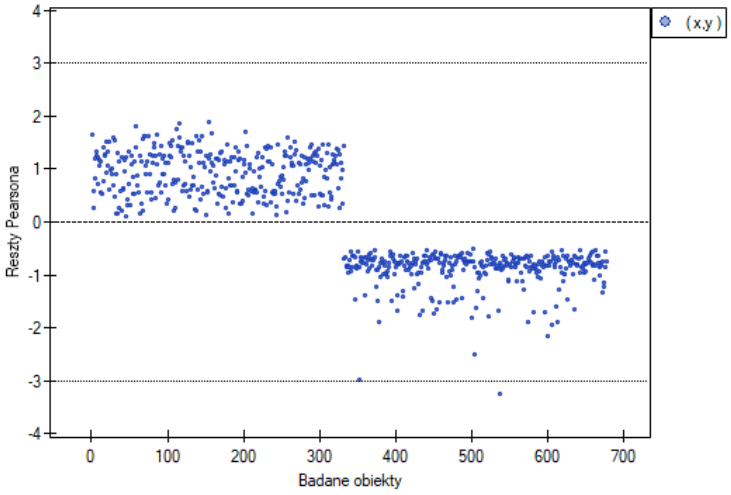


Copyright ©2010-2026 U-smile Stats – All rights reserved

- **ROC curve** – a plot built on the basis of the value of the dependent variable and the predicted probability of the event occurring.



- **Pearson residuals plot** – a plot allowing an assessment of whether outlying data are present. The residuals are the differences between the observed value and the probability predicted by the model. Plots of raw residuals from logistic regression are difficult to interpret, so they are standardised by calculating Pearson residuals (the raw residual divided by the square root of the variance function). The sign (positive or negative) indicates whether the observed value is higher or lower than the value fitted by the model, and the magnitude indicates the degree of deviation. Pearson residuals smaller than -3 or greater than 3 suggest too great a deviation for the given object.



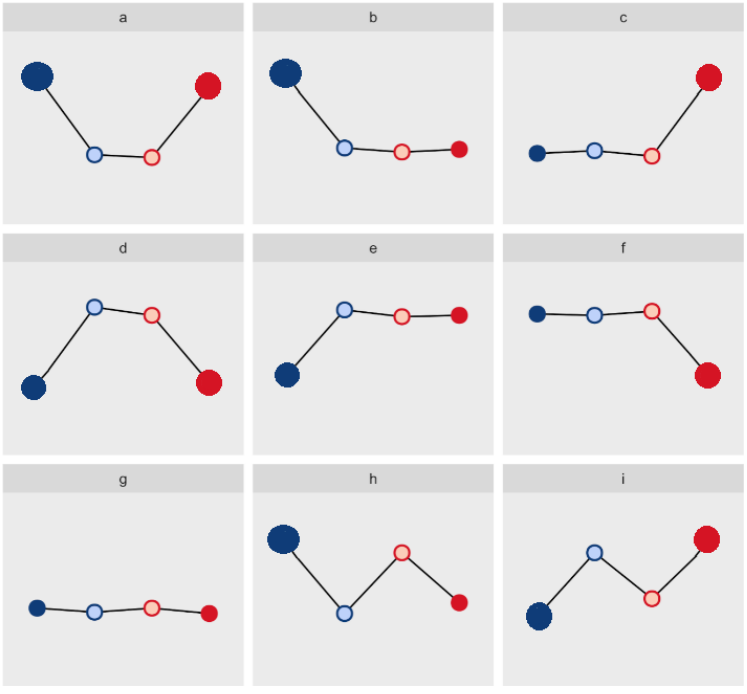
- **Unit changes in the odds ratio** – a plot showing a series of odds ratios, together with a confidence interval, determined for every possible cut-off point of the variable placed on the X axis. It allows

Copyright ©2010-2026 U-smile Stats – All rights reserved

the user to select one good cut-off point and then build, on this basis, a new two-valued variable. This plot is dedicated to the evaluation of continuous variables in univariate analysis.

- **Odds ratio profiles** – a plot showing a series of odds ratios, together with a confidence interval, determined for a given window size, i.e. comparing frequencies inside the window with frequencies outside the window. It works best when we are looking for a U-shaped function, i.e. high risk at low and high values of the variable studied, and low risk at average values. The wider the window, the greater the generalisation of the results and the smoother the odds ratio function; the narrower the window, the more detailed but more erratic the results. This plot is dedicated to the evaluation of continuous variables in univariate analysis.
- **U-smile plot** – on the U-smile plot the four subclasses form a "smile" when the model improves prediction in both classes:
  - (Y axis) the size of the smile shows the magnitude of the improvement (coefficients rLR, BA, RB to choose from),
  - (X axis) the left-hand side represents the non-event class (blue), the right-hand side the event class (red); the outer points are the subclass: improved prediction, and the inner points the subclass: worsened prediction,
  - symmetry (the left side of the plot similar to the right) means that the improvement/deterioration of prediction concerns both classes in the same way,
  - the point size corresponds to the coefficient I, i.e. it shows how large a percentage of objects falls into each subclass,
  - the class line is solid when the improvement in that class is significant (rLR only).

Examples of possible shapes of the U-smile plot. Improved prediction: (a) for both outcome classes, (b) only for non-events, (c) only for events. Worsened prediction: (d) for both outcome classes, (e) only for non-events, (f) only for events. Circles lying at a roughly constant level indicate neither improvement nor deterioration of prediction compared with the reference model (g). A zigzag indicates improved prediction in one outcome class and worsened prediction in the other (h and i).



Copyright ©2010-2026 U-smile Stats – All rights reserved

PERFORMING AND INTERPRETING THE ANALYSIS

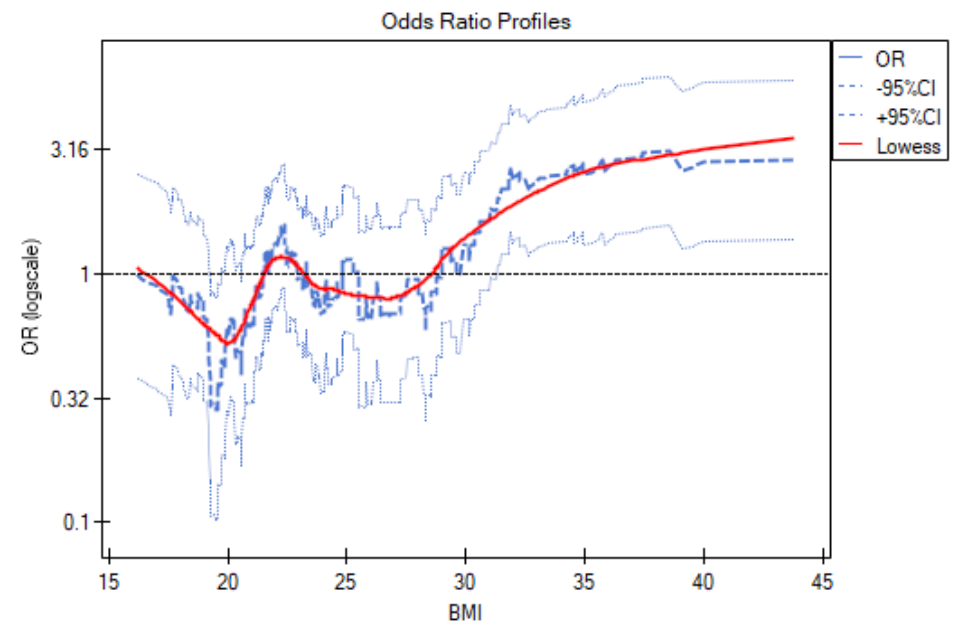
EXAMPLE 25.3. (OR profiles.pqs file)
We examine the risk of disease A and disease B depending on the patient's BMI. Since BMI is a continuous variable, its inclusion in the model results in a unit odds ratio determining a linear trend of increasing or decreasing risk. We do not know whether a linear model will be a good model for analysing this risk, so before building multivariate logistic regression models, we will build several univariate models presenting this variable on plots, in order to assess the shape of the relationship studied and, on this basis, decide how the variable should be prepared for analysis. For this purpose, we will use plots of unit changes in the odds ratio and odds ratio profiles, and for the profiles we will choose a window size of 100, since almost every patient has a different BMI, so about 100 patients will fall within each window.

Disease A
Unit changes in the odds ratio show that when the BMI cut-off point is chosen somewhere between 27 and 37, we obtain a statistically significant and positive odds ratio, showing that people with a BMI above this value have a significantly higher risk of the disease than people below this value.



The odds ratio profiles show that the red curve remains close to 1, with only the end of the curve slightly higher, indicating that it may be difficult to divide BMI into more than 2 categories and choose a good reference category.

Copyright ©2010-2026 U-smile Stats – All rights reserved

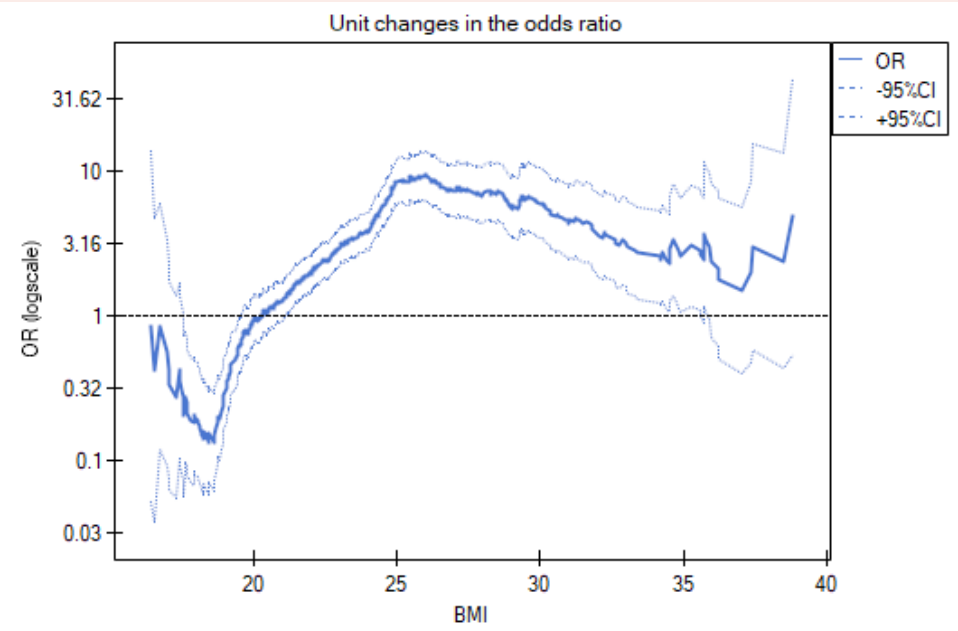


split into 2 values: $OR[95\%CI] = [1.41, 4.90]$**,** $p = 0.0024$**; unit OR:** $OR[95\%CI] = 1.07[1.02, 1.13]$**,** $p = 0.0052$

People with a BMI above 30 have a significantly higher risk of disease A; alternatively, one can keep the unit odds ratio, indicating a constant increase in risk with each unit increase in BMI.

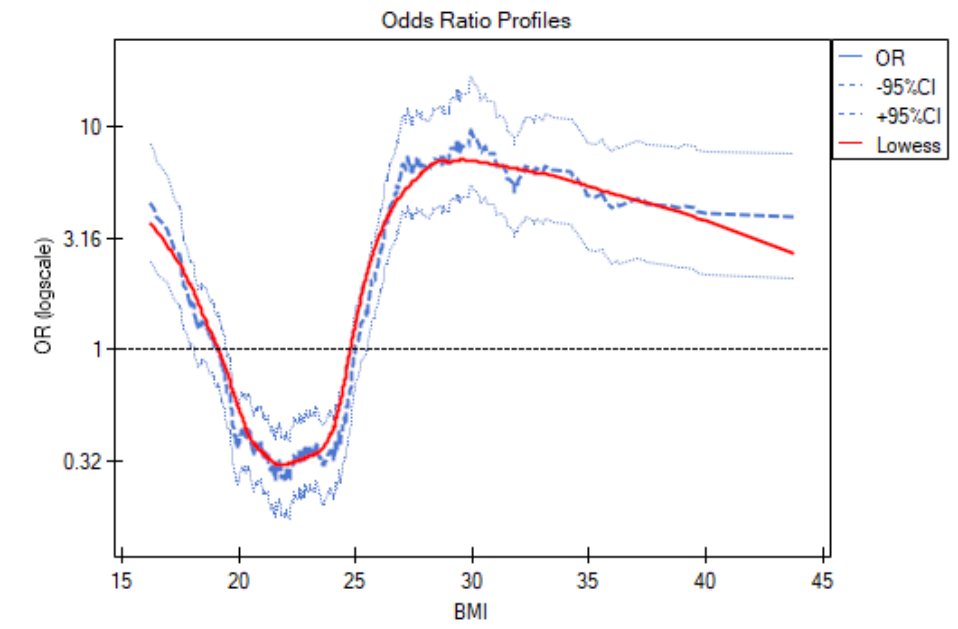
Disease B

Unit changes in the odds ratio show that when the BMI cut-off point is chosen somewhere between 22 and 35, we obtain a statistically significant and positive odds ratio, showing that people with a BMI above this value have a significantly higher risk of the disease than those below this value.

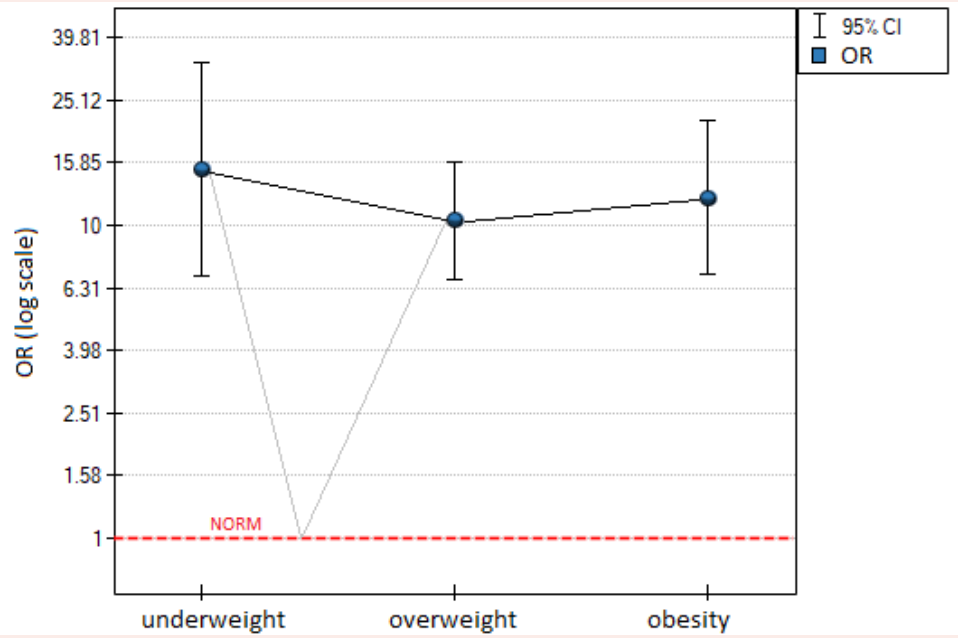


Copyright ©2010-2026 U-smile Stats – All rights reserved

The odds ratio profiles show that it would be much better to divide BMI into 2 or 4 categories, with the reference category being the one covering a BMI somewhere between 19 and 25, since this category is the lowest and is far removed from the results for BMIs to the left and right of this range. We see a distinct U-shape, meaning that disease risk is high at both low and high BMI.



In summary, although the relationship for the unit odds ratio (linear relationship) is statistically significant, it is not worth building such a model. It is much better to divide BMI into categories, with the reference value being the average BMI. Using the standard division of BMI and setting the reference category as BMI within the normal range, we obtain more than a 15-fold higher risk for underweight people (OR[95%CI]=15.14[6.93, 33.10]), more than tenfold for overweight people (OR[95%CI]=10.35[6.74, 15.90]), and more than twelvefold for people with obesity (OR[95%CI]=12.22[6.94, 21.49]).



Copyright ©2010-2026 U-smile Stats – All rights reserved

On the odds ratio plot, the BMI norm is shown at level 1, as the reference category. We have drawn lines connecting the obtained ORs, as well as the norm, to show that the shape of the relationship obtained is identical to that determined earlier through the odds ratio profile.

Example 25.4. (anomaly.pqs file)
A study was conducted to identify risk factors for a certain rare congenital anomaly in children. The study included 395 mothers of children with this anomaly and 375 mothers of healthy children. The data collected were: address of residence, child's sex, child's birth weight, mother's age, pregnancy number, previous spontaneous abortions, respiratory tract infections, smoking, mother's education.

We build a logistic regression model to check which variables may have a significant effect on the occurrence of the anomaly. As the dependent variable, we set the column GROUP, where the distinguished value 1 corresponds to the "cases"group, i.e. mothers of children with the congenital anomaly. The following 9 variables are independent variables: AddressOfRes (2=city/1=village), Sex (1=male/0=female), BirthWeight (in kilograms, to the nearest 0.5 kg), MAge (in years), PregNo (which pregnancy the child is from), SponAbort (1=yes/0=no), RespTInf (1=yes/0=no), Smoking (1=yes/0=no), MEdu (1=primary or lower/2=vocational/3=secondary/4=tertiary).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Logistic Regression	
Analysed variables	GROUP;AddressOfRes;Sex;E
Number of unspecified	0
Number of missing data	89
Significance level	0.05
Sample size assumption n ≥ 10(k+1)	Yes
Sample size assumption n ≥ 10v/p	Yes
Size	678
Number of variables in the model	9
Number of iteration for convergence	895
Convergence criterion met	
Frequency 0 (control)	346
Frequency 1 (case)	332
Likelihood ratio test	
Log Likelihood	-418.0009
-2 Log Likelihood	836.0017
Log Likelihood (intercept)	-469.8092
-2 Log Likelihood (intercept)	939.6185
Chi-square statistic	103.6167
Degrees of freedom	9
p-value	<0.0001
AIC - Akaike criterion	854.0017
AICc - corrected Akaike criterion	854.2712
BIC - Bayesian criterion	894.6741
Pseudo R2	0.1103
R2(Nagelkerke)	0.189
R2(Coxa-Snella)	0.1417
Hosmer-Lemeshow test	
Chi-square statistic	9.8557
Degrees of freedom	8
p-value	0.2753

Model

	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	odds ratio	-95% CI	+95% CI
intercept	1.4739	0.6649	0.1707	2.7771	4.9138	0.0266	4.3662	1.1861	16.0723
AddressOfRes	-0.0409	0.1715	-0.377	0.2953	0.0568	0.8116	0.9599	0.6859	1.3435
Sex	0.4647	0.1701	0.1314	0.798	7.4662	0.0063	1.5915	1.1404	2.2211
BirthWeight	-0.3079	0.1311	-0.5648	-0.051	5.5167	0.0188	0.735	0.5685	0.9503
MAge	-0.0338	0.0187	-0.0704	0.0028	3.2689	0.0706	0.9668	0.9321	1.0028
PregNo	0.2931	0.1004	0.0963	0.49	8.5172	0.0035	1.3406	1.1011	1.6323
SponAbort	-0.4337	0.3032	-1.0279	0.1606	2.0461	0.1526	0.6481	0.3577	1.1742
RespTInf	1.4958	0.2778	0.9514	2.0402	28.9974	<0.0001	4.4628	2.5892	7.6922
Smoking	1.491	0.4119	0.6837	2.2982	13.1048	0.0003	4.4415	1.9813	9.9565
MEdu	-0.1834	0.1012	-0.3818	0.0149	3.2866	0.0698	0.8324	0.6827	1.015

Copyright ©2010-2026 U-smile Stats – All rights reserved

$R^2_{Pseudo} = 0.11$, **likelihood ratio test:** $p < 0.0001$, **Hosmer-Lemeshow test:** $p = 0.2753$

The model fit is not high, but the model is statistically significant, and the result of the Hosmer-Lemeshow test (a lack of significance is desirable here) indicates good agreement between the observed frequencies and the predicted probability.

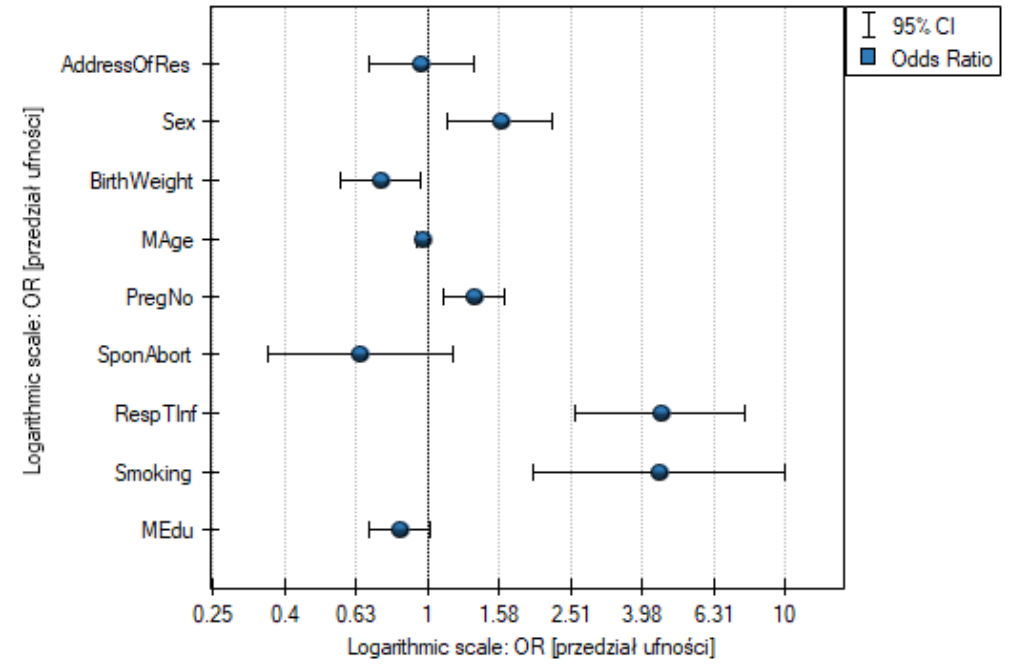
Interpretation of the individual variables in the model begins with checking their significance. In this case, the variables significantly related to the occurrence of the anomaly are: Sex ($p = 0.0063$), BirthWeight ($p = 0.0188$), PregNo ($p = 0.0035$), RespTInf ($p < 0.0001$), Smoking ($p = 0.0003$).

The congenital anomaly studied is rare, but the odds of its occurrence depend on the variables listed above, in the manner described by the odds ratio:

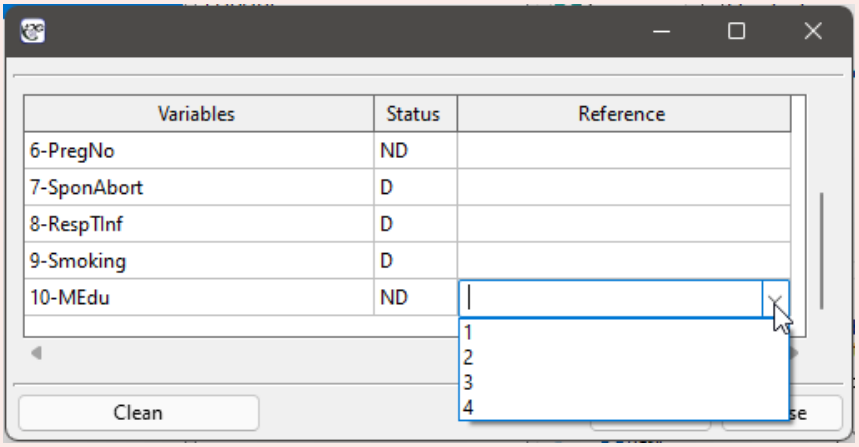
- variable Sex: $OR[95\%CI] = 1.60[1.14; 2.22]$ – the odds of the anomaly occurring in a boy is 1.6 times greater than in a girl;
- variable BirthWeight: $OR[95\%CI] = 0.74[0.57; 0.95]$ – the higher the birth weight, the lower the odds of the anomaly occurring in the child;
- variable PregNo: $OR[95\%CI] = 1.34[1.10; 1.63]$ – the odds of the anomaly occurring in the child increase 1.34-fold with each subsequent pregnancy;
- variable RespTInf: $OR[95\%CI] = 4.46[2.59; 7.69]$ – the odds of the anomaly occurring in the child, when the mother had a respiratory tract infection during pregnancy, are 4.46 times greater than if she did not;
- variable Smoking: $OR[95\%CI] = 4.44[1.98; 9.96]$ – a mother who smokes during pregnancy increases the odds of the anomaly occurring in her child 4.44-fold.

For statistically non-significant variables, the confidence interval for the Odds Ratio contains 1, which means these variables neither increase nor decrease the odds of the anomaly studied occurring. We therefore cannot interpret the resulting ratio in the same way as for statistically significant variables. The effect of individual independent variables on the occurrence of the anomaly can also be described using a plot of the odds ratio:

Copyright ©2010-2026 U-smile Stats – All rights reserved



Continuation of Example 25.4 (anomaly.pqs file): let us build a logistic regression model once again, but this time we will split the mother's education variable into dummy variables (dummy coding). We thereby lose the information about the ordering of the education categories, but we gain the ability to analyse the individual categories in more depth. The split into dummy variables is performed by selecting Dummy var. in the analysis window:



Primary education is selected as the reference category.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Logistic Regression	
Analysed variables	GROUP;AddressOfRes;Sex;E
Number of unspecified	0
Number of missing data	89
Significance level	0.05
Sample size assumption n ≥ 10(k+1)	Yes
Sample size assumption n ≥ 10v/p	Yes
Size	678
Number of variables in the model	11
Number of iteration for convergence	904
Convergence criterion met	
Frequency 0 (control)	346
Frequency 1 (case)	332
Likelihood ratio test	
Log Likelihood	-416.0607
-2 Log Likelihood	832.1214
Log Likelihood (intercept)	-469.8092
-2 Log Likelihood (intercept)	939.6185
Chi-square statistic	107.4971
Degrees of freedom	11
p-value	<0.0001
AIC - Akaike criterion	854.1214
AICc - corrected Akaike criterion	854.5178
BIC - Bayesian criterion	903.832
Pseudo R2	0.1144
R2(Nagelkerke)	0.1955
R2(Coxa-Snella)	0.1466
Hosmer-Lemeshow test	
Chi-square statistic	6.7209
Degrees of freedom	8
p-value	0.567

Model

	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	odds ratio	-95% CI	+95% CI
intercept	1.6651	0.6933	0.3062	3.024	5.7675	0.0163	5.2863	1.3582	20.5744
AddressOfRes	-0.0466	0.173	-0.3856	0.2925	0.0725	0.7878	0.9545	0.68	1.3398
Sex	0.4381	0.171	0.1029	0.7733	6.5635	0.0104	1.5498	1.1084	2.1669
BirthWeight	-0.2959	0.1316	-0.5538	-0.0381	5.06	0.0245	0.7438	0.5748	0.9626
MAge	-0.0341	0.0188	-0.071	0.0028	3.277	0.0703	0.9665	0.9315	1.0028
PregNo	0.2997	0.101	0.1017	0.4976	8.7992	0.003	1.3494	1.107	1.6448
SponAbort	-0.4918	0.3069	-1.0933	0.1098	2.5675	0.1091	0.6116	0.3351	1.116
RespTInf	1.4878	0.2777	0.9436	2.032	28.7124	<0.0001	4.4274	2.5692	7.6294
Smoking	1.4579	0.4145	0.6455	2.2704	12.3713	0.0004	4.2971	1.907	9.6828
MEdu[2]	-0.6823	0.3448	-1.3581	-0.0065	3.9152	0.0479	0.5055	0.2571	0.9936
MEdu[3]	-0.8715	0.3327	-1.5236	-0.2195	6.8633	0.0088	0.4183	0.2179	0.8029
MEdu[4]	-0.7908	0.3599	-1.4963	-0.0853	4.8272	0.028	0.4535	0.224	0.9182

Copyright ©2010-2026 U-smile Stats – All rights reserved

As a result, the variables describing education become statistically significant. The model fit does not change substantially, but the way the odds ratio for education is interpreted does change:

Variable	$OR[95\%CI]$
Primary education	reference category
Vocational education	$0.51[0.26; 0.99]$
Secondary education	$0.42[0.22; 0.80]$
Tertiary education	$0.45[0.22; 0.92]$

The odds of the anomaly studied occurring in each education category are always compared to the odds of the anomaly occurring with primary education. We can see that for more educated mothers, the odds ratio is lower. For a mother with vocational education, the odds of the anomaly occurring in the child is 0.51 of the odds for a mother with primary education; with secondary education, 0.42; with tertiary education, 0.45.

Example 25.5. (task.pqs file)
An experiment was conducted to study the ability of a group of adults to concentrate in uncomfortable situations. 190 people took part in the experiment (130 people formed the training set, 40 people the test set). Each person tested was given a task that required concentration to complete. During the experiment, some people were subjected to a disturbing factor, namely an increase in air temperature to 32 degrees Celsius. Participants were also asked about their address of residence, sex, age, and education. The time allowed to complete the task was limited to 45 minutes. For participants who finished before the time limit, the actual time spent completing the task was recorded. All our calculations will be performed only for those belonging to the training set.

The variable SOLUTION (yes/no) contains the result of the experiment, i.e. information on whether the task was solved correctly or not. The remaining variables that might have influenced the result of the experiment are: ADDRESSOFRES (1=city/0=village), SEX (1=female/0=male), AGE (in years), EDUCATION (1=primary, 2=vocational, 3=secondary, 4=tertiary), TIME needed to complete the task (in minutes), DISTURBANCES (1=yes/0=no). On the basis of all these variables, a logistic regression model was built, in which the distinguished state of the SOLUTION variable was set to “yes”.

Copyright ©2010-2026 U-smile Stats – All rights reserved

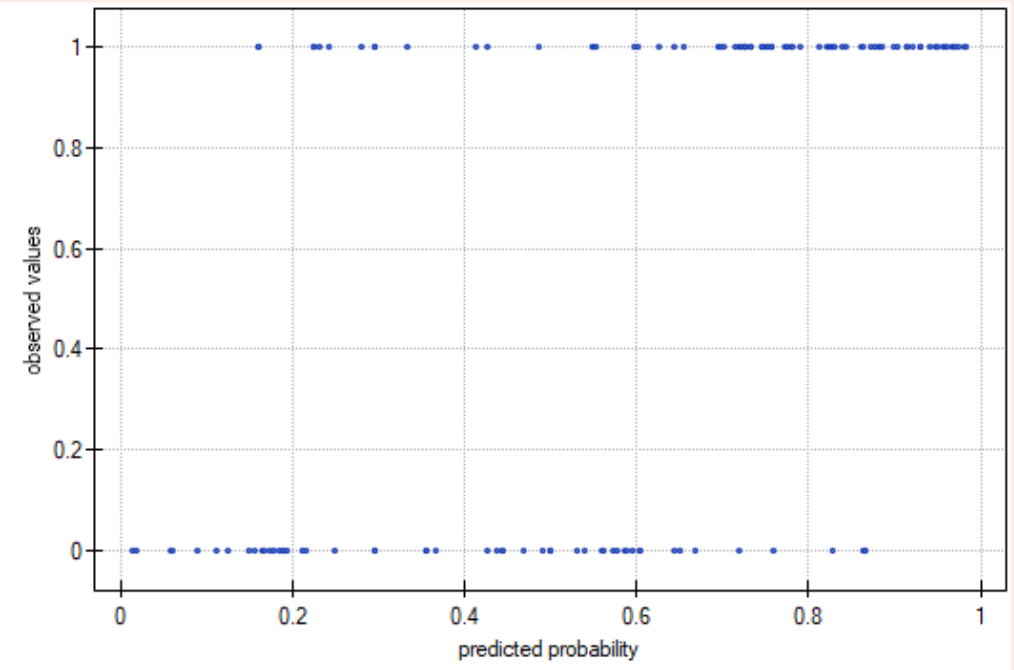
Logistic Regression	
Analysed variables	SOLUTION
	ADDRESSOFRES
	SEX
	AGE
	EDUCATION
	TIME
	DISTURBANCES
Data Filter	set=teaching
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Sample size assumption n ≥ 10(k+1)	Yes
Sample size assumption n ≥ 10v/p	No
Size	130
Number of variables in the model	6
Number of iteration for convergence	158
Convergence criterion met	
Frequency 0 (no)	53
Frequency 1 (yes)	77
Likelihood ratio test	
Log Likelihood	-64.3541
-2 Log Likelihood	128.7082
Log Likelihood (intercept)	-87.881
-2 Log Likelihood (intercept)	175.762
Chi-square statistic	47.0537
Degrees of freedom	6
p-value	<0.0001
AIC - Akaike criterion	140.7082
AICc - corrected Akaike criterion	141.3912
BIC - Bayesian criterion	157.9134
Pseudo R2	0.2677
R2(Nagelkerke)	0.4097
R2(Coxa-Snella)	0.3037
Hosmer-Lemeshow test	
Chi-square statistic	11.5486
Degrees of freedom	8
p-value	0.1725

$R^2_{Pseudo} = 0.27$, $R^2_{Nagelkerke} = 0.41$, $R^2_{Cox-Snell} = 0.30$; **Hosmer-Lemeshow test:** $p = 0.1725$; **likelihood ratio test:** $p < 0.0001$

The model is statistically significant and sufficiently well fitted.

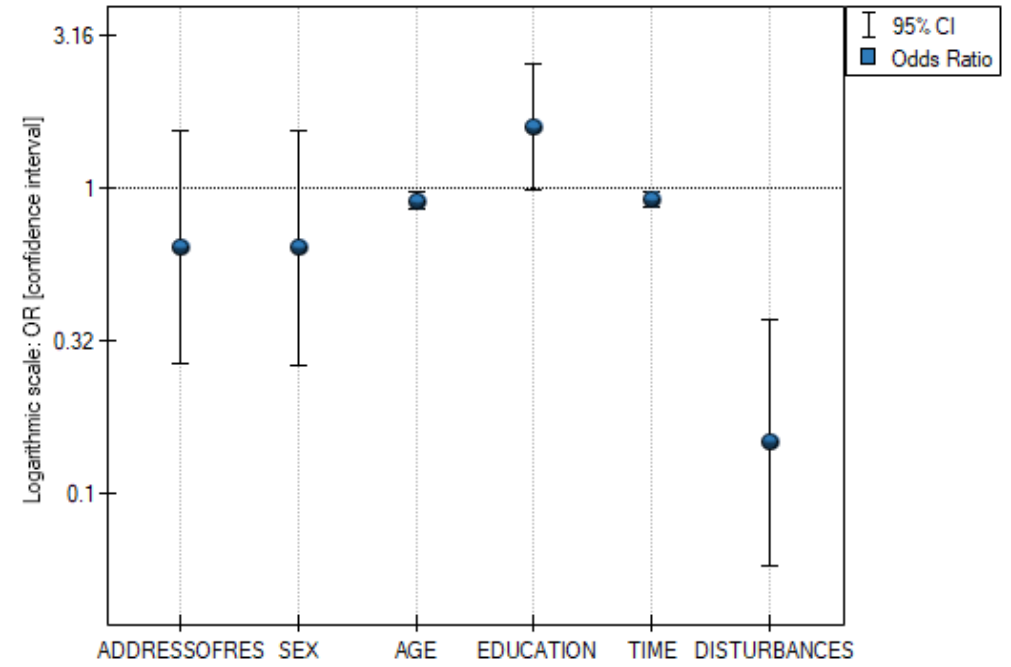
Copyright ©2010-2026 U-smile Stats – All rights reserved

Model									
	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	odds ratio	-95% CI	+95% CI
intercept	7.2306	1.8701	3.5652	10.896	14.9487	0.0001	1381.0527	35.3467	53959.8804
ADDRESSOFRES	-0.4532	0.4505	-1.3363	0.4298	1.0121	0.3144	0.6356	0.2628	1.5369
SEX	-0.4548	0.4513	-1.3393	0.4298	1.0155	0.3136	0.6346	0.262	1.5369
AGE	-0.1009	0.0316	-0.1628	-0.039	10.2009	0.0014	0.904	0.8498	0.9618
EDUCATION	0.4559	0.2418	-0.018	0.9299	3.5552	0.0594	1.5776	0.9822	2.5341
TIME	-0.0894	0.0276	-0.1435	-0.0353	10.4839	0.0012	0.9145	0.8663	0.9653
DISTURBANCES	-1.924	0.4751	-2.8551	-0.9929	16.4029	0.0001	0.146	0.0576	0.3705

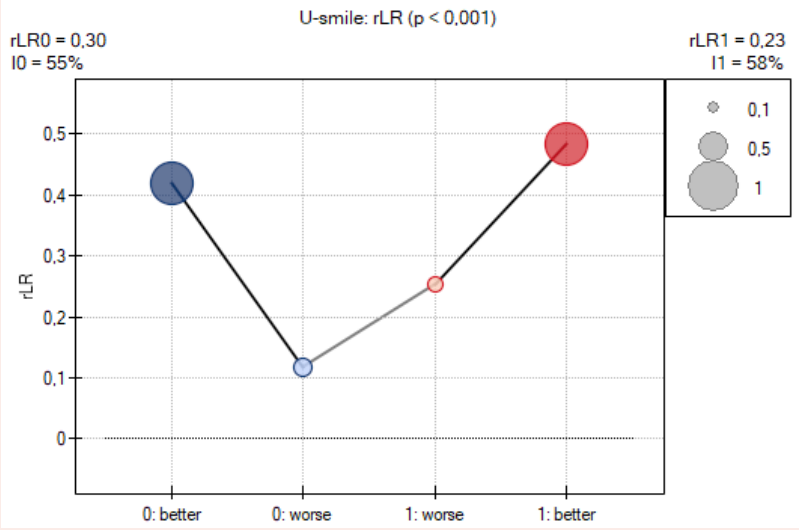


In the model, the variables that have a significant effect on the result are: AGE ($p = 0.0014$), TIME ($p = 0.0012$), DISTURBANCES ($p = 0.0001$). The younger the person solving the task, the shorter the completion time, and the absence of a disturbing factor, are all associated with a greater probability of a correct solution: AGE $OR[95\%CI] = 0.90[0.85; 0.96]$, TIME $OR[95\%CI] = 0.91[0.87; 0.97]$, DISTURBANCES $OR[95\%CI] = 0.15[0.06; 0.37]$.

Copyright ©2010-2026 U-smile Stats – All rights reserved



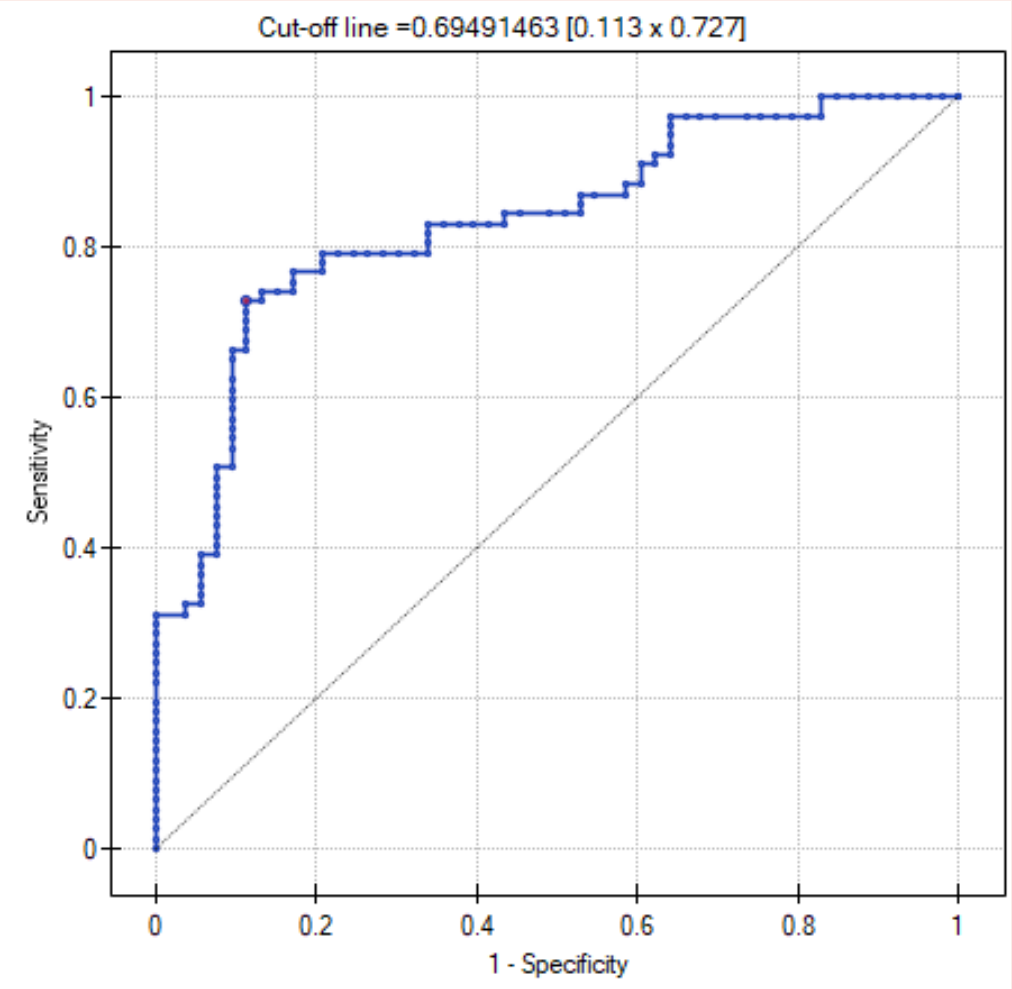
If the model is to be used for forecasting, we should assess whether it will be better at this than random classification. We draw a U-smile plot based on the goodness-of-fit coefficient rLR. The plot "smiles". The likelihood of the prediction indicating that a person will solve the task (the right-hand part of the plot) is 0.23 times greater than in the random model and, net (i.e. after subtracting the people whose likelihood decreased), this concerns 58% of the people. The plot is symmetrical and similar results apply to the left-hand side of the plot, i.e. not solving the task: $rLR_0 = 0.3$, $I_0 = 55\%$.



The classification quality will be assessed on the basis of the ROC curve.

Copyright ©2010-2026 U-smile Stats – All rights reserved

ROC curves (DeLong's method)	
AUC	0.8346
SE(AUC)	0.0355
-95% CI	0.765
+95% CI	0.9042
Z statistic	6.4694
p-value	<0.0001
Cut-off line	0.6949



$AUC = 0.83$, $p < 0.0001$ **(significantly greater than 0.5)**

Classification is possible on the basis of the model built.

The suggested cut-off point for the ROC curve is 0.6949, slightly higher than the standard level used in regression, 0.5. The classification determined on the basis of this cut-off point gives 79.23% of cases correctly classified, of which the correctly classified "yes"values are 72.73% (sensitivity), and "no"values are 88.68% (specificity). The classification obtained from the standard value gives slightly less, 73.85% of cases correctly classified, but yields more correctly classified "yes"values (83.12%), although fewer correctly classified "no"values (60.38%).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Classification		Observed value	
Predicted value		1	0
	1	56	6
	0	21	47
Cut-off line	0.6949		
% correct	79.231%		
Sensitivity (% c	72.727%		
-95% CI	61.38%		
+95% CI	82.259%		
Specificity (% o	88.679%		
-95% CI	76.971%		
+95% CI	95.73%		

Classification		Observed value	
Predicted value		1	0
	1	64	21
	0	13	32
Cut-off line	0.5		
% correct	73.846%		
Sensitivity (% c	83.117%		
-95% CI	72.861%		
+95% CI	90.693%		
Specificity (% o	60.377%		
-95% CI	46.004%		
+95% CI	73.548%		

At this stage, we can end the classification analysis, or, if the result is not satisfactory, we can carry out a more detailed analysis of the ROC curve in the ROC curve module.

However, let us remember that all the assessments were made only on the training set. If the model is really to be used to analyse data on which it was not trained, we should evaluate it on an independent test set. This was done in the Validation and prediction for logistic regression module, where the results were confirmed both on the U-smile plot and on the ROC curve.

Since we have concluded that the classification based on the model is satisfactory, we can calculate the predicted value of the dependent variable for any given conditions. Let us check what odds of solving the task a person has for whom: ADDRESSOFRES (1=city), SEX (1=female), AGE (50 years), EDUCATION (1=primary), TIME needed to complete the task (20 minutes), DISTURBANCES (1=yes). For this purpose, on the basis of the value of coefficient b, the predicted probability is calculated:

$$P(Y = yes|ADDRESSOFRES, SEX, AGE, EDUCATION, TIME, DISTURBANCES) =$$
$$= \frac{e^{7.23-0.45ADDRESSOFRES-0.45SEX-0.1AGE+0.46EDUCATION-0.09TIME-1.92DISTURBANCES}}{1+e^{7.23-0.45ADDRESSOFRES-0.45SEX-0.1AGE+0.46EDUCATION-0.09TIME-1.92DISTURBANCES}} =$$
$$= \frac{e^{7.231-0.453\cdot 1-0.455\cdot 1-0.101\cdot 50+0.456\cdot 1-0.089\cdot 20-1.924\cdot 1}}{1+e^{7.231-0.453\cdot 1-0.455\cdot 1-0.101\cdot 50+0.456\cdot 1-0.089\cdot 20-1.924\cdot 1}}$$

As a result of these calculations, the program returns the result:

Prediction	
ADDRESSOFRES	1
SEX	1
AGE	50
EDUCATION	1
TIME	20
DISTURBANCES	1
Cut-off line	0.5
pred. prob.	0.1215
Pred. Y	0

The predicted probability of solving the task is 0.1215, so, based on the cut-off point 0.5, the predicted outcome is 0 – meaning the task was not solved correctly.

FOR THOSE INTERESTED

Odds Ratio – general formula
Formulas:
For the independent variable X_1, with Z expressed by a linear relationship, we calculate the odds for

Copyright ©2010-2026 U-smile Stats – All rights reserved

the first category:

$$Odds(1) = \frac{P(1)}{1-P(1)} = e^Z(1) = e^{\beta_0+\beta_1 X_1(1)+\beta_2 X_2+...+\beta_k X_k},$$

and the odds for the second category:

$$Odds(2) = \frac{P(2)}{1-P(2)} = e^Z(2) = e^{\beta_0+\beta_1 X_1(2)+\beta_2 X_2+...+\beta_k X_k}.$$

The Odds Ratio for variable X_1 is then expressed by the formula:

$$\begin{aligned} OR_1(2)/(1) &= \frac{Odds(2)}{Odds(1)} = \frac{e^{\beta_0+\beta_1 X_1(2)+\beta_2 X_2+...+\beta_k X_k}}{e^{\beta_0+\beta_1 X_1(1)+\beta_2 X_2+...+\beta_k X_k}} \\ &= e^{\beta_0+\beta_1 X_1(2)+\beta_2 X_2+...+\beta_k X_k-[\beta_0+\beta_1 X_1(1)+\beta_2 X_2+...+\beta_k X_k]} \\ &= e^{\beta_1 X_1(2)-\beta_1 X_1(1)} = e^{\beta_1[X_1(2)-X_1(1)]} = \\ &= \left(e^{\beta_1}\right)^{[X_1(2)-X_1(1)]}. \end{aligned}$$

Example: if the independent variable is age expressed in years, the difference between neighbouring age categories, e.g. 25 and 26 years, is 1 year $(X_1(2) - X_1(1) = 26 - 25 = 1)$. We then obtain the unit Odds Ratio $OR = \left(e^{\beta_1}\right)^1$, which tells us how the odds of the distinguished value occurring change when age changes by 1 year. The odds ratio calculated for non-neighbouring categories, e.g. 25 and 30 years, is the five-year Odds Ratio, $OR = \left(e^{\beta_1}\right)^5$, since the difference $X_1(2) - X_1(1) = 30 - 25 = 5$.

Wald test – significance of individual variables
Hypotheses:

$$\begin{array}{ll} \mathcal{H}_0: & \beta_i = 0, \\ \mathcal{H}_1: & \beta_i \neq 0. \end{array} \quad \text{or, equivalently:} \quad \begin{array}{ll} \mathcal{H}_0: & OR_i = 1, \\ \mathcal{H}_1: & OR_i \neq 1. \end{array}$$

Formulas:
The Wald test statistic is calculated using the formula:

$$\chi^2 = \left(\frac{b_i}{SE_{b_i}}\right)^2.$$

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with 1 degree of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Akaike information criterion

$$AIC = -2 \ln L_{FM} + 2k,$$

an asymptotic criterion – suitable for large samples.

Corrected Akaike information criterion

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}.$$

Bayesian information criterion (Schwarz criterion)

$$BIC = -2 \ln L_{FM} + k \ln(n).$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

Pseudo R^2

$$R^2_{Pseudo} = 1 - \frac{\ln L_{FM}}{\ln L_0}.$$

$$R^2_{Nagelkerke} = \frac{1 - e^{-(2/n)(\ln L_{FM} - \ln L_0)}}{1 - e^{(2/n)\ln L_0}} \quad \text{or} \quad R^2_{Cox-Snell} = 1 - e^{\frac{(-2\ln L_0)-(-2\ln L_{FM})}{n}}.$$

Likelihood ratio test
Hypotheses:

$$\begin{array}{ll} \mathcal{H}_0: & \text{all } \beta_i = 0, \\ \mathcal{H}_1: & \text{there exists } \beta_i \neq 0. \end{array}$$

Formulas:
The test statistic is defined by:

$$\chi^2 = -2\ln(L_0/L_{FM}) = -2\ln(L_0) - (-2\ln(L_{FM})).$$

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with k degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Hosmer-Lemeshow test
Hypotheses:

$$\begin{array}{ll} \mathcal{H}_0: & O_g = E_g \text{ for all categories}, \\ \mathcal{H}_1: & O_g \neq E_g \text{ for at least one category}. \end{array}$$

Formulas:
The test statistic is defined by:

$$H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - \frac{E_g}{N_g})},$$

where N_g – the number of observations in group g.
This statistic asymptotically (for large sample sizes) has the χ^2 distribution with $G - 2$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Test for the significance of AUC
Interpretation:
When the ROC curve coincides with the diagonal $y = x$, the decision to assign a case to the chosen class, based on the model, is exactly as good as a random split of the cases studied into these groups. The classification quality of the model is good when the area under the ROC curve is considerably greater than 0.5.

Hypotheses:

$$\begin{array}{ll} \mathcal{H}_0: & AUC = 0.5, \\ \mathcal{H}_1: & AUC \neq 0.5. \end{array}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

Formulas:
The test statistic is defined by:

$$Z = \frac{AUC - 0.5}{SE_{0.5}},$$

where $SE_{0.5}$ – the standard error of the area. The Z statistic asymptotically (for large sample sizes) has the normal distribution.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

U-smile coefficients
Notation: $y_i \in \{0, 1\}$ – the outcome of the observation, p_i^{ref} and p_i – the event probabilities from the reference model and the new model, n_0, n_1 – the numbers of non-events and events, $n = n_0 + n_1$. Subclasses: 0^+ (non-events, $p_i < p_i^{ref}$), 0^- (non-events, $p_i > p_i^{ref}$), 1^- (events, $p_i < p_i^{ref}$), 1^+ (events, $p_i > p_i^{ref}$); observations with $p_i = p_i^{ref}$ do not belong to any subclass.

Coefficients I**:** $I_0^+ = n_0^+/n_0$, $I_0^- = n_0^-/n_0$, $I_1^- = n_1^-/n_1$, $I_1^+ = n_1^+/n_1$; net $I_0 = I_0^+ - I_0^-$, $I_1 = I_1^+ - I_1^-$; overall $I = I_0 + I_1$.

Coefficients BA **and** RB**:** for the squared residuals $r_i = (y_i - p_i)^2$ and $r_i^{ref} = (y_i - p_i^{ref})^2$ and the differences of the sums of squares in subclass S: $\Delta SS_S = \sum_{i \in S} r_i^{ref} - \sum_{i \in S} r_i$

$$BA_0^+ = \frac{\Delta SS_{0^+}}{n_0}, \quad BA_0^- = \frac{-\Delta SS_{0^-}}{n_0}, \quad BA_1^- = \frac{-\Delta SS_{1^-}}{n_1}, \quad BA_1^+ = \frac{\Delta SS_{1^+}}{n_1}.$$

The RB coefficients are calculated in the same way, but with the denominators $SS_0^{ref} = \sum_{y_i=0} r_i^{ref}$ and $SS_1^{ref} = \sum_{y_i=1} r_i^{ref}$. Net: $BA_0 = BA_0^+ - BA_0^-$, $BA_1 = BA_1^+ - BA_1^-$; overall $BA = \frac{n_0}{n} BA_0 + \frac{n_1}{n} BA_1$ (analogously for RB).

Coefficient rLR**:** for the log-likelihood contributions $\ell_i = y_i \ln p_i + (1 - y_i) \ln(1 - p_i)$ and ℓ_i^{ref} (analogously for p_i^{ref}) and $\Delta_S = \sum_{i \in S} \ell_i^{ref} - \sum_{i \in S} \ell_i$:

$$LR_0^+ = -2\Delta_{0^+}, \quad LR_0^- = 2\Delta_{0^-}, \quad LR_1^- = 2\Delta_{1^-}, \quad LR_1^+ = -2\Delta_{1^+}, \quad LR_0 = LR_0^+ - LR_0^-, \quad LR_1 = LR_1^+ - LR_1^-$$

These values are normalised by the maximum attainable likelihood ratio of the class: $\max LR_0 = -2\sum_{y_i=0} \ell_i^{ref}$ (when $LR_0^+ \geq LR_0^-$; otherwise from the ℓ_i of the new model), and analogously $\max LR_1$. Hence $rLR_0^{\pm} = LR_0^{\pm}/\max LR_0$, $rLR_1^{\pm} = LR_1^{\pm}/\max LR_1$, $rLR_0 = LR_0/\max LR_0$, $rLR_1 = LR_1/\max LR_1$ and $rLR = (LR_0 + LR_1)/\max LR$, where $\max LR = -2\sum_i \ell_i^{ref}$ (the sum over the observations belonging to the subclasses).

Tests: overall – the likelihood ratio test with the statistic $2\left|\sum_i \ell_i - \sum_i \ell_i^{ref}\right|$, which asymptotically has the χ^2 distribution with df equal to the difference in the number of parameters of the compared models (at least 1). For rLR, the significance of the change within a class is checked using the gamma distribution: $|LR_1| \sim \Gamma\left(\frac{df}{2}, \frac{1}{2c_1}\right)$ and $|LR_0| \sim \Gamma\left(\frac{df}{2}, \frac{1}{2c_0}\right)$, where $c_1 = n_1/n$, $c_0 = 1 - c_1$ (shape and rate parameters). The determined p values are compared with the significance level α: if $p \leq \alpha$, the change in prediction is considered significant.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Model-based prediction and test set validation

Validation and prediction for logistic regression

WHEN TO USE:

Validation of a model is a check of its quality. It is first performed on the data on which the model was built (the so-called **training set**), that is, it is returned in the report describing the resulting model. In order to be able to judge with greater confidence how suitable the model is for forecasting new data, an important part of validation is applying the model to data that were not used in estimating the model. If the summary based on the training data is satisfactory, i.e. the calculated errors, R^2 coefficients and information criteria are at a satisfactory level, and the summary based on the new data (the so-called **test set**) is equally favourable, then it can be concluded with high probability that such a model is suitable for prediction. The test data should come from the same population from which the training data were selected. It is often the case that, before building the model, we collect the data and then randomly split them into a training set, i.e. the data that will be used to build the model, and a test set, i.e. the data that will be used for the additional validation of the model.

Prediction – most often the last stage of regression analysis is the use of the built and previously verified model for prediction. On the basis of new data, the value of the probability predicted by the model is determined, followed by the prediction of the occurrence of the event (1) or its absence (0).

DEFINITIONS OF MEASURES

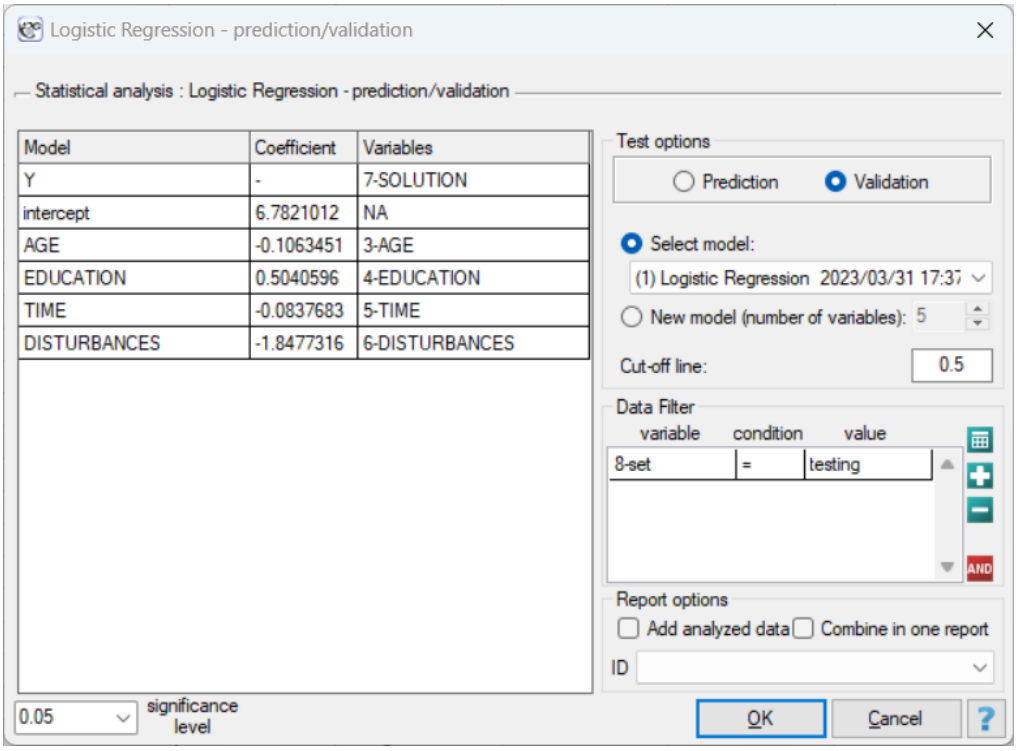
The prediction of the occurrence of the event (1) or its absence (0) is based on comparing the value of the probability predicted by the model with the so-called **cut-off point**. By default, the cut-off point is set to 0.5, but the user can change it to any value from the range $(0-1)$, e.g. the value suggested by the ROC curve.

U-smile on the validation set – when, in validation mode, a variable with the observed outcome is indicated, the coefficients and the U-smile plot of the selected model against the null model (the event frequency in the validation set) can additionally be determined. This corresponds to evaluating the model on data independent of the training data: a smile confirms that the model improves the prediction of both classes also outside the training sample, while asymmetry indicates the class for which the improvement does not hold.

HOW TO SET UP THE ANALYSIS

Advanced statistics $\rightarrow$ Multivariate models $\rightarrow$ Logistic regression – prediction/validation.

Copyright ©2010-2026 U-smile Stats – All rights reserved



To perform validation or prediction, it is necessary to indicate the model on the basis of which we want to perform it. This can be done on the basis of:

- a logistic regression model built in PQStat – simply select a model from those assigned to the given sheet, and the number of variables and model coefficients will be set automatically; the test set should be in the same sheet as the training set;
- a model not built in PQStat but obtained from another source (e.g. described in a scientific paper we have read) – in the analysis window, enter the number of variables and the coefficients for each of them.

In the analysis window, indicate the new variables that should be used for validation or prediction. Prediction for a single object can be performed together with building the model, that is, in the analysis window Advanced statistics→Multivariate models→Logistic regression, while prediction for a larger group of new data is performed via the menu Advanced statistics→Multivariate models→Logistic regression – prediction/validation.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 25.6. continuation of Example 25.5 (task.pqs file)
In the experiment studying the ability to concentrate, for the group of 130 people in the training set, a logistic regression model was built based on the following variables:

dependent variable: SOLUTION (yes/no) – information about whether the task was solved correctly or not;

independent variables:

Copyright ©2010-2026 U-smile Stats – All rights reserved

ADDRESSOFRES (1=city/0=village),
SEX (1=female/0=male),
AGE (in years),
EDUCATION (1=primary, 2=vocational, 3=secondary, 4=tertiary),
TIME needed to complete the task (in minutes),
DISTURBANCES (1=yes/0=no).

However, only four variables – AGE, EDUCATION, TIME needed to complete the task, and DISTURBANCES – contribute significant information to the model. We will build the model for the training set data based on these four variables, and then, to make sure it works correctly, we will validate it on the test data set. If the model passes this test, we will apply it to predict outcomes for new individuals. To use the appropriate sets, we set a data filter each time.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Logistic Regression - prediction/validation	
Analysed variables	AGE
	EDUCATION
	TIME
	DISTURBANCES
	SOLUTION
Data Filter	set=traning
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	130
Validation	
Likelihood ratio test	
Log Likelihood	-65.5411368
-2 Log Likelihood	131.0822737
Log Likelihood (intercept)	-87.8809897
-2 Log Likelihood (intercept)	175.7619793
Chi-square statistic	44.6797056
Degrees of freedom	4
p-value	4.6347E-9
AIC - Akaike criterion	139.0822737
AICc - corrected Akaike criterion	139.4022737
BIC - Bayesian criterion	150.5524115
Pseudo R2 (McFadden)	0.2542057
R2(Nagelkerke)	0.3923627
R2(Coxa-Snella)	0.2908513
Hosmer-Lemeshow test	
Chi-square statistic	15.0231507
Degrees of freedom	8
p-value	0.0586969
ROC curves	
AUC	0.8230826
SE(AUC)	0.036776
-95% CI	0.751003
+95% CI	0.8951622
Z statistic	6.2467167
Two sided p-value	4.1917E-10
Cut-off line	0.5649961

training set: $Pseudo\ R^2 = 0.26$, $R^2(Nagelkerke) = 0.39$; $AUC[95\%CI] = 0.82[0.75, 0.90]$, **sensitivity** $= 82\%$, **specificity** $= 60\%$

The values describing the model fit are not very high, but the quality of its prediction is already satisfactory.

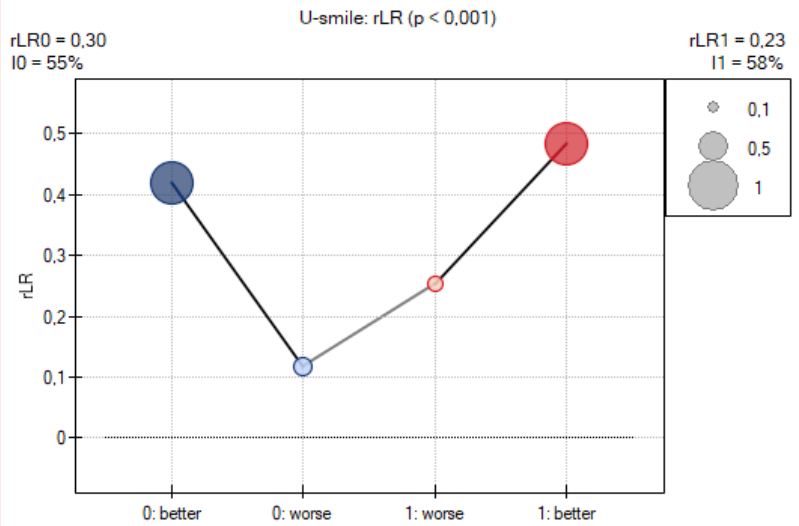
Copyright ©2010-2026 U-smile Stats – All rights reserved

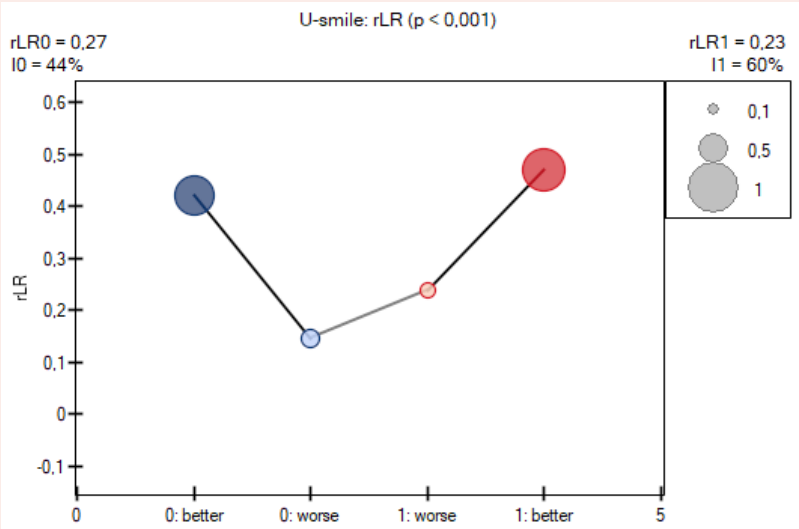
Logistic Regression - prediction/validation	
Analysed variables	AGE
	EDUCATION
	TIME
	DISTURBANCES
	SOLUTION
Data Filter	set=testing
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	60
Validation	
Likelihood ratio test	
Log Likelihood	-30.3865669
-2 Log Likelihood	60.7731337
Log Likelihood (intercept)	-87.8809897
-2 Log Likelihood (intercept)	175.7619793
Chi-square statistic	114.9888456
Degrees of freedom	4
p-value	0
AIC - Akaike criterion	68.7731337
AICc - corrected Akaike criterion	69.5004065
BIC - Bayesian criterion	77.150512
Pseudo R2 (McFadden)	0.6542305
R2(Nagelkerke)	0.9010184
R2(Coxa-Snella)	0.8528762
Hosmer-Lemeshow test	
Chi-square statistic	7.4731153
Degrees of freedom	8
p-value	0.4865496
ROC curves	
AUC	0.8171429
SE(AUC)	0.0552507
-95% CI	0.7088534
+95% CI	0.9254323
Z statistic	4.160875
Two sided p-value	0.0000317
Cut-off line	0.6132695

test set: $Pseudo\ R^2 = 0.65$, $R^2(Nagelkerke) = 0.90$; $AUC[95\%CI] = 0.82[0.71, 0.93]$, **sensitivity** $= 73\%$, **specificity** $= 64\%$

The values describing the model fit on the test set are even higher than for the training data, and the prediction quality is still satisfactory – the model is suitable for prediction. The similarity of the assessments on the training and test sets can also be seen on the U-smile plot, where the first figure comes from the training set and the second from the test set.

Copyright ©2010-2026 U-smile Stats – All rights reserved





For this purpose, we will use the data for three new individuals added at the end of the set. We select the Prediction option, set the filter to the new data set, and use our model to predict whether a given person will solve the task correctly (obtaining a value of 1) or incorrectly (obtaining a value of 0).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Logistic Regression - prediction/validation	
Analysed variables	AGE
	EDUCATION
	TIME
	DISTURBANCES
Data Filter	set=new
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	3
Prediction	
Cut-off line	0.5
frequency of 0	1
frequency of 1	2
percentage of 0	33.333%
percentage of 1	66.667%

Coefficients of model	
intercept	6.7821012
AGE	-0.1063451
EDUCATION	0.5040596
TIME	-0.0837683
DISTURBANCES	-1.8477316

Prediction	
Pred. of probabi	Pred. Y
0.1744817	0
0.6393883	1
0.962964	1

It turns out that the prognosis for the first person is negative, and for the next two, positive. The prognosis for a 50-year-old woman with primary education, solving the test under disturbances in 20 minutes, is 0.17, meaning we predict she will solve the task incorrectly, while the prognosis for a woman 20 years younger is already favourable – the probability of her solving the task is 0.64. The highest probability (equal to 0.96) of solving the task correctly is for the third woman, who solved the test in 10 minutes and without disturbances.

If we wanted to make a prediction based on a different model (e.g. one obtained during another scientific study: SOLUTION$= 6 - 0.1\cdot$AGE$+0.5\cdot$EDUCATION$-0.1\cdot$TIME$-2\cdot$DISTURBANCES) – it is enough to select the new model in the analysis window, set its coefficients, and the prediction for the selected individuals can be repeated based on this model.

Logistic Regression - prediction/validation	
Analysed variables	AGE
	EDUCATION
	TIME
	DISTURBANCES
Data Filter	set=new
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	3
Prediction	
Cut-off line	0.5
frequency of 0	2
frequency of 1	1
percentage of 0	66.667%
percentage of 1	33.333%

Coefficients of model	
intercept	6
age	-0.1
education	0.5
time	-0.1
disturbances	-2

Prediction	
Pred. of probabi	Pred. Y
0.0758582	0
0.3775407	0
0.9241418	1

This time, according to the prediction of the new model, the predictions for the first and second indivi-

Copyright ©2010-2026 U-smile Stats – All rights reserved

duals are negative, and for the third, positive.

Copyright ©2010-2026 U-smile Stats – All rights reserved

25.4 COMPARISON OF LOGISTIC REGRESSION MODELS

Comparing logistic regression models

WHEN TO USE:

Because it is possible to analyse many independent variables simultaneously in a single logistic regression model, just as in multiple linear regression, the problem of choosing the optimal model arises. When selecting independent variables, you should keep in mind that the model should contain variables strongly correlated with the dependent variable and weakly correlated with one another.

Comparing logistic regression models lets you answer the question of whether adding (or removing) independent variables significantly changes the quality of the model. If two models do not differ significantly, it is worth choosing the simpler one – the one with fewer variables. In addition, when we are interested in the practical usefulness of a new risk factor, we can compare how much the model's ability to predict the event under study improves once this factor is added.

ASSUMPTIONS

- the models being compared are built on the same sample (the same observations),
- in the comparison based on the likelihood ratio test, the reduced model is "nested" within the full model, i.e. it is obtained from the full model by removing some of the independent variables,
- in the comparison of predictive value (ROC curves, NRI, IDI), the predicted probability for both models was determined for the same objects.

DEFINITIONS OF MEASURES

When comparing models with a different number of independent variables, we pay attention to the model fit and to the information criteria. We decide which model to choose on the basis of the values of R^2_{Pseudo}, $R^2_{Nagelkerke}$, $R^2_{Cox-Snell}$ and the result of the likelihood ratio test, which compares successive (neighbouring) models as they are created. For every model the maximum of the likelihood function is calculated, and the likelihood ratio test checks whether the full and reduced models differ significantly in this respect. If the compared models do not differ significantly, we should choose the one with fewer variables – the lack of a difference means that the variables present in the full model but absent from the reduced model do not carry significant information. If, on the other hand, the difference is statistically significant, this means that one of the models (the one with more variables, a larger R^2 and a lower value of the AIC, AICc or BIC information criterion) is significantly better than the other.

Comparing the predictive value of models – the regression models that are built let us predict the probability of the event under study occurring, based on the independent variables analysed. When many factors that increase the risk of the event are already known, an important criterion for a new candidate risk factor is how much it improves prediction performance once it is added to the model. For example, when studying risk factors for coronary heart disease (age, systolic and diastolic blood pressure, obesity, cholesterol, smoking), we can check whether adding a further factor significantly

Copyright ©2010-2026 U-smile Stats – All rights reserved

improves the estimation of disease risk. The predictive value of a model follows from the predicted probability determined for every person studied – the closer this value is to 1, the more likely the event is to occur. Based on the predicted probability, we can determine and compare between models the area AUC under the ROC curve, as well as the NRI and IDI coefficients.

- **Change in the area under the ROC curve** – in logistic regression models the ROC curve is built based on the classification of cases into the group experiencing the event or not, and on the predicted probability of the dependent variable. The larger the area under the curve, the more accurately the probability determined by the model predicts the actual occurrence of the event. By comparing the area under the ROC curve for the full and the reduced model, we therefore check whether adding the factors significantly improved the model's prediction (DeLong's method, described in more detail in the Comparing ROC curves chapter).
- **Net Reclassification Improvement (NRI)** – a measure based on the reclassification table, describing the upward or downward shift of the predicted probability once a new factor is added to the model. It is determined separately for objects experiencing the event and for those that did not experience it. NRI can be determined for a given split of the predicted probability into categories (categorical NRI) or without the need to determine categories (continuous NRI).
  - **Categorical** NRI requires the split points of the predicted probability to be set arbitrarily (a maximum of 9 points, i.e. a maximum of 10 categories; most often one or two points are used). The values of the categorical NRI can only be compared with one another when they are based on the same split points. For example, for the split points 0.1 and 0.3: if, in the "old" (smaller) model, a person obtained a probability below 0.1, and in the "new" model – between 0.1 and 0.3, then that person was reclassified upward. If the value from both models falls within the same interval, the person was not reclassified, and if the probability from the "new" model is lower than in the "old" model, we are dealing with a downward reclassification.

regression models	reclassification upward		no reclassification		reclassification downward	
predicted probability	"old"	"new"	"old"	"new"	"old"	"new"
[0.3 to 1]					⊕	
[0.1; 0.3)		⊕	⊕	⊕		⊕
[0; 0.1)	⊕					

  - **Continuous** NRI does not require categories to be determined arbitrarily – every change in probability, even the smallest, upward or downward relative to the probability from the "old" model, is treated as a move into the next category.

Note! Using continuous NRI does not require the arbitrary definition of probability split points, but even small changes in risk (with no counterpart in clinical observation) can influence the value of this coefficient. Categorical NRI, on the other hand, reflects only those changes that the researcher judged to be important when defining the probability split points. The overall NRI coefficient lies in the range from -2 to 2, while the partial coefficients NRI_{events} and $NRI_{nonevents}$ lie in the range from -1 to 1 (from -100% to 100%). Positive values indicate a favourable reclassification, and negative values – an unfavourable reclassification, as a result of adding the new variable to the model. The statistical significance of the obtained NRI value is checked with a dedicated Z test.

- **Integrated Discrimination Improvement (IDI)** – a coefficient showing the difference between the mean change in predicted probability (between the "old" and "new" model) in the group of

Copyright ©2010-2026 U-smile Stats – All rights reserved

objects experiencing the event and the group of objects that did not experience the event. As with NRI, the statistical significance of IDI is checked with a dedicated Z test.

U-smile – the U-smile method described in the Logistic regression chapter can also be used to compare successive (neighbouring) models: the reference model is the earlier model on the list, and the new model is the next one. The report contains the rLR, BA or RB coefficients (to choose from in the AUC-IDI-NRI window) for the four subclasses, the net coefficients for the classes, the overall coefficient, the shares I and the significance tests; a U-smile plot can be included for each pair. The p values of the tests assume nested models – for non-nested pairs (the all subsets method) they should be treated as indicative only.

In PQStat, model comparison can be carried out manually or automatically.

- **Manual** model comparison consists of building 2 models:
  - the full model – the model with more variables,
  - the reduced model – the model with fewer variables; such a model is obtained from the full model by removing the variables that, from the point of view of the phenomenon under study, are superfluous.

  The choice of independent variables in the compared models, and then the choice of the better model on the basis of the results of the comparison, is up to the researcher.

- **Automatic model selection** – the program builds successive models itself and indicates the best one according to the chosen criterion.

  Model terms. The unit that the automatic methods add to or remove from the model is a term: a single variable, a whole dummy variable (all its 0–1 columns at once) or an interaction. The hierarchy principle applies: an interaction can be in the model only when its main effects are in it, and only the interaction can be removed before them.

  The selection criterion is chosen in the analysis window:
  - AIC, AICc or BIC – information criteria; the model with the lower value of the criterion is better. A move (adding or removing a term) is accepted only when it lowers the criterion;
  - p of the nested test – the likelihood ratio test comparing the model before and after the change, with a given p threshold. A term is removed when p is greater than the threshold, and added when p is smaller than the threshold. The same p value is shown in the report as the comparison of neighbouring models;
  - validation error – the mean deviance determined on observations that did not take part in fitting the model (described below: Out-of-sample evaluation of models). A move is accepted only when it reduces this error. The criterion requires the Validation option to be switched on – choosing it while validation is off switches cross-validation on.

  Selection methods:
  - **backward elimination** – starts from the model with all the variables; in each step the term whose removal lowers the criterion the most is removed; it ends when no removal lowers the criterion any further;
  - **forward selection** – starts from the null model (intercept only); in each step the term that lowers the criterion the most is added; it ends when no addition lowers the criterion any further;

Copyright ©2010-2026 U-smile Stats – All rights reserved

- **bidirectional stepwise** – starts from the null model; in each step both additions and removals of terms are considered and the best move is chosen; it ends when no move improves the criterion;
- **all subsets** – all models consistent with the hierarchy principle are fitted (a maximum of 15 terms), and the best model of each size is presented in the report; this method uses the information criteria or the validation error (the p criterion is not available in it).

All the models are computed on the same set of observations – rows with missing values in any of the candidate variables are discarded once, before the analysis. Thanks to this, the information criteria, the likelihood ratio tests and the comparisons of ROC curves, NRI and IDI are comparable between models.

The report contains: the selection method and criterion, the number of terms and of fitted models, the information criteria of the null model, the successive models of the selection path (each with the value of $-2 \ln L$, the AIC, AICc and BIC criteria, the R^2_{Pseudo}, $R^2_{Nagelkerke}$ and $R^2_{Cox-Snell}$ coefficients, and the table of coefficients and odds ratios) and the comparison of each neighbouring pair of models with the likelihood ratio test (the χ^2 statistic, degrees of freedom, p value). If the evaluation of predictive value was selected in the analysis window, the comparison of ROC curves, NRI and IDI are also determined for each neighbouring pair of models – regardless of whether the models are nested. The best model according to the chosen criterion is marked in the report. If no term improves the null model, the report states that the null model is the best. With the all subsets method, neighbouring models may not be nested – in that case the likelihood ratio test is not performed.

Note! The p values and confidence intervals in an automatically selected model are too optimistic, because the same data set was used both to select the variables and to estimate their effects. Automatic selection should be treated as an aid in searching for a model, and the final choice should be based on subject-matter knowledge. The selected model should be built again in the Logistic regression window and its quality assessed.

Out-of-sample evaluation of models

The fit measures, the information criteria and the tests, as well as the ROC curves, NRI and IDI, are determined on the same observations on which the coefficients of the models were estimated. Such an assessment is too optimistic – a model usually predicts new cases worse than its fit would suggest, and the difference grows with the number of variables. For this reason, as in comparing multiple regression models, the analysis window makes it possible to evaluate the models on observations that did not take part in fitting them (the Validation option): **k-fold cross-validation** (each model is additionally fitted k times, each time without one of the k parts of the data, and evaluated on the part left out) or a **random test set** (the models are fitted on the training part and evaluated on the test part set aside). The split is random, determined by the Seed, and **stratified by the dependent variable**: the observations with and without the event are split separately, so that the percentage of events is similar in every part.

For each model the report then gives three additional measures:

- **mean deviance** ($-2 \ln L/n$) – the value of $-2 \ln L$ determined from the probabilities predicted out of sample, divided by the number of evaluated observations. The lower it is, the better the model; the measure strongly "penalises" predictions that are confident but wrong. This is the **validation error** – the best model by validation is indicated according to it, and the validation error selection criterion is based on it;

Copyright ©2010-2026 U-smile Stats – All rights reserved

- **AUC** – the area under the ROC curve built from the probabilities predicted out of sample; it describes the ability of the model to distinguish new objects with the event from objects without the event (0.5 – a useless model, 1 – perfect discrimination);
- **Brier score** – the mean squared difference between the predicted probability and the actual outcome (1 – event, 0 – no event)[29]; the lower it is, the better the predictions.

The point of reference is the **null model**, which assigns to every observation the frequency of the event from the training part – its mean deviance and Brier score are given in the model selection block of the report (the AUC of such a model is 0.5). The same block indicates the **best model by validation** – regardless of the selection criterion chosen. Out-of-sample evaluation also works in the manual comparison of models.

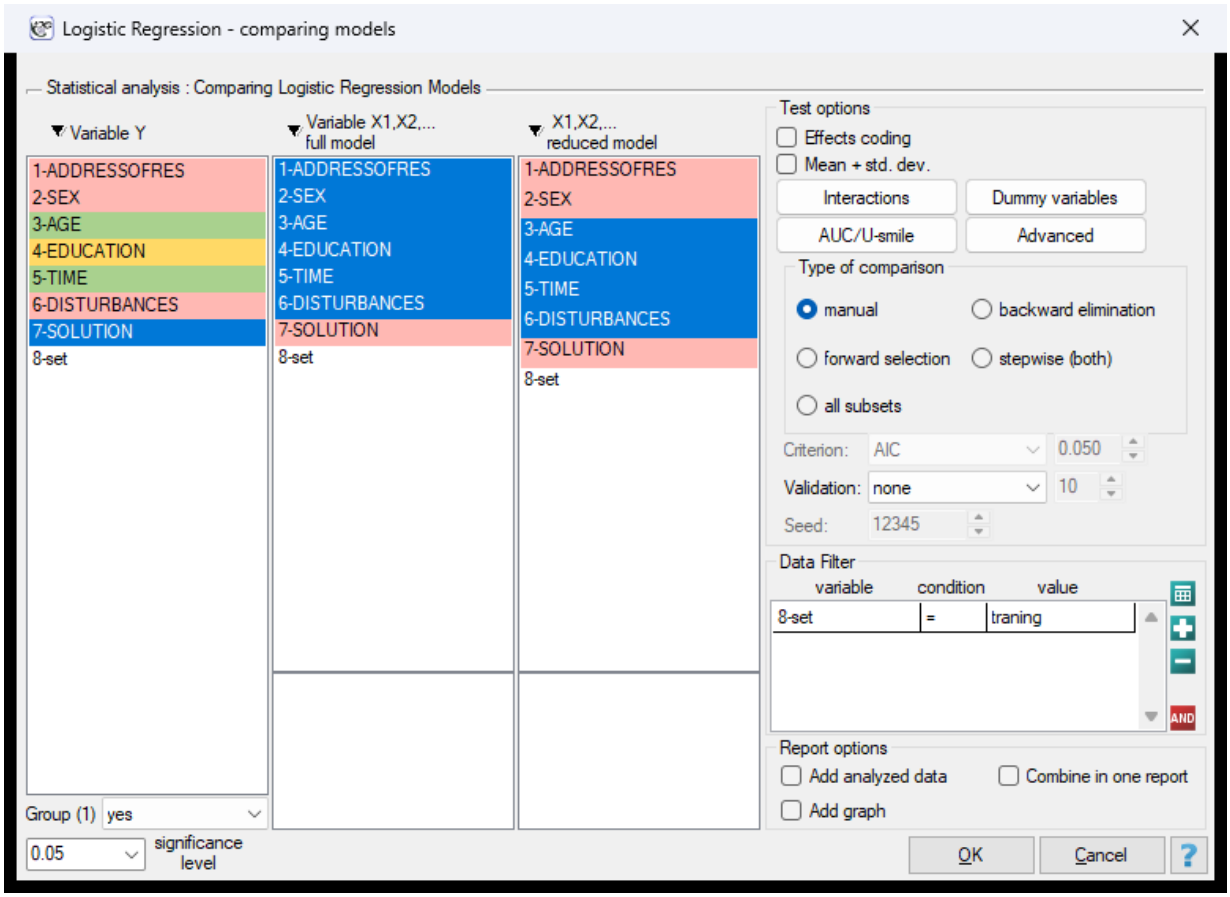
When validation is switched on, the **comparison of the predictive value of the models** (ROC curves, NRI, IDI, U-smile) is also determined from the probabilities predicted out of sample: in cross-validation for all the observations, and with a test set – for the observations of the test set. The report marks this with the row ROC, IDI, NRI, U-smile: validation predictions. These results then show whether the additional variables improve the prediction of new cases, and not only the fit to the data on which the model was built.

Note! When the validation error is the selection criterion, the validation error of the selected model is slightly too optimistic – out of many models, the one that performed best on the given split was chosen. With a small number of events, the validation results depend on the random split – it is then worth checking them for several values of the seed.

HOW TO SET UP THE ANALYSIS

Statistics → Multivariate models → Logistic regression – comparing models.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In the window we indicate the dependent variable Y with the value denoting the event, and the independent variables of the full and the reduced model (in automatic selection – the candidate variables). Model selection options:

- Type of comparison – manual or one of the automatic selection methods,
- Criterion – AIC, AICc, BIC, p (LR / F test) with the threshold given in the adjacent field, or validation error; in the manual mode the criterion is not used,
- Validation – none (evaluation on the training set only), k-fold CV, k = with the number of folds in the adjacent field (from 2 to 100, 10 by default) or test set, % = with the share of the test set as a percentage of the observations (from 5 to 90, 30 by default),
- Seed – the seed of the random split of the data (12345 by default); the same seed means the same split,
- AUC/U-smile – the choice of the comparisons of predictive value (ROC curves, NRI, IDI, U-smile) and of the related plots.

PERFORMING AND INTERPRETING THE ANALYSIS

Continuing example 25.5 (task.pqs file): in an experiment studying concentration ability, a logistic regression model was built for 130 people from the training set, based on the following variables:

Copyright ©2010-2026 U-smile Stats – All rights reserved

dependent variable: SOLUTION (yes/no) – information about whether the task was solved correctly or not;

independent variables:

ADDRESSOFRES (1=city/0=village),
SEX (1=female/0=male),
AGE (in years),
EDUCATION (1=primary, 2=vocational, 3=secondary, 4=tertiary),
TIME needed to solve the task (in minutes),
DISTURBANCES (1=yes/0=no).

We will check whether all the independent variables are needed in the model.

Manual model comparison

Based on the full model built earlier, we can suspect that the variables ADDRESSOFRES and SEX have little influence on the model built (that is, we cannot successfully classify cases on the basis of them). We will check whether, from a statistical point of view, the full model is better than the model with these two variables removed.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Comparing Logistic Regression Models	
Analysed variables	SOLUTION
	ADDRESSOFRES
	SEX
	AGE
	EDUCATION
	TIME
	DISTURBANCES
Data Filter	set=teaching
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	130
Convergence criterion met	
Number of variables in the model 1	6
AIC - Akaike criterion	140.7082
AICc - corrected Akaike criterion	141.3912
BIC - Bayesian criterion	157.9134
-2 Log Likelihood	128.7082
Pseudo R2	0.2677
R2(Nagelkerke)	0.4097
R2(Coxa-Snella)	0.3037
Convergence criterion met	
Number of variables in the model 2	4
AIC - Akaike criterion	139.0823
AICc - corrected Akaike criterion	139.4023
BIC - Bayesian criterion	150.5524
-2 Log Likelihood	131.0823
Pseudo R2	0.2542
R2(Nagelkerke)	0.3924
R2(Coxa-Snella)	0.2909
Comparison - Model 1 vs Model 2	
Chi-square - comparing	2.374
Degrees of freedom	2
p-value	0.3051

Model 1	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	Odds Ratio	-95% CI	+95% CI
intercept	7.2306	1.8701	3.5652	10.896	14.9487	0.0001	1381.0527	35.3467	53959.8804
ADDRESSOFRES	-0.4532	0.4505	-1.3363	0.4298	1.0121	0.3144	0.6356	0.2628	1.5369
SEX	-0.4548	0.4513	-1.3393	0.4298	1.0155	0.3136	0.6346	0.262	1.5369
AGE	-0.1009	0.0316	-0.1628	-0.039	10.2009	0.0014	0.904	0.8498	0.9618
EDUCATION	0.4559	0.2418	-0.018	0.9299	3.5552	0.0594	1.5776	0.9822	2.5341
TIME	-0.0894	0.0276	-0.1435	-0.0353	10.4839	0.0012	0.9145	0.8663	0.9653
DISTURBANCES	-1.924	0.4751	-2.8551	-0.9929	16.4029	0.0001	0.146	0.0576	0.3705

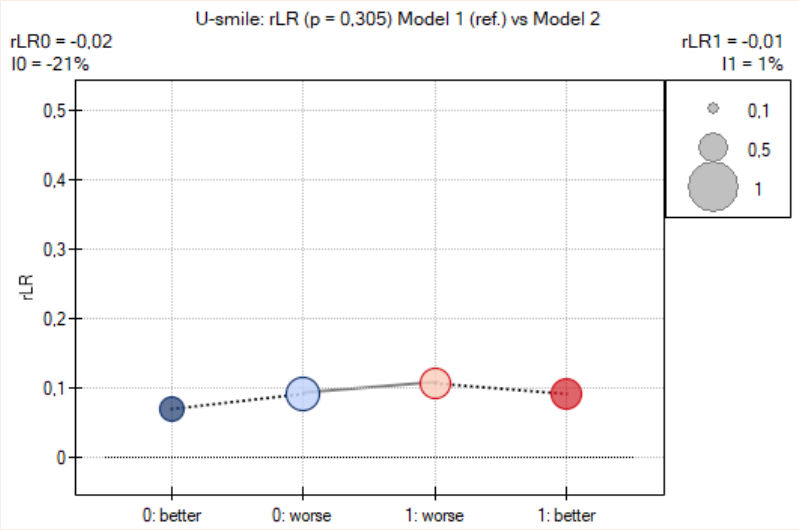
Copyright ©2010-2026 U-smile Stats – All rights reserved

Model 2	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	Odds Ratio	-95% CI	+95% CI
intercept	6.7821	1.8219	3.2113	10.3529	13.8575	0.0002	881.9198	24.8105	31348.9418
AGE	-0.1063	0.0316	-0.1683	-0.0444	11.3179	0.0008	0.8991	0.8451	0.9566
EDUCATION	0.5041	0.2373	0.0389	0.9692	4.5116	0.0337	1.6554	1.0397	2.6358
TIME	-0.0838	0.0268	-0.1364	-0.0312	9.742	0.0018	0.9196	0.8725	0.9693
DISTURBANCES	-1.8477	0.462	-2.7532	-0.9423	15.9967	0.0001	0.1576	0.0637	0.3897

likelihood ratio test: $p = 0.3051$

There is no reason to think that the full model is better than the reduced model. Therefore, at the cost of a negligible loss of model quality, place of residence and sex can be omitted.

A good illustration of how the likelihood ratio test works is the U-smile plot. We can see that the likelihood of the predictions of the compared models is quite similar – the plot is almost flat (a slightly "sad face"). The smaller model is only non-significantly worse than the larger one – the dashed line in both parts of the plot indicates that the deterioration is statistically non-significant both in the group of people who did not solve the task (the left-hand part of the plot) and in the group of people who solved the task (the right-hand part of the plot).



The two models can also be compared in terms of classification ability by comparing their ROC curves and the NRI and IDI values – for this, we select the appropriate option in the analysis window. The resulting report, like the previous one, indicates that the models do not differ in prediction quality, that is, the p values both for the comparison of ROC curves and for the NRI and IDI indices are statistically non-significant. We therefore decide to omit sex and place of residence from the final model.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Comparison - Model 1 vs Model 2	
Comparison of ROC curves [AUC difference]	
AUC Model 1	0,834599
SE(AUC)	0,035518
-95% CI	0,764986
+95% CI	0,904213
AUC Model 2	0,823083
SE(AUC)	0,036776
-95% CI	0,751003
+95% CI	0,895162
AUC1-AUC2 [DeLong's method]	0,011517
SE(AUC1-AUC2)	0,010864
-95% CI	0
+95% CI	0,032809
Z statistic	1,06013
p-value	0,289085
Chi-square - models comparison Model 1 vs Model 2	2,374039
Degrees of freedom	2
p-value	0,305129
U-smile Model 1 (ref.) vs Model 2	
Coefficient	rLR
Observations with unchanged prediction	0
rLR	
non-events (0), better prediction	0,069448
non-events (0), worse prediction	0,090564
events (1), worse prediction	0,105699
events (1), better prediction	0,090821
net non-events (0)	-0,021116
net events (1)	-0,014878
overall	-0,018445
p (non-events (0))	0,17212
p (events (1))	0,452516
p (overall LR)	0,305129
I (subclass proportion)	
non-events (0), better prediction	0,396226 (21)
non-events (0), worse prediction	0,603774 (32)
events (1), worse prediction	0,493506 (38)
events (1), better prediction	0,506494 (39)
net non-events (0)	-0,207547
net events (1)	0,012987
overall	-0,19456

Out-of-sample evaluation

The comparisons above concern the same 130 people on whom both models were fitted. We will check how the models predict the result of people who did not take part in fitting them. In the analysis window we keep the manual comparison of the same two models, and in the Validation field we choose k-fold CV, k = and enter 10 folds (seed 12345).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Out-of-sample evaluation	cross-validation
Number of folds (k)	10
Seed of the random split	12345
Null model: mean deviance	1.3532
Null model: Brier score	0.2418
Best model by validation (mean deviance)	Model 2

Convergence criterion met	
Number of variables in the model 2	4
AIC - Akaike criterion	139.0823
AICc - corrected Akaike criterion	139.4023
BIC - Bayesian criterion	150.5524
-2 Log Likelihood	131.0823
Pseudo R2	0.2542
R2(Nagelkerke)	0.3924
R2(Coxa-Snella)	0.2909
Cross-validation: mean deviance (-2 ln L / n)	1.0921
Cross-validation: AUC	0.7932
Cross-validation: Brier score	0.1859

		training set		cross-validation ($k = 10$)		
model	variables	AIC	$-2 \ln L/n$	mean deviance	AUC	Brier
null	0			1.3532		0.2418
Model 1 (full)	6	140.7082	0.9901	1.1246	0.7856	0.1920
Model 2 (reduced)	4	139.0823	1.0083	1.0921	0.7932	0.1859

Both models predict the result much better than the null model (a mean deviance of about 1.1 versus 1.35). At the same time, the values obtained out of sample are clearly worse than the values computed on the training set (for the full model 1.1246 versus $128.7082/130 = 0.9901$) – the assessment on the training data was too optimistic. More importantly, the order of the models is reversed: on the training set the full model is better fitted (like any model with more variables), whereas out of sample the reduced model is better – it has a lower mean deviance (1.0921 versus 1.1246), a larger AUC (0.7932 versus 0.7856) and a lower Brier score (0.1859 versus 0.1920). Place of residence and sex not only do not significantly improve the fit, but they even slightly worsen the predictions for new people.

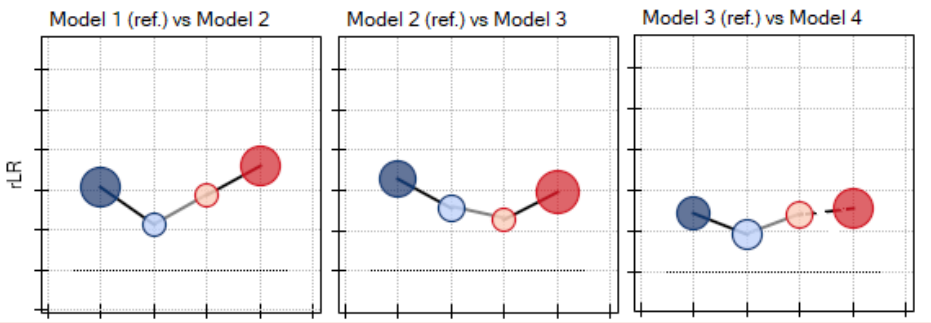
cross-validation ($k = 10$): mean deviance 1.1246 (full model) and 1.0921 (reduced model); null model: 1.3532

The best model by validation is the reduced model – omitting place of residence and sex improves out-of-sample prediction.

Automatic model selection

Automatic selection (bidirectional selection, AIC criterion) gives very similar results. The best model is the one built on the independent variables: AGE, EDUCATION, TIME needed to solve the task, DISTURBANCES. The U-smile plots show the individual steps of building the model. In the last step education was added – we can see that it significantly improved only one part of the plot – the one concerning the people who did not solve the task (solid line).

Copyright ©2010-2026 U-smile Stats – All rights reserved



Based on the analyses above, from a statistical point of view, the optimal model is the one containing the 4 most important independent variables: AGE, EDUCATION, TIME needed to solve the task, DISTURBANCES. A detailed analysis of it can be carried out in the Logistic Regression module. However, the final decision as to which model to choose belongs to the researcher.

EXAMPLE 25.7. (Heart.pqs file)
Risk factors for a certain heart disease were studied, such as age, BMI, smoking, LDL cholesterol, HDL cholesterol and hypertension. From the researcher's point of view it was of interest to determine how much information about smoking could improve the prediction of the occurrence of the disease under study.

We compare a logistic regression model describing the risk of heart disease based on all the variables studied with a model lacking the smoking information. In the analysis window we select the options related to the evaluation of prediction, that is, the ROC curve and the NRI coefficients. In addition, we indicate that all the proposed plots should be included in the report.

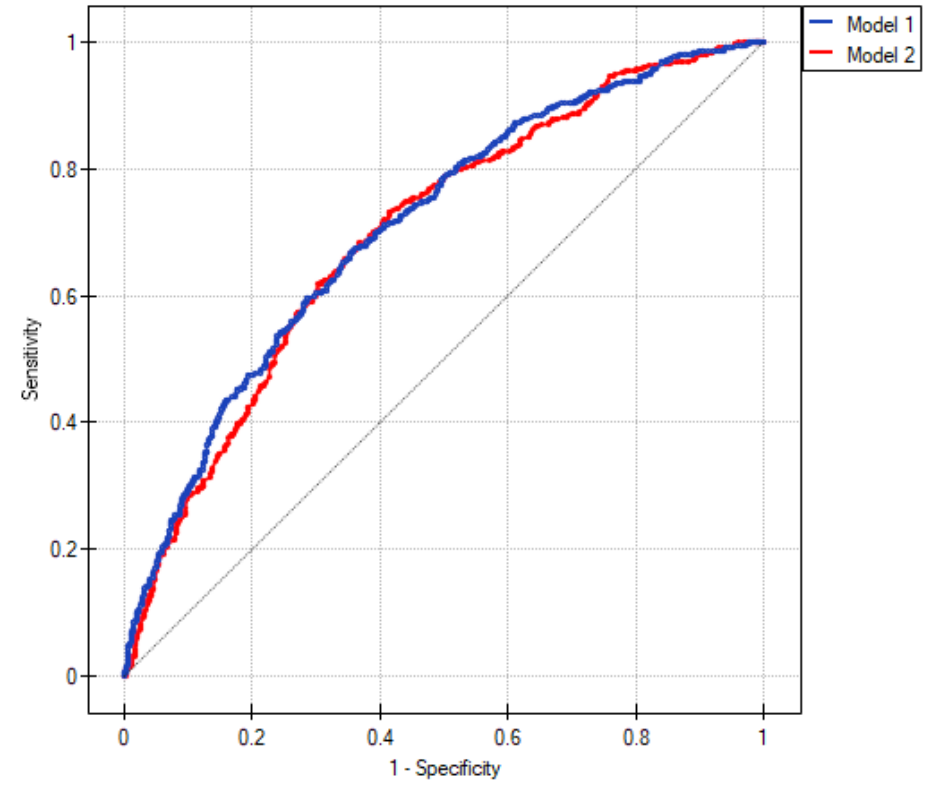
Copyright ©2010-2026 U-smile Stats – All rights reserved

Comparison - Model 1 vs Model 2	
Comparison **ROC [AUC difference]**	
AUC Model 1	0.709
SE(AUC)	0.0138
-95% CI	0.682
+95% CI	0.736
AUC Model 2	0.699
SE(AUC)	0.014
-95% CI	0.6716
+95% CI	0.7265
AUC1-AUC2 [DeLong's method]	0.01
SE(AUC1-AUC2)	0.0053
-95% CI	0
+95% CI	0.0203
Z statistic	1.9033
p-value	0.057
IDI Integrated Discrimination Improvement Model 1	0.0142
-95% CI	0.0078
+95% CI	0.0206
Z statistic	4.3486
p-value	<0.0001
NRI Net Reclassification Improvement (continuous) M	0.1855
-95% CI	0.0805
+95% CI	0.2905
Z statistic	3.4634
p-value	0.0005
NRI (1)	0.0522
NRI (0)	0.1333
NRI Net Reclassification Improvement (categorial) M	-0.0029
-95% CI	-0.0388
+95% CI	0.033
Z statistic	-0.1581
p-value	0.8743
NRI (1)	0.0043
NRI (0)	-0.0072
Probability cut-off points	0.5

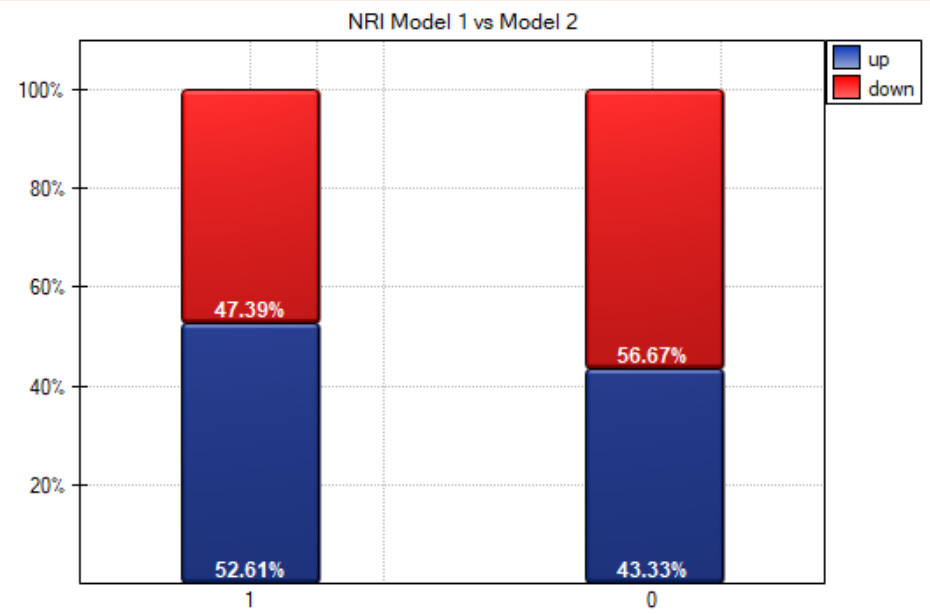
ROC curve: $p = 0.057$**; continuous** NRI **and** IDI**:** $p < 0.05$**,** $NRI(diseased) = 0.0522$**,** $NRI(healthy) = 0.1333$

Adding the smoking information to the model is associated with an important difference in prediction, although this difference is not statistically significant when described by the ROC curve. The continuous NRI and IDI values, on the other hand, indicate a statistically significant and favourable change: the prognosis for people with heart disease improved by more than 5%, and for people without this disease – by more than 13%.

Copyright ©2010-2026 U-smile Stats – All rights reserved

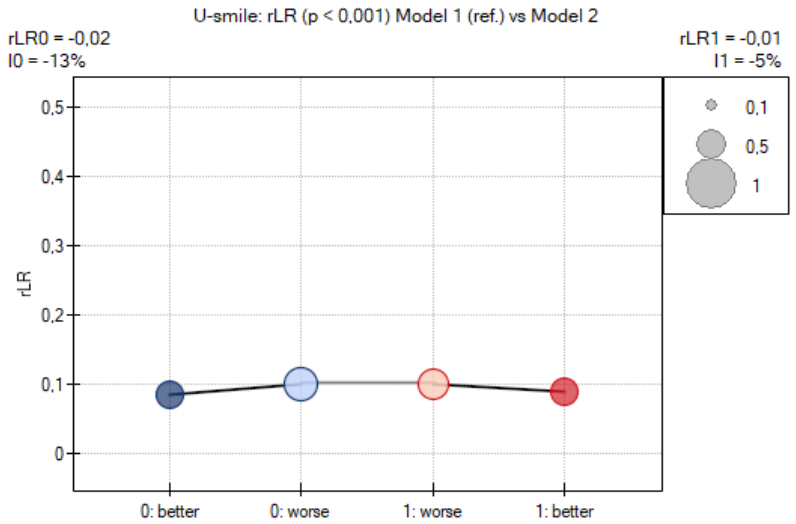


The conclusions drawn from NRI can also be seen on the plot. An increase in the probability of disease predicted by the model is observed in diseased people (more people were reclassified upward than downward: 52.61% vs 47.39%), while a decrease in probability applies to a greater extent to healthy people (more people were reclassified downward than upward: 56.67% vs 43.33%).



The statistically significant deterioration of the model resulting from removing “smoking” from it is visible in both branches of the U-smile plot and in the fact that all its coefficients are negative.

Copyright ©2010-2026 U-smile Stats – All rights reserved



It is also possible to determine a categorical NRI, but for this we would first need to establish the probability cut-off points, determined by the model, that are accepted in the heart disease literature.

FOR THOSE INTERESTED

Likelihood ratio test
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad L_{FM} = L_{RM}, \\ \mathcal{H}_1: &\quad L_{FM} \neq L_{RM}, \end{aligned}$$

where L_{FM}, L_{RM} – the maximum of the likelihood function in the compared models (full and reduced).

Formulas:
The test statistic has the form:

$$\chi^2 = -2\ln(L_{RM}/L_{FM}) = -2\ln(L_{RM}) - (-2\ln(L_{FM}))$$

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with $df = k_{FM} - k_{RM}$ degrees of freedom, where k_{FM} and k_{RM} are the number of estimated parameters in the compared models.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Out-of-sample evaluation
Let $\hat{p}_{(-i)}$ denote the probability of the event predicted for observation i by the model fitted without the part of the data to which observation i belongs (in cross-validation – without its fold, with a test set – on the training set), and y_i – the outcome (1 – event, 0 – no event). For the n_v evaluated observations:

$$\text{mean deviance} = -\frac{2}{n_v}\sum_{i=1}^{n_v}\left[y_i \ln \hat{p}_{(-i)} + (1-y_i)\ln\left(1-\hat{p}_{(-i)}\right)\right],$$

$$\text{Brier score} = \frac{1}{n_v}\sum_{i=1}^{n_v}\left(y_i - \hat{p}_{(-i)}\right)^2.$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

AUC is the area under the ROC curve determined jointly from all the probabilities $\hat{p}_{(-i)}$. In the null model $\hat{p}_{(-i)}$ is the frequency of the event in the corresponding training part. The split is made separately among the observations with and without the event (stratification), as described in the chapter Comparing multiple regression models.

Change in the area under the ROC curve
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad AUC_{FM} = AUC_{RM}, \\ \mathcal{H}_1: &\quad AUC_{FM} \neq AUC_{RM}. \end{aligned}$$

The method of determining the test statistic, based on DeLong's method, is described in the Comparing ROC curves chapter.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Net Reclassification Improvement (NRI)
Formulas:
We define:

$$\hat{p}_{up,events} = \frac{\#events_{up}}{\#events}, \quad \hat{p}_{down,events} = \frac{\#events_{down}}{\#events}, \quad \hat{p}_{up,nonevents} = \frac{\#nonevents_{up}}{\#nonevents}, \quad \hat{p}_{down,nonevents} = \frac{\#non}{\#no}$$

where:
$\#events_{up}$ – the number of objects in the group experiencing the event, for whom the predicted probability changed by at least one category upward,
$\#events_{down}$ – the number of objects in the group experiencing the event, for whom the predicted probability changed by at least one category downward,
$\#events$ – the number of objects in the group experiencing the event,
$\#nonevents_{up}$ – the number of objects in the group not experiencing the event, for whom the predicted probability changed by at least one category upward,
$\#nonevents_{down}$ – the number of objects in the group not experiencing the event, for whom the predicted probability changed by at least one category downward,
$\#nonevents$ – the number of objects in the group not experiencing the event.

The overall NRI coefficient and the coefficients expressing the percentage change of classification are determined from the formula:

$$NRI = (\hat{p}_{up,events} - \hat{p}_{down,events}) - (\hat{p}_{up,nonevents} - \hat{p}_{down,nonevents})$$

$$NRI_{events} = \hat{p}_{up,events} - \hat{p}_{down,events}, \quad NRI_{nonevents} = \hat{p}_{down,nonevents} - \hat{p}_{up,nonevents},$$

Interpretation:
The NRI_{events} coefficient can be interpreted as the **net percentage** of correctly reclassified people who experienced the event, and $NRI_{nonevents}$ as the **net percentage** of correctly reclassified people who did not experience the event. The overall NRI coefficient is the sum of the NRI_{events} and $NRI_{nonevents}$ coefficients, which makes it a coefficient implicitly weighted by the frequency of the event, and it cannot be interpreted as a percentage.

Z test for checking the significance of the NRI coefficient
Hypotheses:

Copyright ©2010-2026 U-smile Stats – All rights reserved

$$\begin{aligned}\mathcal{H}_0: &\quad NRI = 0,\\ \mathcal{H}_1: &\quad NRI \neq 0.\end{aligned}$$

Formulas:
The test statistic has the form:

$$Z = \frac{NRI}{SE(NRI)}$$

where:

$$\begin{aligned}SE(NRI) = &\left[\left(\frac{\#events_{up}+\#events_{down}}{\#events^2} - \frac{(\#events_{up}+\#events_{down})^2}{\#events^3}\right) + \right.\\ &\left. + \left(\frac{\#nonevents_{down}+\#nonevents_{up}}{\#nonevents^2} - \frac{(\#nonevents_{down}+\#nonevents_{up})^2}{\#nonevents^3}\right)\right]^{1/2}\end{aligned}$$

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Integrated Discrimination Improvement (IDI)
Formulas:

$$IDI = \overline{p(diff)}_{events} - \overline{p(diff)}_{nonevents}$$

where:
$\overline{p(diff)}_{events}$ – the mean difference in predicted probability between the regression models (”old” and ”new”) for the objects that experienced the event,
$\overline{p(diff)}_{nonevents}$ – the mean difference in predicted probability between the regression models (”old” and ”new”) for the objects that did not experience the event.

Z test for checking the significance of the IDI coefficient
Hypotheses:

$$\begin{aligned}\mathcal{H}_0: &\quad IDI = 0,\\ \mathcal{H}_1: &\quad IDI \neq 0.\end{aligned}$$

Formulas:
The test statistic has the form:

$$Z = \frac{IDI}{SE(IDI)}$$

where:

$$SE(IDI) = \sqrt{\frac{sd(diff)^2_{events}}{\#events} + \frac{sd(diff)^2_{nonevents}}{\#nonevents}}$$

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Factorial ANOVA – GLM

WHEN TO USE:

Factorial analysis of variance GLM is an extension of one-way analysis of variance (ANOVA) for independent groups and multiple linear regression. The acronym GLM (general linear model) stands for General Linear Model. GLM analysis typically involves the use of linear regression models to calculate various composite ANOVA comparisons.

Factorial ANOVA is a type of analysis of variance in which we can use either one or many factors to separate the groups being compared. Variables that are interactions of the indicated factors may also take part in the analysis. When ANOVA includes more than one factor, these factors are confounded with each other. Factorial ANOVA requires that the factors be divided into individual categories (i.e. independent populations), e.g. factor A: sex, divided into male and female; factor B: education, divided into primary, vocational, secondary, and higher education. The interaction of factor $A * B$ is also divided into categories – in this case, we obtain eight categories.

Example of equivalent analyses that can be performed through GLM. The analyses in each row of the table are equivalent in the sense that their results are the same, although they need not be identical. The study concerns the *income* of a certain group of people, about whom we have some additional information, such as sex and education.

income and...	ANOVA	Multiple regression
sex	**t-test for independent groups** ⋆ comparing income for women with income for men	**dependence of income (Y) on sex (X)** ⋆ sex as a column with two categories
education	**One-way ANOVA** ⋆ comparing income for people with different education (primary, vocational, secondary, higher)	**dependence of income (Y) on education (X)** ⋆ education broken down into dummy variables
sex, education	**Multi-factor ANOVA** ⋆ comparing income for women with income for men taking education into account as a confounding variable ⋆ comparing income for people with different education (primary, vocational, secondary, higher) taking sex into account as a confounding variable	**dependence of income (Y) on sex (X_1) and education (X_2)** ⋆ sex as a column with two categories ⋆ education broken down into dummy variables both variables are mutually confounding
sex, education, sex*education	**Multi-factor ANOVA** ⋆ comparing income for women with income for men taking other variables in the model into account ⋆ comparing income for people with different education (primary, vocational, secondary, higher) taking other variables in the model into account ⋆ comparing income for people with different education and sex (in interaction) taking other variables in the model into account	**dependence of income (Y) on sex (X_1), on education (X_2), and on the interaction of sex and education ($X_1 * X_2$)** ⋆ sex as a column with two categories ⋆ education broken down into dummy variables ⋆ interaction is the product of sex and education the variables in the model are mutually confounding

GLM analysis can be used in any of the above cases, but since multiple regression analysis, like one-way ANOVA, has been discussed in a separate chapter, in this chapter we will present the use of GLM in multi-factor ANOVA.

ASSUMPTIONS

Copyright ©2010-2026 U-smile Stats – All rights reserved

- measurement on an interval scale,
- the samples come from a population with a normal distribution (normality of the variables or model residuals),
- an independent model,
- equality of variances of the variables studied in the populations compared.

DEFINITIONS OF MEASURES

The influence of confounding factors – although all factors participating in the analysis are confounded with each other, their influence on the significance of individual factors can be controlled. There are three ways in which, when examining the significance of individual factors, the influence of confounding variables can be taken into account. They depend on how the sum of squares is calculated:

Type I sums of squares – depend on the order in which the individual factors appear in the model. This type of sum of squares means that the significance of the factor being interpreted is adjusted for those variables whose order in the model was earlier, while the remaining variables in the model only indirectly affect the result of the analysis.
Use: recommended when the study is fully balanced, with equal or proportional counts in each category, including when interactions are present. Not recommended when the study is unbalanced (different counts in each category) and/or interactions are present.

Type II sums of squares – the significance of the factor being interpreted is adjusted for those variables whose order is the same or lower, while the remaining variables in the model only indirectly affect the result of the analysis.
Use: recommended when the study is fully balanced, with equal or proportional counts in each category, including when interactions are present. Not recommended when the study is unbalanced (different counts in each category) and/or interactions are present.

Type III sums of squares – the significance of the factor being interpreted is adjusted for all other variables in the model. We recommend using this type of sum of squares when effect coding is selected.
Use: recommended when the study is balanced or unbalanced, including when interactions are present. Not recommended when the study contains subclasses with missing observations.

Note! In PQStat, Type III sums of squares are selected by default because of their universality. The effect coding option, described in the Preparation of variables for analysis in multivariate models section, is also selected by default. Note that the choice of appropriate coding affects both the interpretation of the model coefficients and the significance of individual factors in factorial ANOVA – especially with unbalanced designs.

ANOVA approach vs. regression approach – ANOVA-type analysis and regression models are treated equivalently, and in the general case their hypotheses converge. When interpreting these hypotheses, it should be remembered that the hypotheses for a given factor are adjusted for those other factors that a given analysis includes. The model approach assumes the following regression model:

$$Y = \mu + \alpha_i A_i + \beta_j B_j + (\alpha\beta)_k B_k + \epsilon,$$

where:
Y – the dependent variable, explained by the model,

Copyright ©2010-2026 U-smile Stats – All rights reserved

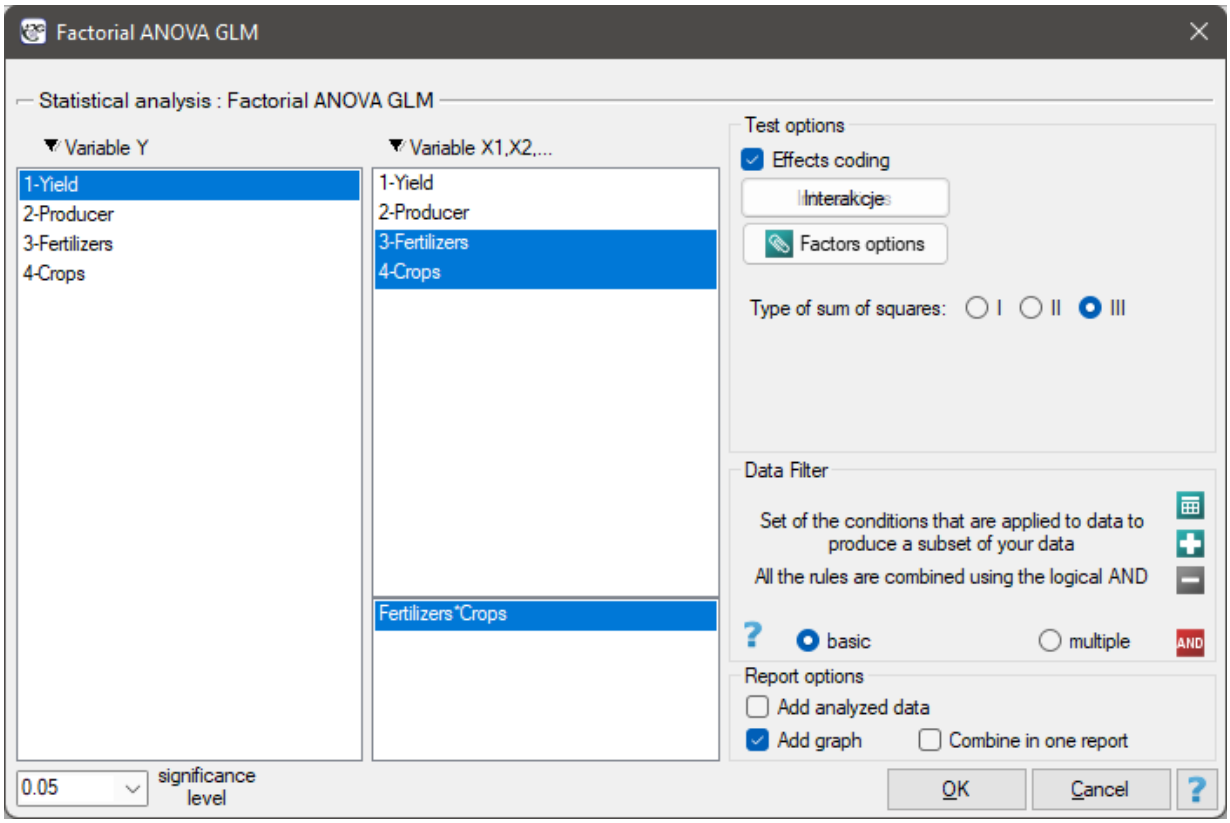
μ – the overall mean of variable Y (if effect coding was used),
A_i, B_j, AB_k – factors – independent, explanatory variables,
$\alpha_i, \beta_i, \alpha\beta_i$ – parameters,
ϵ – the random component (model residual).

Coding – the analysis results obtained (in particular the constructed regression model) and the interpretation of the hypotheses also depend on the coding method. PQStat offers **dummy coding** and effect coding. By default, the program selects effect coding – unchecking this option is equivalent to selecting dummy coding.

Note! When using Type III sums of squares, when interactions are present, it is advisable to use effect coding.

HOW TO SET UP THE ANALYSIS

Statistics → Multivariate models → Factorial ANOVA GLM.



PERFORMING AND INTERPRETING THE ANALYSIS

Example 25.8. (yield.pqs file)
To increase crop yields, fertilizers are being developed using increasingly newer technologies. On the basis of an experiment, the researchers want to find out which of three blends of new fertilizers is the

Copyright ©2010-2026 U-smile Stats – All rights reserved

most effective. The crops were grown on two different farms and involved sowing wheat, rye, oats, and barley. Yield was reported as a % (compared to the yield obtained without fertilization).

First, we want to check whether: (1) the average yields obtained when fertilizing with blend X are the same as those obtained with blend Y and the same as those obtained with blend Z (regardless of the farm growing the crop); (2) the average yields obtained on farm 1 are the same as on farm 2 (regardless of the fertilizer blend used). Equivalently, using a regression approach: (1) the coefficients indicating the change in yield obtained when the fertilizer applied changes are zero (regardless of the farm growing the crop); (2) the coefficient indicating the change in yield obtained when the farm growing the crop changes is zero. In a second application of GLM, we will check whether: (3) the average yields obtained from growing individual cereals are the same when different fertilizer treatments are used.

We will conduct the analysis using the third type of sum of squares and effect coding.

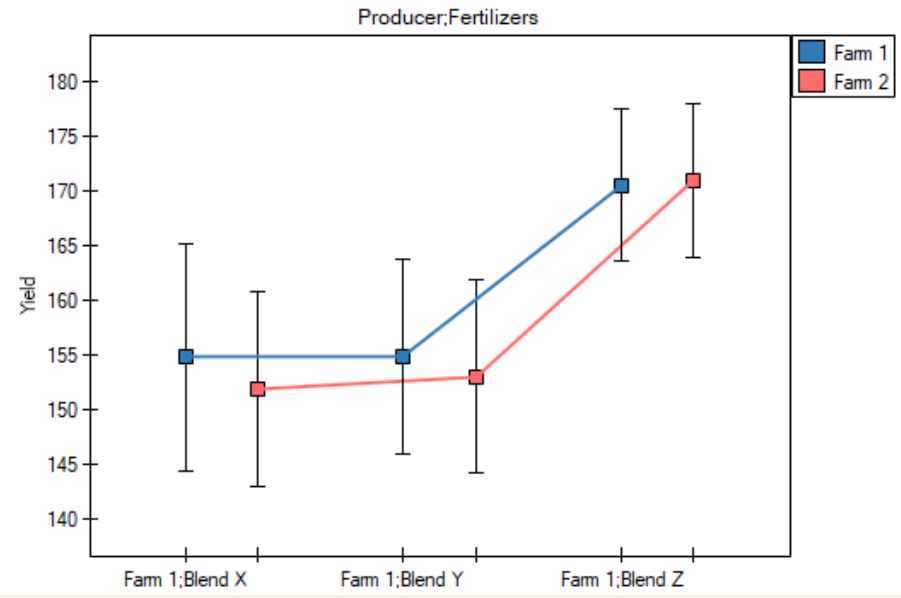
Between-groups (effects) test

	SS	MS	DF	F	p	Partial Eta-squared
intercept	4262167.148	4262167.148	1	8777.5923	<0.0001	0.9817
Producer	90.0536	90.0536	1	0.1855	0.6673	0.0011
Fertilizers	10957.7262	5478.8631	2	11.2833	<0.0001	0.121
within(błąd)	79634.0714	485.5736	164			
total	90681.8512					

fertilizer blend: $p = 0.000026$, partial Eta-squared ≈ 12%; farm: $p = 0.667289$, partial Eta-squared = 0.1%

Yield depends significantly on the fertilizer blend applied, but does not depend on the farm on which the crop was grown.

After selecting the observed or expected averages in the Factor Options window, we can represent these differences graphically on plots showing the average yields for each fertilizer blend applied. The exact values of the averages can be read from the descriptive statistics table.



Copyright ©2010-2026 U-smile Stats – All rights reserved

Descriptive statistics (observed)

Fertilizers	frequency	mean	stand. dev.	stand. error	-95% CI	+95% CI
Blend X	56	153.2857	24.7942	3.3133	146.6458	159.9256
Blend Y	56	153.8571	22.6165	3.0223	147.8004	159.9139
Blend Z	56	170.6964	17.9797	2.4026	165.8814	175.5114
Crops	frequency	mean	stand. dev.	stand. error	-95% CI	+95% CI
Barley	42	155.3095	25.3459	3.911	147.4112	163.2079
Oat	42	158.6667	19.6899	3.0382	152.5309	164.8025
Wheat	42	170.5238	21.5749	3.3291	163.8006	177.247
Rye	42	152.619	22.9059	3.5345	145.4811	159.757

We can check where the differences are located using post-hoc tests. Fisher's LSD post-hoc test indicates that the most favourable results are obtained using blend Z – the yield obtained is on average 170.7% of the yield that would be obtained without fertilization. The remaining blends do not differ statistically significantly in the yield obtained. Since the model analysed the farm on which the crops were grown simultaneously, we can say that the advantage of blend Z is independent of the farm on which the sowing was done.

POST-HOC Fisher LSD

	NIR [F]	Statistic [F]	p-value [F]
Producer			
Farm 1;Farm 2	6.7138	0.4306	0.6673
Fertilizers			
Blend X;Blend Y	8.2227	-0.1372	0.891
Blend X;Blend Z	8.2227	-4.1809	<0.0001
Blend Y;Blend Z	8.2227	-4.0437	0.0001

Regression approach

An analogous interpretation can be obtained using the regression model, although the interpretation here is somewhat more difficult – the difficulty arises from the need to determine the coding method and the choice of reference category. Let us first look at the results obtained with dummy coding, which we can obtain by unchecking the effect coding option. The analysis automatically took the alphabetically first level as the reference level – for fertilizers, this level was blend X; for farms, it was farm 1.

Factor levels Producer	2
Size 0 [Farm 1] REF	84
Size 1 [Farm 2]	84
Factor levels Fertilizers	3
Size 0 [Blend X] REF	56
Size 1 [Blend Y]	56
Size 2 [Blend Z]	56

Model

	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. erro
intercept	154.0179	3.4002	147.3041	160.7316	45.2969	<0.0001		
Producer	-1.4643	3.4002	-8.1781	5.2495	-0.4306	0.6673	-0.0315	0.0732
Fertilizers[1]	0.5714	4.1644	-7.6512	8.7941	0.1372	0.891	0.0116	0.0845
Fertilizers[2]	17.4107	4.1644	9.188	25.6334	4.1809	<0.0001	0.3533	0.0845

The analysis of the model coefficients resembles the analysis of post-hoc tests, with the difference that we compare only to the reference category. So if we compare all fertilizer blends to blend X, we can see

Copyright ©2010-2026 U-smile Stats – All rights reserved

that only using blend Z produced significantly higher results (p=0.000047) – these results are higher by 17.410714 (recall that the means were, respectively, 153.285714 for blend X and 170.696429 for blend Z). Comparing farms is simple, since we only have two farms to compare, and the result obtained is the result of comparing farm 2 with farm 1, which was the reference category. This time the difference obtained was small (-1.464286) and not statistically significant (0.667289).

Using effect coding, we also choose a reference category, but the magnitude of the coefficients and their significance is related not to the chosen reference category but to the overall average yield obtained, recorded in the model as the intercept (159.279762).

Model								
	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. erro
intercept	159.2798	1.7001	155.9229	162.6367	93.6888	<0.0001		
Producer	-0.7321	1.7001	-4.089	2.6248	-0.4306	0.6673	-0.0315	0.0732
Fertilizers[1]	-5.4226	2.4043	-10.17	-0.6753	-2.2554	0.0254	-0.1906	0.0845
Fertilizers[2]	11.4167	2.4043	6.6693	16.164	4.7484	<0.0001	0.4012	0.0845

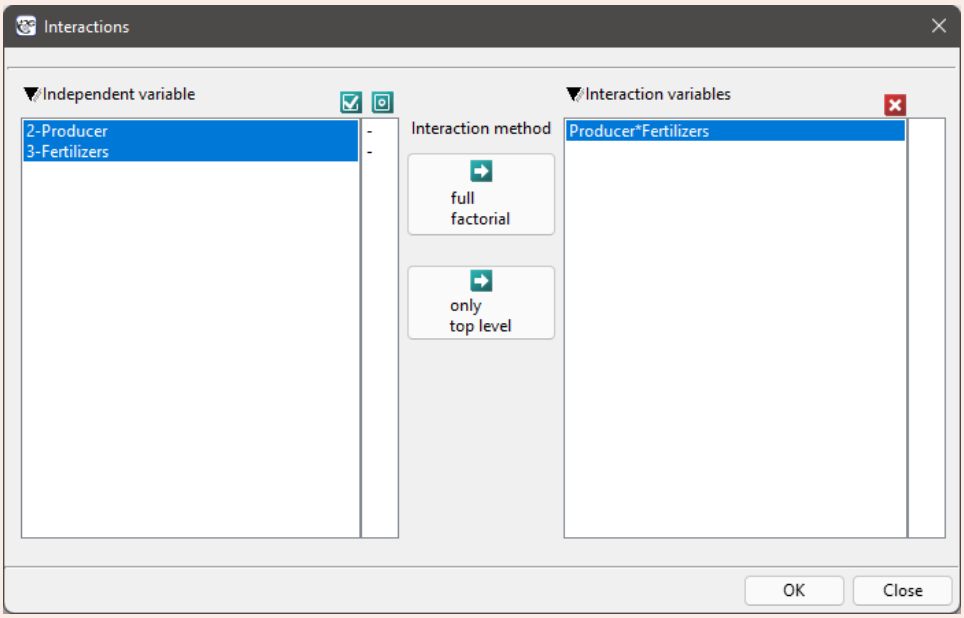
Compared to the overall average, we find quite a few differences: the yield obtained when fertilizing with blend Y is 5.422619 lower than the overall average, and with blend Z, 11.416667 higher. Both differences are statistically significant.

An undeniable advantage of building a regression model is the possibility of using its formula to predict the yields obtained. The models built are as follows:
For dummy coding: $yield = 154.017857 + 0.571429 \cdot blendY + 17.410714 \cdot blendZ - 1.464286 \cdot farm2$
For effect coding: $yield = 159.279762 - 5.422619 \cdot blendY + 11.416667 \cdot blendZ - 0.732143 \cdot farm2$

To be able to use the selected model for forecasting, go to the multiple regression – prediction menu and make a prediction based on the new data. The preparation of the data depends on how it is coded.

Based on all the results obtained, we do not suspect that yield depends on the interaction between the types of fertilizer used and the farm growing the crop. Most often, the presence of an interaction can be seen on the plot as clearly intersecting lines. Here, the two lines were almost parallel and close enough to each other that the difference between farms was not statistically significant. Although intersecting lines usually indicate the presence of an interaction, it should be remembered that when the lines are close to each other, their accidental crossing is very likely, meaning the interaction will not be statistically significant. However, to be sure, we will check whether there is an interaction in our case. To do this, we will select both variables once again in the interaction window and move them to the list of interactions on the right side of the window, then repeat the analysis.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Between-groups (effects) test

	SS	MS	DF	F	p	Partial Eta-s
intercept	4262167.14	4262167.14	1	8679.3183	<0.0001	0.9817
Producer	90.0536	90.0536	1	0.1834	0.6691	0.0011
Fertilizers	10957.7262	5478.8631	2	11.157	<0.0001	0.1211
Producer*Fertilizers	80.4643	40.2321	2	0.0819	0.9214	0.001
within(błąd)	79553.6071	491.0716	162			
total	90681.8512					

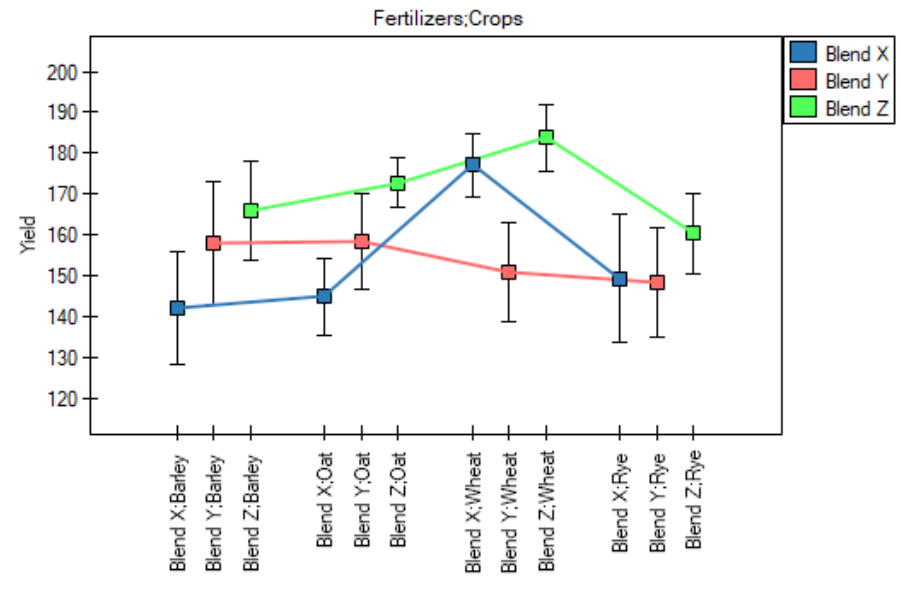
interaction blend*farm: $p = 0.921377$ (not statistically significant)

We confirmed our suspicion that there is no significant interaction – in this case, a simpler model, i.e. without interaction, is recommended.

Hypothesis 3)

We encounter a different situation when examining the yield obtained depending on the fertilizer dose applied and the type of cereal grown.

Copyright ©2010-2026 U-smile Stats – All rights reserved



We will perform an analysis that, besides the main effects, also takes interactions into account.

Between-groups (effects) test

	SS	MS	DF	F	p	Partial Eta-s
intercept	4262167.14	4262167.14	1	10483.1113	<0.0001	0.9853
Fertilizers	10957.7262	5478.8631	2	13.4757	<0.0001	0.1473
Crops	7851.1607	2617.0536	3	6.4368	0.0004	0.1102
Fertilizers*Crops	8447.3214	1407.8869	6	3.4628	0.0031	0.1175
within(błąd)	63425.6429	406.5746	156			
total	90681.8512					

interaction fertilizer*cereal type: $p = 0.003066$ (statistically significant)

We should use the model with interactions, and the description of the results obtained should focus on this interaction.

In the ANOVA approach, the hypothesis concerning the interaction concerns all possible pairs of means, i.e. the average yields obtained when fertilizing each type of cereal with each fertilizer blend are equal. In the regression approach, we would say that the coefficients determining the change in yield obtained when the fertilizer applied and the type of crop change are zero.

Based on the plot (and the averages recorded in the table), we can see that by far the best yields are obtained with blend Z, regardless of the type of cereal grown.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Descriptive statistics (observed)						
Fertilizers	frequency	mean	stand. dev.	stand. error	-95% CI	+95% CI
Blend X	56	153.2857	24.7942	3.3133	146.6458	159.9256
Blend Y	56	153.8571	22.6165	3.0223	147.8004	159.9139
Blend Z	56	170.6964	17.9797	2.4026	165.8814	175.5114
Crops	frequency	mean	stand. dev.	stand. error	-95% CI	+95% CI
Barley	42	155.3095	25.3459	3.911	147.4112	163.2079
Oat	42	158.6667	19.6899	3.0382	152.5309	164.8025
Wheat	42	170.5238	21.5749	3.3291	163.8006	177.247
Rye	42	152.619	22.9059	3.5345	145.4811	159.757

On the other hand, blend X and blend Y produce worse yields than blend Z, and there is also an interaction effect between them. This is manifested by the fact that wheat crops yield an unusually high yield when blend X is applied, compared to the wheat yield obtained when blend Y is applied, while barley and oat crops yield better when blend Y is applied. We can check the differences obtained in more detail by performing post-hoc tests. An excerpt from this report is given below:

POST-HOC Fisher LSD			
	NIR [F]	Statistic [F]	p-value [F]
Fertilizers;Crops			
Blend X;Barley<>Ble	15.054	-0.3655	0.7152
Blend X;Barley<>Ble	15.054	-4.6018	<0.0001
Blend X;Barley<>Ble	15.054	-0.956	0.3406
Blend X;Barley<>Ble	15.054	-2.0994	0.0374
Blend X;Barley<>Ble	15.054	-2.1557	0.0326
Blend X;Barley<>Ble	15.054	-1.1528	0.2508
Blend X;Barley<>Ble	15.054	-0.8154	0.4161
Blend X;Barley<>Ble	15.054	-3.1398	0.002
Blend X;Barley<>Ble	15.054	-4.0395	0.0001
Blend X;Barley<>Ble	15.054	-5.4735	<0.0001
Blend X;Barley<>Ble	15.054	-2.4087	0.0172

The result of Fisher's post-hoc test is extensive and confirms the large and statistically significant advantage in yield obtained when using blend Z for any crop, and blend Y for the wheat crop.

The coefficients of the regression model can be used for prediction via the multiple regression – prediction menu, remembering to code the new data appropriately depending on the model selected.

Copyright ©2010-2026 U-smile Stats – All rights reserved

dummy coding

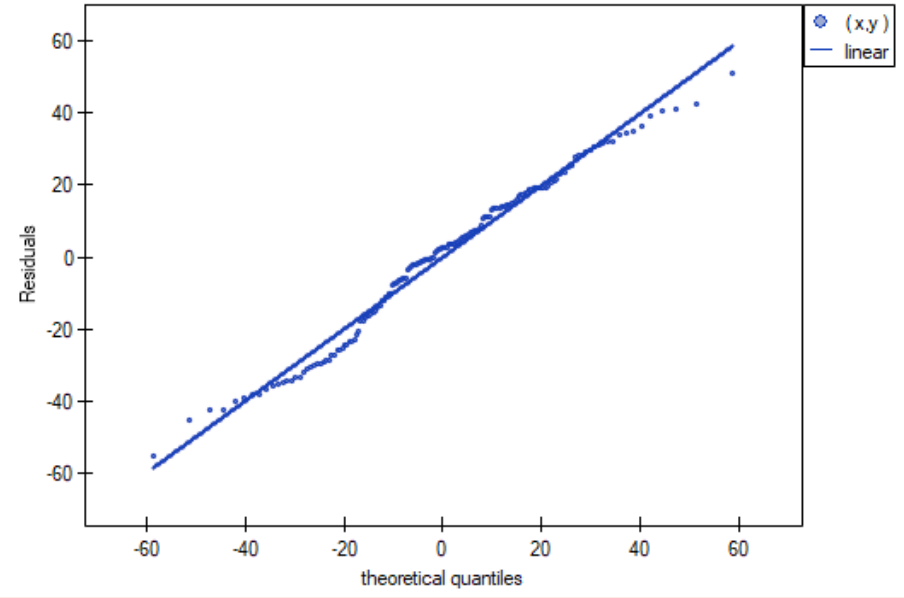
Model	b coeff.
intercept	142
Fertilizers[1]	16
Fertilizers[2]	23.9286
Crops[1]	2.7857
Crops[2]	35.0714
Crops[3]	7.2857
Fertilizers[1]Crops[1	-2.3571
Fertilizers[1]Crops[2	-42.2857
Fertilizers[1]Crops[3	-17.0714
Fertilizers[2]Crops[1	4.0714
Fertilizers[2]Crops[2	-17.2857
Fertilizers[2]Crops[3	-12.8571

effect coding

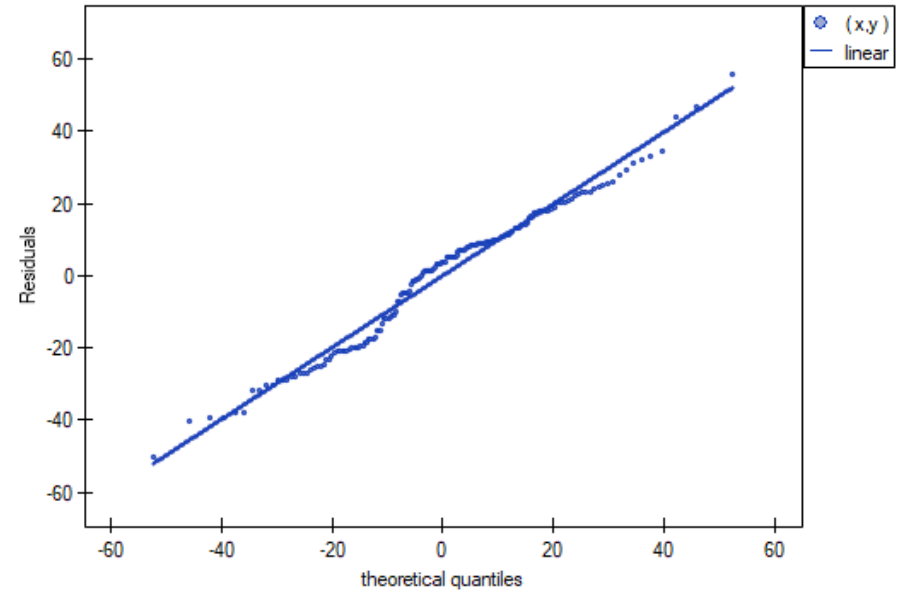
Model	b coeff.
intercept	159.2798
Fertilizers[1]	-5.4226
Fertilizers[2]	11.4167
Crops[1]	-0.6131
Crops[2]	11.244
Crops[3]	-6.6607
Fertilizers[1]Crops[1	5.1845
Fertilizers[1]Crops[2	-14.3155
Fertilizers[1]Crops[3	1.0179
Fertilizers[2]Crops[1	2.7024
Fertilizers[2]Crops[2	1.7738
Fertilizers[2]Crops[3	-3.6786

Checking assumptions

Checking the main assumptions will consist of comparing variances and visually determining the normality of the model residuals.



Copyright ©2010-2026 U-smile Stats – All rights reserved



The Q-Q normality plot of the model residuals for the first and second analyses shows the model residuals well distributed around a straight line, indicating a good fit of the residuals to the normal distribution. The comparison of variances is performed using Levene's test or the Brown-Forsythe test. For these tests, we can assume that the results obtained are inconclusive and are on the borderline of equality of variances.

Equality of variance				
	Levene test		Brown-Forsythe test	
	F statistic	p-value	F statistic	p-value
Producer	0.1266	0.7224	0.2196	0.6399
Fertilizers	2.9548	0.0549	3.2231	0.0424

Equality of variance				
	Levene test		Brown-Forsythe test	
	F statistic	p-value	F statistic	p-value
Fertilizers	2.9548	0.0549	3.2231	0.0424
Crops	2.3834	0.0713	2.2981	0.0795
Fertilizers*Crops	3.2663	0.0005	1.7547	0.0664

FOR THOSE INTERESTED

Formulas – sums of squares

Type I sums of squares

For example: if we place the factors in the model in the order indicated: A, B, $A*B$, C, $A*C$, $B*C$, $A*B*C$, D, then the significance for factor $A*C$ takes the whole model into account (through the sums of squares for the error), but only the factors A, B, $A*B$, C are used explicitly as confounding

Copyright ©2010-2026 U-smile Stats – All rights reserved

variables. The sums of squares for factor $A * C$ are then calculated as follows:

$$SS(A * C) = SS(A, B, A * B, C, A * C, B * C, A * B * C, D) - SS(A, B, A * B, C)$$

Type II sums of squares
For example: if we include the factors A, B, $A * B$, C, $A * C$, $B * C$, $A * B * C$, D in the model, then the significance for factor $A * C$ takes the whole model into account, but the first-order variables A, B, C, D and all other second-order variables $A * B$, $B * C$ are used explicitly as confounding variables. The sums of squares for factor $A * C$ are then calculated as follows:

$$SS(A * C) = SS(A, B, A * B, C, A * C, B * C, A * B * C, D) - SS(A, B, A * B, C, B * C, D)$$

Type III sums of squares
For example: if we include the factors A, B, $A * B$, C, $A * C$, $B * C$, $A * B * C$, D in the model, then the significance for variable $A * C$ takes the whole model into account, and all factors except the one under study, i.e. A, B, $A * B$, C, $B * C$, $A * B * C$, D, are used explicitly as confounding variables. The sums of squares for factor $A * C$ are then calculated as follows:

$$SS(A*C) = SS(A, B, A*B, C, A*C, B*C, A*B*C, D) - SS(A, B, A*B, C, B*C, A*B*C, D)$$

Hypotheses – ANOVA approach
Hypotheses for factor A:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2 = ... = \mu_a, \\ \mathcal{H}_1: &\quad \text{not all } \mu_i \text{ are equal to each other } (i = 1, 2, ..., a), \end{aligned}$$

where $\mu_1, \mu_2, ..., \mu_a$ – the means of factor A for its individual categories.

Hypotheses for factor B:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2 = ... = \mu_b, \\ \mathcal{H}_1: &\quad \text{not all } \mu_j \text{ are equal to each other } (j = 1, 2, ..., b), \end{aligned}$$

where $\mu_1, \mu_2, ..., \mu_b$ – the means of factor B for its individual categories.

Hypotheses for the interaction of factors $A * B$:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2 = ... = \mu_{ab}, \\ \mathcal{H}_1: &\quad \text{not all } \mu_k \text{ are equal to each other } (k = 1, 2, ..., ab), \end{aligned}$$

where $\mu_1, \mu_2, ..., \mu_{ab}$ – the means of the interaction of factors $A * B$ for their individual categories.

Hypotheses – regression approach
Hypotheses for factor A:

$$\begin{aligned} \mathcal{H}_0: &\quad \alpha_1 = \alpha_2 = ... = \alpha_a = 0, \\ \mathcal{H}_1: &\quad \text{not all } \alpha_i = 0, \end{aligned}$$

Hypotheses for factor B:

Copyright ©2010-2026 U-smile Stats – All rights reserved

$$\begin{array}{ll} \mathcal{H}_0: & \beta_1 = \beta_2 = ... = \beta_b = 0, \\ \mathcal{H}_1: & \text{not all } \beta_j = 0\ , \end{array}$$

Hypotheses for the interaction of factors $A * B$:

$$\begin{array}{ll} \mathcal{H}_0: & (\alpha\beta)_1 = (\alpha\beta)_2 = ... = (\alpha\beta)_{ab} = 0, \\ \mathcal{H}_1: & \text{not all } (\alpha\beta)_k = 0, \end{array}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

ANCOVA

WHEN TO USE:

Analysis of covariance (ANCOVA) is a method for testing the hypothesis that the means of two or more populations are equal, after adjusting for other continuous variables. These adjustments produce effects that researchers tend to find more appealing than those obtained through ANOVA, i.e. narrower confidence intervals and greater statistical power.

Suppose an experiment is carried out to evaluate the effects of two treatments. The groups to which the treatments were randomly assigned differ slightly in mean age, which also affects the treatment outcome. Differences between the groups in outcomes would be quite ambiguous to interpret, since the groups differ both in age and in treatment conditions. Analysis of covariance provides ”adjusted means”, which estimate what the mean outcomes would be if the groups were exactly identical in terms of age. At the same time, the within-group variability in outcomes attributable to the variable (age) is removed from the error variability, increasing the precision of the test of differences between the adjusted means.

The label ”analysis of covariance” is now regarded as somewhat anachronistic by some research methodologists and statisticians, since this analysis is not a separate technique but a variant of the general linear model (GLM). Nevertheless, the term remains useful, since it immediately conveys to most researchers the idea that a categorical variable (e.g. treatment condition) and a continuous variable (e.g. age) are both involved in a single analysis determining the treatment outcome.

ASSUMPTIONS

- the dependent variable Y is measured on an interval scale,
- the samples come from a normally distributed population (normality of the variables, or of the model residuals),
- an independent model with respect to the factors,
- equality of variance of the residuals between the groups formed by the factors,
- equal slopes of the regression lines (the regression coefficients between each confounding variable and the dependent variable) for every possible level of the factor.

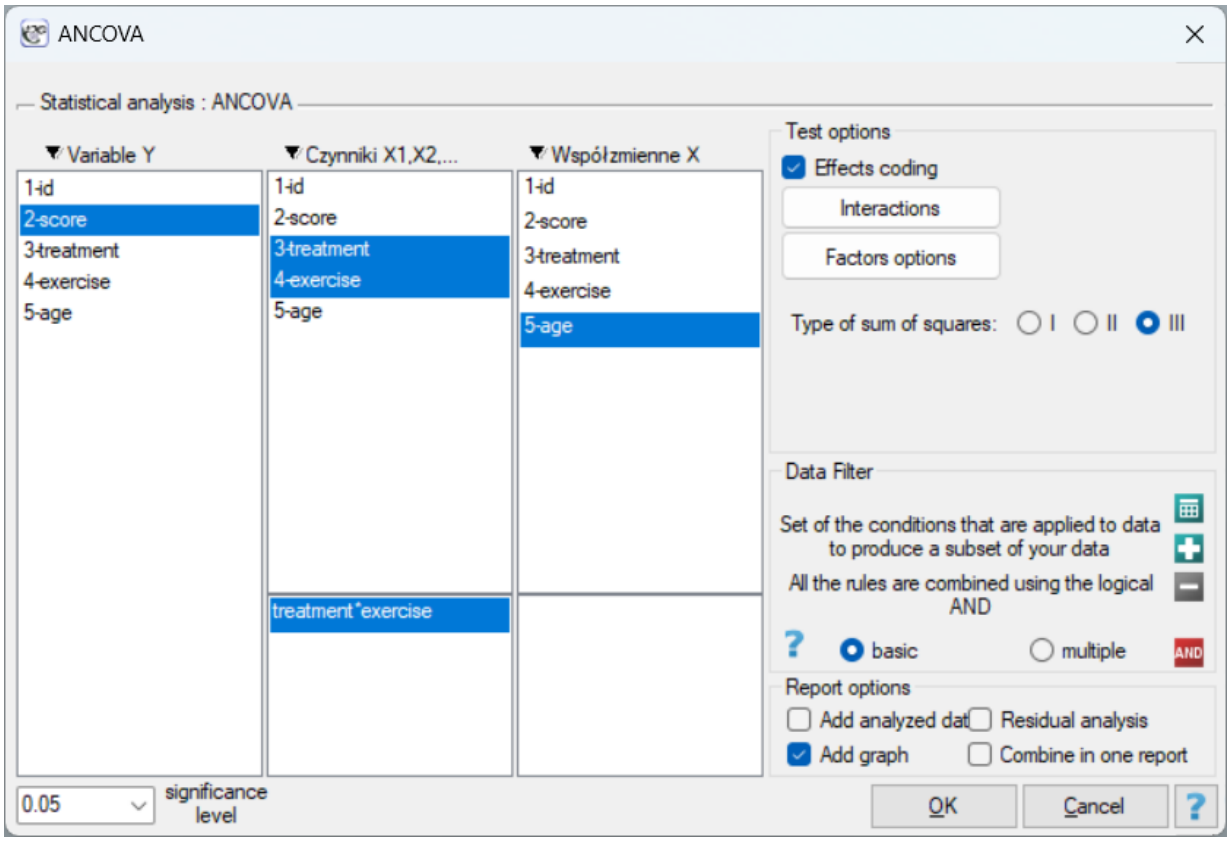
Note! The way in which the studied factors and confounding variables are taken into account is described in the chapter on multi-factor ANOVA (The influence of confounding factors). The recommended approach is to choose Type III sums of squares together with effect coding.

Equality of the slopes of the regression lines is tested using an F test comparing the model containing the analysed factors with the same model extended to include interactions with the confounding factors. A statistically significant result means that the assumption of equal slopes has been violated, because the interaction becomes significant, indicating different slopes of the lines.

HOW TO SET UP THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

Advanced statistics → Multivariate models → ANCOVA.



PERFORMING AND INTERPRETING THE ANALYSIS

Example 25.9. (drug_cholesterol.pqs file)
Imagine that a researcher was conducting a study of a new cholesterol-lowering drug. The study was designed so that the drug dose occurred at three levels: high, low, and placebo. The researcher checked (using an independent ANOVA) whether cholesterol after treatment differed depending on the dose of the drug.

Copyright ©2010-2026 U-smile Stats – All rights reserved

One-way ANOVA for independent groups	
Analysed variables	post
	Dose
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	Dose
ANOVA for independent groups	
Eta-square	0.15233
Total sum of squares (SS[T])	90.14667
Between-groups sum of squares (SS[BG])	13.73167
Within-groups sum of squares (SS[WG])	76.415
Mean square between-groups (MS[BG])	6.86583
Mean square within-groups (MS[WG])	2.83019
Between-groups degrees of freedom (df[BG])	2
Within-groups degrees of freedom (df[WG])	27
Total degrees of freedom (df[T])	29
F statistic	2.42593
p-value	0.10742

Unfortunately, the researcher found no confirmation of differences between the results. Suppose the researcher then realized that whether a given drug changes cholesterol levels might be related to the patient's baseline cholesterol level and to their age. For this reason, they decided to perform a one-way ANCOVA (with drug dose as the factor), including pre-treatment cholesterol level and age as covariates.

Between-groups (effects) test

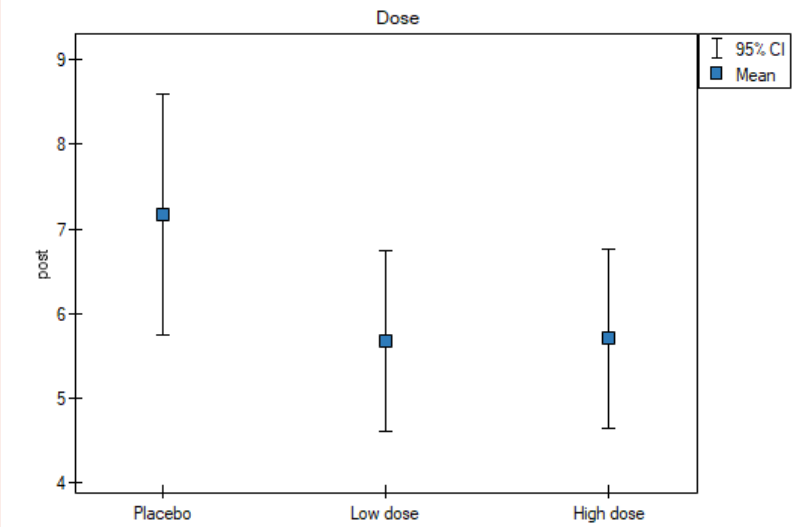
	SS	MS	DF	F	p	Partial Eta-s
intercept	0.04193	0.04193	1	0.04026	0.8426	0.00161
Dose	33.87685	16.93842	2	16.26451	0.00003	0.56544
pre	45.79979	45.79979	1	43.9776	<0.00001	0.63756
age	1.21765	1.21765	1	1.16921	0.28988	0.04468
within(błąd)	26.03586	1.04143	25			
total	90.14667					

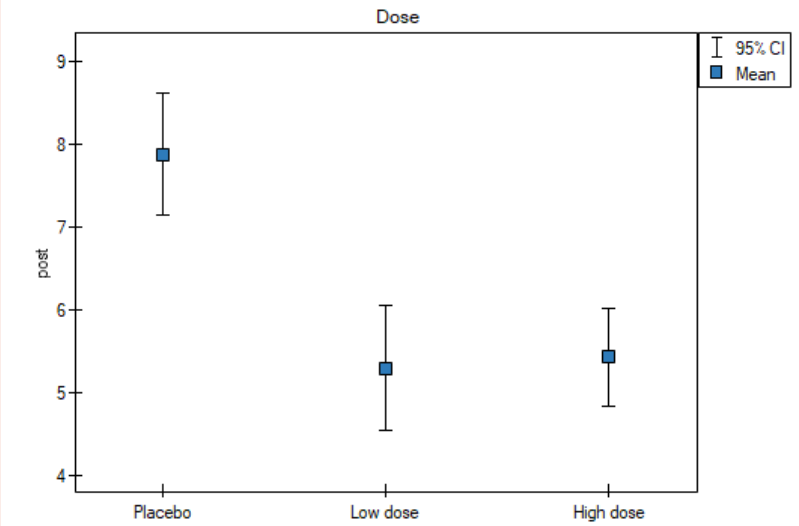
ANCOVA: $p = 0.00003$ (statistically significant, $\alpha = 0.05$)

After adjusting for pre-treatment cholesterol level and age, cholesterol levels differ significantly across the different drug doses – unlike the result of the plain ANOVA.

Including the pre-treatment cholesterol level reduced the errors obtained for the means and narrowed the confidence intervals. To display the observed or expected means, we choose the appropriate settings via Factor options, checking the error plot. The first plot shows the observed means with their confidence interval, i.e. not accounting for the effect of age and pre-treatment cholesterol level; the second plot shows the means expected from the fitted model, together with their confidence intervals, i.e. after accounting for the effect of these two covariates:

Copyright ©2010-2026 U-smile Stats – All rights reserved





As a result, by accounting for pre-treatment cholesterol level, the researcher was able to demonstrate the effectiveness of the new treatment. Pre-treatment cholesterol level and age explain part of the change in cholesterol level after treatment, but the remaining part of the change – 57% – can be attributed to the drug dose used (partial Eta-squared = 0.565437). Post-hoc tests (selected via Factor options) suggested the formation of two homogeneous groups: the placebo group and the group of patients receiving the drug, indicating that raising the dose to the high level makes no difference, since the cholesterol levels obtained would be similar.

Copyright ©2010-2026 U-smile Stats – All rights reserved

POST-HOC Fisher LSD			
	NIR [F]	Statistic [F]	p-value [F]
Dose			
Placebo;Low dose	1.02128	3.00814	0.00592
Placebo;High dose	0.91139	3.31434	0.0028
Low dose;High dose	0.94445	-0.05452	0.95696

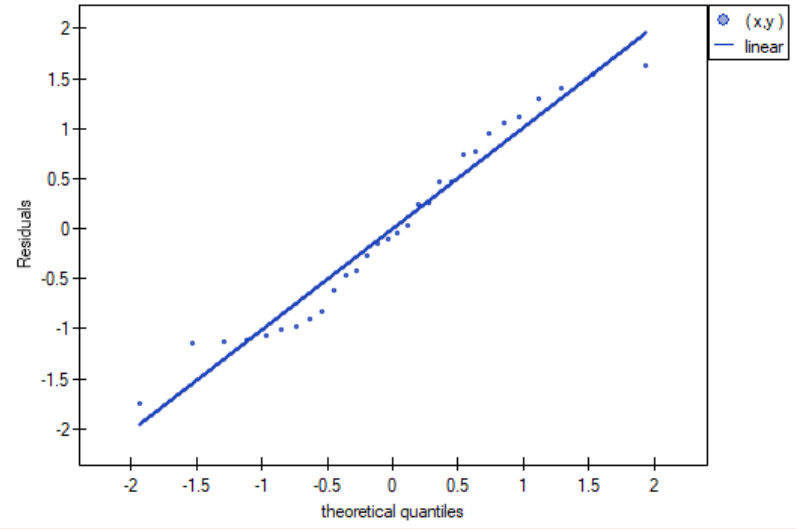
Homogeneous groups		
Dose	A	B
Placebo(b)		*
Low dose(a)	*	
High dose(a)	*	

The assumptions of ANCOVA remained to be checked. Homogeneity of variance and the constancy of the regression line slopes were confirmed using tests:

Homogeniczność nacheleń	
DF1	4
DF2	21
p-value	0.80795

Equality of variance				
	Levene test		Brown-Forsythe test	
	F statistic	p-value	F statistic	p-value
Dose	0.69296	0.50877	0.3326	0.71995

Normality of the distribution of the residuals was assessed visually by plotting Q-Q plots:



Copyright ©2010-2026 U-smile Stats – All rights reserved

EXAMPLE 25.10. (stress.pqs file)
This example comes from the Datarium R-Cran package. Researchers want to assess the effect of a new treatment and of exercise on stress reduction, after accounting for differences in age. The stress measurement is the interval outcome variable Y. Because the variables "treatment" and "exercise" have 2 and 3 categories respectively, we will carry out a two-way ANCOVA to determine whether the interaction between exercise and treatment, while simultaneously accounting for the subjects' age, is related to stress.

Factors	
treatment X_1	exercise X_2
yes	low
	moderate
	high
no	low
	moderate
	high

In the analysis window we set "stress" as the dependent variable, "treatment" and "exercise" as factors, and add the interaction of these two variables; the continuous covariate is "age".

Between-groups (effects) test

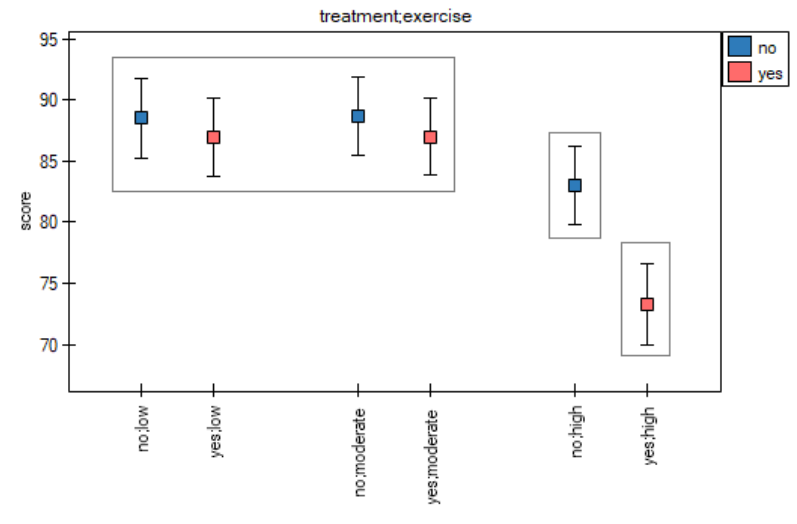
	SS	MS	DF	F	p	Partial Eta-s
intercept	731.73136	731.73136	1	29.4484	<0.00001	0.35717
treatment	274.95918	274.95918	1	11.06568	0.0016	0.17272
exercise	1029.43256	514.71628	2	20.71466	<0.00001	0.43873
treatment*exercise	220.93738	110.46869	2	4.44579	0.01641	0.14366
age	226.35626	226.35626	1	9.10967	0.0039	0.14667
within(błąd)	1316.93974	24.84792	53			
total	3888.26733					

treatment*exercise interaction: $p = 0.016409$ **(statistically significant,** $\alpha = 0.05$**)**

The effect of treatment on stress varies depending on the intensity of exercise.

We will plot the expected mean stress levels for each of the six subgroups into which the interaction has divided our data, and determine the post-hoc tests.

Copyright ©2010-2026 U-smile Stats – All rights reserved



treatment;exercise	A	B	C
no;low(b)		*	
no;moderate(b)		*	
no;high(c)			*
yes;low(b)		*	
yes;moderate(b)		*	
yes;high(a)	*		

According to the results of the post-hoc test, we can speak of three distinct homogeneous groups: (B) the group with the highest stress level is the group that exercises little or moderately (regardless of whether they are treated or not); (C) the group with a lower stress level is the group that exercises a lot and is not treated; (A) the group with the lowest stress level is the group that exercises a lot and is treated. The values of the individual means, together with their confidence intervals, are shown in the table:

Descriptive statistics (expected)

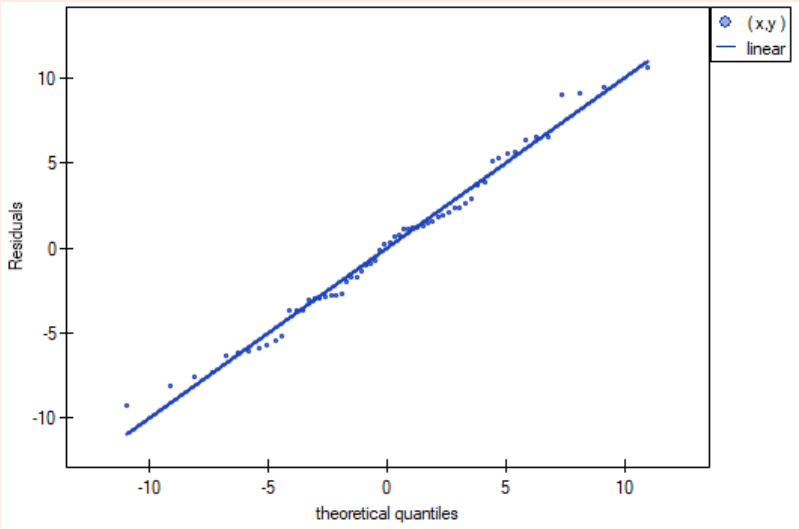
treatment	frequency	mean	stand. error	-95% CI	+95% CI
no	30	86.7365	0.91416	84.90292	88.57008
yes	30	82.41683	0.91416	80.58326	84.25041
exercise	frequency	mean	stand. error	-95% CI	+95% CI
low	20	87.74308	1.16114	85.41413	90.07202
moderate	20	87.83805	1.1184	85.59482	90.08127
high	20	78.14888	1.19011	75.76182	80.53594
treatment*exercise	frequency	mean	stand. error	-95% CI	+95% CI
no;low	10	88.50737	1.61662	85.26484	91.74989
no;moderate	10	88.67985	1.59478	85.48113	91.87857
no;high	10	83.02228	1.613	79.78701	86.25755
yes;low	10	86.97879	1.60313	83.76331	90.19426
yes;moderate	10	86.99624	1.5774	83.83237	90.16011
yes;high	10	73.27548	1.65137	69.96325	76.5877

The assumptions of equality of variance, equality of the regression line slopes, and normality of the model residuals are met:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Homogeniczność nacheleń	
DF1	5
DF2	48
p-value	0.98578

Equality of variance	Levene test		Brown-Forsythe test	
	F statistic	p-value	F statistic	p-value
treatment	0.33456	0.56522	0.30407	0.58346
exercise	0.32361	0.72485	0.35321	0.70396
treatment*exercise	0.95521	0.4534	0.82873	0.53491



FOR THOSE INTERESTED

Hypotheses – for a single factor A:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2 = ... = \mu_a, \\ \mathcal{H}_1: &\quad \text{not all } \mu_i \text{ are equal } (i = 1, 2, ..., a), \end{aligned}$$

where $\mu_1, \mu_2, ..., \mu_a$ – the expected means of factor A for its individual categories (adjusted for the covariates).

Hypotheses – for the interaction of factors $A * B$:

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_1 = \mu_2 = ... = \mu_{ab}, \\ \mathcal{H}_1: &\quad \text{not all } \mu_k \text{ are equal } (k = 1, 2, ..., ab), \end{aligned}$$

where $\mu_1, \mu_2, ..., \mu_{ab}$ – the expected means of the interaction of factors $A * B$ for their individual categories (adjusted for the covariates).

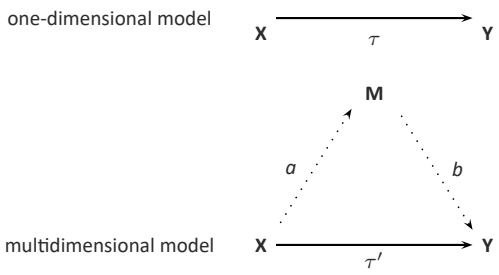
Copyright ©2010-2026 U-smile Stats – All rights reserved

Mediation effect

WHEN TO USE:

We use mediation analysis when we want to check whether the influence of one independent variable (X) on an outcome variable (Y) occurs, at least in part, "through" a third variable, called the **mediator** (M). In other words, we are looking for an answer to the question of whether the relationship between X and Y can be partly explained by the fact that X influences M, and M in turn influences Y.

Baron and Kenny (1986)[16] defined a mediator (M) as a variable that significantly explains the relationship between an independent variable (X) and an outcome variable (Y). In mediation, the relationship between the independent variable and the dependent variable is assumed to be, at least in part, an indirect effect that exists thanks to the influence of a third variable (the mediator).



First we build a **one-dimensional model**, in which the mediator is not taken into account at all – it describes the "raw" relationship of X with Y with the coefficient τ. Next we build a **multidimensional model**, in which both X and M are used simultaneously to explain Y – in this model the relationship of X with Y is described by the coefficient τ'. If, after the mediator is added to the model, the influence of X on Y clearly decreases (that is, τ' is noticeably smaller than τ), this means that part of the effect of X on Y indeed passes through the mediator M.

ASSUMPTIONS

- variables X, Y and M should be measured on a scale that allows the construction of linear regression models (that is, they should be quantitative),
- it is necessary to determine three regression models: the relationship of X with Y (without the mediator), the relationship of X and M with Y (with the mediator), and the relationship of X with M,
- the tests assessing the significance of the mediation effect (Sobel, Aroian, Goodman) are conservative tests and give reliable results only for large sample sizes (above 100 observations).

DEFINITIONS OF MEASURES

Comparing the one-dimensional model (τ) with the multidimensional model that includes the mediator (τ'), we distinguish three kinds of effects:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Total effect – the entire influence of variable X on variable Y, determined in the one-dimensional model (without taking the mediator into account) and described by the coefficient τ.

Direct effect – the part of the influence of X on Y that remains once the mediator M is included in the model. This is the coefficient τ' – it describes the influence of X on Y that is independent of the mediation by M.

Indirect (mediation) effect – the part of the influence of X on Y that passes "through" the mediator, i.e. the difference between the total effect and the direct effect ($\tau - \tau'$). This quantity can also be expressed as the product of the coefficients describing the successive legs of the path X→M→Y (coefficient *a*, describing the influence of X on M, and coefficient *b*, describing the influence of M on Y while simultaneously taking X into account).

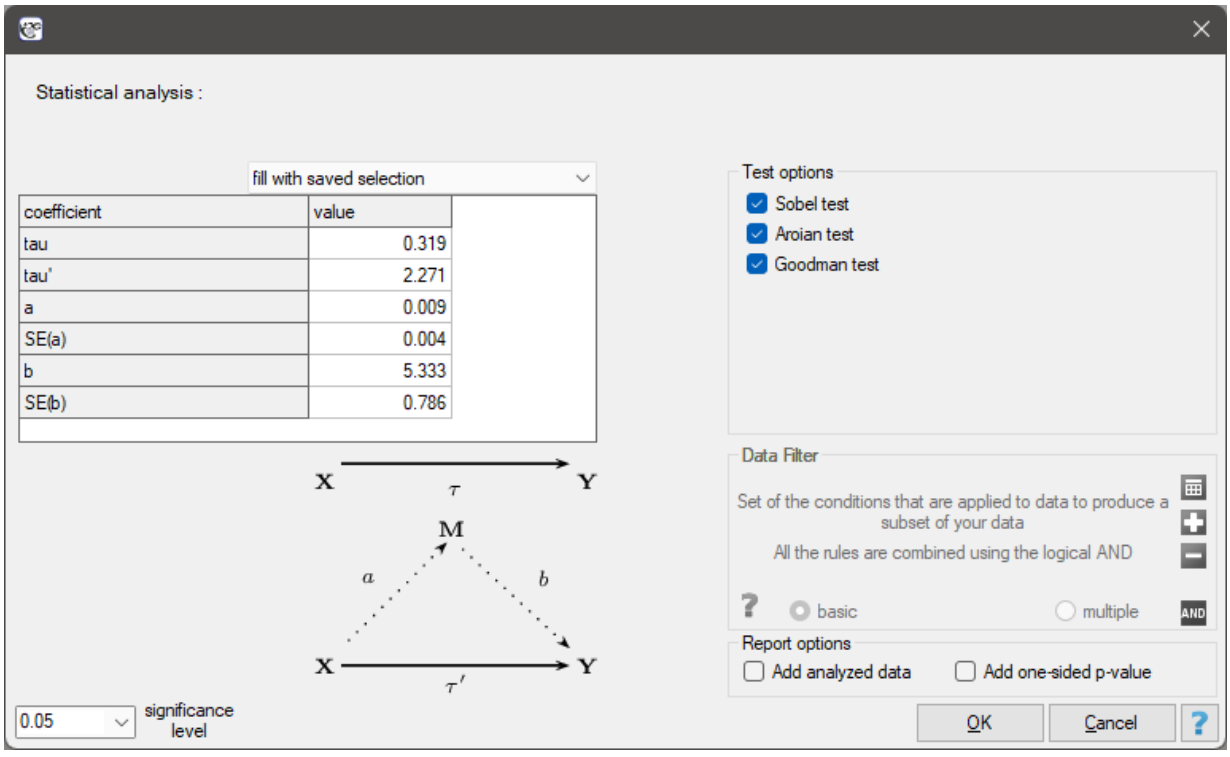
To assess how large a part of the relationship of X with Y is "taken over" by the mediator, the **percentage mediation effect** is determined, that is, the share of the indirect effect in the total effect. The higher this value, the more important the mediator becomes in explaining the relationship under study.

To check whether the indirect effect determined in this way is statistically significant (that is, whether the reduction in the influence of X on Y after including the mediator is not due to chance), the **Sobel (1982)[159], Aroian (1947)[7] (popularized by Baron and Kenny[16]) and Goodman (1960)[72] tests** are used.

Note! The Sobel test, as well as the Aroian and Goodman tests, are very conservative tests intended only for large samples (exceeding 100 observations).

HOW TO SET UP THE ANALYSIS

The mediation effect analysis window is opened via: Advanced Statistics→Multivariate models→Mediation effect.



Copyright ©2010-2026 U-smile Stats – All rights reserved

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 25.11. Based on the paper by Mimar Sinan Fine (2017)[126].
The study includes 300 adults living in Istanbul. The dependent variable Y is systolic blood pressure, and the independent variable X is age. The mediating variable M is the frequency of alcohol consumption. The aim of the study is to examine the relationship between age and systolic blood pressure and to present the influence of the frequency of alcohol consumption on this relationship.

A one-dimensional model was built, in which the potential mediator was not taken into account:
$Y = 0.319 \cdot X + c$.
The size of the influence of variable X (age) on variable Y (systolic blood pressure) was **tau**=0.319.

A multidimensional model was built, in which the potential mediator was taken into account:
$Y = 2.271 \cdot X + 5.333 \cdot M + c$.
The size of the influence of variable X (age) on variable Y (systolic blood pressure) was **tau'**=2.271. We also know from this model that **b**=5.333, with error SE_b=0.786.

The difference between the coefficients is **tau-tau'**=a*b=0.048. The mediation effect is **(tau-tau')/tau**=(0.319-0.271)/0.371=0.15047, which means that M (frequency of alcohol consumption) modifies the relationship under study, reducing the coefficient by about 15%.

A further one-dimensional model was built, examining the influence of variable X on the mediator:
$M = 0.009 \cdot X + c$.
We know from this model that the coefficient **a**=0.009, with error SE_a=0.004. All this information is entered in the analysis window, giving the following report:

Mediation effect	
tau-tau'	-1.952
Mediation effect	-6.1191
Sobel Test	
Z statistic	2.1356
Two sided p-value	0.0327
Aroian Test	
Z statistic	2.115
Two sided p-value	0.0344
Goodman Test	
Z statistic	2.1568
Two sided p-value	0.031

On the basis of the coefficients a and b and their standard errors, the results of the Sobel, Aroian and Goodman tests are determined.

Sobel $p = 0.0327$, Aroian $p = 0.0344$, Goodman $p = 0.0310$ (statistically significant, $\alpha = 0.05$)

The frequency of alcohol consumption significantly mediates the relationship between age and systolic blood pressure.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The p values obtained indicate a statistically significant role of the mediator. We have therefore confirmed that the frequency of alcohol consumption influences the relationship between age and systolic blood pressure noticeably enough that it is worth explaining why this effect occurs.

FOR THOSE INTERESTED

Formulas:
We determine the size of the change through the difference between the coefficients describing the relationship of variable X with variable Y in the one-dimensional model:
$Y = \tau \cdot X + c$
and in the multidimensional model, that is, including the variable M:
$Y = \tau \cdot X + b \cdot M + c.$

Difference:

$$\tau - \tau' = a \cdot b$$

Mediation effect:

$$\frac{\tau - \tau'}{\tau} \cdot 100\%$$

As a result, when the mediator (M) is included in the regression model describing the relationship between variable X and Y, the influence of the independent variable τ is reduced to τ'.

Tests assessing the mediation effect
Interpretation:
The Sobel (1982)[159] test, the Aroian (1947)[7] test popularized by Baron and Kenny[16], and the Goodman (1960)[72] test are tests that determine whether the reduction in the influence of the independent variable on the outcome variable, after including the mediator in the model, is a substantial reduction, and hence whether the mediation effect is statistically significant.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \quad \tau = \tau' \\ \mathcal{H}_1 &: \quad \tau \neq \tau', \end{aligned}$$

Formulas:
The test statistic for the Sobel test has the form:

$$Z = \frac{a \cdot b}{\sqrt{b^2 \cdot SE_a^2 + a^2 \cdot SE_b^2}}$$

The test statistic for the Aroian test has the form:

$$Z = \frac{a \cdot b}{\sqrt{b^2 \cdot SE_a^2 + a^2 \cdot SE_b^2 + SE_a^2 \cdot SE_b^2}}$$

The test statistic for the Goodman test has the form:

$$Z = \frac{a \cdot b}{\sqrt{b^2 \cdot SE_a^2 + a^2 \cdot SE_b^2 - SE_a^2 \cdot SE_b^2}}$$

These statistics asymptotically (for large sample sizes) have the normal distribution.

The p value, determined on the basis of the test statistic, is compared with the significance level α:

Copyright ©2010-2026 U-smile Stats – All rights reserved

if $p \leq \alpha \implies$ we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$,
if $p > \alpha \implies$ there is no reason to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

26 MIXED MODELS

Introduction: dependent data, fixed effects and random effects

WHY MIXED MODELS

The classical regressions available in the program – multiple regression, logistic regression and Cox regression – assume that all observations are independent of one another. In many studies this assumption does not hold, because observations naturally fall into groups:

- **repeated measurements** – the same patient examined several times (before treatment, after a week, after a month), the same person tested on consecutive days;
- **clustered data** – patients treated in the same centre, pupils from the same class, animals from one litter or herd;
- **hierarchical data** – measurements of a patient, patients within a centre, centres within a region (several levels of nesting);
- **recurrent events and paired objects** – successive infections of the same patient, the two eyes of the same person.

Observations from the same group are more alike than observations from different groups: a patient with high blood pressure usually has high pressure at every visit, and a centre with poorer outcomes has poorer outcomes for most of its patients. If we ignore this dependence and fit an ordinary regression, the program treats e.g. 10 measurements of one patient as 10 independent people. The consequences are twofold:

- **standard errors are underestimated** and p-values are too small – we declare effects significant too easily (so-called pseudoreplication);
- we do not know **what share of the variability** comes from differences between groups (people, centres) and what share from variability within groups, although this is often the most interesting information in the study.

A mixed model solves both problems: it accounts for the group structure when computing errors, and it explicitly estimates how much the groups differ from one another.

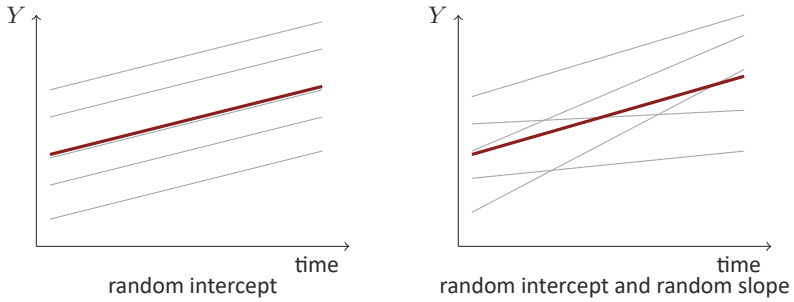
FIXED EFFECTS AND RANDOM EFFECTS

The name “mixed model” comes from the fact that the model contains two kinds of effects:

- **fixed effects** – the ordinary regression coefficients $\beta_0, \beta_1, \ldots$ of the independent variables. They describe the population as a whole and are interpreted exactly as in classical regression: by how much Y changes when X increases by one unit;
- **random effects** – deviations of individual groups from what the fixed effects predict. We are not interested in the individual values of these deviations for each group, but in their **variance**, i.e. a measure of how much the groups differ. We assume that the groups in the study are a random sample from a larger population of groups (patients, centres), and that their effects are normally distributed with mean 0.

Copyright ©2010-2026 U-smile Stats – All rights reserved

The simplest random effect is a **random intercept**: each group has its own level of Y, shifted relative to the population level. If, in addition, the effect of some variable (most often time) differs between groups, we introduce a **random slope** of that variable. The figure below shows the difference: on the left the groups have different levels but the same trend; on the right they also differ in the rate of change.



Thick line – fixed effects (population); thin lines – individual groups.

THE THREE MIXED MODELS IN THE PROGRAM

Each classical regression has a mixed counterpart in the program. The choice depends solely on the type of the dependent variable:

Dependent variable	Classical model	Mixed model	Random effects
continuous	multiple regression	mixed multiple regression	intercept, slopes
binary (0/1)	logistic regression	mixed logistic regression	intercept, slopes
time and status	Cox PH regression	Cox regression with random effect	group frailty

Mixed multiple regression replaces repeated measures ANOVA when measurements are incomplete, unevenly spaced in time, or when continuous variables are to be included in the model. Mixed logistic regression plays the same role with respect to Cochran's Q ANOVA. Cox regression with random effect is the counterpart of Cox regression for patients from many centres or for recurrent events.

DATA PREPARATION: LONG FORMAT

All three analyses require data in **long format**: each row of the sheet is one observation (one measurement), and a separate column – the **grouping variable** (ID) – says which group the observation belongs to. The grouping variable can be a number, text or date; the program treats it purely as a label. Groups may have different numbers of observations, and missing data are removed row by row, not group by group.

wide format (repeated measures ANOVA)

Patient	Day 0	Day 1	Day 2
308	249.6	258.7	250.8
309	222.7	205.3	203.0
310	199.1	194.3	234.3

long format (mixed models)

Patient	Day	Reaction
308	0	249.6
308	1	258.7
308	2	250.8
309	0	222.7
309	1	205.3
...	...	...

Data stored in wide format (successive measurements in successive columns, as in repeated measures ANOVA) must be rearranged into long format before the analysis: the measurement columns are copied

Copyright ©2010-2026 U-smile Stats – All rights reserved

one below another into a common column of values, next to it we enter the object identifier (repeated as many times as there are measurements) and the number or time of the measurement. This rearrangement is performed automatically by the command Data→Reshape: wide → long... (see reshaping data between the wide and the long format); the reverse command Reshape: long → wide... returns to the wide layout. The commands Create raw data and Create table serve a different purpose – they convert a contingency table (counts) into raw data and back.

Note! An independent variable that is to receive a random slope must vary within groups (e.g. measurement time, a dose given at successive visits). A variable constant within a group (e.g. the patient's sex when the group is the patient) can only be a fixed effect.

WHAT TO READ IN THE REPORT AND IN WHAT ORDER

The reports of the three models share a common layout. Recommended reading order:

1. **Convergence and singular fit.** If the algorithm did not converge, or the report flags a singular fit (a random-effect variance practically equal to zero, or a random-effects correlation equal to ± 1), the remaining numbers are unreliable. The model should be simplified: remove a random slope, choose the independent covariance, or reduce the number of variables.
2. **ICC from the null model.** The intraclass correlation coefficient says what share of the total variability of Y lies between groups. An ICC close to 0 means that observations within a group are almost independent; an ICC of 0.4 means that 40% of the variability is due to differences between groups, and two observations from the same group are correlated at 0.4. A small ICC is not, however, a reason to abandon the mixed model if the study design is hierarchical – the mixed model remains the correct method, and ordinary regression does not.
3. **Random effects test.** The likelihood ratio (LR) test compares the mixed model with a model without random effects. The null hypothesis (variance equal to 0) lies on the boundary of the admissible values, so the program corrects the p-value (a mixture of χ^2 distributions; in the simplest case the p-value is halved) and says so in the report.
4. **Random effects.** The standard deviation of the random intercept is expressed in units of Y (in logistic and Cox regression – on the logit or log-hazard scale) and says how far a typical group departs from the mean. The standard deviation of a random slope says how much the groups differ in their rate of change.
5. **Fixed effects.** Read as in classical regression, with one difference: in logistic and Cox regression the odds ratio and hazard ratio are **conditional** – they describe the effect of a variable "within the same group" (for the same patient, the same centre). In absolute value they are larger than the population OR and HR from ordinary regression, which averages over groups, and the two should not be compared directly.
6. **Group random effects (BLUP).** Predicted deviations of the individual groups. They serve to point out atypical groups (a patient with an outlying trajectory, a centre with a higher risk), not to test significance. The "caterpillar" plot (groups ordered by effect, with intervals) shows them most clearly.

HOW MANY GROUPS ARE NEEDED

The variance of a random effect is estimated from the number of groups, not the number of observations. The program requires at least 3 groups, but with fewer than 10 groups it warns that the variance

Copyright ©2010-2026 U-smile Stats – All rights reserved

estimate is uncertain. In practice reliable estimates begin at ten to several dozen groups, and each random slope raises these requirements. With 3–5 centres it is more sensible to treat the centre as a dummy variable (fixed effect) than as a random effect.

LIMITATIONS

Mixed models are not included in Test summaries or in the Wizard. Two levels of nesting (e.g. patient within centre) and the AR(1) and exchangeable residual correlation structures are available in mixed multiple regression; in Cox regression with random effect only a random intercept (gamma frailty) is available.

Copyright ©2010-2026 U-smile Stats – All rights reserved

26.1 MIXED MULTIPLE REGRESSION

Mixed multiple regression (linear mixed model)

WHEN TO USE:

Mixed multiple regression (linear mixed model, LMM) extends multiple linear regression to data in which observations are not independent but form groups: repeated measurements of the same patient, patients treated in the same centre, pupils from one class, animals from one litter (see the introduction to mixed models). Observations from the same group are more alike than observations from different groups. Ordinary regression ignores this dependence, which underestimates the standard errors of the coefficients and overstates significance. A mixed model splits the variability of the dependent variable into a between-group part and a within-group part, and every coefficient receives a proper error.

The model contains two kinds of effects:

- **fixed effects** – the ordinary regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ describing the population,
- **random effects** – deviations of individual groups from the population: a random intercept (each group has its own level of Y) and optionally random slopes (the effect of a chosen variable X differs between groups).

A model with a random intercept and a random slope of X_1 has the form:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \ldots + \beta_k X_{kij} + u_{0j} + u_{1j} X_{1ij} + \epsilon_{ij},$$

where:
Y_{ij} – the i-th observation in group j,
$\beta_0, \beta_1, \ldots, \beta_k$ – fixed effects,
u_{0j} – the random intercept of group j, u_{1j} – the random slope of group j; (u_{0j}, u_{1j}) are assumed normal with mean 0 and covariance matrix Σ (variances τ_0^2, τ_1^2 and correlation ρ_{01}),
ϵ_{ij} – the residual, normal with mean 0 and variance σ^2.

For group j the model thus describes its own straight line: its intercept is $\beta_0 + u_{0j}$ and its slope $\beta_1 + u_{1j}$. The fixed effects describe the line of the "average" group, and the random-effect variances – how widely the lines of the individual groups are scattered around it.

This model replaces repeated measures ANOVA when the repeated measurements are incomplete or unevenly spaced in time, when the number of measurements differs between subjects, or when, besides the repeated factor, the model is to contain continuous variables (e.g. age, dose) or variables describing the group (e.g. sex, centre).

ASSUMPTIONS

- the dependent variable Y is measured on an interval scale,
- a linear relationship between the dependent variable and the independent variables,
- random effects and residuals are normally distributed; residuals have constant variance,

Copyright ©2010-2026 U-smile Stats – All rights reserved

- groups are independent of one another – dependence occurs only within groups,
- a sufficient number of groups – at least 3, and usually a dozen or more for a reliable estimate of the random-effect variances,
- missing data are random (the model does not require complete or balanced measurements).

Note! The data must be prepared in long format: each row of the sheet is one observation, and a separate column (the grouping variable, e.g. patient ID) indicates which group the observation belongs to. Data in wide format (measurements in successive columns) must first be rearranged into long format (see the introduction).

DEFINITIONS OF MEASURES

Random-effect variances – τ_0^2 (random intercept) says how much the groups differ in the level of Y; τ_1^2 (random slope) – how much they differ in the effect of X_1. Their square roots, the standard deviations, are easier to read because they are expressed in units of Y (or Y per unit of X_1): a typical group departs from the population by about one standard deviation, and about 95% of groups fall within ± 2 deviations.

Random-effects correlation ρ_{01} – the relationship between a group's level and its slope. A negative correlation means that groups starting higher change more slowly. Note: its value depends on where the zero of X_1 lies – the intercept is the level at $X_1 = 0$. When X_1 is e.g. age 8–14 years, it is worth centring it (subtracting the middle value) so that the correlation describes the level in the middle of the range rather than a hypothetical level at age 0.

Intraclass correlation coefficient (ICC) – computed from the **null model** (no independent variables, only a random intercept):

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2},$$

where τ^2 is the between-group variance and σ^2 the within-group (residual) variance. The ICC is the share of between-group variability in the total variability and, at the same time, the correlation of two observations from the same group. Its confidence interval is approximate (delta method).

Marginal and conditional R^2 (Nakagawa and Schielzeth 2013[196]) – the marginal R^2 is the share of variance explained by the fixed effects alone, the conditional R^2 – by the fixed and random effects together. Their difference says how much variability is "explained" by group membership.

LR test of random effects – a likelihood ratio test comparing the model with and without a given random effect. The null hypothesis (variance equal to 0) lies on the boundary of the parameter space, so the p-value from the χ^2 distribution is corrected (a mixture of distributions; for a single parameter the p-value is halved). The program always computes this test by ML, regardless of the estimation method chosen for the main model.

Group random effects (BLUP) – predicted values of u_{0j}, u_{1j} for each group (best linear unbiased predictor). They are "shrunk" towards zero: a group with few observations or a large scatter receives an effect closer to 0 than the plain mean of its observations would suggest. They serve to identify atypical groups and to predict for a specific group, not to test significance.

Singular fit – a situation in which the estimated variance of some random effect is practically 0, or the random-effects correlation equals ± 1. The model is then too complex for the data and should be simplified (remove the random slope or choose the independent covariance).

Copyright ©2010-2026 U-smile Stats – All rights reserved

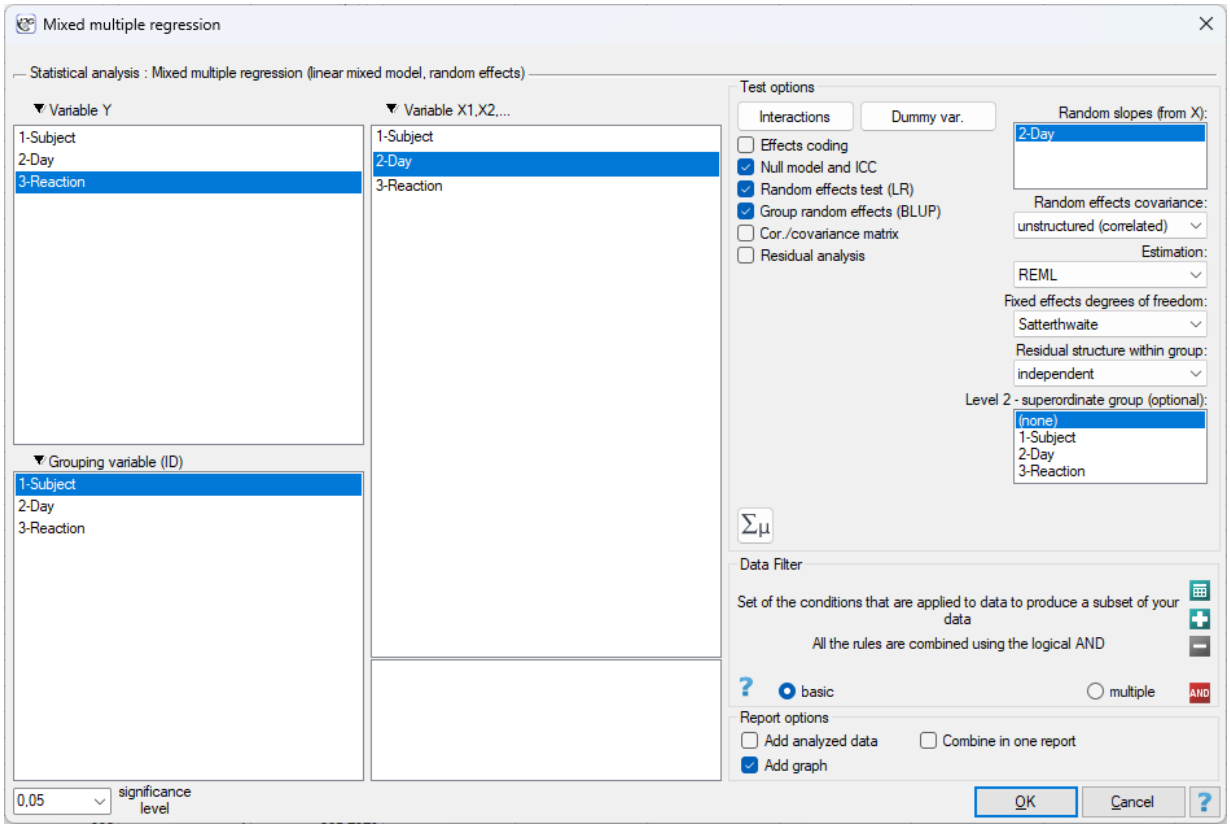
Satterthwaite degrees of freedom – in a mixed model there is no single simple number of degrees of freedom for the test of a fixed effect. A variable describing the group (e.g. the patient's sex) effectively has as many independent observations as there are groups, while a variable varying within groups has more. The Satterthwaite method assigns each coefficient its own, usually non-integer, degrees of freedom, so that the t-test is reliable even with few groups.

Residual structure within group – by default the residuals ϵ_{ij} are independent and all within-group correlation comes from the random effects. For measurements over time one may additionally assume that residuals of neighbouring measurements are correlated: AR(1) (the correlation ρ decays exponentially with the distance in the row order of the group: $\rho, \rho^2, \rho^3, \ldots$) or exchangeable (CS) (equal correlation of all pairs of residuals within the group).

Level 2 – a second, superordinate level of grouping, e.g. patient (level 1) nested within centre (level 2). The model then receives two random intercepts, and the report – a variance and an ICC separately for each level.

HOW TO SET UP THE ANALYSIS

Statistics → Mixed models → Mixed multiple regression.



The analysis window has the layout known from multiple regression. In its left part we indicate the variables:

- Dependent variable Y – the continuous variable to be explained,
- Grouping variable (ID) – the column identifying the group (patient, centre, litter); numeric or text,

Copyright ©2010-2026 U-smile Stats – All rights reserved

– Independent variables X1, X2, … – predictors; dummy variables and interactions are prepared with the Dummy variables and Interactions buttons as in multiple regression (chapter Preparing variables for analysis in multidimensional models).

The right part of the window contains the mixed-model options. Below we describe what each of them changes.

Random effects

– Random slopes (from X) – the list is filled with the variables chosen as X. Selecting a variable means that its effect is to differ between groups (e.g. the rate of change over time differs for each patient). The random intercept is always in the model and does not need to be selected. A variable with a random slope must vary within groups; a variable constant within a group (e.g. the patient's sex) cannot have a random slope – the program will report an error. Each random slope adds a variance and (with the unstructured covariance) correlations to the model, so it requires more groups; it is sensible to start with a single slope, most often for time.

– Random effects covariance – unstructured (correlated) also estimates the correlation between the random intercept and the slope (and between slopes); independent sets these correlations to 0. The independent variant has fewer parameters, is more stable with few groups, and is recommended when the correlation in the full model comes out close to 0 or ± 1 (singular fit). Without random slopes the option has no effect.

Estimation

– Estimation – REML (restricted maximum likelihood, default) gives unbiased variance estimates and is recommended for the final model; ML (maximum likelihood) slightly underestimates the variances, but only the ML likelihood allows models differing in fixed effects to be compared with information criteria. The choice of method does not noticeably change the fixed-effect coefficients, but it does change the random-effect variances, their errors and the values of $-2LL$, AIC, BIC. Rule: variable selection – ML, reporting the model – REML.

– Fixed effects degrees of freedom – Satterthwaite (default) assigns each coefficient its own degrees of freedom and a t-test; always recommended, especially with few groups. Wald z test treats the statistic as in a large sample (normal distribution, no degrees of freedom); it gives smaller p-values and is appropriate only with many groups.

– Residual structure within group – independent (default), AR(1) (row order) or exchangeable (CS). AR(1) requires the rows of each group to be ordered in time in the sheet and roughly equally spaced. The parameter ρ appears in the report. Exchangeable together with a random intercept is not identifiable (both describe the same constant within-group correlation), so it makes sense mainly as a comparison with the AR(1) model.

– Level 2 – superordinate group (optional) – the column indicating the higher-level group in which the grouping variable is nested (e.g. centre for patients). It adds a level-2 random intercept and, in the report, a separate variance, ICC and BLUP table for that level.

Report elements

– Null model and ICC – a model without independent variables is additionally fitted and the ICC with its confidence interval is derived from it; worth leaving on, because the ICC is the first number to look at.

Copyright ©2010-2026 U-smile Stats – All rights reserved

- Random effects test (LR) – likelihood ratio tests for each random component separately and for all of them jointly (mixed model versus ordinary multiple regression). Lengthens the computation, because additional models must be fitted.
- Group random effects (BLUP) – adds a table of the predicted effects of each group with errors; with hundreds of groups the table is long, and the same values are shown by the caterpillar plot.
- Cor./covariance matrix – correlation and covariance matrices of the fixed-effect estimates.
- Residual analysis – fitted values (conditional and population), residuals and standardized residuals for every observation.
- Add graphs – the caterpillar plot of group random effects and (with a random slope) group trajectories against the population line.

The remaining elements of the window (significance level, data filter, merging reports) work as in other analyses.

RESULTS AND INTERPRETATION

Counts – number of observations, number of groups and the minimum, median and maximum number of observations per group. With fewer than ten groups the report warns that the random-effect variance estimates are uncertain.

Estimation – method (REML/ML), number of iterations, information on convergence and on **singular fit**. Non-convergence or a singular fit means the model must be simplified before the remaining results are interpreted.

Model fit – log-likelihood, $-2LL$, the information criteria AIC, AICc and BIC (the smaller, the better the model) and the marginal and conditional R^2. Information criteria may only be compared between models fitted by the same method; models differing in fixed effects – only under ML.

LR test of random effects (vs multiple regression, ML) – whether accounting for the group structure is needed at all. A significant result means that ordinary multiple regression is not valid for these data. A non-significant result with a hierarchical design does not, however, exempt one from using the mixed model.

Null model – between-group variance τ^2, residual variance σ^2 and the **ICC** with its confidence interval.

Fixed effects (conditional) – coefficient b, its error, confidence interval, the t statistic with Satterthwaite degrees of freedom (may be non-integer) or the z statistic, and the p-value. Interpretation as in multiple regression: b is the change of Y per unit of X with the other variables held fixed, within the same group. The degrees of freedom hint at the level at which a variable "operates": a number close to the number of groups indicates a variable describing groups, a number close to the number of observations – a variable varying within groups.

Random effects – for each component the variance and standard deviation. The columns LR χ^2, df and p contain the test of whether the component is needed (model with it versus without it, ML, boundary-corrected). The last row is the residual variance. With an AR(1) or CS structure the report gives the parameter ρ, and with level 2 – the variance and ICC of that level.

Random effects correlations – with more than one component, the correlation matrix between the random intercept and the slopes. A value close to ± 1 signals a singular fit.

Group random effects (BLUP) – the predicted deviation of each group from the population with its error (for the intercept and each random slope). The sum of the fixed effect and the BLUP gives the group's

Copyright ©2010-2026 U-smile Stats – All rights reserved

own coefficient. On the caterpillar plot, groups whose interval does not include zero clearly depart from the average.

Residual analysis – conditional fitted values (with the group's random effects, i.e. "for this patient") and population fitted values (fixed effects only, i.e. "for the average patient"), residuals and standardized residuals; observations with $|residual| \geq 3$ are flagged as outliers.

Graphs – Group random effects (BLUP $\pm$ error): groups ordered by the value of the effect, with an error interval and a zero line; a separate panel for the intercept and for each slope. Group trajectories: for a model with a random slope, thin lines fitted for the individual groups against the thick population line – the most vivid summary of repeated measurements.

EXECUTION AND INTERPRETATION

Example 26.1. (sleepDeprivation.pqs file)
In a study of the effect of sleep loss on psychomotor performance (Belenky et al. 2003[190]), 18 subjects were allowed only 3 hours of sleep per night for 10 consecutive days. Each day their average reaction time (in milliseconds) in a vigilance test was measured. The data are in long format: the column Subject identifies the person, Day is the day of deprivation (0–9) and Reaction – the reaction time. In total 180 observations, 10 per subject.

Subject	Day	Reaction
308	0	249,56
308	1	258,7047
308	2	250,8006
308	3	321,4398
308	4	356,8519
308	5	414,6901
308	6	382,2038
308	7	290,1486
308	8	430,5853
308	9	466,3535
309	0	222,7339
309	1	205,2658
309	2	202,9778
309	3	204,707
309	4	207,7161
309	5	215,9618

We want to answer three questions: (1) by how much does the reaction time lengthen with each day without sleep, (2) do people differ in their resistance to sleep loss, i.e. is the rate of deterioration the same for everyone, (3) which subjects react to deprivation most strongly.

If we fitted an ordinary multiple regression of reaction time on day, the program would treat the 180 measurements as 180 independent people, whereas in fact we have 18 people, each measured ten

Copyright ©2010-2026 U-smile Stats – All rights reserved

times. We therefore build a mixed model: Reaction as the dependent variable Y, Subject as the grouping variable (ID), Day as the independent variable X_1. Since we ask whether the rate of change differs between people, we select Day in the Random slopes list; we leave the covariance unstructured (correlated), the estimation REML, the degrees of freedom Satterthwaite. We switch on Null model and ICC, Random effects test (LR), Group random effects (BLUP) and Add graphs.

Mixed multiple regression (linear mixed model, random effects)	
Analysed variables	Reaction
	Day
	Subject
Number of unspecified	0
Number of missing data	0
Significance level	0,05
Grouping variable	Subject
Size	180
Number of groups	18
Observations per group (min / median / max)	10 / 10 / 10
Number of variables in the model	1
Estimation method	REML
Number of iterations	169
Convergence	achieved
Singular fit	no
Log Likelihood (REML)	-871,814136
-2LL	1743,628272
Number of parameters	6
AIC - Akaike criterion	1755,628272
AICc - corrected Akaike criterion	1756,113821
BIC - Bayesian criterion	1774,786013
Marginal R2 (fixed effects)	0,278652
Conditional R2 (fixed + random)	0,799219
Residual variance	654,941024
Residual std. dev.	25,591816
LR test of random effects (vs multiple regression, ML)	
Chi-square	148,353712
degrees of freedom	3
p-value	<0,000001
Null model	
Between-group variance	1278,338204
Residual variance	1958,865115
ICC	0,39489
-95% CI	0,200158
+95% CI	0,589621

Step 1 – is the model properly fitted. The report gives 180 observations in 18 groups (10 observations per group), convergence achieved and no singular fit. We can read on.

Step 2 – how much variability lies between subjects (null model and ICC). In the null model the

Copyright ©2010-2026 U-smile Stats – All rights reserved

between-group variance is 1278.34 and the residual variance 1958.87, hence ICC $= 0.395$ (95% CI: 0.200–0.590). About 39% of the total variability of reaction time consists of stable differences between people (some have naturally faster reactions, others slower), and two measurements of the same person are correlated at 0.39. Far too much to treat the measurements as independent.

Step 3 – are random effects needed. The LR test of random effects against ordinary multiple regression gives $\chi^2 = 148.35$, $df = 3$, $p < 0.000001$ – the group structure is indispensable.

Fixed effects (conditional)

	b coeff.	b error	-95% CI	+95% CI	t stat.	df	p-value
intercept	251,405105	6,824555	237,006565	265,803645	36,838314	17,000193	<0,000001
Day	10,467286	1,545789	7,205957	13,728615	6,771483	17,000087	0,000003

Step 4 – fixed effects (answer to question 1). The intercept $b_0 = 251.41$ ms (95% CI: 237.01–265.80) is the average reaction time on day 0, i.e. before deprivation. The coefficient of day $b_1 = 10.47$ ms (95% CI: 7.21–13.73), $t = 6.77$, $df = 17.0$, $p = 0.000003$: each successive day without sleep lengthens the reaction time of an average person by about 10.5 ms. Note the degrees of freedom: they equal 17 (18 subjects minus 1), not 178, because it is 18 people who inform us about the rate of change, not 180 measurements. For comparison, ordinary multiple regression would give the same coefficient 10.47, but with an error of 0.80 instead of 1.55 – underestimated twofold.

Random effects

	component	variance	std. dev.	LR chi2	degrees of fi	p-value
Subject	intercept	612,089596	24,740445	22,140971	2	0,000009
	Day	35,071683	5,922135	42,139299	2	<0,000001
Residual		654,941024	25,591816			

LR test: model v

Random effects correlations

	intercept	Day
intercept	1	0,065551
Day	0,065551	1

Step 5 – random effects (answer to question 2). The standard deviation of the random intercept is 24.74 ms: a typical person on day 0 differs from the average by about 25 ms, and most people (95%) fall within $251 \pm 2 \cdot 25$, i.e. from about 200 to 300 ms. The standard deviation of the random slope is 5.92 ms/day: since the average rate is 10.5 ms/day, in some people the reaction lengthens by only 0–5 ms per day, in others by 15–20 ms. The LR test for the slope ($\chi^2 = 42.14$, $df = 2$, $p < 0.000001$) confirms that the rate of deterioration differs significantly between people. The correlation between the intercept and the slope is only 0.066 – how fast someone reacts at the start says nothing about how fast they will deteriorate. The residual variance 654.94 (standard deviation 25.59 ms) is the day-to-day variability of the same person's measurements, unexplained by their own line.

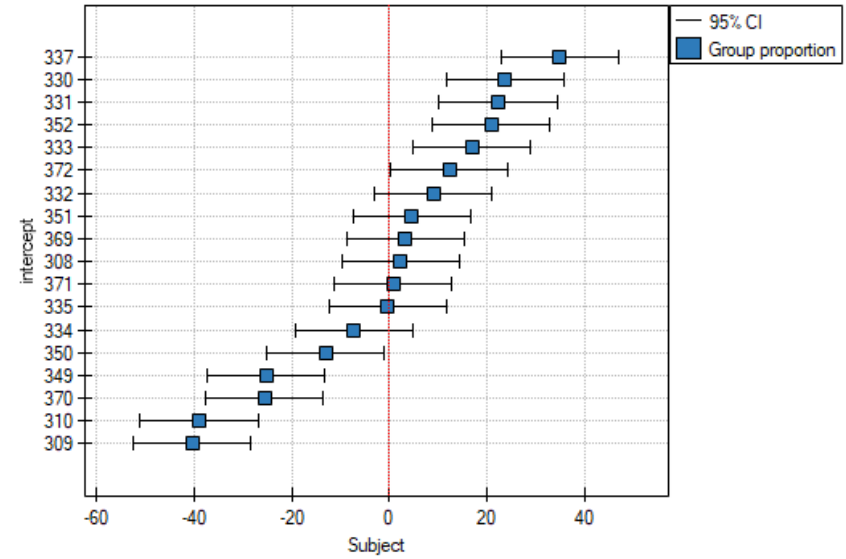
Marginal $R^2 = 0.279$: the day of deprivation alone explains 28% of the variability of reaction time. Conditional $R^2 = 0.799$: day together with the individual levels and slopes explains 80%. The difference (52 percentage points) is the variability attributed to differences between people.

Copyright ©2010-2026 U-smile Stats – All rights reserved

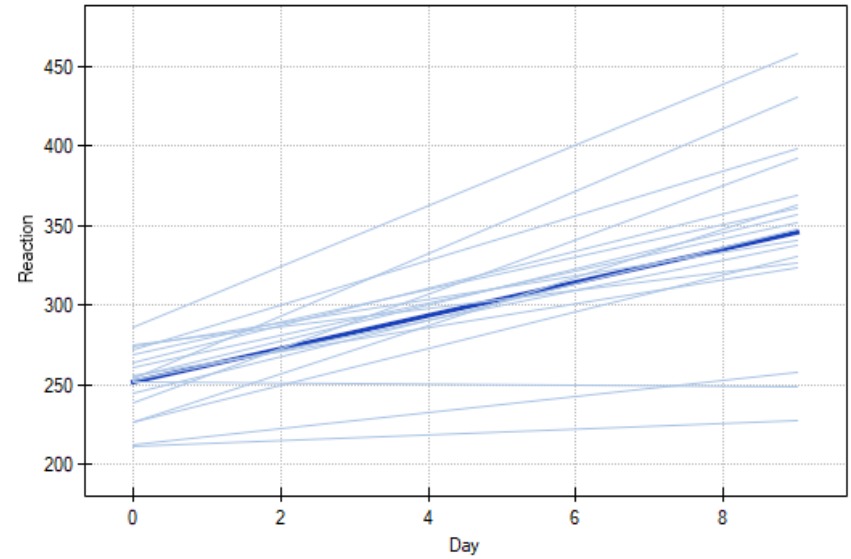
Group random effects (BLUP)

group	intercept	error	Day	error
308	2,258558	12,070837	9,198973	2,304837
309	-40,398571	12,070837	-8,619704	2,304837
310	-38,960242	12,070837	-5,44888	2,304837
330	23,690503	12,070837	-4,814332	2,304837
331	22,260206	12,070837	-3,069895	2,304837

Step 6 – which subjects are atypical (answer to question 3). The BLUP table gives each subject's deviations from the population. Subject 308 has intercept $+2.26$ (average start) and slope $+9.20$: their reaction time grows by $10.47 + 9.20 \approx 19.7$ ms per day, almost twice as fast as average – this is the person most sensitive to sleep loss. Subjects 309 and 310 start low (-40.4 and -39.0 ms respectively, i.e. they react fast) and have slopes -8.62 and -5.45: their reaction lengthens by only about 2 and 5 ms per day – they are almost immune to deprivation. The same information is shown by the graphs: the caterpillar plot (subjects ordered by effect, with an error interval) and the group trajectories (thin lines of the individual subjects against the thick population line).



Copyright ©2010-2026 U-smile Stats – All rights reserved



$b_{Day} = 10.47$ **ms (**$p = 0.000003$**), ICC** $= 0.395$**, slope** $SD = 5.92$ **ms/day (LR** $p < 0.000001$**)**

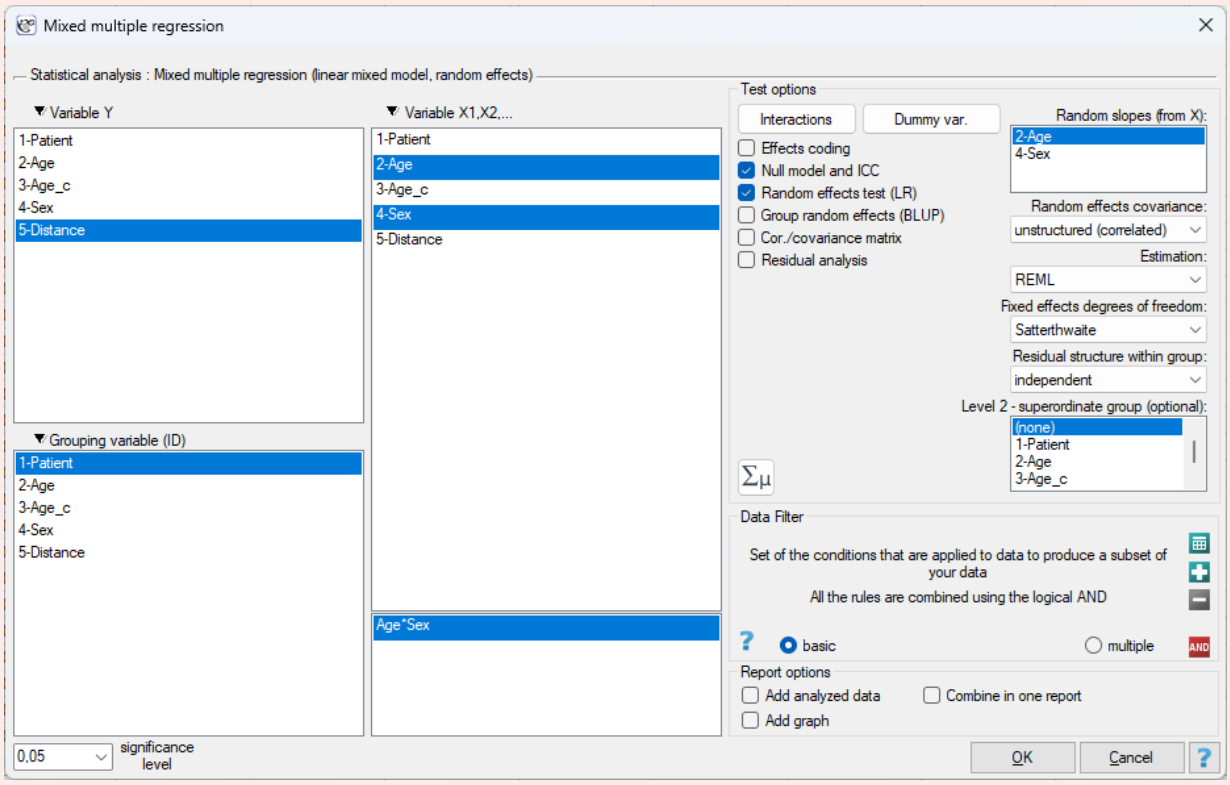
Each day of sleep deprivation lengthens the reaction time by 10.5 ms on average, but the rate of this deterioration differs significantly between people: from almost zero to about 20 ms per day.

What other settings change. The same model fitted by ML gives identical fixed-effect coefficients but smaller random-effect variances (565.5 and 32.7 instead of 612.1 and 35.1) – the well-known ML bias, which is why we use REML for reporting. Choosing the independent covariance removes the correlation (close to zero anyway): the variances hardly change, and the AIC drops from 1755.63 to 1753.67 because the model has one parameter fewer – the simpler model is equally good here. Choosing Wald z test instead of Satterthwaite gives for day $z = 6.77$ and $p < 0.000001$ instead of $p = 0.000003$; with 18 groups the difference is small, but with a handful of groups the Wald test would be too liberal.

EXAMPLE 26.2. (orthodontics.pqs file)
In a classical orthodontic study (Potthoff and Roy 1964[197]) the distance between the pituitary and the pterygomaxillary fissure (in mm) was measured on X-ray images of 27 children (16 boys and 11 girls) at ages 8, 10, 12 and 14. To show how the model handles unbalanced data, 20 measurements were removed at random from the full set – 88 observations remain, from 2 to 4 per child. Columns: Patient, Age, Age_c (age minus 11 years, i.e. centred), Sex (0 – boy, 1 – girl) and Distance. We ask whether girls and boys grow in this dimension at a different rate.

We build a model with Distance as Y, Patient as the grouping variable and the independent variables Age, Sex and the interaction Age×Sex created with the Interactions button. We give a random slope to Age (each child grows at its own rate); Sex cannot have a random slope, because it is constant within a child. The interaction is the key to the research question: its coefficient says by how much the slope (growth rate) of girls differs from that of boys.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Fixed effects (conditional)

	b coeff.	b error	-95% CI	+95% CI	t stat.	df	p-value
intercept	15,641529	1,08197	13,407767	17,87529	14,456528	23,861652	<0,000001
Age	0,842648	0,095888	0,644103	1,041193	8,787821	22,618417	<0,000001
Sex	2,111642	1,777548	-1,55245	5,775733	1,187952	24,58156	0,246204
Age*Sex	-0,394188	0,156384	-0,716839	-0,071537	-2,520636	24,156631	0,018721

The model converges without a singular fit, even though some children have only 2 measurements – a mixed model does not require complete data. Fixed effects: the intercept 15.64 is the distance of a boy at age 0 (a clinically meaningless extrapolation); age $b = 0.843$ mm/year ($p < 0.000001$) is the growth rate of boys; sex $b = 2.11$ ($p = 0.246$) is the difference of girls relative to boys at age 0, again meaningless; the interaction $b = -0.394$ mm/year ($t = -2.52$, $df = 24.2$, $p = 0.019$): girls grow 0.39 mm/year more slowly than boys, i.e. at a rate of $0.843 - 0.394 = 0.449$ mm/year. This is the answer to the research question.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Random effects

	component	variance	std. dev.	LR chi2	degrees of fı	p-value
Patient	intercept	6,799717	2,607627	2,043173	2	0,256457
	Age	0,055702	0,236013	2,093144	2	0,24955
Residual		1,546741	1,24368			

LR test: model v

Random effects correlations

	intercept	Age
intercept	1	-0,780067
Age	-0,780067	1

Random effects: the standard deviation of the intercept is 2.61 mm and of the slope 0.24 mm/year, but the correlation between them is -0.78. This does not mean that "larger" children grow more slowly – the intercept refers to age 0, far outside the data range (8–14 years), and with the zero placed there the correlation is an artefact of the geometry of the lines. The LR tests for both components are not significant ($p = 0.256$ and $p = 0.250$): with 27 children and 2–4 measurements per child, individual growth rates cannot be reliably distinguished, and a model with only a random intercept would be just as good.

Centring age. Let us repeat the analysis, replacing Age with the column Age_c (age -11) among the X variables, in the interaction and in the random-slopes list. The model fit ($-2LL$, AIC) and the coefficients of age and of the interaction remain identical, but the meaning of the intercept and of sex changes: the intercept 24.91 mm is the distance of an average boy at age 11, sex $b = -2.22$ ($p = 0.006$) is the difference of girls at age 11 (they are 2.2 mm smaller), and the random-effects correlation changes from -0.78 to $+0.33$. General conclusion: when the intercept is to be interpretable and the random-effects correlation meaningful, the time (age) variable should be centred in the middle of the observed range.

interaction Age×Sex: $b = -0.39$ mm/year, $p = 0.019$

Girls grow significantly more slowly than boys in this dimension (0.45 vs 0.84 mm/year); at age 11 they are 2.2 mm smaller.

EXAMPLE 26.3. (centresPatients.pqs file)
Simulated data: in 8 centres, 6 patients each had 5 measurements over time ($t = 0, 1, \ldots, 4$) of some index Y, with a continuous variable X also recorded. Patients are nested within centres, so two sources of dependence overlap: measurements of the same patient are alike, and patients of the same centre are alike too.

We indicate Y, the grouping variable Patient, the independent variables t and X, and in the field Level 2 – superordinate group we choose Centre. The grouping variable must be unique across the whole sheet (patient 1 in centre 1 and patient 1 in centre 2 are different people, so they must have different identifiers).

Copyright ©2010-2026 U-smile Stats – All rights reserved

ICC	0,620058
Level-2 ICC (Centre)	0,429007
-95% CI	NA
+95% CI	NA

Fixed effects (conditional)

	b coeff.	b error	-95% CI	+95% CI	t stat.	df	p-value
intercept	10,753816	1,26561	7,775658	13,731973	8,496944	7,172563	0,000054
t	1,367874	0,069479	1,230825	1,504923	19,68763	190,021452	<0,000001
X	-2,070848	0,11767	-2,302893	-1,838803	-17,59876	198,300312	<0,000001

Random effects

	component	variance	std. dev.	LR chi2	degrees of fi	p-value
Patient	intercept	5,333983	2,309542			
	intercept lev	11,692803	3,419474			
Residual		2,316493	1,522003			

The report shows three variances: patients (5.33), centres (11.69) and residuals (2.32). The null model gives a total ICC of 0.620 (62% of the variability consists of differences between patients and centres together) and a level-2 ICC of 0.429 (43% of the variability consists of differences between centres alone). Fixed effects: time $b = 1.37$ ($p < 0.000001$), X: $b = -2.07$ ($p < 0.000001$). The degrees of freedom for the intercept are only 7.2 – it is effectively 8 centres that inform us about the overall level, which is why its confidence interval is wide (7.78–13.73) despite 240 observations. A separate level-2 BLUP table identifies centres with an atypical level of Y (e.g. centre 1: -3.86).

EXAMPLE 26.4. continuation of example 26.1 (sleepDeprivation.pqs file)
The reaction-time measurements were taken day after day, so the residuals of neighbouring days may be correlated (a bad day drags the next one along). We check this with a model with a random intercept only (no random slope) and the $\mathrm{AR}(1)$ residual structure; the rows of each subject are ordered by day in the sheet, which is a requirement of this option.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Residual structure within group	AR(1), rho = 0,653133
Log Likelihood (REML)	-868,095532
-2LL	1736,191064
Number of parameters	5
AIC - Akaike criterion	1746,191064
AICc - corrected Akaike criterion	1746,535891
BIC - Bayesian criterion	1762,155848
Marginal R2 (fixed effects)	0,272839
Conditional R2 (fixed + random)	0,574758
Residual variance	1416,569316
Residual std. dev.	37,637339

Random effects

	component	variance	std. dev.	LR chi2	degrees of fi	p-value
Subject	intercept	1005,752667	31,713604	3,283727	1	0,034985
Residual		1416,569316	37,637339	49,248553	1	<0,000001

The parameter $\rho = 0.653$: the residuals of two neighbouring days are correlated at 0.65, of days two apart – $0.65^2 \approx 0.43$, and so on. The AR(1) model with a random intercept has $-2LL = 1736.19$ and AIC $= 1746.19$, whereas the same model without a residual structure has $-2LL = 1786.47$ and AIC $= 1794.47$ – an enormous improvement. Interestingly, the AIC of the AR(1) model is also lower than that of the random-slope model of example 26.1 (1755.63; the comparison is allowed because both models have the same fixed effects and are fitted by REML). Both models describe the same phenomenon in a different language: the random slope says "each person has their own constant rate", AR(1) says "deviations from the line persist for several days". The effect of day remains almost the same ($b = 10.47$, error 1.34).

FOR THOSE INTERESTED

Estimation: The parameters are estimated by maximum likelihood (ML) or restricted maximum likelihood (REML; Laird and Ware 1982[193], Bates et al. 2015[189]). The covariance of a group's random effects is parameterized as $\sigma^2 \Lambda\Lambda^T$, where Λ is a lower-triangular matrix (diagonal for the independent covariance). For a fixed Λ the fixed effects and σ^2 have a closed form (generalized least squares), and the profiled deviance

$$d(\Lambda) = \sum_j \log\det A_j + (n-p)\left[1 + \log\frac{2\pi r^2}{n-p}\right] + \log\det(X^T W X)$$

(for REML; for ML without the last term and with n instead of $n-p$) is minimized numerically by the Nelder-Mead method. $A_j = I + \Lambda^T Z_j^T Z_j \Lambda$, W is the inverse of the marginal covariance matrix, and r^2 is the weighted residual sum of squares. The computations are carried out blockwise, group by group. With an AR(1) or CS residual structure, the products $Z_j^T Z_j$ and the sums of squares are computed with weights R_j^{-1}, where R_j is the residual correlation matrix of the group ($\rho^{|i-i'|}$ for AR(1), a constant ρ off the diagonal for CS), and ρ is an additional optimization parameter. A second level of nesting adds a random intercept of the superordinate group with variance $\tau^2_{(2)}$.

Satterthwaite degrees of freedom: for the j-th fixed effect $df_j = 2\,\mathrm{Var}(\hat\beta_j)^2 / \mathrm{Var}[\mathrm{Var}(\hat\beta_j)]$, where the variance of the variance is computed by the delta method from the Hessian of the deviance with respect

Copyright ©2010-2026 U-smile Stats – All rights reserved

to the variance parameters (Kuznetsova et al. 2017[192]).

ICC and its confidence interval: $ICC = \tau^2/(\tau^2 + \sigma^2)$ from the null model; the confidence interval by the delta method from the Hessian of the deviance with respect to (τ^2, σ^2).

Nakagawa's R^2: $R^2_{marg} = \sigma^2_f/(\sigma^2_f + \sigma^2_u + \sigma^2)$ and $R^2_{cond} = (\sigma^2_f + \sigma^2_u)/(\sigma^2_f + \sigma^2_u + \sigma^2)$, where σ^2_f is the variance of the fixed-effects prediction and σ^2_u the mean random-effects variance (for random slopes averaged over the values of X).

LR test on the boundary: the statistic $2(\ell_1 - \ell_0)$ for the hypothesis of zero variance asymptotically follows a 50:50 mixture of χ^2_{q-1} and χ^2_q (Self and Liang 1987[198]); for one parameter the p-value from χ^2_1 is halved.

Group random effects: $\hat{u}_j = \Lambda A_j^{-1} \Lambda^T Z_j^T (y_j - X_j \hat{\beta})$ with conditional covariance $\sigma^2 \Lambda A_j^{-1} \Lambda^T$ (best linear unbiased predictors, BLUP).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Validation and prediction for mixed multiple regression

WHEN TO USE:

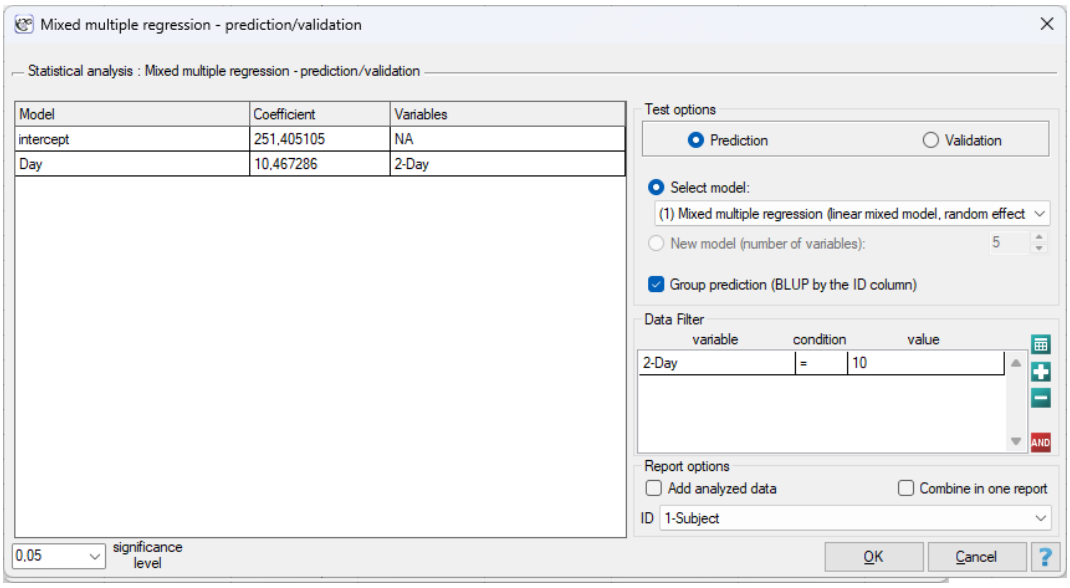
A built and verified mixed model can be used for **prediction** of Y for new data and for **validation** on a test set, as in validation and prediction for multiple regression. A mixed model, however, allows two kinds of prediction:

- **population prediction** – from the fixed effects alone (random effects equal to 0); it answers the question "what value do we expect for an average, new group (patient, centre)";
- **group prediction** – from the fixed effects and the random effect (BLUP) of a specific group that took part in building the model; it answers the question "what value do we expect for this patient (in this centre)". It is more accurate, but available only for groups known to the model.

For new observations whose identifier did not occur in the training data, the program always uses population prediction.

HOW TO SET UP THE ANALYSIS

Statistics → Mixed models → Mixed multiple regression – prediction/validation.



Prediction requires a **model saved in the program** (a mixed multiple regression report assigned to the sheet), because besides the coefficients the group random effects and the covariance matrix are needed; coefficients cannot be entered by hand. After the model is chosen, the program fills in the list of variables, and we indicate the sheet columns with the new values of X. Options:

- Group prediction (BLUP by the ID column) – after ticking it we indicate the column with the group identifier; for identifiers known to the model the prediction includes the group's BLUP, for the others it is population prediction. Without this option all predictions are population predictions;

Copyright ©2010-2026 U-smile Stats – All rights reserved

– validation – when the new data contain an observed Y, we indicate it, and the report contains measures of the model's fit to the test set.

The data for prediction are placed in the same sheet as the training data (e.g. in subsequent rows), and the analysis is restricted with a filter.

RESULTS AND INTERPRETATION

The report states which kind of prediction was used (population or for groups), the number of observations for which the group BLUP was used, and for each observation the predicted Y with its confidence interval. The interval of the population prediction is wider than that for a known group, because it includes the uncertainty about the level of the new group. In validation mode the prediction errors and fit measures are additionally computed.

Example 26.5. continuation of example 26.1 (sleepDeprivation.pqs file)
Based on the model with a random intercept and a random slope of day, we want to predict the reaction time on the tenth day of deprivation (day $= 10$, i.e. one day beyond the study range) for subject 308, for subject 309 and for a new, unknown person. We add three rows to the sheet with the value 10 in the Day column and the identifiers 308, 309 and 999, then run the prediction with a filter on these rows and the Group prediction option on.

Prediction	
Mean Pred. Y	345,295525
Standard Deviation Pred. Y	110,816079
Minimum Pred. Y	229,482352
Maximum Pred. Y	450,326257

Coefficients of model	
intercept	251,405105
Day	10,467286

Prediction

	Pred. Y	-95% CI (of	+95% CI (ol	-95% CI (of	+95% CI (ol
308 (BLUP)	450,326257	390,035301	510,617213	371,678359	528,974155
309 (BLUP)	229,482352	169,191396	289,773308	150,834454	308,13025
999	356,077965	324,473111	387,682819	296,501466	415,654464

For the new person (999) the prediction is the population one: $251.41 + 10.47 \cdot 10 \approx 356$ ms. For subject 308 the program adds their random effects (intercept $+2.26$, slope $+9.20$): $(251.41 + 2.26) + (10.47 + 9.20) \cdot 10 \approx 450$ ms, and for subject 309 (effects -40.40 and -8.62): $(251.41 - 40.40) + (10.47 - 8.62) \cdot 10 \approx 230$ ms. The difference between subjects 308 and 309 reaches 220 ms, although at the population level both "should" have 356 ms – this illustrates how much information the random effects carry when the group is known. The confidence interval of the population prediction is at the same time the widest, because it includes the scatter between people.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Comparing mixed multiple regression models

WHEN TO USE:

Model comparison serves to choose the set of independent variables (fixed effects) that describes the data best with as few parameters as possible. It works like comparing multiple regression models: we compare a full model with a reduced model (without some of the variables), or let the program select the model automatically (backward, forward, stepwise or subset selection).

In mixed models one additional rule applies: **the random structure is shared by all compared models**. Selection concerns the fixed effects only, and the random intercept, random slopes and covariance type set in the window remain in every model. A variable that has been given a random slope stays in every compared model as a fixed effect (one cannot remove the fixed effect of a variable while keeping its random slope). Whether the random effects themselves are needed is decided by the LR tests in the mixed multiple regression report.

DEFINITIONS OF MEASURES

Models are compared with the information criteria AIC, AICc and BIC (the smaller, the better) and, for nested models, with the **likelihood ratio test** (LR): the statistic $2(\ell_{full} - \ell_{reduced})$ follows a χ^2 distribution with degrees of freedom equal to the number of removed variables. A non-significant result means that the removed variables carry no information and the simpler model should be chosen; a significant result – that the full model is significantly better.

Note! Information criteria and the LR test for models differing in fixed effects are valid only under maximum likelihood (ML) estimation. Therefore **model selection is always carried out by ML**, regardless of the Model coefficients setting. That setting decides only which method (REML or ML) is used to compute the coefficient tables of the presented models – REML is recommended, because it gives unbiased variances.

Out-of-sample evaluation of models

The information criteria and the LR test are determined on the same data on which the models were fitted, which makes the assessment of the models too optimistic. As in comparing multiple regression models, the analysis window therefore makes it possible to evaluate the models on observations that did not take part in fitting them (the Validation option): k**-fold cross-validation** or a **random test set**, with the split determined by the Seed. For each model the report then gives RMSE, MAE and the predictive R^2, and in the model selection block – the error of the null model and the best model by validation (the one with the smallest RMSE); validation error can also be chosen as the automatic selection criterion. In mixed models two additional rules apply:

- **whole groups are split** – all the observations of one group (e.g. one patient) go to the same part of the data. Observations of the same group are correlated with each other, so dividing them between the training part and the evaluated part would understate the error. The number of folds k is limited to the number of groups, and the share of the test set means the percentage of groups;
- **the prediction for a new group is based on the fixed effects only** – the random effect of a group that was not in the training set is unknown, so its expected value, i.e. zero, is assumed. The validation measures therefore tell how well the fixed effects of the model predict the observations

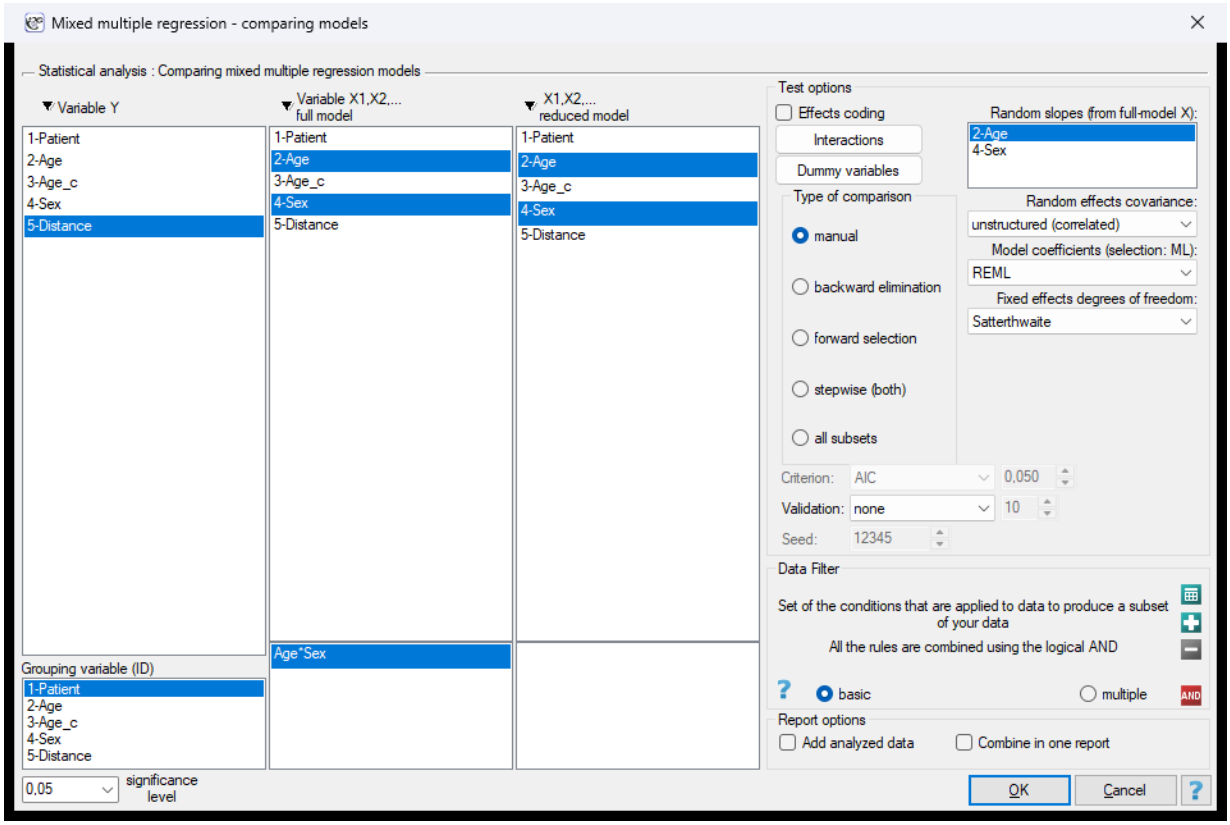
Copyright ©2010-2026 U-smile Stats – All rights reserved

of a new group (e.g. a new patient), and they correspond to the marginal, not the conditional, R^2. RMSE also includes the variability between groups, so it is larger than the residual standard deviation of the model.

The null model here is the model with the intercept, the shared random structure and the variables that were given a random slope (they stay in every model); it is fitted and evaluated in the same way as the other models.

HOW TO SET UP THE ANALYSIS

Statistics → Mixed models → Mixed multiple regression – comparing models.



In the window we indicate the dependent variable Y, the variables of the full model and of the reduced model (or the automatic selection criterion), as in comparing multiple regression models. Mixed-model options:

- Grouping variable (ID) – the column identifying the group; shared by all models,
- Random slopes (from full-model X) – variables whose effect differs between groups; such a variable stays in every compared model,
- Random effects covariance – independent or unstructured (correlated), shared by all models,
- Model coefficients (selection: ML) – REML or ML for the coefficient tables; the selection itself is always ML,

Copyright ©2010-2026 U-smile Stats – All rights reserved

- Fixed effects degrees of freedom – Satterthwaite or Wald z test in the coefficient tables,
- Criterion – the automatic selection criterion: AIC, AICc, BIC, p (LR / F test) with a threshold, or validation error,
- Validation – none, k-fold CV, k = with the number of folds (from 2 to 100, 10 by default) or test set, % = with the share of the test set as a percentage of the groups (from 5 to 90, 30 by default); whole groups are split,
- Seed – the seed of the random split (12345 by default); the same seed means the same split.

RESULTS AND INTERPRETATION

The report contains a frame describing the shared random structure (grouping variable, number of groups, random components, covariance), the table **Information criteria and LR tests** with $-2LL$ (ML), AIC, AICc, BIC, marginal and conditional R^2 of each model and the LR test between successive models, and the coefficient tables of the compared models (conditional fixed effects, with the random-effect variances of each model). If a singular fit occurred for any model, it is marked in the table. When validation is switched on, the report additionally contains the **Out-of-sample evaluation** block (the method of splitting, the seed, the split unit, the error of the null model, the best model by validation), and for each model – RMSE, MAE and the predictive R^2.

Example 26.6. continuation of example 26.2 (orthodontics.pqs file)
The growth model of the pituitary–pterygomaxillary fissure distance contained: age, sex and the interaction age×sex. We check whether the interaction (different growth rates of girls and boys) is really needed, comparing the full model with a reduced model without the interaction. Random structure: random intercept and random slope of age, unstructured covariance, shared by both models.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Model selection	manual (selected models)
Number of variables in the model 1	3
Convergence criterion met	
-2 Log Likelihood (ML)	348,704807
AIC - Akaike criterion	364,704807
AICc - corrected Akaike criterion	366,527592
BIC - Bayesian criterion	384,523501
Marginal R2 (fixed effects)	0,465299
Conditional R2 (fixed + random)	0,818385
Variance: intercept	6,799717
Variance: Age	0,055702
Variance: residual	1,546741
Number of variables in the model 2	2
Convergence criterion met	
-2 Log Likelihood (ML)	354,986352
AIC - Akaike criterion	368,986352
AICc - corrected Akaike criterion	370,386352
BIC - Bayesian criterion	386,32771
Marginal R2 (fixed effects)	0,42409
Conditional R2 (fixed + random)	0,811562
Variance: intercept	9,855281
Variance: Age	0,078336
Variance: residual	1,568481
Chi-square - models comparison Model 1 vs Model 2	6,281545
Degrees of freedom	1
p-value	0,0122

The full model has $-2LL = 348.70$ (ML), AIC $= 364.70$, BIC $= 384.52$; the model without the interaction $-2LL = 354.99$, AIC $= 368.99$, BIC $= 386.33$. LR test: $\chi^2 = 354.99 - 348.70 = 6.28$, $df = 1$, $p = 0.012$ – removing the interaction significantly worsens the fit, and both information criteria point to the full model. We keep the model with the interaction: girls and boys grow at different rates. If we additionally compared the model with age only ($-2LL = 360.72$, AIC $= 372.72$), it would turn out that sex carries information only together with the interaction – in the model without the interaction its effect is the average difference over the whole age range ($b = -2.04$, $p = 0.008$), which hides the fact that this difference grows with age.

LR test full vs without interaction: $\chi^2 = 6.28$, $df = 1$, $p = 0.012$; **AIC** 364.70 **vs** 368.99

The age×sex interaction stays in the model.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Out-of-sample evaluation. We will also check whether the interaction improves the prediction of growth in children who did not take part in fitting the model. We repeat the same comparison, choosing in the Validation field the k-fold CV, k = option with 10 folds (seed 12345). The 27 patients are split, not the 88 measurements – all the measurements of one child go to the same part.

Out-of-sample evaluation	cross-validation
Number of folds (k)	10
Seed of the random split	12345
Split unit	whole group
Null model: RMSE	2,463707
Null model: MAE	1,842072
Best model by validation (RMSE)	Model 1
New-group prediction: fixed effects only	

	training set	cross-validation ($k = 10$)		
model	AIC	RMSE	MAE	predictive R^2
null (age only)		2.46	1.84	
Model 1 (with interaction)	364.70	2.27	1.80	0.41
Model 2 (without interaction)	368.99	2.31	1.80	0.38

The model with the interaction has a smaller out-of-sample error (RMSE $= 2.27$ mm versus 2.31 mm) and a larger predictive R^2 (0.41 versus 0.38); both models predict better than the null model, which – because of the random slope of age – contains age only (RMSE $= 2.46$ mm). Validation therefore confirms the result of the LR test and of the information criteria, although the advantage of the model with the interaction is small. It is worth noting that the predictive R^2 (0.41) is close to the marginal R^2 of the model (0.47), not to the conditional one (0.82): for a new child we do not know its random effect, so the prediction is based on the fixed effects alone.

cross-validation ($k = 10$, whole groups): RMSE $= 2.27$ (with interaction), 2.31 (without interaction), 2.46 (null model)

The best model by validation is the model with the age×sex interaction.

Copyright ©2010-2026 U-smile Stats – All rights reserved

26.2 MIXED LOGISTIC REGRESSION

Mixed logistic regression (generalized linear mixed model)

WHEN TO USE:

Mixed logistic regression (generalized linear mixed model, GLMM) extends logistic regression to data in which observations form groups: repeated assessments of the same patient (e.g. presence of a symptom at successive visits), patients from the same centre, animals from the same herd, pupils from one class (see the introduction to mixed models). The binary (0/1) outcomes of observations from one group are correlated, and ordinary logistic regression, which ignores this correlation, underestimates standard errors and overstates significance.

The model describes the logit of the probability of the event:

$$\ln \frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)} = \beta_0 + \beta_1 X_{1ij} + \ldots + \beta_k X_{kij} + u_{0j} + u_{1j} X_{1ij},$$

where:
Y_{ij} – the outcome of the i-th observation in group j (0 or 1),
$\beta_0, \beta_1, \ldots, \beta_k$ – fixed effects,
u_{0j} – the random intercept of group j, u_{1j} – the random slope of group j; the random effects are assumed normal with mean 0 and covariance matrix Σ on the logit scale.

A random intercept means that each group has its own "baseline" odds of the event: a group with $u_{0j} = 1$ has odds $e^1 \approx 2.7$ times higher than an average group, at the same values of the variables. Unlike mixed multiple regression, the model has no residual variance: on the logit scale it is replaced by the constant $\pi^2/3 \approx 3.29$.

This model replaces Cochran's Q ANOVA when the repeated measurements are incomplete or when, besides the repeated factor, the model is to contain continuous variables and variables describing the group.

ASSUMPTIONS

- a binary dependent variable,
- random effects normally distributed on the logit scale; groups independent of one another,
- at least 3 groups, and usually a dozen or more for a reliable variance estimate; with random slopes also several observations in each group,
- a sufficient number of events (ones) and non-events (zeros) – as in logistic regression,
- missing data are random.

Note! The data must be in long format: each row is one observation (one 0/1 outcome), and a separate column indicates the group. Aggregated data (number of events and number of trials per group) must first be expanded into individual rows.

Copyright ©2010-2026 U-smile Stats – All rights reserved

DEFINITIONS OF MEASURES

Conditional odds ratio $OR = e^b$ – says how many times the odds of the event change when X increases by one unit **within the same group** (for the same patient, in the same centre). Ordinary logistic regression gives a **population** odds ratio, averaged over all groups. The conditional OR is always larger in absolute value (further from 1) than the population one, and the difference grows with the random-effect variance; with variance τ^2 the population coefficient is approximately $b/\sqrt{1 + 0.346\,\tau^2}$. Both are correct, but they answer different questions and should not be compared directly.

Random-effect variance τ^2 on the logit scale and its square root – the standard deviation. A random-intercept standard deviation of 1 means that a group one deviation above average has odds of the event $e^1 \approx 2.7$ times higher than an average group, and the extreme groups (± 2 deviations) differ in odds by $e^4 \approx 55$ times.

ICC on the logit scale – from the null model:

$$ICC = \frac{\tau^2}{\tau^2 + \pi^2/3},$$

interpreted as the correlation of the latent propensities to the event within a group (the share of between-group variability in the variability of the latent continuous variable behind the 0/1 outcome).

Marginal and conditional R^2 (Nakagawa) – on the logit scale, with $\pi^2/3$ as the residual variance: the share of variance explained by the fixed effects, and by the fixed and random effects together.

Likelihood approximation – the likelihood of the model requires integration over the random effects, which cannot be done exactly. Laplace approximates the integral by a single function around its maximum – fast and sufficiently accurate with larger groups. Gauss-Hermite quadrature (adaptive, 7, 15 or 25 nodes) computes the integral numerically at several points and is more accurate, especially with small groups (a few observations per group) and a large variance; the cost is computation time. Quadrature is available only for a model with a random intercept alone; with random slopes the program uses the Laplace approximation. The results of the different approximations usually differ in the third significant digit.

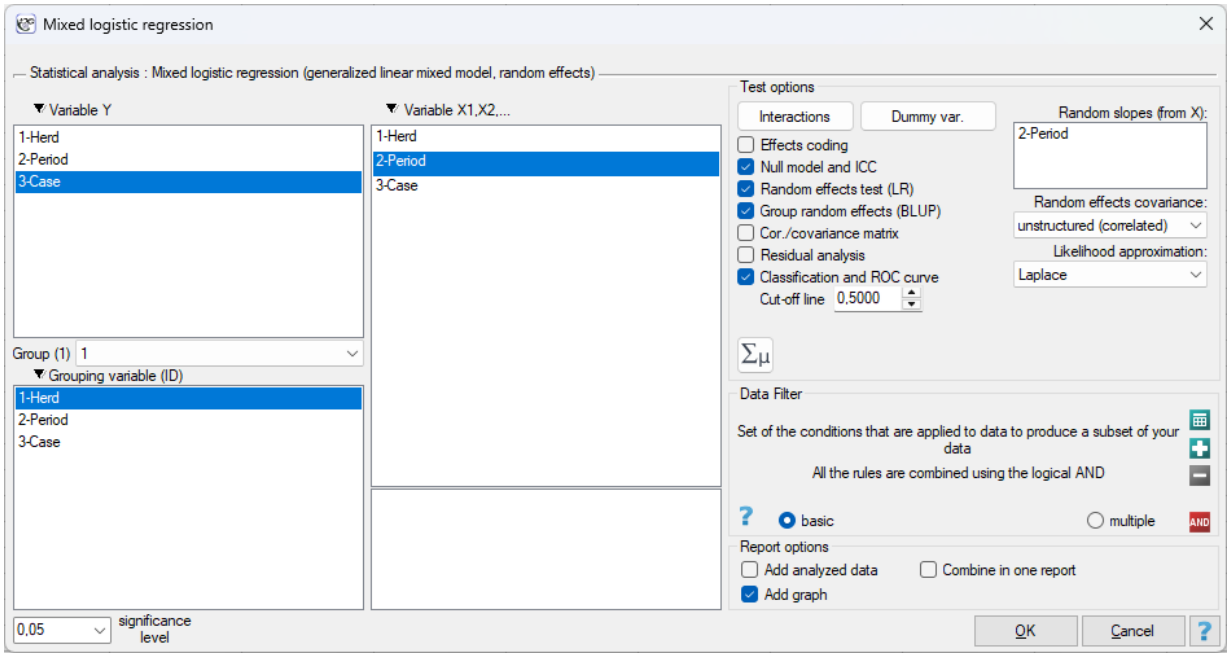
LR test of random effects, BLUP, singular fit – as in mixed multiple regression; the BLUPs are group deviations on the logit scale.

Conditional and population prediction – the predicted probability with the group's random effects ("for this patient") and from the fixed effects alone ("for the average patient"). Classification and the ROC curve can be computed for both; the conditional prediction is always more accurate, because it uses information about the group, but it cannot be used for new, unknown groups.

HOW TO SET UP THE ANALYSIS

Statistics → Mixed models → Mixed logistic regression.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In the window we indicate the Dependent variable Y and, in the Group (1) field, the value denoting the event, the Grouping variable (ID) and the Independent variables X1, X2, ...; dummy variables and interactions are prepared as in logistic regression. Mixed-model options:

- Random slopes (from X) – variables among X whose effect is to differ between groups; they must vary within groups. The random intercept is always in the model. Random slopes in a logistic model require more observations per group than in a linear model, because each observation carries only one bit of information,
- Random effects covariance – unstructured (correlated) or independent; the latter is more stable with few groups,
- Likelihood approximation – Laplace (default, always available) or G-H quadrature with 7, 15 or 25 nodes (random intercept only). Changing the approximation slightly changes all numbers in the report; for publication, with small groups, quadrature with 15–25 nodes is worth using,
- Null model and ICC, Random effects test (LR), Group random effects (BLUP), Cor./covariance matrix, Residual analysis – as in mixed multiple regression,
- Classification and ROC curve – sensitivity, specificity, fraction correct and AUC for the indicated cut-off, separately for the conditional and population prediction,
- Add graphs – the caterpillar plot of group random effects.

RESULTS AND INTERPRETATION

Counts – number of observations, count of $Y = 1$, number of groups, minimum, median and maximum of observations per group. With a median below 3 observations per group the Laplace approximation is least accurate – quadrature is then worth using.

Approximation and convergence – method (Laplace or Gauss-Hermite with the number of nodes), number of iterations, convergence, singular fit.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Model fit – log-likelihood, $-2LL$, AIC, AICc, BIC (ML estimation), marginal and conditional R^2 and – for comparison – the $-2LL$ of ordinary logistic regression (no random effects) and the $-2LL$ of the null model.

LR test of random effects (vs logistic regression) – whether the group structure is needed; the p-value is boundary-corrected.

Null model – between-group variance τ^2 and the **ICC (logit scale)** with its confidence interval.

Fixed effects (conditional) – b, error, confidence interval, the z statistic, the p-value and the **odds ratio (conditional)** with its confidence interval. The test is a Wald z-test (no degrees of freedom), as in logistic regression.

Random effects – the variance and standard deviation of each component on the logit scale, LR tests of the need for the components; the last row is the constant $\pi^2/3$. With more than one component – the random-effects correlation matrix.

Group random effects (BLUP) – deviations of the groups from the population on the logit scale with errors. A group with BLUP $= 0.5$ has odds of the event $e^{0.5} \approx 1.65$ times higher than an average group. Groups with few observations or without events have their BLUPs shrunk towards zero.

Classification and ROC curve – for the conditional prediction (with BLUP) and the population prediction: sensitivity, specificity, fraction correct, AUC with its error and confidence interval. The AUC difference between them shows how much discriminating ability knowledge of the group adds.

Residual analysis – the conditional and population predicted probability and Pearson residuals for each observation.

EXECUTION AND INTERPRETATION

Example 26.7. (herdsCBPP.pqs file)
In a study of contagious bovine pleuropneumonia (CBPP) in Ethiopia (Lesnoff et al. 2004[194]), 15 zebu herds were observed over four consecutive periods. For each animal in each period it was recorded whether a new case of the disease appeared. The data are in long format: Herd (1–15), Period (1–4) and Case (1 – new case, 0 – none); 842 observations, including 99 cases, from 26 to 96 observations per herd. We ask whether the incidence changes over successive periods of the epidemic and how much the herds differ from one another.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Herd	Period	Case
1	1	1
1	1	1
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0

Animals from one herd share conditions, contacts and immunity, so their outcomes are not independent – the herd is the natural grouping variable. We build the model: Case as Y with the value 1 in the Group (1) field, Herd as the grouping variable, Period as the independent variable. Period is a qualitative variable with four categories, so with the Dummy variables button we convert it into three dummy variables with period 1 as the reference category. We keep the random intercept only (period has no random slope), the Laplace approximation, and switch on Null model and ICC, Random effects test (LR), Group random effects (BLUP), Classification and ROC curve and Add graphs.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Size	842
Count of Y = 1 (1)	99
Number of groups	15
Observations per group (min / median / max)	26 / 61 / 96
Number of variables in the model	1
Likelihood approximation	Laplace
Number of iterations	154
Convergence	achieved
Singular fit	no
Log Likelihood	-278,760323
-2LL	557,520645
Number of parameters	3
AIC - Akaike criterion	563,520645
AICc - corrected Akaike criterion	563,549285
BIC - Bayesian criterion	577,727985
Marginal R2 (fixed effects)	0,09339
Conditional R2 (fixed + random)	0,200452
Logit-scale variance (n²/3)	3,289868
LR test of random effects (vs logistic regression)	
-2LL of logistic regression (no random effects)	573,155871
Chi-square	15,635226
degrees of freedom	1
p-value	0,000038
Null model	
-2LL of the null model	580,613947
Between-group variance	0,659343
ICC (logit scale)	0,166956
-95% CI	0,036797
+95% CI	0,297114

Step 1 – fit. 842 observations, 99 events, 15 groups (26/61/96 observations per group), convergence achieved, no singular fit. The $-2LL$ of the model is 555.00 (AIC $= 565.00$), whereas ordinary logistic regression with the same variables has $-2LL = 569.01$. LR test of random effects: $\chi^2 = 14.01$, $df = 1$, $p = 0.00009$ (boundary-corrected) – the herds differ significantly in incidence, and a model without the random effect would be invalid.

Step 2 – ICC. In the null model the between-herd variance is 0.659, hence ICC $= 0.659/(0.659+3.29) = 0.167$ (95% CI: 0.037–0.297): about 17% of the latent variability of the propensity to fall ill consists of differences between herds.

Fixed effects (conditional)

	b coeff.	b error	-95% CI	+95% CI	z stat.	p-value	odds ratio (c	-95% CI	+95% CI
intercept	-0,932359	0,30419	-1,52856	-0,336157	-3,065054	0,002176	0,393624	0,216848	0,714511
Period	-0,559401	0,123377	-0,801216	-0,317586	-4,534061	0,000006	0,571551	0,448783	0,727904

Step 3 – fixed effects. The intercept $b_0 = -1.399$ corresponds to a probability of falling ill in period 1 in an average herd of $1/(1+e^{1.399}) = 0.198$. Conditional odds ratios for the successive periods relative to

Copyright ©2010-2026 U-smile Stats – All rights reserved

period 1: period 2 – $OR = 0.371$ (95% CI: 0.203–0.676, $p = 0.0012$), period 3 – $OR = 0.324$ (0.171–0.614, $p = 0.0005$), period 4 – $OR = 0.206$ (0.089–0.476, $p = 0.0002$). Within the same herd the odds of a new case in period 2 are 63% lower than in period 1, and in period 4 – 79% lower: the epidemic is dying out. The word "conditional" reminds us that we compare periods within a herd; the odds ratio averaged over all herds (from ordinary logistic regression) would be somewhat closer to 1.

Random effects

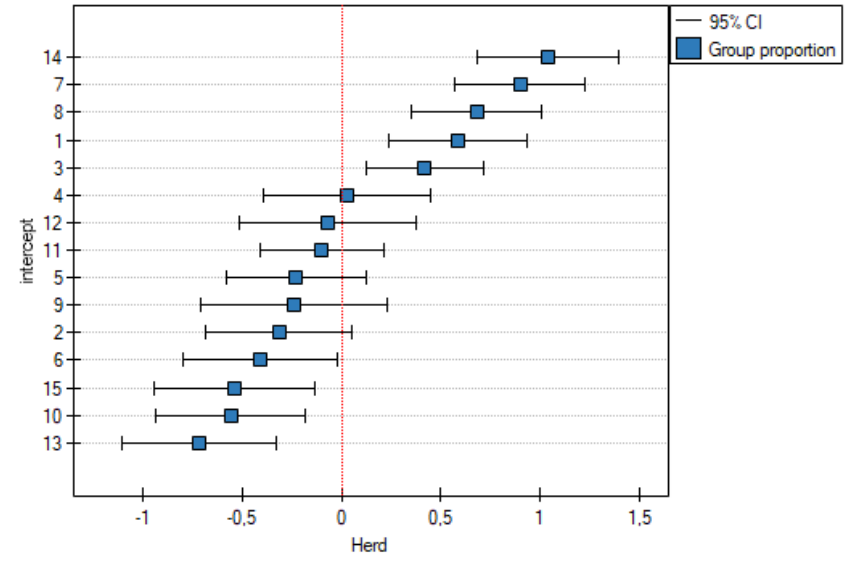
	component	variance	std. dev.	LR chi2	degrees of fi	p-value
Herd	intercept	0,440526	0,663721	15,635226	1	0,000038
Logit scale	π²/3	3,289868	1,813799			

Step 4 – random effects. The random-intercept variance is 0.413, the standard deviation 0.642 on the logit scale. A herd one deviation above average has odds of disease $e^{0.642} \approx 1.9$ times higher than an average herd, and the difference between the extreme herds (± 2 deviations) is a $e^{2.57} \approx 13$-fold difference in odds. Marginal $R^2 = 0.088$, conditional $= 0.190$: period explains less than 9% of the variability on the logit scale, and together with herd – 19%.

Group random effects (BLUP)

group	intercept	error
1	0,584387	0,348839
2	-0,315111	0,36858
3	0,420441	0,29854
4	0,029539	0,420429
5	-0,227296	0,351872
6	-0,409099	0,384834
7	0,900907	0,32688
8	0,68009	0,324263
9	-0,23751	0,470164
10	-0,557189	0,375212
11	-0,099301	0,310526
12	-0,070944	0,446874
13	-0,715275	0,390102
14	1,038322	0,358252
15	-0,539435	0,403825

Copyright ©2010-2026 U-smile Stats – All rights reserved



Step 5 – atypical herds. The BLUP of herd 1 is $+0.590$ (error 0.348): its odds of disease are $e^{0.59} \approx 1.8$ times higher than those of an average herd; in period 1 this gives a probability of $1/(1+e^{1.399-0.590}) = 0.308$ instead of 0.198. Herd 2 has BLUP -0.299 (odds 26% lower). The caterpillar plot orders the herds from the least to the most at risk; herds whose interval does not include zero clearly depart from the average. Bear in mind that herds with few animals have their BLUPs shrunk towards zero.

Classification and ROC curve (cut-off = 0,5)												
	sensitivity	-CI	+CI	specificity	-CI	+CI	fraction corr	AUC	SE(AUC)	-CI	+CI	z stat.	p
conditional pred	0	NA	0,036576	1	0,995047	NA	0,882423	0,769213	0,026922	0,716447	0,821979	8,711338	<0,000001
population predi	0	NA	0,036576	1	0,995047	NA	0,882423	0,677842	0,026898	0,625122	0,730562	5,754696	<0,000001

Step 6 – classification. The classification and ROC table shows two versions: the conditional prediction (with the herd's BLUP) and the population prediction (period only). The AUC of the conditional prediction is higher – knowing the herd improves the discrimination between animals that will and will not fall ill. The population version says how the model will perform in a new, unknown herd.

conditional OR: period 2 – 0.37, period 3 – 0.32, period 4 – 0.21 (all $p < 0.002$); herd $SD = 0.64$ (LR $p = 0.00009$), ICC $= 0.17$

The incidence decreases over the successive periods of the epidemic, and the herds differ significantly in baseline risk.

What the approximation changes. The same model with G-H quadrature, 25 nodes gives $-2LL = 554.92$, a herd variance of 0.419 instead of 0.413 and coefficients differing in the third decimal place (e.g. period 4: $b = -1.579$ instead of -1.580). With herds of 26–96 animals the Laplace approximation is fully sufficient; if the groups had 2–3 observations each, the differences would be larger and quadrature would become the method of choice.

FOR THOSE INTERESTED

Estimation: The marginal likelihood requires integration over the random effects, which has no closed form. The program uses the Laplace approximation (Bates et al. 2015[189]): for each group it finds the

Copyright ©2010-2026 U-smile Stats – All rights reserved

mode $\hat{u}_j$ of the penalized log-likelihood

$$f_j(u) = -\sum_{i \in j} \left[y_{ij}\eta_{ij} - \ln(1 + e^{\eta_{ij}}) \right] + \frac{1}{2} u^T u, \qquad \eta_{ij} = x_{ij}^T \beta + z_{ij}^T \Lambda u,$$

and approximates the deviance by $\sum_j \left[2 f_j(\hat{u}_j) + \ln\det(I + \Lambda^T Z_j^T W_j Z_j \Lambda) \right]$, where W_j is the matrix of weights $p(1-p)$ and $\Sigma = \Lambda\Lambda^T$. For a single random component the integral can be computed by adaptive Gauss-Hermite quadrature around the mode, with nodes shifted and scaled according to the mode and curvature of each group. The parameters (θ, β) are optimized jointly by the Nelder-Mead method, starting from logistic regression.

Standard errors: from the inverse of the Hessian of the deviance with respect to all parameters (θ, β), computed numerically at the optimum, as in the lme4 package.

ICC: $\tau^2/(\tau^2 + \pi^2/3)$ from the null model, with a delta-method confidence interval.

Nakagawa's R^2: as in mixed multiple regression, with $\pi^2/3$ in place of the residual variance.

Test on the boundary: the statistic $2(\ell_1 - \ell_0)$ for the hypothesis of zero variance follows a 50:50 mixture of χ^2_{q-1} and χ^2_q (Self and Liang 1987[198]); for one parameter the p-value from χ^2_1 is halved.

Population effect: with a random intercept of variance τ^2 the population (marginal) coefficient is approximately $\beta / \sqrt{1 + 0.346\,\tau^2}$ (Zeger et al. 1988[200]).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Validation and prediction for mixed logistic regression

WHEN TO USE:

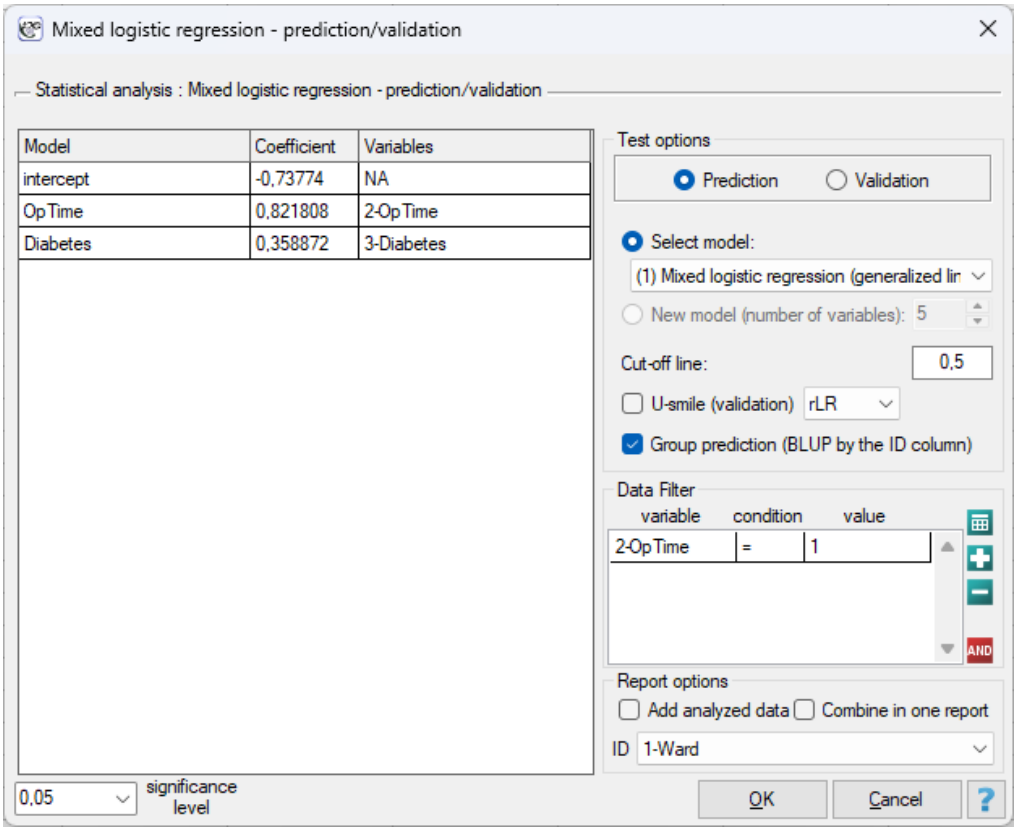
A built model can be used for **prediction** of the probability of the event for new observations and for **validation** on a test set, as in validation and prediction for logistic regression. As in mixed multiple regression, two kinds of prediction are available:

- **population** – from the fixed effects alone; the probability for an average, unknown group,
- **for a group** – with the random effect (BLUP) of a group known to the model; the probability for this specific patient, centre, herd.

Prediction of the occurrence of the event (1) or its absence (0) consists in comparing the predicted probability with the **cut-off point** (0.5 by default).

HOW TO SET UP THE ANALYSIS

Statistics → Mixed models → Mixed logistic regression – prediction/validation.



We indicate a model saved in the program (a mixed logistic regression report assigned to the sheet) – entering coefficients by hand is not possible, because the group random effects are needed. Then we indicate the columns with the new values of the variables, optionally the observed Y (validation)

Copyright ©2010-2026 U-smile Stats – All rights reserved

and the option Group prediction (BLUP by the ID column) with the identifier column. For identifiers unknown to the model the prediction is the population one.

RESULTS AND INTERPRETATION

The report states the kind of prediction, the number of observations for which the group BLUP was used, and for each observation the predicted probability and the predicted class. In validation mode additionally the classification (sensitivity, specificity, fraction correct) and the AUC on the test set.

EXAMPLE 26.8. continuation of example 26.9 (wardsInfections.pqs file)
Based on the surgical site infection model (operation time and diabetes as fixed effects, a random ward intercept and a random slope of operation time), we want to estimate the risk of infection for a diabetic patient whose operation lasted one standard deviation longer than average (OpTime $= 1$, Diabetes $= 1$), if operated in ward 1, in ward 2, or in a new ward unknown to the model.

First we build and save the model in the Mixed logistic regression window: Infection as Y (Group $(1) = 1$), Ward as the grouping variable, OpTime and Diabetes as X, OpTime in the random-slopes list, unstructured covariance, with the Group random effects (BLUP) option on. Then we add three rows to the sheet with the values 1 and 1 in the OpTime and Diabetes columns and ward identifiers 1, 2 and 99, set a filter on these rows, and in the prediction window indicate the saved model and the Group prediction (BLUP by the ID column) option with the Ward column.

Coefficients of model	
intercept	-0,73774
OpTime	0,821808
Diabetes	0,358872

Prediction				
	Pred. of prob	-95% CI	+95% CI	Pred. Y
1 (BLUP)	0,292478	0,101743	0,601388	0
2 (BLUP)	0,842557	0,631994	0,943427	1
99	0,608959	0,496041	0,711302	1

Fixed effects of the model: intercept -0.738, operation time 0.822, diabetes 0.359; BLUP of ward 1: intercept -1.145, slope -0.182; of ward 2: $+1.193$ and $+0.041$. For the new ward (99) the prediction is the population one: logit $= -0.738+0.822+0.359 = 0.443$, probability 0.61. For ward 1 the program adds its intercept and slope BLUPs (the latter multiplied by the operation time): logit $= 0.443-1.145-0.182 \cdot 1 = -0.884$, probability 0.29. For ward 2: logit $= 0.443 + 1.193 + 0.041 = 1.677$, probability 0.84. The same patient thus has a risk of infection from 29% to 84% depending on the ward, and the population prediction (61%) is merely the middle of this range. With a cut-off of 0.5 the patient in ward 1 would be classified as "no infection", in ward 2 and in the new ward – as "infection". In validation mode (with the Infection column indicated) on a test set the report would additionally give sensitivity, specificity and AUC.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Comparing mixed logistic regression models

WHEN TO USE:

Model comparison serves to choose the set of independent variables (fixed effects), as in comparing logistic regression models: we compare a full model with a reduced model, or use automatic selection (backward, forward, stepwise, subsets). **The random structure is shared by all compared models**: the random intercept, random slopes and covariance set in the window apply to every model, and a variable with a random slope stays in every model as a fixed effect. Selection concerns the fixed effects only; whether a random slope is needed is decided by the LR test in the mixed logistic regression report.

DEFINITIONS OF MEASURES

Models are compared with the information criteria AIC, AICc, BIC (the smaller, the better), pseudo-R^2 and the marginal and conditional R^2, and nested models – with the likelihood ratio test: $2(\ell_{full} - \ell_{reduced})$ follows a χ^2 distribution with degrees of freedom equal to the number of removed variables. The likelihood is always approximated by the same method (Laplace or quadrature) in all models, so it is comparable.

Out-of-sample evaluation of models

The information criteria and the LR test are determined on the same data on which the models were fitted, which makes the assessment of the models too optimistic. As in comparing logistic regression models, the analysis window therefore makes it possible to evaluate the models on observations that did not take part in fitting them (the Validation option): **k-fold cross-validation** or a **random test set**, with the split determined by the Seed. For each model the report then gives the mean deviance ($-2 \ln L/n$; this is the validation error), AUC and the Brier score, and in the model selection block – the error of the null model and the best model by validation; validation error can also be chosen as the automatic selection criterion. In mixed models two additional rules apply:

- **whole groups are split** – all the observations of one group (e.g. one ward) go to the same part of the data. Observations of the same group are correlated with each other, so dividing them between the training part and the evaluated part would understate the error. The number of folds k is limited to the number of groups, and the share of the test set means the percentage of groups (the split is not stratified by the dependent variable here);
- **the prediction for a new group is based on the fixed effects only** – the random effect of a group that was not in the training set is unknown, so its expected value, i.e. zero, is assumed. The validation measures therefore tell how well the fixed effects of the model predict the event in a new group (e.g. in a new ward).

The null model here is the model with the intercept, the shared random structure and the variables that were given a random slope (they stay in every model); it is fitted and evaluated in the same way as the other models.

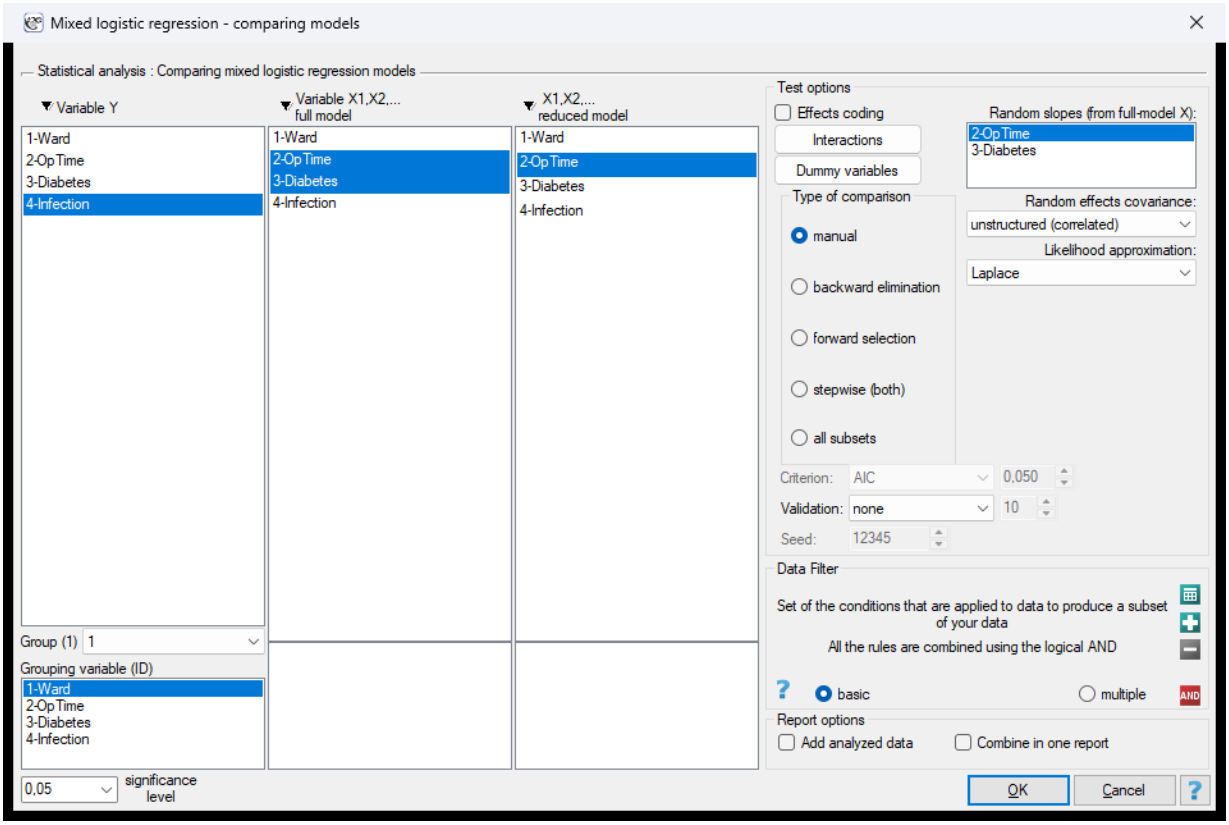
Note! The coefficients of a mixed model are conditional – they describe the effect of the variables within a group. The probability determined from the fixed effects alone refers to an "average" group (with a random effect equal to zero) and, when the variability between groups is large, it is more extreme than the frequency of the event averaged over the groups. The measures that depend on calibration (mean

Copyright ©2010-2026 U-smile Stats – All rights reserved

deviance, Brier score) are therefore less favourable than they would be if the group effect were known; however, all the models are evaluated in the same way, so their comparison remains meaningful.

HOW TO SET UP THE ANALYSIS

Statistics → Mixed models → Mixed logistic regression – comparing models.



We indicate the dependent variable Y with the event value (Group (1)), the variables of the full and reduced models (or the selection criterion) and the mixed-model options: Grouping variable (ID), Random slopes (from full-model X), Random effects covariance and Likelihood approximation – all shared by the compared models. Below the list of selection methods there are the fields: Criterion (AIC, AICc, BIC, p (LR / F test) with a threshold, or validation error), Validation (none, k-fold CV, k = with the number of folds, or test set, % = with the share of the test set as a percentage of the groups) and Seed of the random split (the same seed means the same split).

RESULTS AND INTERPRETATION

The report contains a frame describing the shared random structure, the table **Information criteria and LR tests** with $-2LL$, AIC, AICc, BIC, marginal and conditional R^2 of each model and the LR test between successive models, and the coefficient tables of the models with conditional odds ratios and the random-effect variance of each model. When validation is switched on, the report additionally contains the **Out-of-sample evaluation** block (the method of splitting, the seed, the split unit, the error of the null model, the best model by validation), and for each model – the mean deviance, AUC and the Brier score.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Example 26.9. (wardsInfections.pqs file)
Simulated data: in 40 surgical wards, 25 patients each were observed after surgery. For each patient it was recorded whether a surgical site infection occurred (Infection: 1 – yes), the standardized duration of the operation (OpTime: 0 means an average duration, 1 – one standard deviation longer) and the presence of diabetes (Diabetes: 1 – yes). In total 1000 patients, 404 infections. Patients of the same ward share staff, procedures and theatre conditions, so the ward is the grouping variable. Since the effect of operation time may depend on the ward (in some wards long operations are routine, in others a sign of complications), operation time receives a random slope. We ask whether diabetes adds anything beyond operation time.

Ward	OpTime	Diabetes	Infection
1	0,343	0	0
1	0,004	0	0
1	0,029	1	0
1	-0,393	0	0
1	-0,793	0	0
1	-0,312	0	0
1	-0,346	0	0
1	-0,305	1	0
1	-1,786	0	0
1	0,587	1	0
1	1,636	0	0
1	-0,645	1	0
1	0,619	1	0
1	0,236	0	1
1	0,847	0	0
1	-0,574	0	0
1	1,118	1	0
1	1,540	0	0

We set Infection as Y with the value 1 in the Group (1) field, Ward as the grouping variable, full model: OpTime and Diabetes, reduced model: OpTime only. In the Random slopes (from full-model X) list we select OpTime (this variable will stay in both models), covariance unstructured (correlated), approximation Laplace.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Number of variables in the model 1	2
Convergence criterion met	
-2 Log Likelihood (ML)	1121,206672
AIC - Akaike criterion	1133,206672
AICc - corrected Akaike criterion	1133,291264
BIC - Bayesian criterion	1162,653204
Marginal R2 (fixed effects)	0,125476
Conditional R2 (fixed + random)	0,395786
Variance: intercept	1,324926
Variance: OpTime	0,152284
Variance: logit scale (pi^2/3)	3,289868
Number of variables in the model 2	1
Convergence criterion met	
-2 Log Likelihood (ML)	1126,474737
AIC - Akaike criterion	1136,474737
AICc - corrected Akaike criterion	1136,535099
BIC - Bayesian criterion	1161,013513
Marginal R2 (fixed effects)	0,11888
Conditional R2 (fixed + random)	0,392954
Variance: intercept	1,330535
Variance: OpTime	0,159713
Variance: logit scale (pi^2/3)	3,289868
Chi-square - models comparison Model 1 vs Model 2	5,268065
Degrees of freedom	1
p-value	0,02172

The full model has $-2LL = 1121.21$, AIC $= 1133.21$, BIC $= 1162.65$; the reduced model $-2LL = 1126.47$, AIC $= 1136.47$, BIC $= 1161.01$. LR test: $\chi^2 = 1126.47 - 1121.21 = 5.27$, $df = 1$, $p = 0.022$ – removing diabetes significantly worsens the fit, AIC points to the full model, and BIC (which penalizes parameters more severely) to the reduced model, but the difference is small (1.6). In the coefficient table of the full model diabetes has a conditional $OR = 1.43$ (95% CI: 1.05–1.95, $p = 0.022$): for patients of the same ward, at the same operation time, diabetes increases the odds of infection by 43%. Operation time: $OR = 2.27$ (1.82–2.84) per standard deviation. We keep the full model – the LR test agrees with the Wald test, and the effect of diabetes is clinically meaningful.

Random structure. Both models have a random-intercept variance of about 1.33 (standard deviation 1.15 on the logit scale – the wards differ greatly in baseline risk) and a random-slope variance of about 0.15 (standard deviation 0.39: the effect of operation time varies between wards from about $0.82 - 0.78 = 0.04$ to $0.82 + 0.78 = 1.60$ on the logit scale). The LR test of the slope in the mixed logistic

Copyright ©2010-2026 U-smile Stats – All rights reserved

regression report gives $p = 0.095$, so the evidence for different slopes is weak; the model with a random intercept only has AIC $= 1133.11$, practically equal to the model with the slope. The version with the independent covariance (the correlation is -0.06 anyway) has the lowest AIC $= 1131.24$ and would be the most parsimonious choice.

LR test full vs without diabetes: $\chi^2 = 5.27$, $df = 1$, $p = 0.022$; AIC 1133.21 vs 1136.47

Diabetes stays in the model ($OR = 1.43$); operation time is the strongest factor ($OR = 2.27$ per standard deviation).

Out-of-sample evaluation. We will also check whether diabetes improves the prediction of infections in wards that did not take part in fitting the model. We repeat the same comparison, choosing in the Validation field the k-fold CV, k = option with 10 folds (seed 12345). The 40 wards are split – in each of the 10 steps the models are fitted on 36 wards and evaluated on the remaining 4.

Out-of-sample evaluation	cross-validation
Number of folds (k)	10
Seed of the random split	12345
Split unit	whole group
Null model: mean deviance	1,28897
Null model: Brier score	0,224097
Best model by validation (mean deviance)	Model 1
New-group prediction: fixed effects only	

	training set	cross-validation ($k = 10$)		
model	AIC	mean deviance	AUC	Brier
Model 1 (operation time, diabetes)	1133.21	1.2843	0.6677	0.2227
Model 2 (operation time)	1136.47	1.2890	0.6657	0.2241

Operation time has a random slope, so it stays in every model – the null model is here the same model as Model 2, hence the identical values in the report. The model with diabetes has a lower mean deviance (1.2843 versus 1.2890), a slightly larger AUC (0.6677 versus 0.6657) and a lower Brier score (0.2227 versus 0.2241), so it is indicated as the best by validation. The improvement is small, however: diabetes is a statistically significant factor, but it improves the prediction of infections in a new ward only modestly – the risk is determined mainly by operation time and by the ward itself, whose random effect is unknown for a new ward.

cross-validation ($k = 10$, whole wards): mean deviance 1.2843 (with diabetes) and 1.2890 (without diabetes); AUC 0.668 and 0.666

The best model by validation is the full model, although the gain from adding diabetes in predictions for new wards is small.

Copyright ©2010-2026 U-smile Stats – All rights reserved

26.3 COX PH REGRESSION WITH RANDOM EFFECT

Cox PH regression with random effect (frailty)

WHEN TO USE:

The Cox model with random effect extends Cox proportional hazards regression to data in which observations form groups: patients treated in the same centre, animals from one litter, recurrent events in the same patient (successive infections, successive hospitalizations), the eyes or kidneys of the same patient (see the introduction to mixed models). Survival times within a group are related, because the group shares unobserved risk factors: a patient has their own immunity, a centre – its own quality of care. Ordinary Cox regression ignores this dependence, which underestimates the standard errors of the coefficients and overstates significance.

The random effect, called *frailty* in survival analysis, is an unobserved factor w_j shared by the whole group j, which multiplies the hazard of each of its observations:

$$h_{ij}(t) = w_j \, h_0(t) \, \exp(\beta_1 X_{1ij} + \ldots + \beta_k X_{kij}),$$

where:
$h_{ij}(t)$ – the hazard of the i-th observation in group j,
$h_0(t)$ – the baseline hazard,
$\beta_1, \ldots, \beta_k$ – coefficients (fixed effects),
w_j – the frailty of group j; a gamma distribution with mean 1 and variance θ is assumed (shared gamma frailty).

Groups with $w_j > 1$ have a higher risk than their variables imply, groups with $w_j < 1$ – a lower one. The variance θ says how much the groups differ in risk; $\theta = 0$ means no within-group dependence, and the model reduces to ordinary Cox regression. Frailty is the counterpart of the random intercept of mixed multiple regression, except that it acts multiplicatively on the hazard; random slopes are not available in this model.

ASSUMPTIONS

- proportional hazards, as in Cox regression,
- groups independent of one another; dependence occurs only within groups,
- at least 3 groups, and usually a dozen or more for a reliable estimate of θ,
- a sufficient number of events (complete observations); groups without events carry little information about θ,
- no strong collinearity of the independent variables.

Note! The data must be prepared in long format: each row of the sheet is one observation (time, status, independent variables), and a separate column indicates the group (e.g. centre or patient ID). For recurrent events each episode is a separate row with the time counted from the start of that episode.

Copyright ©2010-2026 U-smile Stats – All rights reserved

DEFINITIONS OF MEASURES

Frailty variance θ – a measure of the differentiation of risk between groups. Frailty has mean 1, so with $\theta = 0.4$ the standard deviation of w_j is $\sqrt{0.4} \approx 0.63$: a typical group has a hazard 60% higher or lower than its variables would imply. The error of θ is approximate (from the profile likelihood) and is not meant for testing – the LR test serves that purpose.

Kendall's τ – a measure of the dependence of the survival times of two observations from the same group, following from the gamma distribution:

$$\tau = \frac{\theta}{\theta + 2}.$$

$\tau = 0$ is independence; $\tau = 0.17$ means moderate concordance: an observation from a group in which another observation survived long has itself a better chance of surviving long.

LR test: $\theta = 0$ – a likelihood ratio test comparing the frailty model with ordinary Cox regression. The hypothesis lies on the boundary of the parameter space, so the p-value from χ^2_1 is halved (boundary correction), which the report states.

Conditional hazard ratio $HR = e^b$ – how many times the hazard changes when X increases by one unit **within the same group** (at the same frailty). The population HR from ordinary Cox regression is averaged over groups and is closer to 1 in absolute value; the two should not be compared directly.

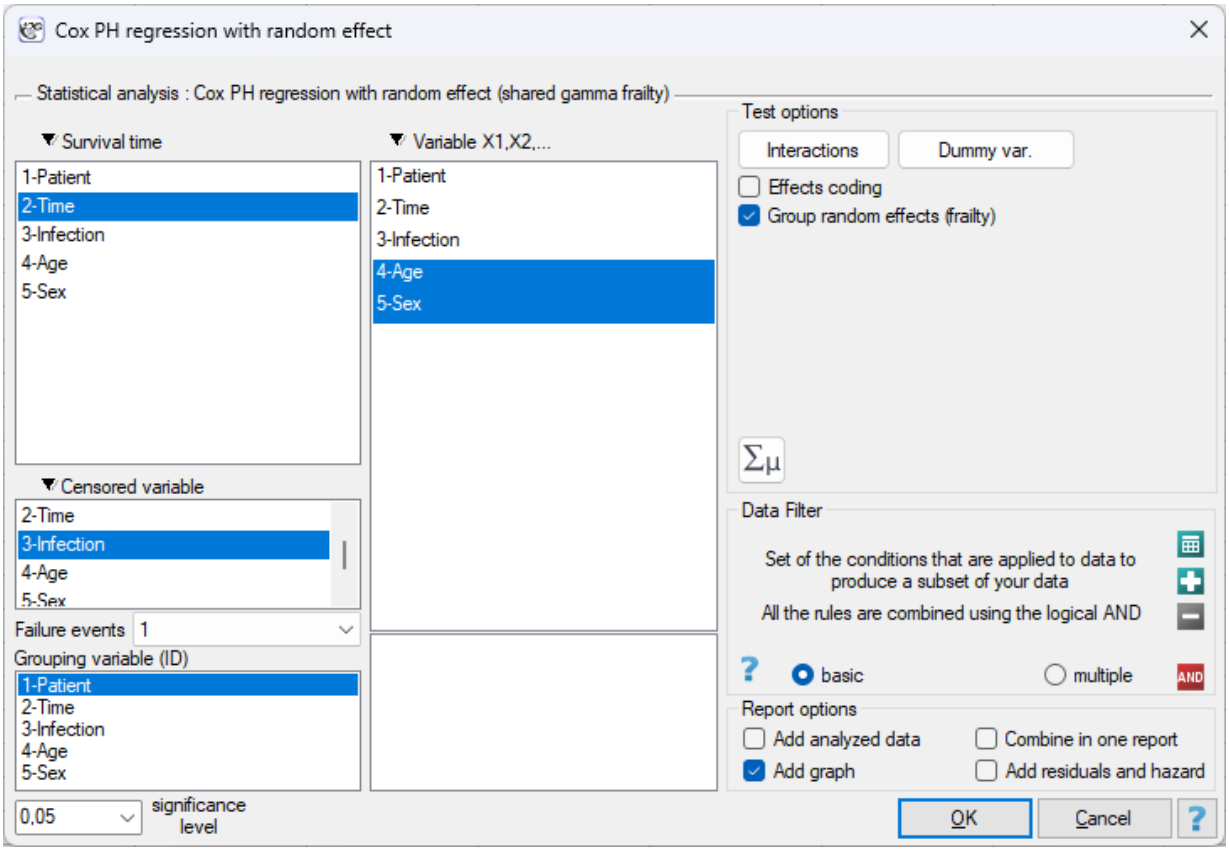
Group random effects – the estimated w_j for each group and $\log w_j$ with its error. Groups with few events have their w_j shrunk towards 1 (we do not know enough about them to call them atypical). Their significance is not tested; they serve to identify groups with elevated or reduced risk.

Integrated likelihood – the log-likelihood of the model after integrating out the frailty; the information criteria and LR tests (both the test of $\theta = 0$ and the test of the joint effect of the independent variables) are based on it.

HOW TO SET UP THE ANALYSIS

Statistics → Mixed models → Cox PH regression with random effect.

Copyright ©2010-2026 U-smile Stats – All rights reserved



The window has the layout of the Cox regression window. We indicate the Survival time, the Censoring variable with the value denoting the event in the Complete field, and the Independent variables X1, X2, ...; dummy variables and interactions are prepared exactly as in Cox regression. Options of the random-effect model:

- Grouping variable (ID) – the column indicating the group of the observation (centre, patient, litter); numeric or text. It is mandatory – without it the model does not differ from Cox regression,
- Group random effects (frailty) – adds to the report a table of the estimated w_j for each group; with hundreds of groups the table is long,
- Add residuals and hazard – martingale and deviance residuals accounting for the frailty, and the resulting hazard $w_j H_i$ of each observation; they serve to assess the fit and identify outlying observations,
- Add graphs – group survival curves (frailty).

Tied times are handled by the Breslow method, as in Cox regression, so with $\theta = 0$ the results of both analyses are identical.

RESULTS AND INTERPRETATION

Counts – the number of observations, complete (events) and censored, and the number of groups. The report also recalls the sample-size condition $n \geq 10k/p$ known from Cox regression.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Model fit – the integrated log-likelihood of the model, the same quantity for the model with the random effect only (no variables), and the likelihood ratio test of the joint effect of the independent variables. The information criteria AIC, AICc, BIC and pseudo-R^2 as in Cox regression.

Random effect (gamma frailty) – grouping variable, number of groups, frailty variance θ with an approximate error (profile), Kendall's τ and the **LR test:** $\theta = 0$ **(vs Cox regression)** with boundary correction. A significant result means that the group structure must not be ignored. The number of EM iterations is also given.

Coefficients – b, error, confidence interval, Wald statistic, p-value and the **hazard ratio (conditional)** $HR = e^b$ with its confidence interval. The standard errors come from the penalized information matrix and account for the presence of the random effect, which is why they are larger than in ordinary Cox regression.

Baseline functions – cumulative hazard and survival for an average group ($w_j = 1$): at the mean values of the variables and baseline, as in Cox regression.

Group random effects (gamma frailty) – for each group: w, log w with its error, number of events, size and the cumulative hazard of the group. A group with $w = 1.4$ has a hazard 40% higher than an average group with the same variables; a group with $w = 0.6$ – 40% lower. Groups with many events at a small expected cumulative hazard receive $w > 1$, groups without events despite long observation – $w < 1$.

Graphs – Group survival curves (frailty): the survival curve for an average group ($w = 1$) and for selected groups with their frailty; the higher w, the lower the curve lies.

EXECUTION AND INTERPRETATION

Example 26.10. (kidneyCatheter.pqs file)
In the study of McGilchrist and Aisbett (1991)[195], 38 dialysis patients with an inserted catheter were observed. The time (in days) to infection at the catheter insertion site was recorded; after the infection was cured the catheter was reinserted, so each patient has two episodes. The catheter was sometimes removed for other reasons – such observations are censored. Data in long format: Patient (1–38), Time (days), Infection (1 – infection, 0 – censored observation), Age (years) and Sex (0 – male, 1 – female). In total 76 observations, 58 infections. We ask whether age and sex affect the risk of infection and whether patients differ in their individual susceptibility.

Patient	Time	Infection	Age	Sex
1	8	1	28	0
1	16	1	28	0
2	23	1	48	1
2	13	0	48	1
3	22	1	32	0
3	28	1	32	0
4	447	1	31	1
4	318	1	32	1
5	30	1	10	0
5	12	1	10	0
6	24	1	16	1
6	245	1	17	1
7	7	1	51	0

Two episodes of the same patient are not independent – the patient has their own immunity, hygiene, comorbidities. If we fitted ordinary Cox regression, we would treat 76 episodes as 76 different people.

Copyright ©2010-2026 U-smile Stats – All rights reserved

We therefore build a model with a random effect: Time as the survival time, Infection as the censoring variable with the value 1 as Complete, Age and Sex as independent variables, Patient as the grouping variable. We switch on Group random effects (frailty) and Add graphs.

Model	Cox with random effect (gan
Sample size assumption n ≥ 10k/p	No
Number of variables in the model	2
Number of iteration for convergence	96
Convergence criterion met	
Frequency	76
Failure events (1)	58
Censored (0)	18
Likelihood ratio test	
Log Likelihood	-182,053359
-2 Log Likelihood	364,106718
Log Likelihood (intercept)	-187,945526
-2 Log Likelihood (intercept)	375,891051
Chi-square statistic	11,784333
Degrees of freedom	2
p-value	0,002761
AIC - Akaike criterion	370,106718
AICc - corrected Akaike criterion	370,551163
BIC - Bayesian criterion	376,288047
Pseudo R2 (McFadden)	0,03135
R2 (Nagelkerke)	0,184149
R2 (Coxa-Snella)	0,183867
Random effect (gamma frailty)	
Grouping variable	Patient
Number of groups	38

Random effect (gamma frailty)	
Grouping variable	Patient
Number of groups	38
Frailty variance (theta)	0,397184
Theta error (profile)	0,234386
Kendall's tau (within-group dependence)	0,165688
LR test: theta = 0 (vs Cox regression)	
Chi-square statistic	5,207469
p-value (boundary corrected)	0,011245
Log Likelihood (Cox without random effect)	-184,657094
EM iterations	42

Step 1 – do patients differ in susceptibility. Frailty variance $\theta = 0.397$ (error 0.234), Kendall's $\tau = 0.166$. LR test $\theta = 0$: $\chi^2 = 5.21$, $p = 0.011$ (boundary-corrected) – ordinary Cox regression is not valid for these data; patients differ significantly in their unobserved susceptibility to infection. The frailty standard deviation $\sqrt{0.397} \approx 0.63$: a typical patient has a hazard of infection about 60% higher or lower than their age and sex imply.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Step 2 – fit. The integrated log-likelihood of the model is -182.05, of the model with frailty only (no variables) -187.95; LR test of the joint effect of the variables: $\chi^2 = 11.78$, $df = 2$, $p = 0.0028$. AIC $= 370.11$.

Model

	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
Age	0,005462	0,01158	-0,017234	0,028158	0,222456	0,637175	1,005477	0,982913	1,028558
Sex	-1,55613	0,444803	-2,427928	-0,684332	12,239291	0,000468	0,210951	0,088219	0,504427

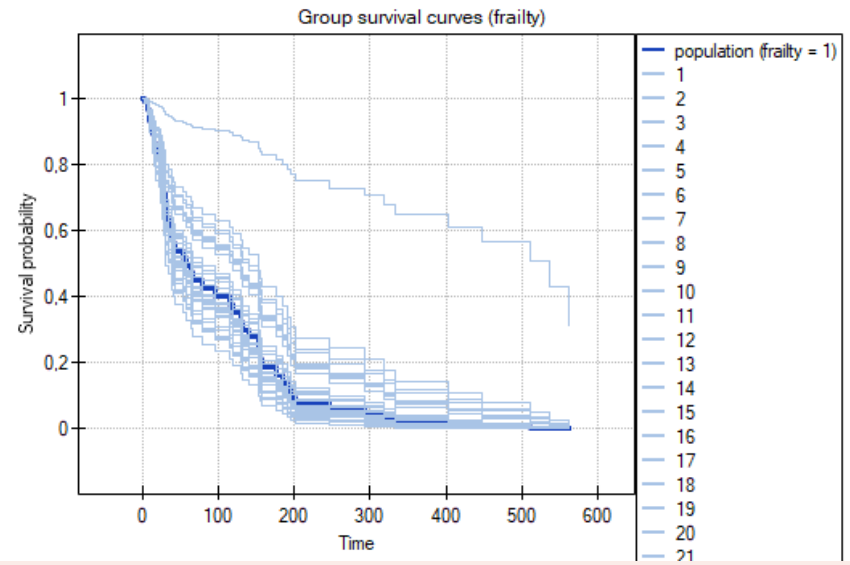
Step 3 – coefficients. Age: $b = 0.0055$, $HR = 1.006$ (95% CI: 0.983–1.029), $p = 0.637$ – age does not affect the risk of infection. Sex: $b = -1.556$ (error 0.445), $HR = 0.211$ (95% CI: 0.088–0.504), $p = 0.0005$. The hazard of infection in women is 79% lower than in men – when compared within the same "susceptibility", i.e. conditionally. The error of the sex coefficient (0.445) is larger than ordinary Cox regression would give, because the model takes into account that the information from two episodes of the same patient partly repeats itself.

Group rand

group	w	log(w)	log(w) error	events	n	cum. hazard
1	1,435732	0,361675	0,478199	2	2	0,628911
2	1,282004	0,248424	0,533957	1	2	0,226203
3	1,102199	0,097307	0,493495	2	2	1,581103
4	0,574402	-0,554426	0,51423	2	2	5,347371
5	1,145763	0,136071	0,506425	2	2	1,42526
6	1,055481	0,053997	0,497397	2	2	1,762527

Step 4 – atypical patients. The group random effects table: patient 1 has $w = 1.44$ ($\log w = 0.36$, error 0.48) – two infections at a small expected cumulative hazard (0.63) mean a susceptibility 44% above average. Patient 4 has $w = 0.57$: also two infections, but at a large cumulative hazard (5.35), i.e. after a long observation – their susceptibility is 43% below average. The errors of $\log w$ (about 0.5) are large, because each patient has only two episodes; the frailties of individual patients should be treated as indicative.

Copyright ©2010-2026 U-smile Stats – All rights reserved



The group survival curves graph shows the curve for an average patient ($w = 1$) and the curves of patients with higher and lower frailty: patient 1 loses "infection-free survival" faster, patient 4 – more slowly.

$\theta = 0.40$**, LR test** $\theta = 0$**:** $p = 0.011$**; sex:** $HR = 0.21$ **(**$p = 0.0005$**); age:** $HR = 1.01$ **(**$p = 0.64$**)**

Patients differ significantly in susceptibility to infection, and once this is accounted for, women have a 79% lower hazard of infection than men; age does not matter.

FOR THOSE INTERESTED

Estimation: The parameters are estimated by the EM algorithm on the profiled partial likelihood (Klein 1992[191], Therneau and Grambsch 2000[199]). In the M step a Cox regression with the offset $\log w_j$ is fitted and the Breslow cumulative baseline hazard $H_0(t)$ is determined; in the E step the frailty of a group is replaced by its expected value

$$w_j = \frac{1/\theta + D_j}{1/\theta + A_j}, \qquad A_j = \sum_{i \in j} H_0(t_i) \exp(\beta^T X_i),$$

where D_j is the number of events in the group and A_j its cumulative hazard. The variance θ is determined by maximizing the integrated log-likelihood

$$\ell = \sum_{i:\delta_i=1} [\eta_i - \log S_0(t_i)] + \sum_j \ell_j + D,$$

$$\ell_j = D_j \log\theta + \log\Gamma(1/\theta + D_j) - \log\Gamma(1/\theta) - (1/\theta + D_j)\log(1 + \theta A_j),$$

which as $\theta \to 0$ turns into the Cox partial likelihood (D – the total number of events). Convergence is accepted when the change of ℓ is smaller than 10^{-9}.

Standard errors: from the penalized information matrix for the parameters $(\beta, \log w_1, \ldots, \log w_G)$, where the penalty follows from the gamma distribution; the β block is inverted through the Schur complement.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Kendall's τ: for shared gamma frailty $\tau = \theta/(\theta + 2)$.

Test on the boundary: the statistic $2(\ell_{frailty} - \ell_{Cox})$ asymptotically follows a 50:50 mixture of a distribution degenerate at zero and χ^2_1 (Self and Liang 1987[198]), which is why the p-value from χ^2_1 is halved.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Comparing Cox PH regression models with random effect

WHEN TO USE:

Model comparison serves to choose the set of independent variables, as in comparing Cox PH regression models: we compare a full model with a reduced model, or use automatic selection. The random effect (gamma frailty for the indicated grouping variable) is **shared by all compared models** – in each of them the variance θ is re-estimated, but the group structure remains the same. Selection concerns the independent variables only; whether the random effect is needed is decided by the LR test $\theta = 0$ in the Cox regression with random effect report.

DEFINITIONS OF MEASURES

Models are compared on the basis of the integrated log-likelihood, the information criteria AIC, AICc, BIC (the smaller, the better) and the likelihood ratio test for nested models: $2(\ell_{full} - \ell_{reduced})$ follows a χ^2 distribution with degrees of freedom equal to the number of removed variables.

Out-of-sample evaluation of models

The information criteria and the LR test are determined on the same data on which the models were fitted, which makes the assessment of the models too optimistic. As in comparing Cox PH regression models, the analysis window therefore makes it possible to evaluate the models on observations that did not take part in fitting them (the Validation option): **k-fold cross-validation** or a **random test set**, with the split determined by the Seed. For each model the report then gives $-2 \ln PL$ per event (the validation error; the lower, the better) and Harrell's concordance index C, and in the model selection block – the error of the null model (without variables) and the best model by validation; validation error can also be chosen as the automatic selection criterion. In models with a random effect two additional rules apply:

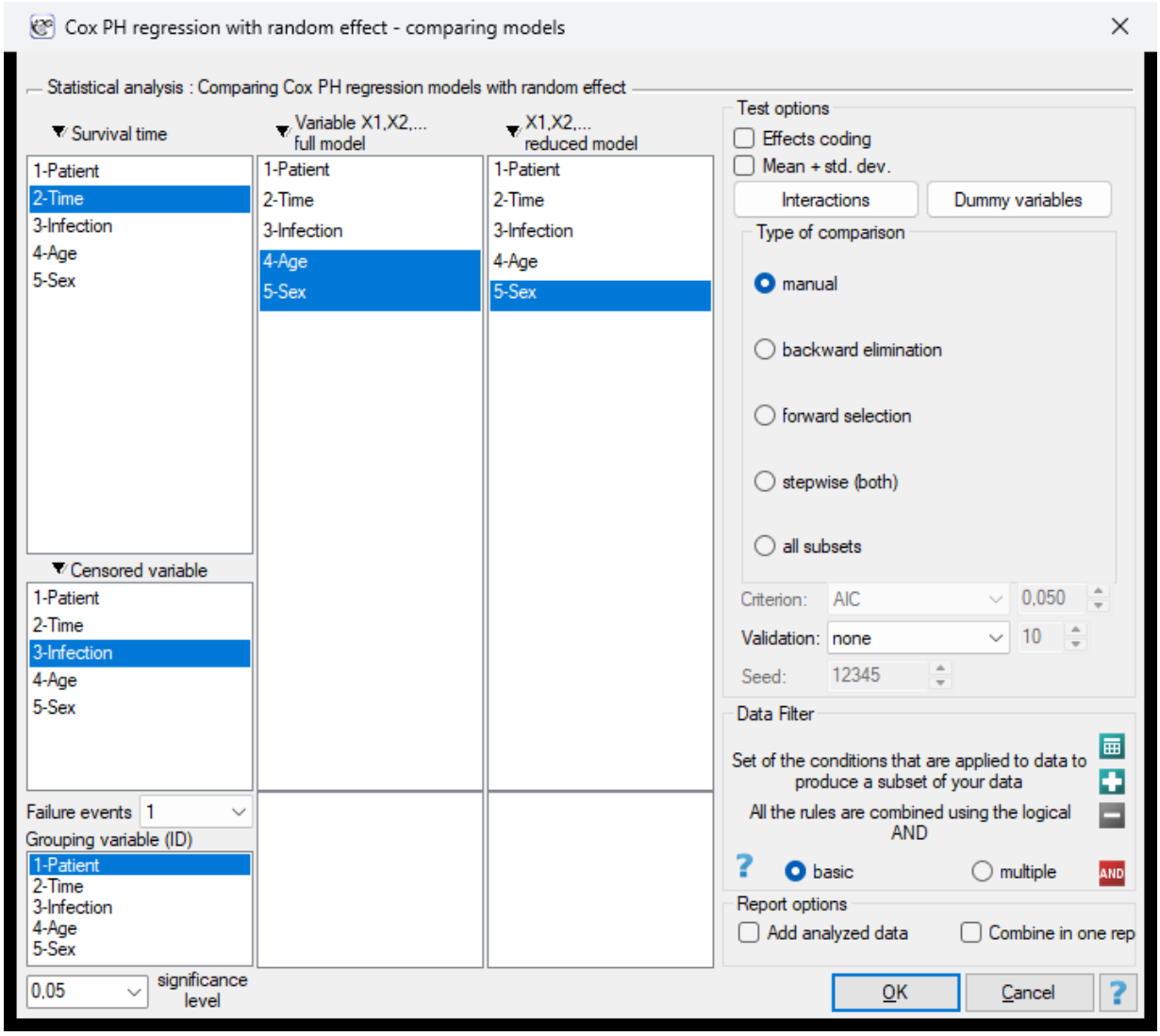
- **whole groups are split** – all the observations of one group (e.g. one patient) go to the same part of the data. The number of folds k is limited to the number of groups, and the share of the test set means the percentage of groups (the split is not stratified by status here);
- **the prediction for a new group is based on the fixed effects only** – the random effect of a group that was not in the training set is unknown, so its expected value is assumed (frailty $= 1$).

Note! The hazard ratios in a model with a random effect are conditional – they compare observations with the same value of the random effect – and are usually stronger than the effects averaged over the population. Applied to a new group, for which frailty $= 1$ is assumed, they give predictions that are more extreme than the data would warrant, which the out-of-sample partial likelihood clearly "penalises". With a large variance θ and a small number of groups, the validation error of the models with variables may therefore be larger than the error of the null model, even though the variables are significant and Harrell's C clearly exceeds 0.5. Here the validation error serves primarily to compare the models with each other.

HOW TO SET UP THE ANALYSIS

Statistics → Mixed models → Cox PH regression with random effect – comparing models.

Copyright ©2010-2026 U-smile Stats – All rights reserved



We indicate the survival time, the censored variable with the Failure events value, the variables of the full and reduced models (or the automatic selection criterion) and the Grouping variable (ID), shared by all models. Below the list of selection methods there are the fields: Criterion (AIC, AICc, BIC, p (LR / F test) with a threshold, or validation error), Validation (none, k-fold CV, k = with the number of folds, or test set, % = with the share of the test set as a percentage of the groups) and Seed of the random split (the same seed means the same split).

RESULTS AND INTERPRETATION

The report contains a fit table of the compared models (integrated log-likelihood, AIC, AICc, BIC, θ of each model) with the LR test between them, and the coefficient tables of the models with conditional hazard ratios. When validation is switched on, the report additionally contains the **Out-of-sample evaluation** block (the method of splitting, the seed, the split unit, the error of the null model, the best model by validation), and for each model – $-2 \ln PL$ per event and Harrell's C.

EXAMPLE 26.11. continuation of example 26.10 (kidneyCatheter.pqs file)

Copyright ©2010-2026 U-smile Stats – All rights reserved

In the model of the risk of catheter infection, age turned out to be non-significant. We compare the full model (age and sex) with the reduced model (sex only); the patient random effect in both.

Number of variables in the model 1	2
Convergence criterion not met	
Frailty variance (theta)	0,397184
Theta error (profile)	0,234387
Kendall's tau	0,165688
-2 Log Likelihood	364,106718
AIC - Akaike criterion	370,106718
AICc - corrected Akaike criterion	370,551163
BIC - Bayesian criterion	376,288047
Pseudo R2 (McFadden)	NA
R2 (Nagelkerke)	NA
R2 (Coxa-Snella)	NA
Number of variables in the model 2	1
Convergence criterion not met	
Frailty variance (theta)	0,387508
Theta error (profile)	0,231994
Kendall's tau	0,162306
-2 Log Likelihood	364,328423
AIC - Akaike criterion	368,328423
AICc - corrected Akaike criterion	368,546605
BIC - Bayesian criterion	372,449309
Pseudo R2 (McFadden)	NA
R2 (Nagelkerke)	NA
R2 (Coxa-Snella)	NA
Chi-square - models comparison Model 1 vs Model 2	0,221705
Degrees of freedom	1
p-value	0,637744

Full model: log-likelihood -182.05, AIC $= 370.11$; model with sex only: -182.16, AIC $= 368.33$. LR test: $\chi^2 = 0.22$, $df = 1$, $p = 0.64$ – removing age does not worsen the fit, and the AIC of the simpler model is lower. We choose the model with sex only: $HR = 0.216$ (95% CI: 0.091–0.512), $\theta = 0.388$. For contrast, the model with age only has a log-likelihood of -187.86 and AIC $= 379.72$, hardly better than the model without variables (-187.95): age alone contributes nothing.

Copyright ©2010-2026 U-smile Stats – All rights reserved

LR test full vs without age: $\chi^2 = 0.22$, $p = 0.64$; AIC 370.11 vs 368.33

Age can be removed from the model; sex and the patient random effect remain.

Out-of-sample evaluation. We repeat the same comparison, choosing in the Validation field the k-fold CV, k = option with 10 folds (seed 12345). The 38 patients are split – both episodes of the same patient go to the same part of the data.

Out-of-sample evaluation	cross-validation
Number of folds (k)	10
Seed of the random split	12345
Split unit	whole group
Null model: -2 ln PL / event	8,295639
Best model by validation (-2 ln PL / event)	null model
New-group prediction: fixed effects, frailty = 1	

	training set	cross-validation ($k = 10$)	
model	AIC	$-2 \ln PL$ per event	Harrell's C
null		8.2956	
Model 1 (age, sex)	370.11	8.5322	0.6723
Model 2 (sex)	368.33	8.4458	0.6186

The model with sex only has a smaller validation error than the model with age and sex (8.4458 versus 8.5322) – the out-of-sample evaluation confirms that age can be removed. At the same time, the null model has the smallest error (8.2956), although sex is significant in the model and Harrell's C of both models (0.62–0.67) shows that they order new patients better than at random. This is the situation described in the note above: the effect of sex is conditional (strong: $HR = 0.216$ at $\theta = 0.388$), the sample is small (38 patients, 3–4 in each evaluated part), and frailty $= 1$ is assumed for a new patient, which makes the predictions too extreme. This result does not undermine the importance of sex, but it is a reminder that the predictions of this model for new patients – without knowing their random effect – are uncertain.

cross-validation ($k = 10$, whole groups): $-2 \ln PL$ per event $= 8.53$ (age and sex), 8.45 (sex), 8.30 (null model)

Validation confirms the removal of age; with such a small sample, however, none of the models predicts the time to infection in new patients better than the null model.

Copyright ©2010-2026 U-smile Stats – All rights reserved

27 DIMENSION REDUCTION AND GROUPING

As the number of variables subjected to a statistical analysis grows, the precision of the description of the phenomenon under study increases, but so does the level of complexity and the difficulty of interpreting the results obtained. Too large a number of variables carries the risk that some of them are correlated with one another, and therefore carry information that is partly repeated – already contained in other variables. The need to reduce the number of variables (dimensionality reduction) has given rise to a whole family of methods devoted to this issue, such as: factor analysis, principal component analysis, cluster analysis and discriminant analysis. These methods make it possible to detect relationships among variables and, on that basis, to distinguish groups of similar variables and select, for further analysis, either a single representative of each group, or a new variable whose values are computed from the remaining variables in the group. In this way we can be sure that the information carried by every group of variables is taken into account in the analysis – with only a small loss of information, a set of p variables can be reduced to a set of k variables, where $k < p$.

Just as we may want to group variables, we may also want to group objects. Given information about certain characteristics of objects, we are often able to distinguish groups of objects that are similar with respect to these characteristics – for example, based on the amounts customers of a certain retail chain spend on particular products, customers can be divided into segments with similar purchasing preferences. As a result, an offer or advertisement can be prepared not for the general body of customers, but separately for each segment, so as to more precisely address the needs of a potential customer.

Principal component analysis

WHEN TO USE:

Principal component analysis (PCA) is used to reduce the number of variables describing the objects under study. Instead of many, often mutually correlated, original variables, a small number of completely new variables – **principal components** – are computed, each being a linear combination of the original variables. Successive principal components are mutually uncorrelated (orthogonal), and their number (k) is less than or equal to the number of original variables (p).

A detailed principal component analysis makes it possible to identify which original variables have a strong influence on a given component, i.e. which together with it form a homogeneous group – the principal component then becomes the representative of that group. The new variables obtained in this way are usually fewer than the original variables, yet carry the same amount of information, or only slightly less. Moreover, because they are not correlated with one another, they can be successfully used in most subsequent multivariate analyses.

ASSUMPTIONS

- the variables are measured on an interval scale,
- at least a few original variables are included in the analysis,
- the original variables are significantly correlated – if the variables are not correlated with one another (Pearson correlation coefficients close to 0), performing a principal component analysis makes no sense, since each variable would then already constitute a separate component;

Copyright ©2010-2026 U-smile Stats – All rights reserved

the degree of correlation among the variables can be assessed with Bartlett's test and the Kaiser-Meyer-Olkin (KMO) coefficient, both described below.

DEFINITIONS OF MEASURES

Each principal component is a linear combination of the original variables $X_1, X_2, ..., X_p$. Every component explains a certain part of the variability of the original variables and is therefore naturally based on measures of variability such as covariance (when the original variables are of similar size and expressed in the same units) or correlation (when the conditions needed to use covariance are not met). Each principal component is described by the following quantities.

Eigenvalue

Indicates what part of the total variability of the original variables is explained by a given principal component. The first component explains the largest part of the variance, the second explains the largest part of the variance not explained by the first, and so on – each subsequent component explains the largest part of the variance not explained by the preceding components. As a result, each successive principal component explains a smaller and smaller part of the variance, i.e. the successive eigenvalues get smaller and smaller. Since the total variance is the sum of the eigenvalues, the percentage of variability explained by each component can be calculated, and, for successive components, the cumulative variability and the cumulative percentage of variability.

Eigenvector

Reflects the influence of the individual original variables on a given principal component – it contains the coefficients of the linear combination that defines the component. The sign of these coefficients indicates the direction of the influence and is arbitrary, which does not change the value of the information carried.

Factor loadings

Like the eigenvector, factor loadings reflect the influence of the individual original variables on a given principal component – they are values showing what part of the variance of a given component is accounted for by the original variables. When the analysis is based on the correlation matrix, these loadings are interpreted as correlation coefficients between the original variables and the given principal component.

Variable contributions

Based on the coefficients of determination between the original variables and a given principal component. They show what percentage of the variability of a given principal component can be explained by the variability of the individual original variables.

Communalities

Also based on the coefficients of determination between the original variables and a given principal component, but showing the reverse relationship – what percentage of the variability of a given original variable can be explained by the variability of the first several principal components. For example, the result for the second variable in the column for the fourth principal component tells us what percentage of the variability of the second variable can be explained by the variability of the first four principal components.

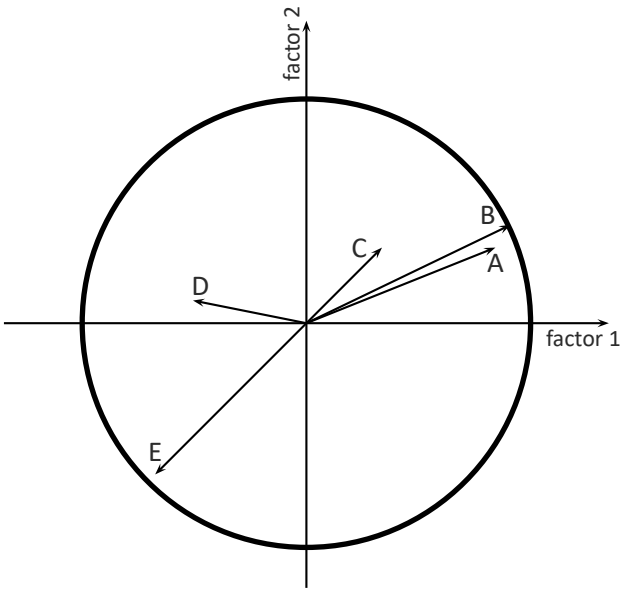
Graphical interpretation

Much of the information carried by the coefficients returned in the tables can be presented on a single

Copyright ©2010-2026 U-smile Stats – All rights reserved

chart. Such charts bring together, in one place, information about the mutual relationships among the components, the original variables and the cases, giving an overall picture of the analysis and providing a very good summary of it.

The *factor loadings plot* shows vectors, attached to the origin of the coordinate system, representing the original variables, placed on the plane defined by two selected principal components.



The coordinates of the tip of the vector are the corresponding factor loadings of the variable.

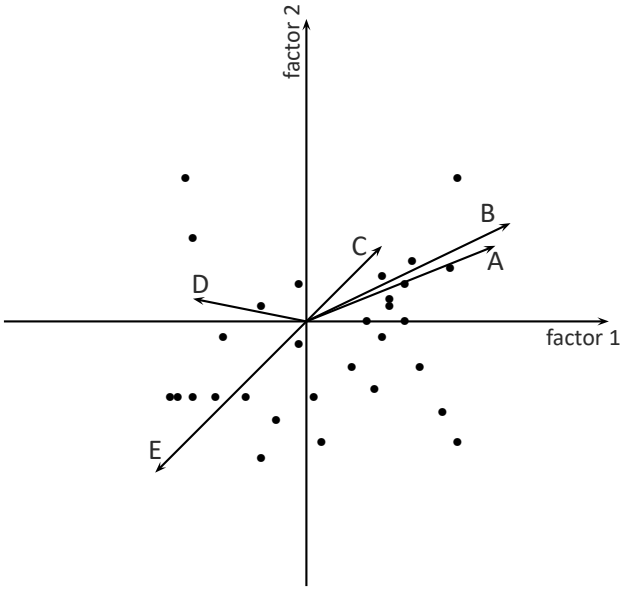
The length of the vector represents the amount of information about the original variable carried by the principal components that define the coordinate system – the longer the vector, the greater the contribution of the original variable to the components. In an analysis based on the correlation matrix, the loadings are correlations between the original variables and the components, so the points fall within the unit circle (a correlation coefficient cannot exceed one). The closer a given original variable lies to the edge of this circle, the better it is represented by the principal components shown.

The sign of the coordinates of the tip of the vector , i.e. the sign of the factor loading, indicates a positive or negative correlation between the original variable and the principal components forming the coordinate system. Considering both axes together (2 components), the original variables can be classified into one of four categories, depending on the combination of signs ($+/-$) of their factor loadings.

The angle between vectors indicates the correlation between the original variables:
$0 < \alpha < 90^0$: the smaller the angle between the vectors representing the original variables, the stronger the positive correlation between these variables.
$\alpha = 90^0$ – the vectors are perpendicular, i.e. the original variables are not correlated.
$90^0 < \alpha < 180^0$ – the larger the angle between the vectors representing the original variables, the stronger the negative correlation between these variables.

The *biplot* shows two series of data placed in the coordinate system defined by two principal components: the first series is the data from the factor loadings plot (the vectors of the original variables), and the second series consists of points representing the individual cases.

Copyright ©2010-2026 U-smile Stats – All rights reserved



The coordinates of the points are interpreted like standardized values, i.e. positive coordinates indicate a value higher than the mean value of the principal component, negative ones indicate a lower value, and the higher the absolute value, the further the point lies from the mean. If atypical (outlying) observations appear on the plot, they may distort the analysis and should be removed, with the analysis then repeated.

The distances between points indicate the similarity of the cases – the closer two points lie to each other (in the sense of Euclidean distance), the more similar the information carried by the compared cases.

The orthogonal projections of points onto the vectors are interpreted in the same way as the coordinates of the points, except that they concern the original variables rather than the principal components. Values on the side of the tip of the vector are greater than the mean value of the original variable, while values on the extension of the vector in the opposite direction are smaller than the mean.

Criteria for dimensionality reduction

There is no single universal criterion for choosing the number of principal components, so it is good practice to be guided by several methods at once.

Percentage of explained variance

All the principal components together carry 100% of the variance of the original variables. If the sum of the variances of the first few components accounts for a substantial part of the total variance, these components can satisfactorily replace the original variables. It is generally assumed that the selected principal components should account for over 80% of the variance.

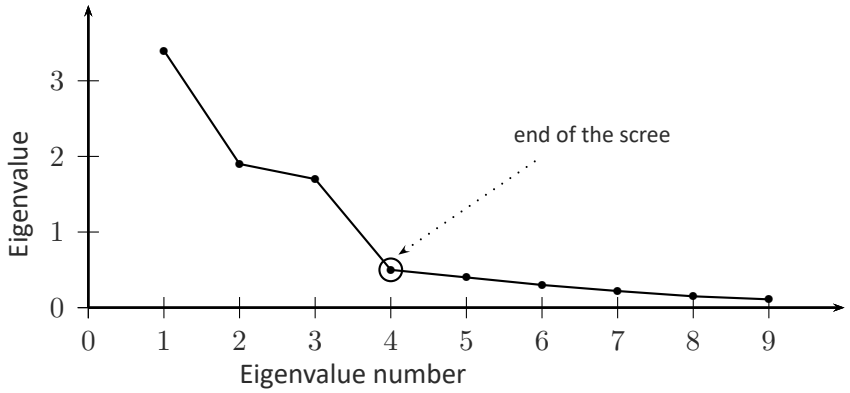
Kaiser criterion

According to this criterion, only those principal components that have at least as much variance as any standardized original variable (i.e. variance equal to 1) are retained for interpretation. In other words, only components whose eigenvalue exceeds, or is close to, 1 are considered important.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Scree plot

Shows the rate at which the eigenvalues (the percentage of explained variance) decrease for successive components.



The point on the plot where this process stabilizes and the decreasing line becomes horizontal is the so-called end of the scree. The components to the right of this point represent negligible variance and, for the most part, represent random noise.

Computing the principal components

Once we have decided how many principal components we need, we can proceed to generate them. Principal components derived from the correlation matrix are computed as a linear combination of the standardized original variables, while those derived from the covariance matrix are computed as a linear combination of the original variables centred with respect to their mean.

Checking whether the analysis is warranted

Before proceeding to compute and interpret the principal components, it is worth checking whether the analysis makes sense at all, i.e. whether the original variables are sufficiently correlated with one another. Bartlett's test and the Kaiser-Meyer-Olkin (KMO) coefficient are used for this purpose.

Bartlett's test

Tests the hypothesis that the correlation coefficients between the variables are zero (i.e. that the correlation matrix is the identity matrix). A statistically significant test result (p value lower than the assumed significance level α) shows that the original variables are significantly correlated, and therefore that performing a principal component analysis makes sense.

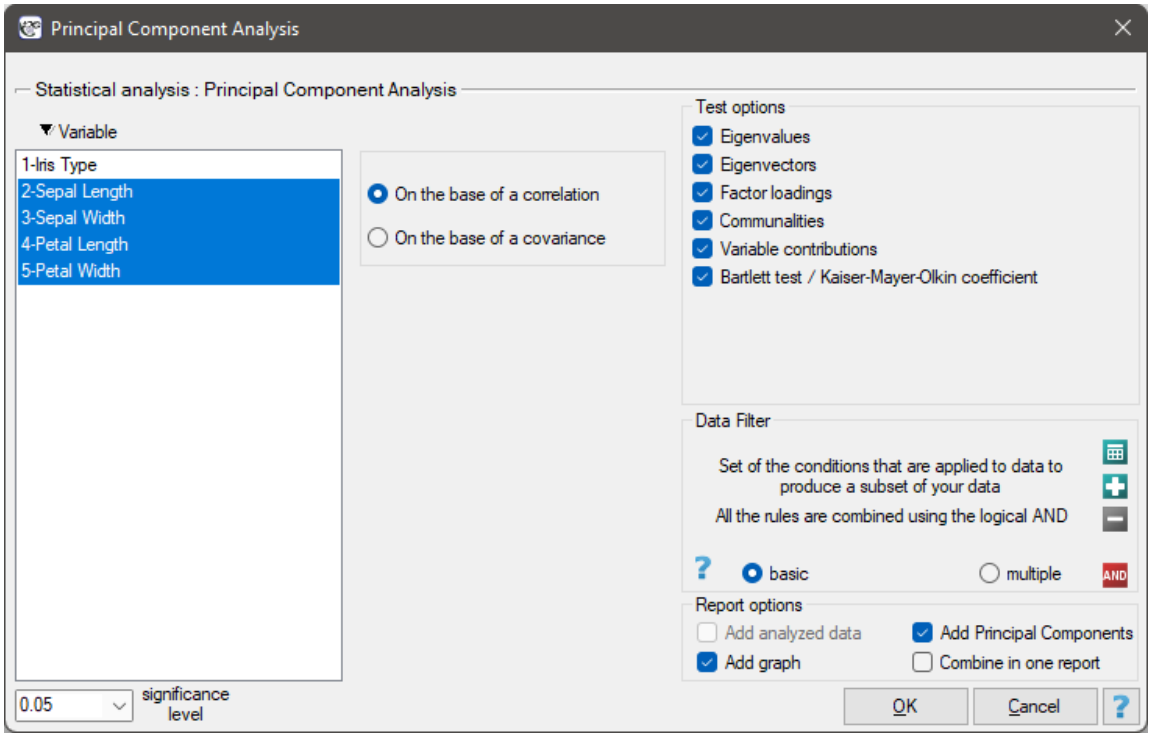
Kaiser-Meyer-Olkin (KMO) coefficient

Assesses the degree of correlation among the original variables, i.e. the strength of the case for performing a principal component analysis. The coefficient's value lies in the range $< 0, 1 >$ – low values indicate that there is no basis for performing the analysis, while high values are a reason to perform it.

HOW TO SET UP THE ANALYSIS

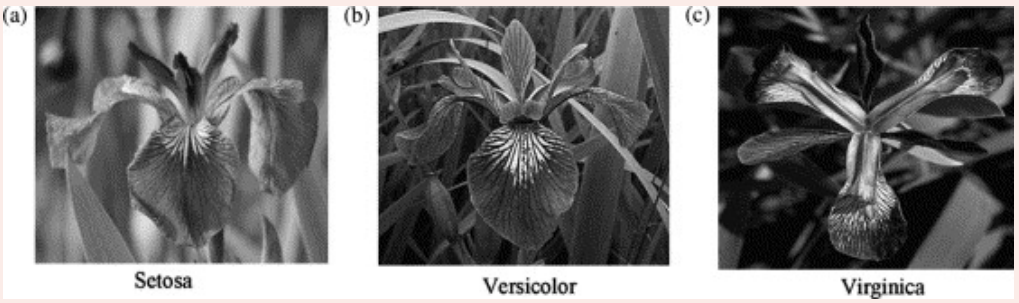
The window with the settings for Principal component analysis is opened via the menu Statistics → Multivariate models → Principal component analysis.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 27.1. (file: iris.pqs)
This classic data set was originally published in R.A. Fisher's 1936 paper[60], which introduced discriminant analysis. The file contains measurements (in centimetres) of the length and width of the petals and sepals for 3 varieties of iris. The species studied are setosa, versicolor and virginica. It is of interest to determine how these species can be distinguished on the basis of the measurements obtained.



The figure comes from Lee et al. (2006): "Application of a noisy data classification technique to determine the occurrence of flashover in compartment fires"

Principal component analysis will let us identify which measurements (the length and width of the petals and sepals) give the researcher the most information about the observed flowers.

The first step, carried out even before computing and analysing the components, is to check whether performing this analysis is worthwhile. We therefore start by computing the correlation matrix of the variables and examining the correlations obtained using Bartlett's test and the KMO coefficient.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Correlation matrix

Variable	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	1	-0.1176	0.8718	0.8179
Sepal Width	-0.1176	1	-0.4284	-0.3661
Petal Length	0.8718	-0.4284	1	0.9629
Petal Width	0.8179	-0.3661	0.9629	1

Principal Component Analysis	
Analysed variables	Sepal Length
	Sepal Width
	Petal Length
	Petal Width
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	150
Analysis of correlation matrix	
Bartlett test	
Chi-square statistic	706.9592
Degrees of freedom	6
p-value	<0.0001
Kaiser-Mayer-Olkin coefficient	
KMO	0.5401

Bartlett's test is statistically significant; KMO $= 0.54$

The original variables are significantly correlated (the correlation matrix differs significantly from the identity matrix), and the average value of the KMO coefficient sufficiently confirms that performing a principal component analysis is worthwhile.

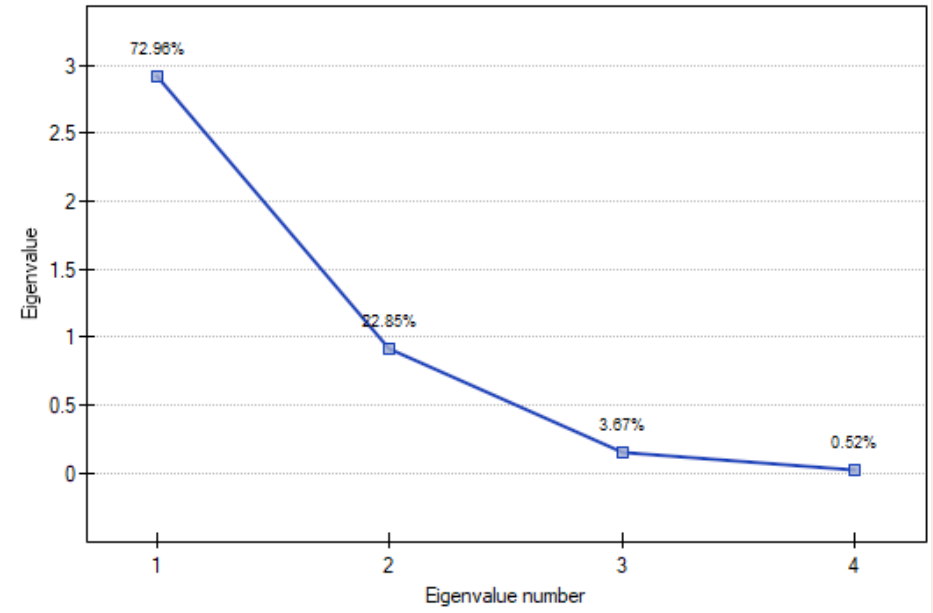
The first result of this analysis that deserves special attention is the eigenvalues:

Eigenvalues

Number	Eigenvalue	% variance	Cumulative	% cumulativ
1	2.9185	72.9624	2.9185	72.9624
2	0.914	22.8508	3.8325	95.8132
3	0.1468	3.6689	3.9793	99.4821
4	0.0207	0.5179	4	100

The eigenvalues obtained show that two, or even a single, principal component will describe our data well. The eigenvalue for the first component is 2.92, and the percentage of variance it explains is 72.96%. The second component explains considerably less variance, 22.85%, and its eigenvalue is 0.91. According to the Kaiser criterion, only a single principal component is sufficient for interpretation, since only the first component has an eigenvalue greater than 1. Looking at the scree plot, however, we can conclude that the decreasing line only becomes horizontal from the third principal component onward.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In line with this, the first two components carry significant information – together they explain as much as 95.81% of the variance (see the cumulative % column).

The communalities for the first component are high for all the original variables except sepal width, for which it is 21.17%. This means that if we were to rely only on the first component, the variable sepal width would be poorly represented.

% communalities				
Variable	Factor1	Factor2	Factor3	Factor4
Sepal Length	79.24	92.2599	99.8586	100
Sepal Width	21.1731	99.0919	99.9684	100
Petal Length	98.3182	98.373	98.6694	100
Petal Width	93.1184	93.528	99.4321	100

For the first two components, the communalities are already at a similarly very high level – for every variable analysed they exceed 90%, i.e. the variance of each variable is represented by these two components to a degree of over 90%. Having gathered all this information, it was decided to extract and interpret 2 components.

To take a closer look at the relationship between the principal components and the original variables, i.e. the length and width of the petals and sepals, we interpret: the eigenvectors, the factor loadings and the variable contributions.

Eigenvectors				
Variable	Factor1	Factor2	Factor3	Factor4
Sepal Length	-0.5211	-0.3774	0.7196	0.2613
Sepal Width	0.2693	-0.9233	-0.2444	-0.1235
Petal Length	-0.5804	-0.0245	-0.1421	-0.8014
Petal Width	-0.5649	-0.0669	-0.6343	0.5236

Factor loadings				
Variable	Factor1	Factor2	Factor3	Factor4
Sepal Length	-0.8902	-0.3608	0.2757	0.0376
Sepal Width	0.4601	-0.8827	-0.0936	-0.0178
Petal Length	-0.9916	-0.0234	-0.0544	-0.1153
Petal Width	-0.965	-0.064	-0.243	0.0754

Copyright ©2010-2026 U-smile Stats – All rights reserved

% variable contributions				
Variable	Factor1	Factor2	Factor3	Factor4
Sepal Length	27.151	14.2444	51.7776	6.8271
Sepal Width	7.2548	85.2475	5.9722	1.5255
Petal Length	33.6879	0.06	2.02	64.2321
Petal Width	31.9063	0.4481	40.2302	27.4154

The individual original variables affect the first principal component to different degrees. Let us rank them by the size of this influence:

1. Petal length is negatively correlated with the first component, so the longer the petal, the lower the value of this component. The eigenvector for petal length is the largest in this component, at -0.58. Its factor loading tells us that the correlation between the first principal component and petal length is very high, at -0.99, giving a 33.69% contribution to the first component;

2. Petal width has only a slightly smaller influence on the first component and is also negatively correlated with it;

3. Sepal length is interpreted similarly to the two previous variables, but its influence on the first component is smaller;

4. Sepal width is the most weakly correlated with the first component, and the sign of this correlation is positive.

The second component mainly represents the original variable sepal width – the remaining original variables are only weakly reflected in it. The eigenvector, factor loading and contribution of sepal width are the highest of all the variables in the second component.

Each principal component defines a homogeneous group of original variables. We will call the first component "petal size", since the variables that matter most for it are those carrying information about the petal, although it should be noted that sepal length also has a considerable influence on the value of this component. In interpreting it, we should remember that the larger the values of this component, the smaller the petals. We will call the second component "sepal width", since only sepal width is reflected in it to any great extent. The larger the values of this component, the narrower the sepal.

Finally, we generate the components by selecting the option Add principal components in the analysis window. An excerpt of the result obtained is shown below:

Principal Components			
Factor1	Factor2	Factor3	Factor4
2.2571	-0.4784	0.1273	0.0241
2.074	0.6719	0.2338	0.1027
2.3563	0.3408	-0.0441	0.0283
2.2917	0.5954	-0.091	-0.0657
2.3819	-0.6447	-0.0157	-0.0358
2.0687	-1.4842	-0.0269	0.0066
2.4359	-0.0475	-0.3344	-0.0367
2.2254	-0.2224	0.0884	-0.0245
2.3268	1.1116	-0.1446	-0.0268
2.177	0.4674	0.2529	-0.0398
2.1591	-1.0402	0.2678	0.0167
2.3184	-0.1326	-0.0934	-0.133
2.211	0.7262	0.2301	0.0024
2.6243	0.9583	-0.1802	-0.0192

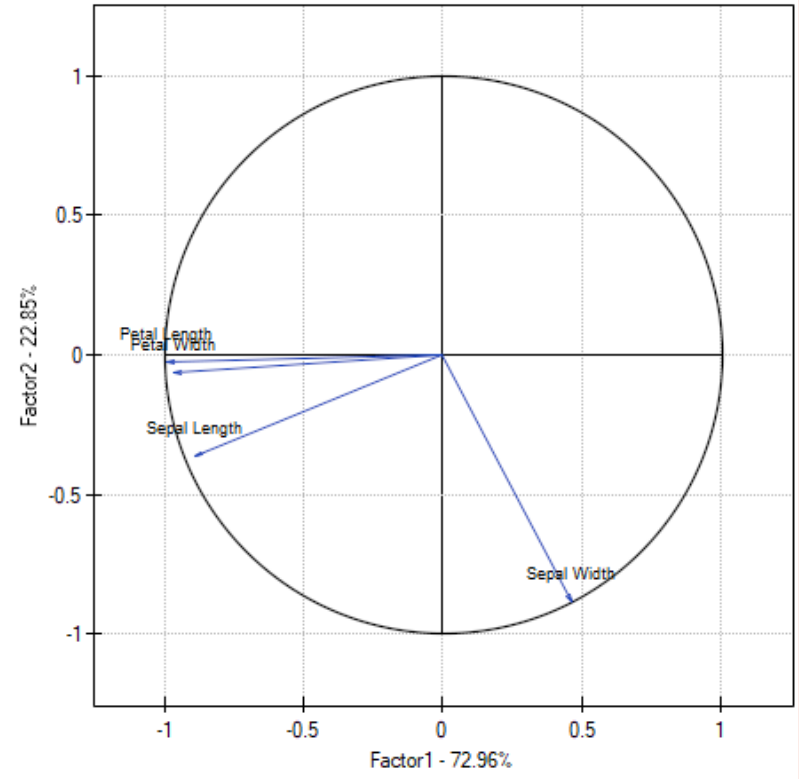
Copyright ©2010-2026 U-smile Stats – All rights reserved

So that the first two components can be used instead of the four original variables, we finally paste them into the data sheet. The researcher can now carry out any further planned statistics on two new, uncorrelated variables.

Analysis of the plots of the first two components

Analysing the plots not only leads the researcher to the same conclusions as analysing the tables, but also makes it possible to assess the results obtained more thoroughly.

Factor loadings plot



The plot shows the first two principal components, which represent 72.96% and 22.85% of the variance, giving a combined total of 95.81% of the variance of the original variables.

The vectors representing the original variables almost reach the edge of the unit circle (a circle of radius 1), so all of them are very well represented by the first two principal components forming the coordinate system.

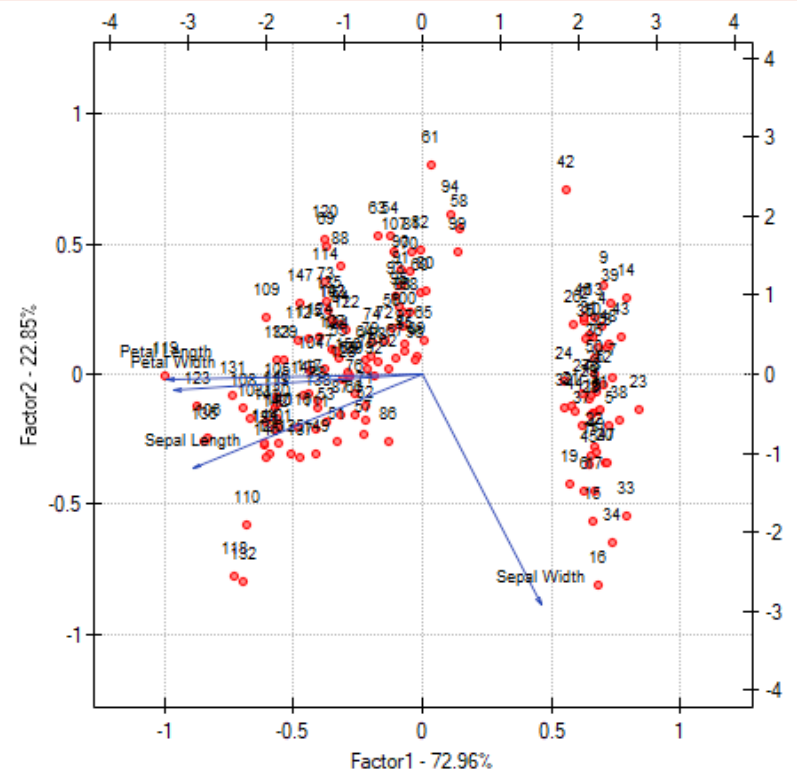
The angle between the vectors depicting petal length, petal width and sepal length is small, meaning these variables are strongly correlated. The correlation of these variables with the components forming the coordinate system is negative – the vectors lie in the third quadrant of the plot, with higher coordinate values for the tip of the vector observed for the first component than for the second. Such an arrangement of vectors indicates that they form a homogeneous group, represented mainly by the first component.

The vector for sepal width points in an entirely different direction; it is only weakly correlated with the remaining original variables, as shown by its angle to them, which is close to a right angle. Its correlation with the first component is positive but not very high (a low value of the first coordinate of the tip of the vector), while its correlation with the second component is negative and high (a high value of the second coordinate of the tip of the vector). We conclude from this

Copyright ©2010-2026 U-smile Stats – All rights reserved

that sepal width is the only original variable that is well represented by the second component.

Biplot



The biplot shows two series of data plotted against the first two components. One series consists of the vectors of the original variables, already shown in the previous plot, and the other consists of points carrying information about the individual flowers, with the values of the second series read from the upper X axis and the right-hand Y axis. The interpretation of the vectors, i.e. the first series, was discussed for the previous plot. To understand how to interpret the points, let us focus on the flowers numbered 33 and 34, and on flower number 109.

Flowers 33 and 34 are similar to each other – the distance between them is small. Both points have a value of the first component considerably higher than average, and a value of the second component considerably lower than average. The average value, i.e. the arithmetic mean of both components, is 0, so this is the centre of the coordinate system. Recalling that the first component mainly represents petal size and the second mainly represents sepal width, we can say that flowers 33 and 34 have small petals and wide sepals. Flower 109, on the other hand, is a point located far away from the other two. It has a negative first component and a positive, but not high, second component, i.e. it is a flower with relatively large petals and a sepal width only slightly smaller than average.

Similar information can be obtained by projecting the points onto the lines extending the vectors of the original variables. For example, flower 33 has a large sepal width (high, positive projection values onto the original variable sepal width), but low values of the remaining original variables (negative projection values onto the extensions of the vectors depicting the remaining original variables).

Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Formulas:
Each principal component is a linear combination of the original variables:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{in}X_p$$

where:
$X_1, X_2, ..., X_p$ – the original variables,
$a_{i1}, a_{i2}, ..., a_{ip}$ – the coefficients of the ith principal component.

The mathematical computations needed to extract the principal components amount to finding the eigenvalues and the corresponding eigenvectors from the matrix equation:

$$(M - \lambda I)a = 0$$

where:
λ – the eigenvalues,
$a_i = (a_{i1}, a_{i2}, ..., a_{ip})$ – the eigenvector corresponding to the ith eigenvalue,
M – the variance or covariance matrix of the original variables $X_1, X_2, ..., X_p$,
I – the identity matrix (1 on the main diagonal, 0 elsewhere).

The percentage of variability explained by the ith principal component is given by the formula:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p} \cdot 100\%.$$

Bartlett's test

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad M = I, \\ \mathcal{H}_1: &\quad M \neq I. \end{aligned}$$

where:
M – the variance or covariance matrix of the original variables $X_1, X_2, ..., X_p$,
I – the identity matrix (1 on the main diagonal, 0 elsewhere).

Formulas:
The test statistic is given by:

$$\chi^2 = -\left(n - 1 - \frac{2p + 5}{6}\right) \sum_{i=1}^{k} \ln \lambda_i,$$

where:
p – the number of original variables,
n – the sample size (number of cases),
λ_i – the ith eigenvalue.

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with $p(p-1)/2$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α:

Copyright ©2010-2026 U-smile Stats – All rights reserved

if $p \leq \alpha \implies$ we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$,
if $p > \alpha \implies$ there is no reason to reject $\mathcal{H}_0$.

The Kaiser-Meyer-Olkin coefficient – formula:

$$KMO = \frac{\sum_{i \neq j}^{p} \sum_{j \neq i}^{p} r_{ij}^2}{\sum_{i \neq j}^{p} \sum_{j \neq i}^{p} r_{ij}^2 + \sum_{i \neq j}^{p} \sum_{j \neq i}^{p} \hat{r}_{ij}^2},$$

where:
r_{ij} – the correlation coefficient between the ith and jth variable,
$\hat{r}_{ij}$ – the partial correlation coefficient between the ith and jth variable.

27.1 CLUSTER ANALYSIS

Cluster analysis is the search, within the data, for groups of similar objects (e.g. customers, patients, species) – such that the objects inside one group (a **cluster**) are as similar to one another as possible, while objects from different clusters differ as much as possible. Two main families of methods are distinguished: **hierarchical methods**, which build an entire hierarchy of possible partitions, and non-hierarchical methods, of which the most popular is the **k-means method**. Both families of methods base their operation on a similarity matrix, i.e. on information about how similar (or how different) the individual objects are.

In exactly the same way that objects (e.g. studied subjects) are grouped, variables (features) can also be grouped – e.g. in order to identify variables carrying similar information. In this chapter, the grouping methods will be described using the grouping of objects as an example.

Hierarchical methods

WHEN TO USE:

Hierarchical cluster analysis is used when we want to see not just a single, final division of the objects into groups, but the entire hierarchy of possible divisions – from single objects, through progressively larger clusters, up to a single cluster containing all the objects. The method builds this hierarchy step by step, joining at each step the two most similar (nearest) objects or clusters, on the basis of the similarity matrix. This lets the researcher later decide for themselves at what level to "cut" the resulting hierarchy in order to obtain the number of clusters they are interested in.

ASSUMPTIONS

- the possibility of calculating the distance/similarity between objects – the choice of an appropriate measure depends on the measurement scale of the analysed variables (see Similarity matrix),
- standardisation of the variables before calculating the similarity matrix, so that none of them (e.g. one expressed by much larger numbers) dominates the others when building the clusters,

Copyright ©2010-2026 U-smile Stats – All rights reserved

– the absence of grossly atypical (outlying) observations, which could significantly distort the distances between the remaining objects.

Note! To ensure a balanced influence of all variables on the elements of the similarity matrix, the data should be standardised by selecting the appropriate option in the analysis window. Without standardisation, variables expressed by larger numbers have a greater influence on the result obtained.

DEFINITIONS OF MEASURES

Similarity of objects

In the course of work related to cluster analysis, similarity or distance measures play a fundamental role. The mutual similarity of the objects is placed in the similarity matrix. The large variety of methods for determining the distance/dissimilarity between objects allows the choice of measures that best reflect the actual relations. Distance and similarity measures are described in more detail in the Similarity matrix section.

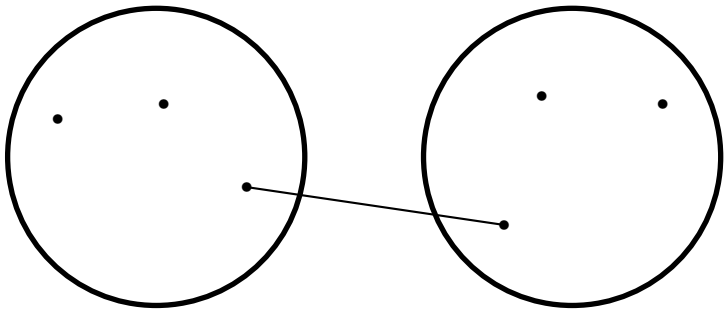
Cluster analysis bases its operation on searching for clusters inside the similarity matrix. Such a matrix is built during the course of performing the cluster analysis. For the cluster analysis to produce the desired results, when choosing the way of calculating the distance it should be remembered that larger numerical values in the similarity matrix should indicate greater dissimilarity of the objects, and smaller values their similarity.

Note! To increase the influence of the selected variables on the elements of the similarity matrix, appropriate weights should be indicated when setting up the way of defining the distance, while at the same time remembering to standardise the data.

For example, for people wanting to take care of a dog, grouping dogs according to size, coat colour, tail length, character, breed, etc. will make the choice easier. However, treating all characteristics identically may result in completely dissimilar dogs being placed in one group. For most of us, on the other hand, a dog's size and character matter more than its tail length, so when grouping, the similarity measures should be set so that it is precisely size and character that carry the greatest weight in building the clusters.

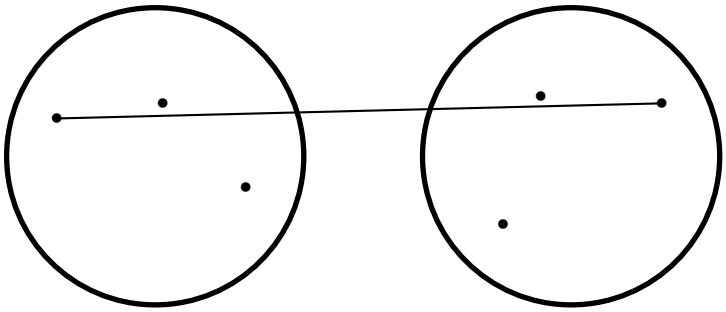
Methods of linking objects and clusters

- Single linkage method (nearest neighbour) – the distance between clusters is determined by the distance between the objects of each cluster that lie closest to each other.

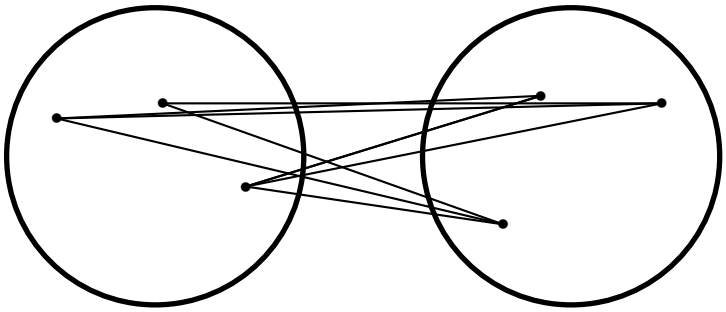


Copyright ©2010-2026 U-smile Stats – All rights reserved

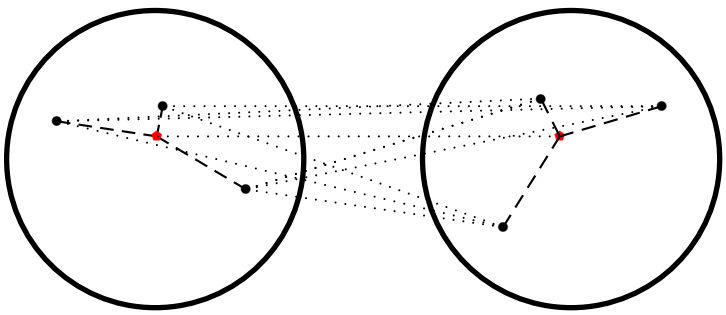
- Complete linkage method (farthest neighbour) – the distance between clusters is determined by the distance between the objects of each cluster that lie farthest from each other.



- Average linkage method – the distance between clusters is determined by the average distance between all pairs of objects located within two different clusters.



- Weighted average linkage method – analogously to the average linkage method, this involves calculating the average distance, but this average is weighted by the number of elements in each cluster. As a result, this method should be chosen when we expect to obtain clusters of similar sizes.
- Ward's method – based on the principle of analysis of variance – it calculates the difference between the sums of squared deviations of the distances of the individual objects from the centroid of the clusters to which those objects belong. This method is chosen most often because of its fairly universal character.



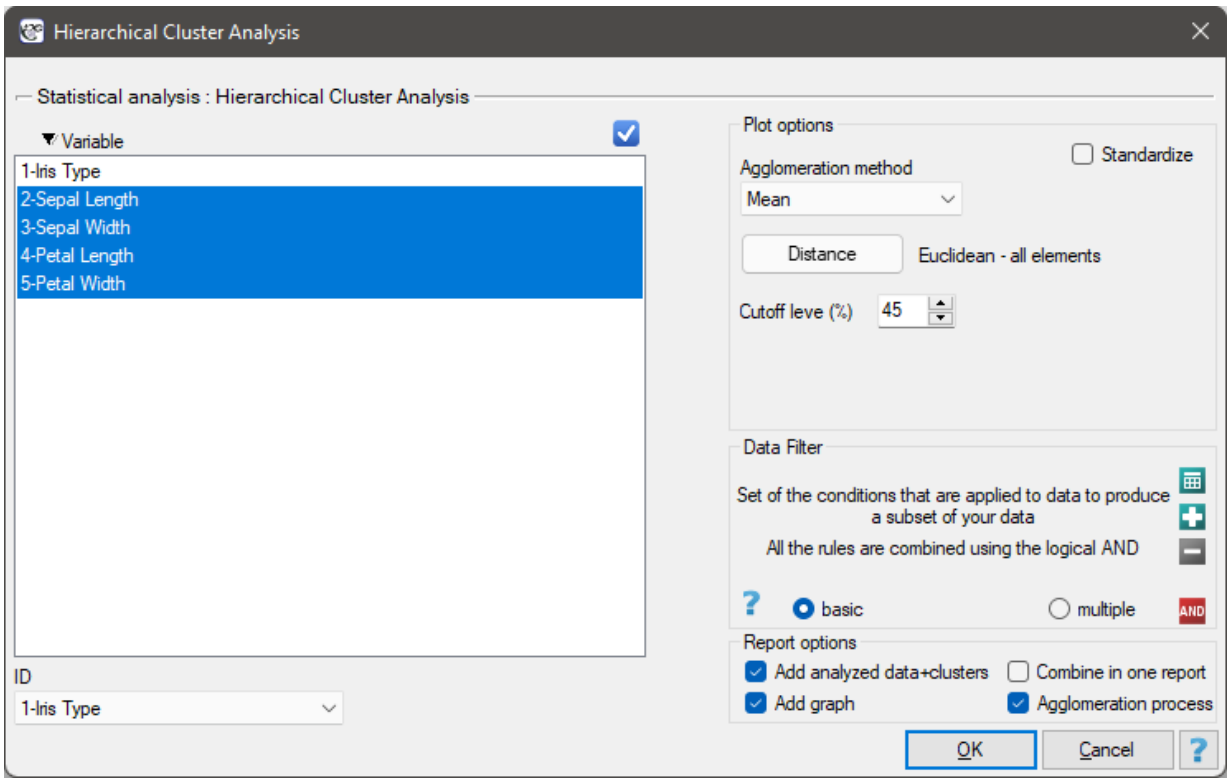
Copyright ©2010-2026 U-smile Stats – All rights reserved

Dendrogram

The result of a cluster analysis carried out by a hierarchical method is presented by means of a dendrogram. A **dendrogram** is a tree-like form showing the relationships between the individual objects, obtained from the analysis of the similarity matrix. The cutoff level of the dendrogram determines the number of clusters into which we want to divide the collected objects. We define the way the cutoff is made by specifying, as a percentage, the length of the linkage at which the cut will occur, where 100% is the length of the last, and at the same time the longest, linkage in the dendrogram.

HOW TO SET UP THE ANALYSIS

The window with the hierarchical cluster analysis options is opened through the Advanced statistics→Grouping and reduction→Hierarchical cluster analysis menu.



PERFORMING AND INTERPRETING THE ANALYSIS

Example (27.1) cont. *(iris.pqs file)*

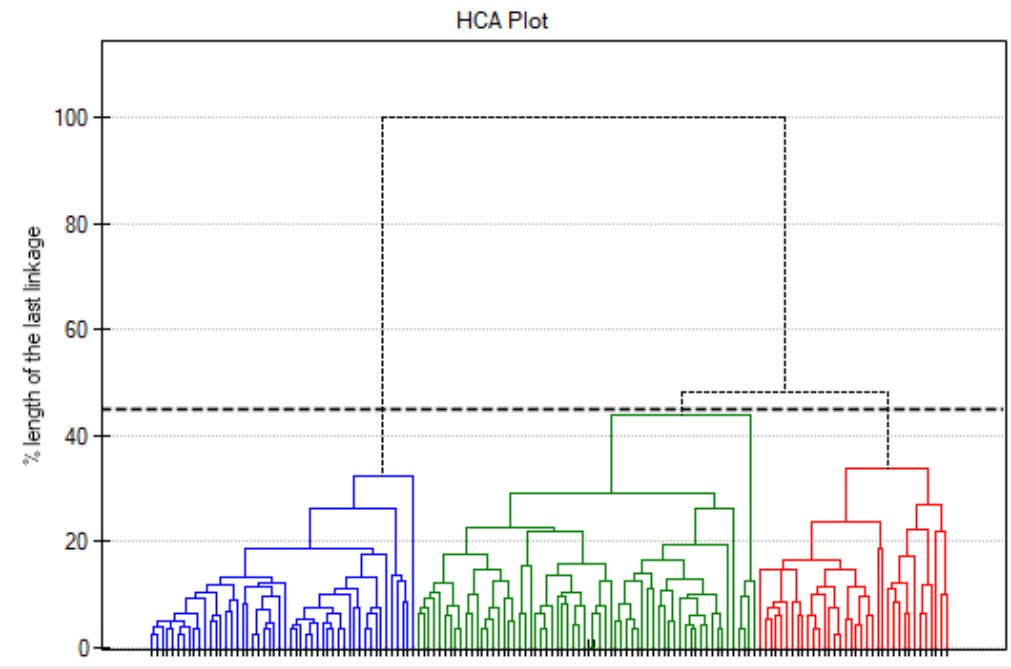
The analysis will be performed on the classic data set concerning the division of iris flowers into 3 varieties, on the basis of the width and length of the petals and sepals (R.A. Fisher, 1936[60]). Because this data set contains information about the actual variety of each flower, once the cluster analysis has been performed, it is possible to determine the accuracy of the division made.

We assign the flowers to particular groups on the basis of columns 2 to 5. We choose the way of calculating the distance, e.g. the Euclidean distance, and the linkage method, e.g. the average. Specifying the cluster cutoff level will let us cut the dendrogram in such a way that clusters are formed – in the case of this analysis we want to obtain 3 clusters, and to achieve this we change the cutoff level to 45%. We also attach data+clusters to the report.

Copyright ©2010-2026 U-smile Stats – All rights reserved

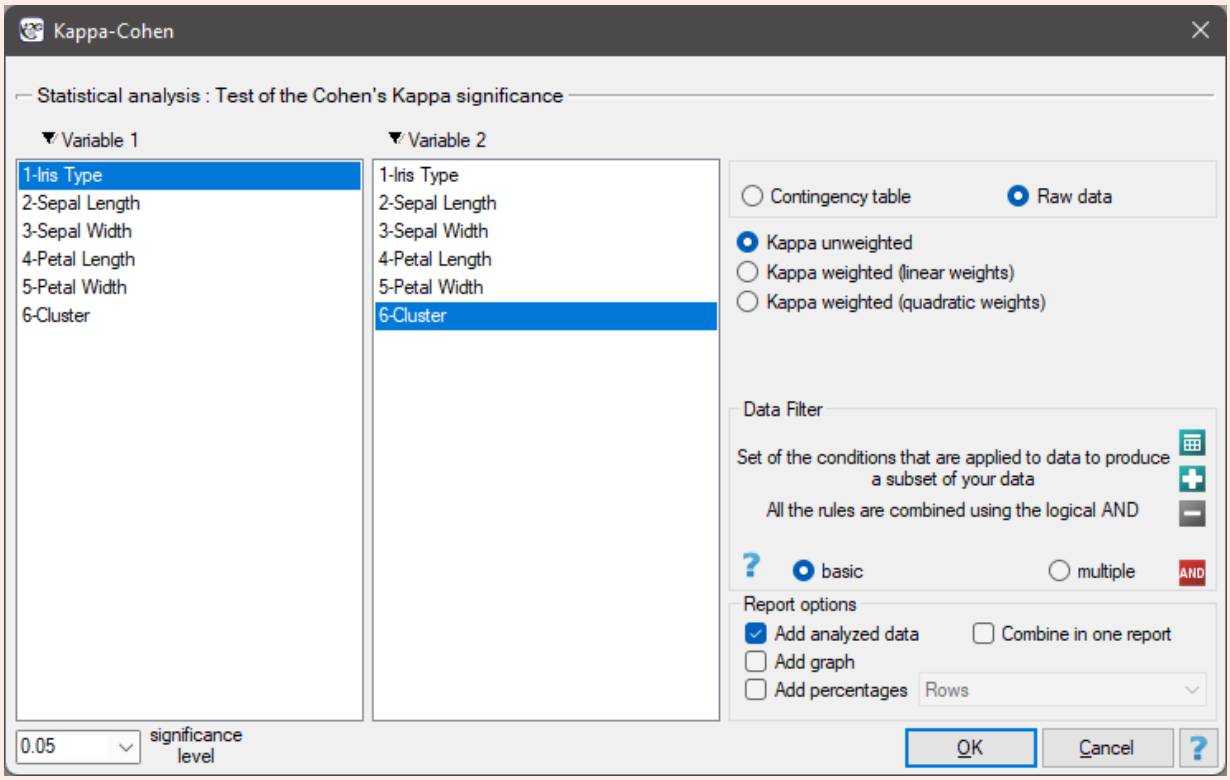
Hierarchical Cluster Analysis	
Analysed variables	Sepal Length
	Sepal Width
	Petal Length
	Petal Width
Number of unspecified	0
Number of missing data	0
Frequency	150
Standardization	No
Agglomeration method	Mean
Cutoff leve (%)	45
Distance	Euclidean - all elements
Number of clusters	3
Number of elements in cluster 1	50
Number of elements in cluster 2	64
Number of elements in cluster 3	36

The dendrogram chart shows the order of the linkages and their length.



To examine whether the extracted clusters represent the 3 actual iris varieties, we can copy the column containing the information on cluster membership from the report and paste it into the datasheet. Like the clusters, the varieties are also described numerically through Codes/Labels/Format, so an agreement analysis can easily be carried out. We will check the agreement of our results with the actual membership of a given flower in the corresponding species using the Cohen's Kappa method.

Copyright ©2010-2026 U-smile Stats – All rights reserved



For the example presented, the observed agreement is shown by the table:

Data :			✓Cluster	
Iris Type	1	2	3	Summary
setosa	50	0	0	50
versicolor	0	50	0	50
virginica	0	14	36	50
Summary	50	64	36	150

Cohen's Kappa $= 0.86$ (statistically significant)

The virginica variety can be confused with versicolor, which is why we observe 14 misclassifications here. The Kappa agreement coefficient itself is, however, statistically significant and equals 0.86, which shows a high level of agreement between the obtained clusters and the actual flower variety.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Test of the Cohen's Kappa significance	
Analysed variables	Iris Type
	Cluster
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of pairs	150
Type	Kappa unweighted
Kappa coefficient	0.86
Std. err. of Kappa	0.0351
-95% CI for Kappa coefficient	0.7911
+95% CI for Kappa coefficient	0.9289
Std. err. of Kappa distribution	0.057
Z statistic	15.0942
p-value (asymptotic)	<0.0001

The k-means method

WHEN TO USE:

The k-means method is used when it is already known (or assumed) in advance into how many groups (k) we want to divide the objects, and we are primarily interested in a single, final division – without building an entire hierarchy, as in the hierarchical methods. The method is modelled on the algorithm initially proposed by Stuart Lloyd and published in 1982[110]. Objects are moved between clusters so that the dissimilarity of the objects inside each cluster is as small as possible, and the distances between clusters as large as possible.

ASSUMPTIONS

- measurement of the variables allowing the Euclidean distance between objects to be calculated (the algorithm is based on the matrix of Euclidean distances),
- a number k of sought clusters assumed in advance,
- standardisation of the variables, so that none of them dominates the others when building the clusters (analogously to the hierarchical methods).

DEFINITIONS OF MEASURES

In this method, the objects are divided into a predetermined number of k clusters. The initial clusters are improved in the course of the agglomeration procedure, by moving objects between them so

Copyright ©2010-2026 U-smile Stats – All rights reserved

that the dissimilarity of the objects within a cluster is as small as possible, and the distances between clusters as large as possible. The algorithm operates on the basis of the matrix of Euclidean distances between objects, and the parameters needed in the agglomeration procedure of the k-means method are: the starting centres and the stopping criterion. The starting centres are the objects from which the algorithm will begin building the clusters, and the stopping criterion defines how the algorithm is to be stopped.

Starting centres
The choice of starting centres has a fundamental effect on the convergence of the k-means algorithm towards obtaining suitable clusters. The starting centres can be selected in two ways:

k-means++ – the optimal selection of starting points, using the k-means++ algorithm proposed in 2007 by David Arthur and Sergei Vassilvitskii[1]. It guarantees that the optimal solution of the k-means algorithm is obtained with as few iterations as possible. The algorithm uses an element of randomness, so the results obtained may differ slightly between successive runs of the analysis. If the data do not form natural clusters, using k-means++ will give quite different results in successive runs of the k-means analysis. High reproducibility of the results, on the other hand, indicates the possibility of achieving a good division into separate clusters.

First n – lets the user indicate the points that are to be the starting centres, by placing these objects in the first positions in the data table.

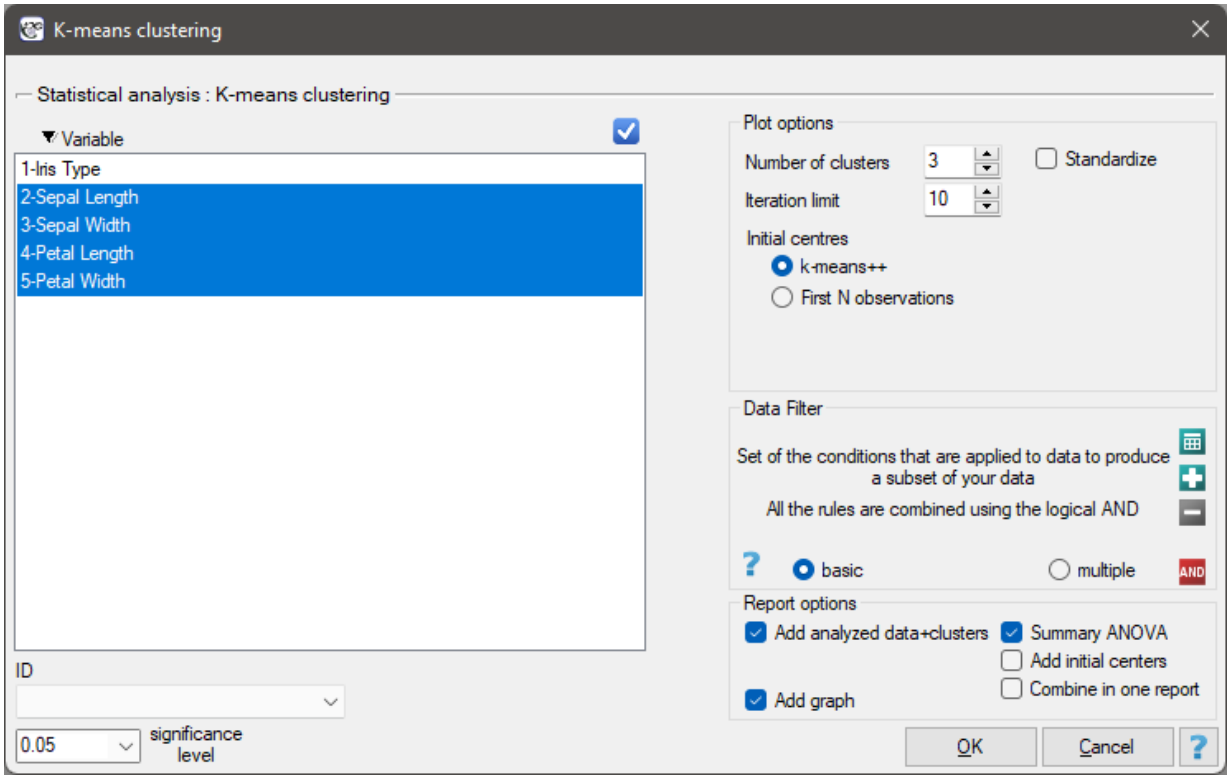
The **stopping criterion** is the moment at which the membership of the points in the classes no longer changes, or the number of repetitions of the steps of the agglomeration procedure reaches the number of iterations set by the user.

Given the way the k-means cluster analysis algorithm works, a natural consequence of it is to compare the resulting clusters by means of a one-way analysis of variance (**ANOVA**) for independent groups.

HOW TO SET UP THE ANALYSIS

The window with the k-means cluster analysis options is opened through the Advanced statistics→Grouping and reduction→K-means cluster analysis menu.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

Example (27.1) cont. *(iris.pqs file)*
The analysis will be performed on the classic data set concerning the division of iris flowers into 3 varieties, on the basis of the width and length of the petals and sepals (R.A. Fisher, 1936[60]).

We assign the flowers to particular groups on the basis of columns 2 to 5. We also indicate that we want to divide the studied flowers into 3 groups.

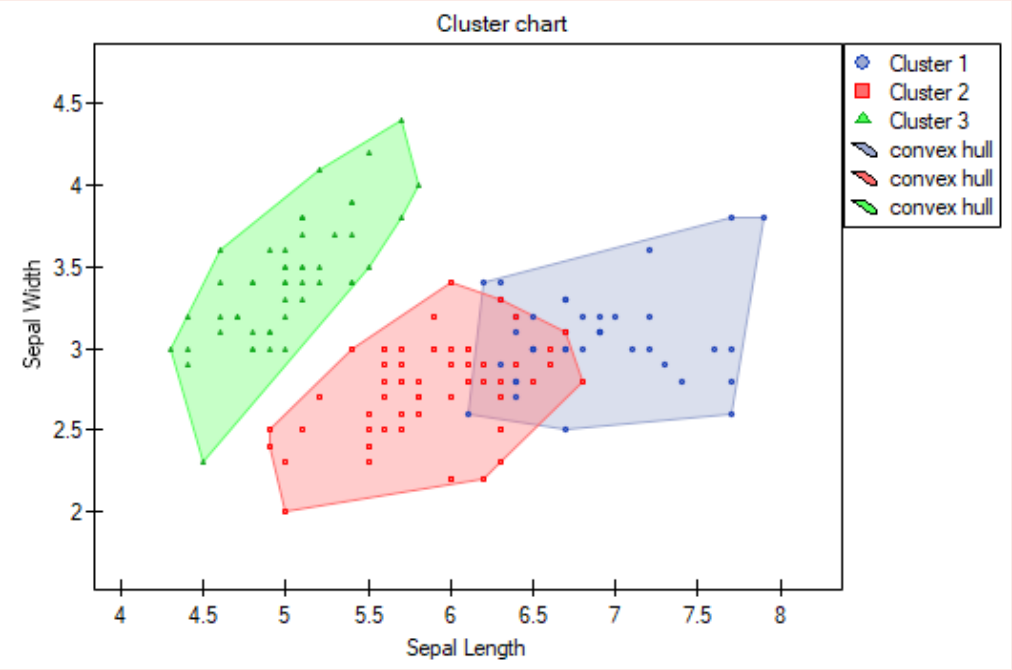
K-means clustering	
Analysed variables	Sepal Length
	Sepal Width
	Petal Length
	Petal Width
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Frequency	150
standardization	No
Iteration limit	10
Initial centers based on	k-means++
Number of clusters	3
Number of elements in cluster 1	39
Number of elements in cluster 2	61
Number of elements in cluster 3	50

Copyright ©2010-2026 U-smile Stats – All rights reserved

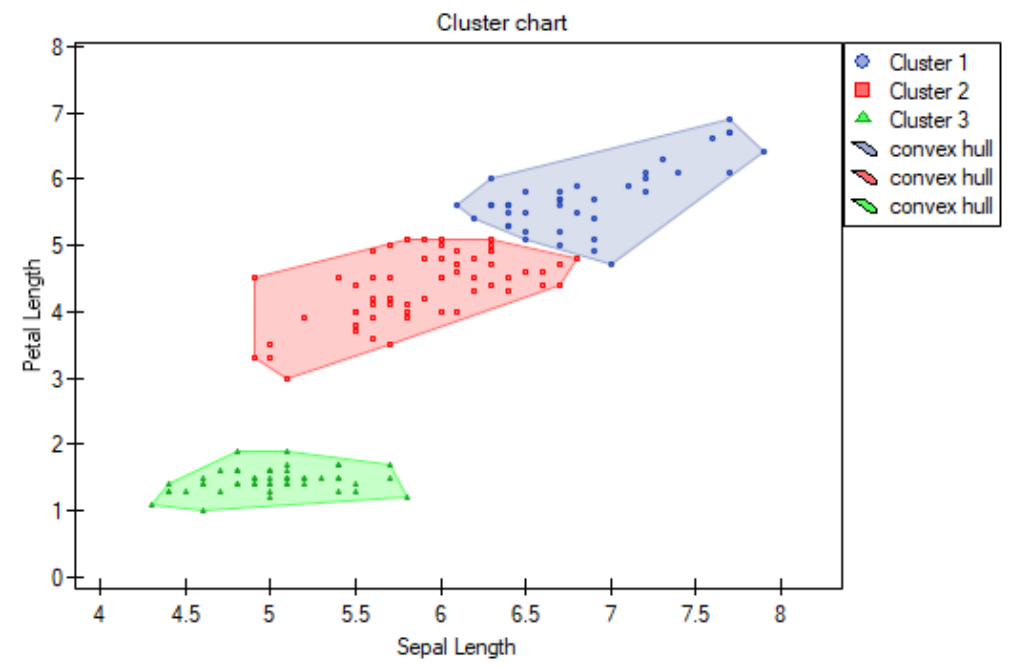
One-way ANOVA				
	Sepal Length	Sepal Width	Petal Length	Petal Width
SS[BG]	74.9796	12.9894	437.2549	77.2911
SS[WG]	27.1887	15.3176	27.0705	9.2788
df[BG]	2	2	2	2
df[WG]	147	147	147	147
F statistic	202.6943	62.3283	1187.2034	612.2429
p-value	<0.0001	<0.0001	<0.0001	<0.0001

ANOVA result statistically significant for each of the 4 dimensions

As a result of the analysis performed, 3 clusters were formed, which differ statistically significantly in every studied dimension, i.e. for petal width, petal length, sepal width, and sepal length. The difference can be observed on the charts, where, for two selected dimensions, we show the membership of the individual points in a cluster:



Copyright ©2010-2026 U-smile Stats – All rights reserved



Repeating the analysis may give slightly different results, but the variability obtained will not be large. This shows that the data submitted to the analysis do form natural clusters, and that carrying out a cluster analysis is justified in this case.

Note! After running the analysis a few times, we can select the result that interests us most, and then set the data that form the starting centres at the beginning of the sheet – the analysis carried out on the basis of starting centres chosen as First N observations will then consistently produce the result we selected.

Note! To find out whether the clusters represent the 3 actual iris varieties, we can copy the information on cluster membership from the report and paste it into the datasheet. The agreement of our results with the actual membership of a given flower in the corresponding species can be checked analogously as for the hierarchical cluster analysis.

FOR THOSE INTERESTED

Agglomeration procedure – hierarchical methods

1. Following the indicated **linkage method**, the algorithm finds a pair of similar objects in the **similarity** matrix and joins them into a cluster;
2. The dimension of the similarity matrix is reduced by one (two objects are replaced by one), and the distances in the matrix are recalculated;
3. Steps 2-3 are repeated until a single cluster containing all objects is obtained.

Agglomeration procedure – the k-means method

Copyright ©2010-2026 U-smile Stats – All rights reserved

1. Selection of the starting centres;
2. On the basis of the similarity matrix, the algorithm assigns each object to the nearest centre;
3. Corrected centres are determined for the resulting clusters;
4. Steps 2-3 are repeated until the stopping criterion is met.

Copyright ©2010-2026 U-smile Stats – All rights reserved

28 SURVIVAL ANALYSIS

Survival analysis allows us to evaluate how long it takes for an event of interest to a researcher to occur – e.g. how long patients live after surgery, how long a disease remission lasts, or after what time a device fails. Outside medicine (where the name originates) it is also used in other fields, where it goes by other names: reliability analysis, duration analysis, or event history analysis.

The basic descriptive tools here are **life tables** and **Kaplan-Meier curves**, which allow us to estimate what proportion of the study group "survives" up to a given point in time. When we are interested in comparing survival time between groups (e.g. between different treatment methods), we use the **comparison of survival curves**. The influence of many variables at once on survival time is studied by means of **regression models**, primarily Cox proportional hazard regression.

The basic concepts of survival analysis will be explained using the example of the life expectancy of heart transplant patients:

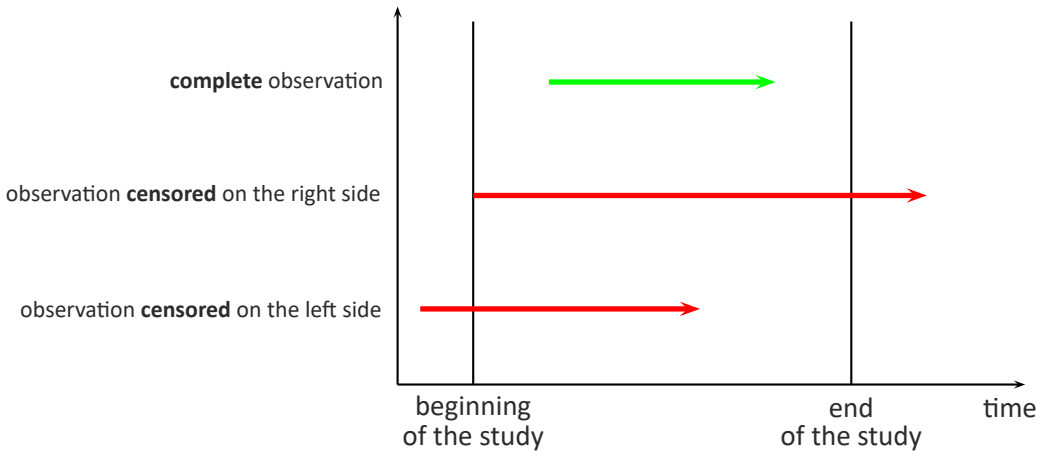
Event – a change of interest to the researcher, e.g. death.

Survival time – the period that elapses between the initial state and the occurrence of the event, e.g. the length of a patient's life after a heart transplantation.

Note! One column with the ready-made, calculated time should be prepared for the analysis. When we have two dates at our disposal – an initial and a final one – the elapsed time between them should be calculated before the analysis, using worksheet formulas.

Censored observations – observations for which we do not have complete information about survival time. We distinguish here:

- **a complete observation** – both the date of the transplantation and the date of the patient's death are known, so the exact survival time after the transplantation can be determined;
- **an observation censored on the right** – the date of the patient's death is not known (the patient is still alive when the study ends), so the exact survival time cannot be determined;
- **an observation censored on the left** – the date of the heart transplantation is not known, although it is known to have occurred before the study began, so here too the exact survival time cannot be determined.



Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! The end of the study means the end of observation of a given patient – it is not always the same moment for all subjects. The end of observation can be, for example, the moment when contact with the patient is lost (so we no longer know their further survival time). Similarly, the start of the study does not have to occur at the same time for all patients.

Life tables

WHEN TO USE:

Life tables (survival tables) are the simplest way of summarizing the survival time of a study group – the timeline is divided into equal intervals (e.g. every 3 years), and for each of them we check how many subjects survived to it, how many died, and how many were lost to follow-up (censored). Their advantage is ease of interpretation; their drawback is the need to arbitrarily divide time into intervals.

DEFINITIONS OF MEASURES

For each time interval PQStat calculates:

- **the number of entering cases** – the number of people who survived to the beginning of the given interval;
- **the number of censored cases** – the number of people in a given interval defined as censored cases;
- **the number of cases at risk** – the number of cases entering the given interval reduced by half of the number of censored cases in that interval;
- **the number of complete cases** – the number of people who experienced the event (e.g. died) in a given interval;
- **the proportion of complete cases** – the ratio of the number of complete cases (deaths) in a given interval to the number of cases at risk in that interval;
- **the proportion of surviving cases** – calculated as 1 minus the proportion of complete cases in a given interval;
- **the cumulative survival proportion (survival function)** – the probability of surviving up to a given time; since surviving another interval requires having survived all the previous ones, this probability is calculated as the product of all the previous proportions of surviving cases

  ± the standard error of the survival function;
- **probability density** – the estimated probability of experiencing the event (death) in a given interval, expressed per unit of time

  ± the standard error of the probability density;
- **the hazard rate** – the probability (per unit of time) that a case which survived to the beginning of a given interval will experience the event (die) precisely in that interval

  ± the standard error of the hazard rate.

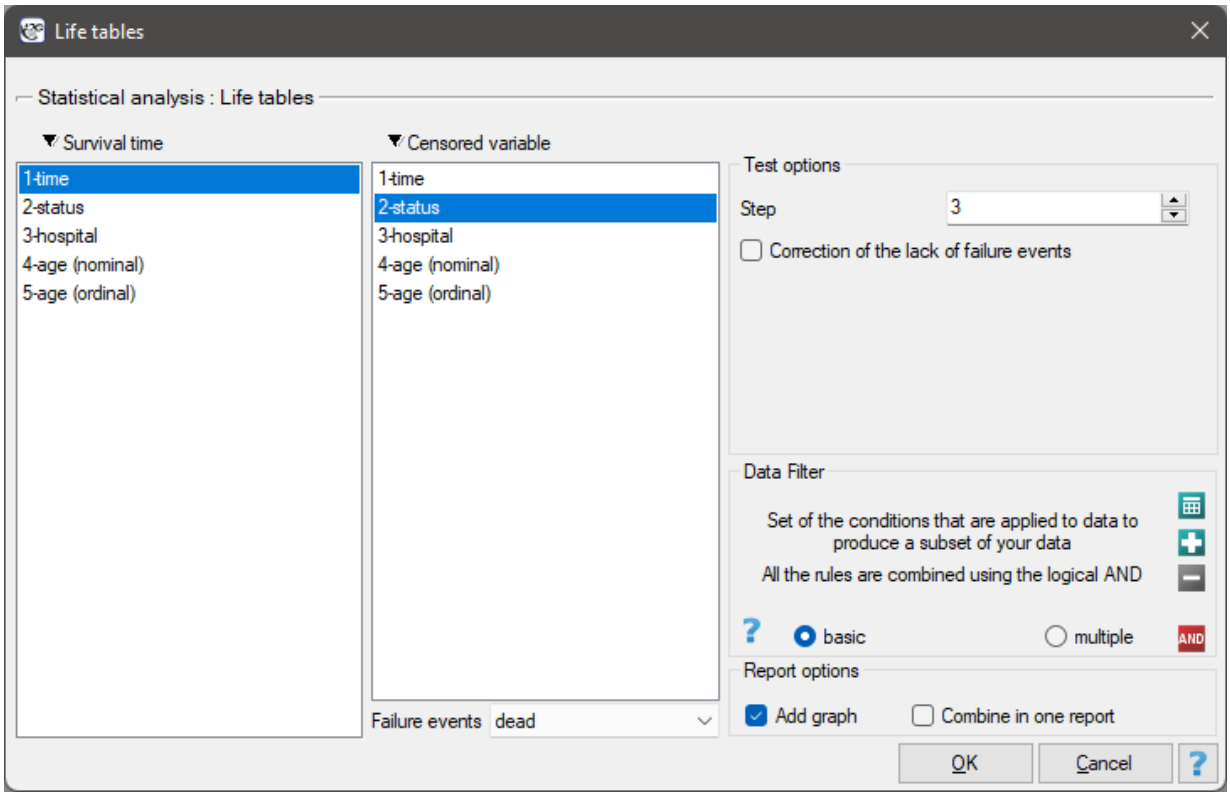
Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! When there is no complete observation in some interval of survival time, the program allows the use of a correction – the zero number of complete cases is then replaced with the value 0.5.

The information obtained can be illustrated with three graphs: a graph of the survival function, a graph of the probability density, and a graph of the hazard rate.

HOW TO SET UP THE ANALYSIS

The window with the settings for life tables is accessed via the menu Statistics→Survival analysis→Life tables.



Life tables are built for survival-time intervals of equal width, given by the researcher – these intervals are defined by specifying their step.

CALCULATION AND INTERPRETATION

EXAMPLE 28.1. (transplant.pqs file)
Patients' survival after liver transplantation was studied. A group of 89 patients was observed over 21 years; their age at the time of the transplantation ranged from $\langle 45\text{years}; 60\text{years})$. A fragment of the collected data is presented in the table below:

Copyright ©2010-2026 U-smile Stats – All rights reserved

time	status	hospital	age (nominal)
14	dead	1	<45; 50)
21	alive	1	<45; 50)
4	dead	1	<50; 55)
4	alive	1	<50; 55)
5	dead	1	<50; 55)
6	alive	1	<50; 55)
6	alive	1	<50; 55)
9	dead	1	<50; 55)
16	alive	1	<50; 55)
1	dead	1	<55; 60)

The complete data in this analysis are those for which we have full information about the length of life after the transplantation, i.e. described as "death" (53 people, i.e. 59.55%). The censored data are those for which we lack that information because the patients were still alive when the study ended (36 people, i.e. 40.45%). We build life tables for these patients, creating 3-year time intervals:

Life tables

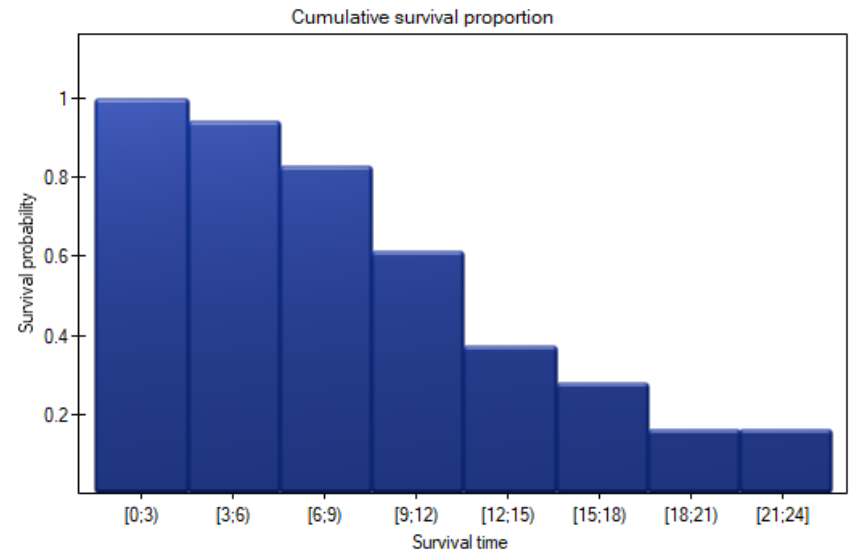
Interval	Censored	Failure event	Entered	At risk	Failure event	Censored pr	Cumulative	Probability d	Hazard rate	Standard err	Standard err	Standard err
[0;3)	0	5	89	89	0.0562	0.9438	1	0.0187	0.0193	0	0.0081	0.0086
[3;6)	5	10	84	81.5	0.1227	0.8773	0.9438	0.0386	0.0436	0.0244	0.0115	0.0137
[6;9)	14	16	69	62	0.2581	0.7419	0.828	0.0712	0.0988	0.0404	0.0157	0.0244
[9;12)	7	14	39	35.5	0.3944	0.6056	0.6143	0.0808	0.1637	0.0549	0.0183	0.0424
[12;15)	3	4	18	16.5	0.2424	0.7576	0.3721	0.0301	0.092	0.0604	0.014	0.0455
[15;18)	3	4	11	9.5	0.4211	0.5789	0.2819	0.0396	0.1778	0.0603	0.0173	0.0857
[18;21)	2	0	4	3	0	1	0.1632	0	0	0.0571	0	0
[21;24]	2	0	2	1	0	1	0.1632			0.0571		

For each 3-year interval we can interpret the results obtained in the table – for example, for people who lived at least 9 years after the transplantation, i.e. who entered the interval [9;12):

- the number of people who survived to 9 years after the transplantation is 39;
- there are 7 people we know lived at least 9-12 years – information about them was collected, but we do not know whether they lived longer, since they were lost to the study after this time;
- the number of people at risk of death in this interval is 36;
- there are 14 people we know died 9 to 12 years after the transplantation;
- 39.4% of the endangered patients died 9-12 years after the transplantation;
- 60.6% of the endangered patients lived 9-12 years after the transplantation;
- the percentage surviving to 9 years after the transplantation is 61.4% $\pm$ 5%;
- 0.08 $\pm$ 0.02 – the probability of death for each year of the 9-12 interval;
- 0.16 $\pm$ 0.04 – the conditional, annual probability of death for a case that survived to the beginning of the 9-12 interval.

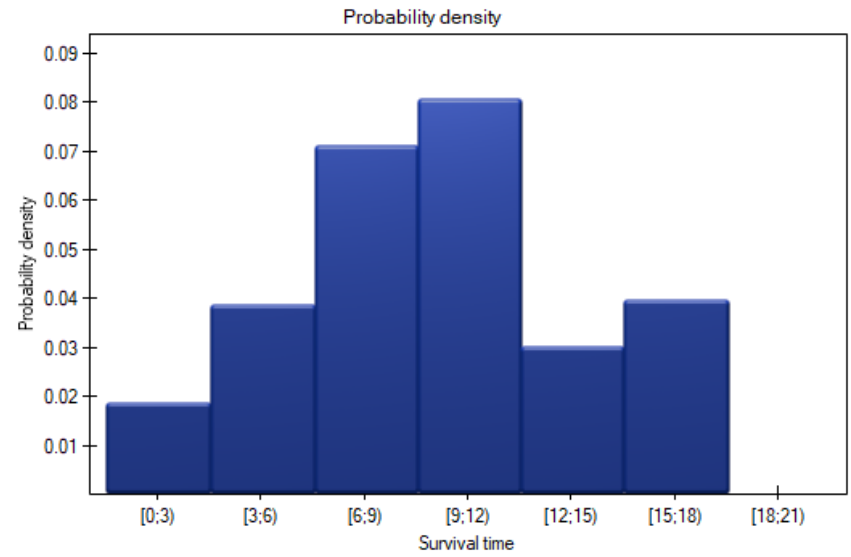
We will present the results on a few graphs:

Copyright ©2010-2026 U-smile Stats – All rights reserved

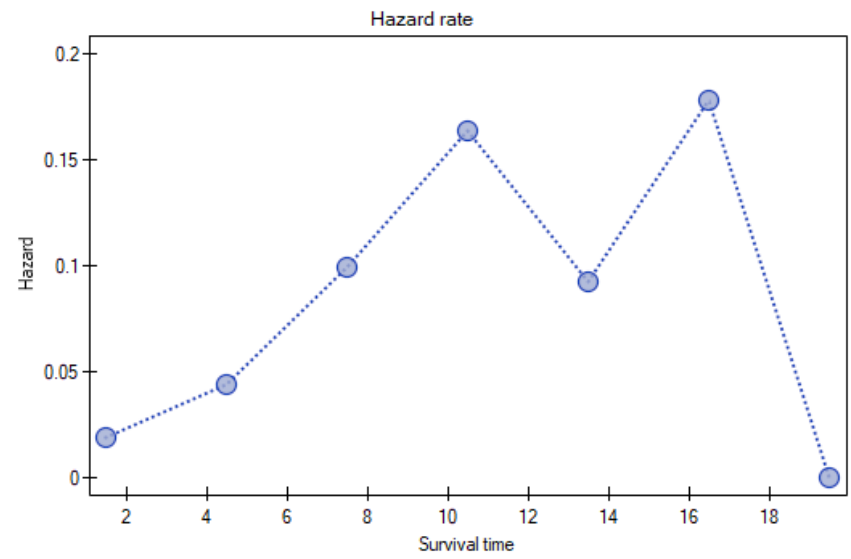


The survival function declines gradually, without a sudden drop

The probability of survival decreases with the time elapsed since the transplantation, but we do not observe a period in which the probability of death would rise sharply.



Copyright ©2010-2026 U-smile Stats – All rights reserved



Kaplan-Meier curves

WHEN TO USE:

Kaplan-Meier curves allow the evaluation of survival time without the need to arbitrarily group observations into intervals, as is the case with life tables. This estimator was introduced by Kaplan and Meier (1958)[94]. The graph of the Kaplan-Meier survival function has the form of a step function.

DEFINITIONS OF MEASURES

As with life tables, we calculate here the **survival function**, i.e. the probability of survival up to a given time. Confidence intervals around this curve are constructed on the basis of the standard error (Greenwood's formula) and a logarithmic transformation (log-log).

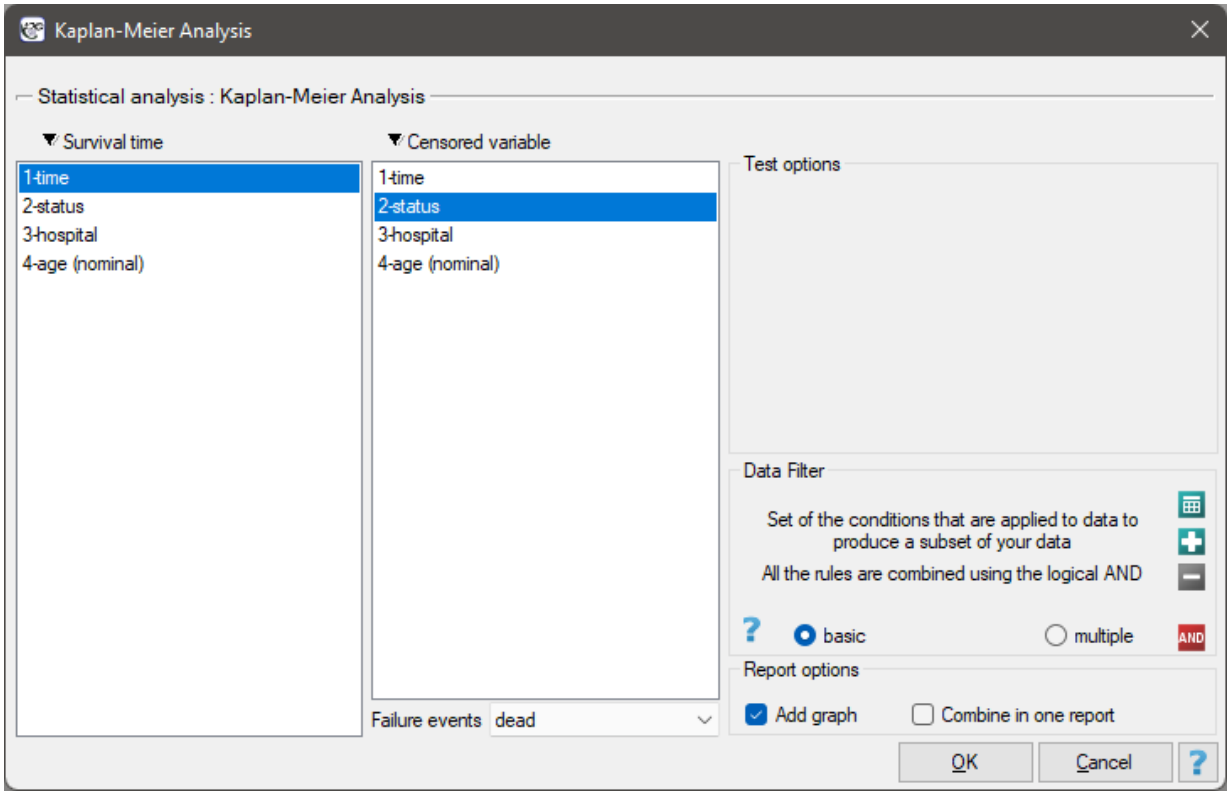
The point in time at which the value of the function equals 0.5 is the **survival time median**. This median indicates a 50% risk of death, i.e. it predicts that half of the patients will die within the indicated time. Both the median and other percentiles are determined as the shortest survival time for which the survival function is less than or equal to the given percentile. For the median, a confidence interval is determined using the "test-based" method of Brookmeyer and Crowley (1982)[30].

The survival time mean is determined as the area under the survival curve. Data concerning survival time are usually strongly skewed, which is why in survival analysis the median is usually a better measure of central tendency than the mean.

HOW TO SET UP THE ANALYSIS

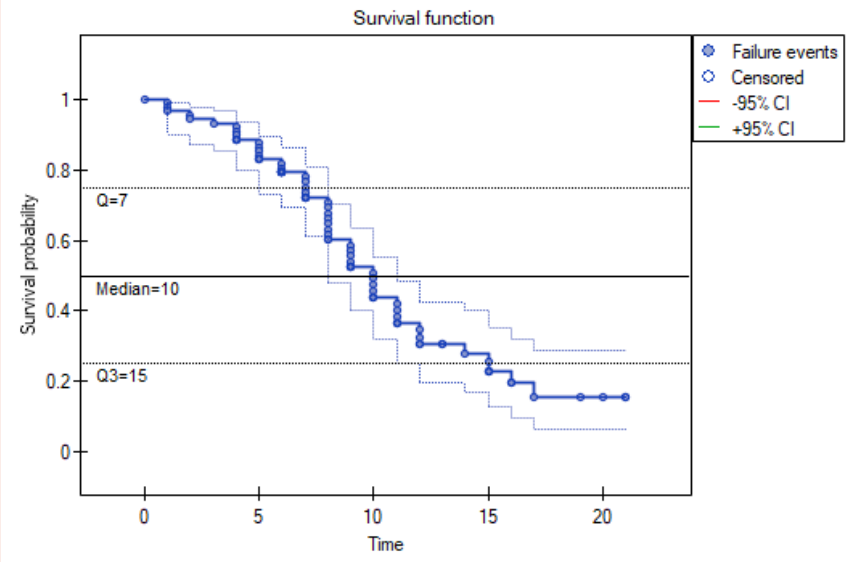
The window with the settings for the Kaplan-Meier curve is accessed via the menu Statistics→Survival analysis→Kaplan-Meier Analysis.

Copyright ©2010-2026 U-smile Stats – All rights reserved



CALCULATION AND INTERPRETATION

Example (28.1) continued *(transplant.pqs file)*
We present the survival time after a liver transplantation using the Kaplan-Meier curve:



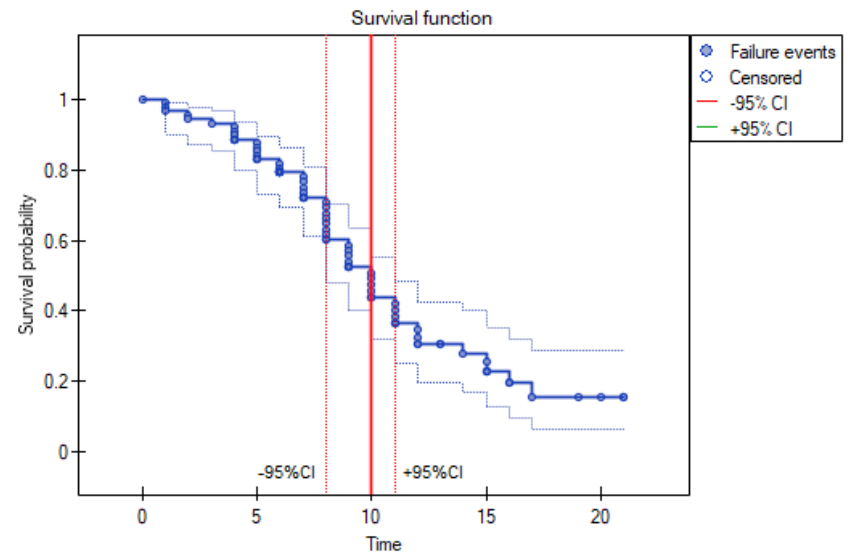
Copyright ©2010-2026 U-smile Stats – All rights reserved

Kaplan-Meier Analysis	
Analysed variables	time
	status
Number of unspecified	0
Number of missing data	0
Censored variable	status
Frequency	89
Failure events	dead
Frequency	53
Percent	59.551%
Censored	alive
Frequency	36
Percent	40.449%
Survival time	
Lower quartile	7
Median	10
-95% CI	8
+95% CI	11
Upper quartile	15
Mean	10.9549

The survival function does not drop sharply right after the transplantation. We conclude from this that the initial period after transplantation does not carry a particular risk of death.

Median survival time $= 10$ years

Within 10 years of the transplantation we expect that death will occur in half of the patients. This value is marked on the graph by drawing a line at the point 0.5 corresponding to the median. The quartiles are marked on the graph in a similar way.



We can visualize the confidence interval for the median on the graph by drawing vertical lines based on the confidence interval around the curve and a line at the 0.5 level.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Comparison of survival curves

WHEN TO USE:

Survival functions can be constructed separately for different subgroups, e.g. separately for women and men, and then compared with one another. Such a comparison may concern two or several curves at once; it can also examine whether the curves follow a natural trend (e.g. with age), and can adjust (correct) the analysis for the effect of an additional factor (a stratum).

A comparison of k survival curves $S_1, S_2, \ldots, S_k$ can be made in the program with the use of three tests:

the log-rank test – the best-known and most widely used, drawing on the Mantel-Haenszel procedure for many 2×2 tables (Mantel-Haenszel 1959[115], Mantel 1966[117], Cox 1972[49]),
Gehan's generalization of the Wilcoxon test – derived from the Wilcoxon test (Breslow 1970, Gehan 1965[70][71]),
the Tarone-Ware test – derived from the Wilcoxon test (Tarone and Ware 1977[162]).

All three tests are based on the same test statistic – they differ only in the **weights** w_j assigned to particular points on the timeline:

the log-rank test: $w_j = 1$ – all points on the timeline have the same weight, which gives later values of the timeline a greater influence on the result;
Gehan's generalization of the Wilcoxon test: $w_j = n_j$ – time points are weighted by the number of observations at each of them, so the initial values of the timeline carry greater weight;
the Tarone-Ware test: $w_j = \sqrt{n_j}$ – time points are weighted by the square root of the number of observations, which situates this test between the two above.

ASSUMPTIONS

- **hazard proportionality** – hazard, defined as the slope of the survival curve, is a measure of how quickly the event of interest occurs; violating this assumption does not entirely disqualify the tests above, but it lowers their power, the more so the earlier on the timeline the curves cross.

DEFINITIONS OF MEASURES

The Hazard Ratio (HR) describes the magnitude of the observed difference between a pair of survival curves, on the basis of the number of events observed and expected in each group:

$$HR = \frac{O_1/E_1}{O_2/E_2}$$

If $HR > 1$ (e.g. $HR = 2$), the degree of exposure to the event of interest in the first group is correspondingly greater than in the second group (twice as great for $HR = 2$); if $HR < 1$, the situation is reversed; $HR \approx 1$ means that both groups are exposed to a similar degree.

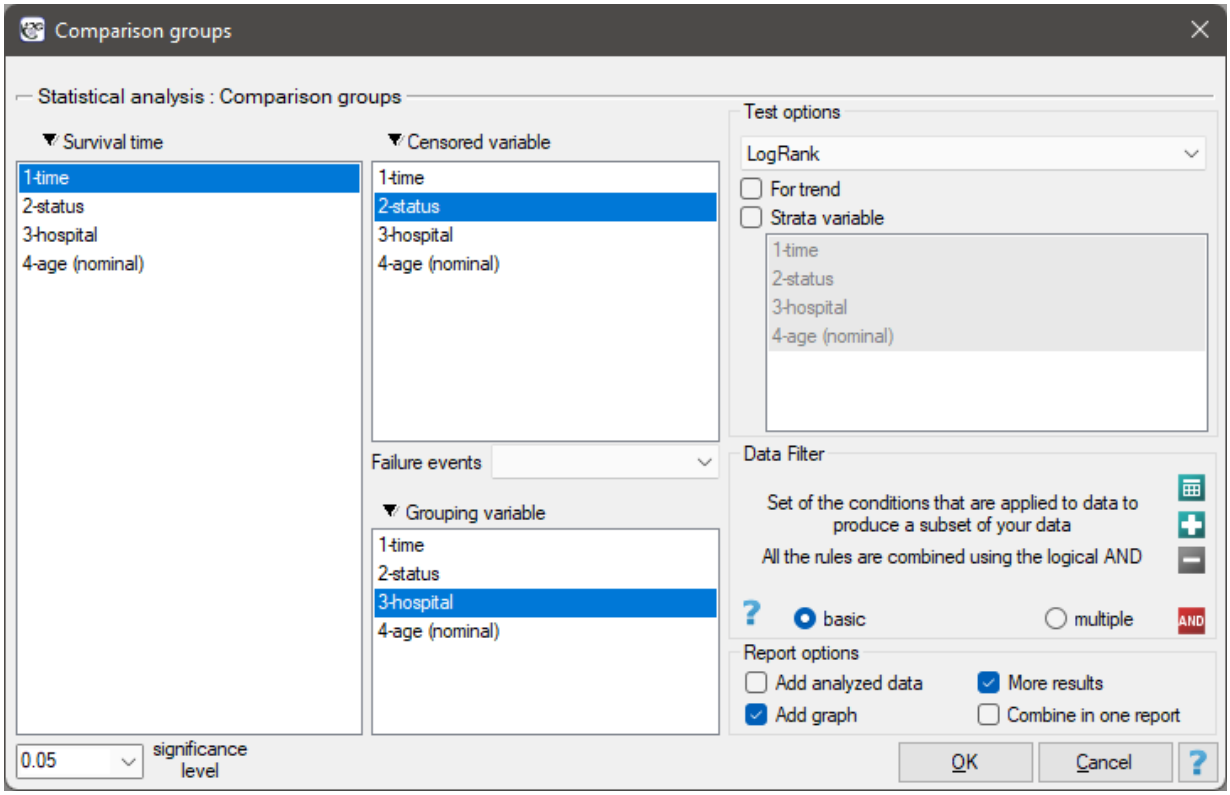
Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! The confidence interval for HR is calculated on the basis of the standard error of the logarithm of HR (Armitage and Berry 1994[11]).

The program also allows us to check whether the curves follow a **trend** consistent with the natural order of the compared groups (e.g. successive age categories), and to compare curves **taking strata into account**, i.e. after adjusting for an additional factor which may affect the result of the comparison (e.g. when comparing survival time with respect to age, separately within strata defined by sex).

HOW TO SET UP THE ANALYSIS

The window with the settings for the comparison of survival curves is accessed via the menu Statistics→Survival analysis→Comparison of groups.



CALCULATION AND INTERPRETATION

Example (28.1) continued *(transplant.pqs file)*
Differences between two survival curves

Liver transplantations were performed in two different hospitals. We will check whether patients' survival time after transplantation depended on the hospital in which the transplantation was performed. The comparison of the survival curves for these hospitals will be made using all the tests offered by the program for this purpose.
Hypotheses:

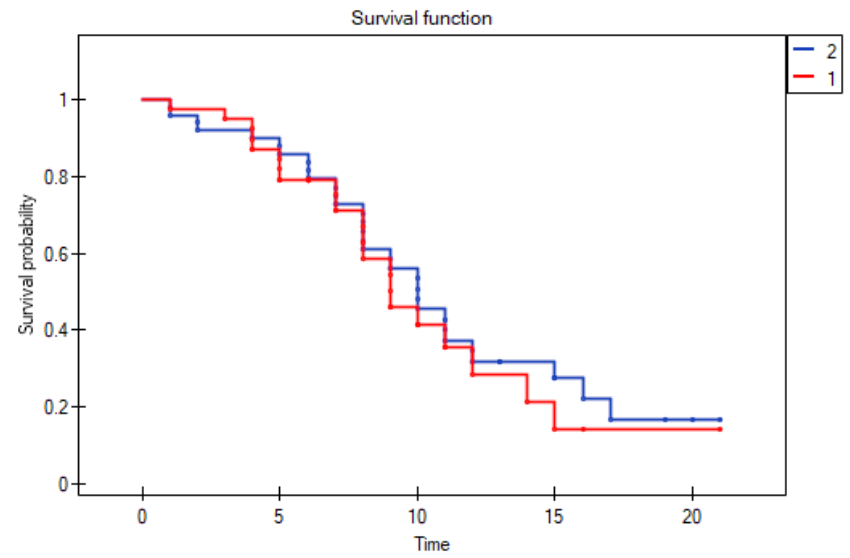
$\mathcal{H}_0$: the survival curve of hospital 1 patients $=$ the survival curve of hospital 2 patients,
$\mathcal{H}_1$: the survival curve of hospital 1 patients $\neq$ the survival curve of hospital 2 patients.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Comparison groups	
Analysed variables	time
	status
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	hospital
Frequency	89
Failure events	dead
Censored	alive
Test: LogRank	
Chi-square statistic	0.2744
Degrees of freedom	1
p-value	0.6004

Logrank			
Group	Obs.	Exp.	Obs./Exp.
2	32	33.7429	0.9483
1	21	19.2571	1.0905

Logrank			
Group	Hazard r.	-95%CI	+95%CI
1/2	1.1499	0.657	2.0126



$p = 0.6004$ for the log-rank test ($p = 0.6959$ for Gehan's, $p = 0.6465$ for Tarone-Ware); $HR = 1.1499$, 95%CI: $\langle$0.6570; 2.0126$\rangle$

There is no reason to reject $\mathcal{H}_0$ – the length of life of the patients in both hospitals is similar (the confidence interval for HR contains the value 1).

Differences among several survival curves

Liver transplantations were performed on people of different ages. Three age groups were distinguished: $\langle 45\text{years}; 50\text{years}\rangle$, $\langle 50\text{years}; 55\text{years}\rangle$, $\langle 55\text{years}; 60\text{years}\rangle$. We will check whether patients' survival time after transplantation depends on their age at the time of the transplantation.
Hypotheses:

Copyright ©2010-2026 U-smile Stats – All rights reserved

$\mathcal{H}_0$: the survival curves of patients aged $\langle 45\text{years}; 50\text{years})$, $\langle 50\text{years}; 55\text{years})$, $\langle 55\text{years}; 60\text{years})$ are similar,

$\mathcal{H}_1$: at least one survival curve, out of the 3 curves above, differs from the others.

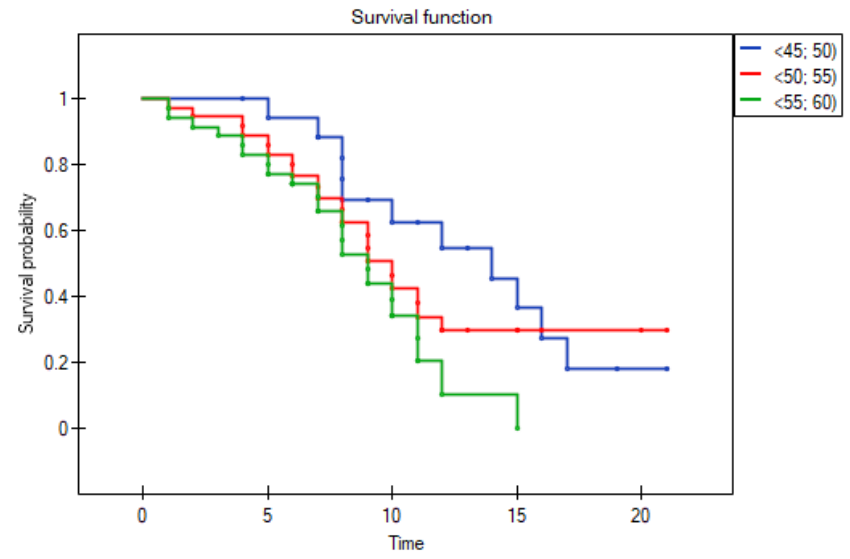
Comparison groups	
Analysed variables	time
	status
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	age (nominal)
Frequency	89
Failure events	dead
Censored	alive
Test: LogRank	
Chi-square statistic	5.3423
Degrees of freedom	2
p-value	0.0692

Logrank

Group	Obs.	Exp.	Obs./Exp.
<45; 50)	11	16.1201	0.6824
<50; 55)	20	21.4904	0.9306
<55; 60)	22	15.3895	1.4295

Logrank

Group	Hazard r.	-95%CI	+95%CI
<50; 55)/<45; !	1.3638	0.715	2.6015
<55; 60)/<45; !	2.0949	1.0419	4.2124
<55; 60)/<50; !	1.5361	0.7983	2.9557



$p = 0.0692$ for the log-rank test ($p = 0.0928$ for Gehan's, $p = 0.0779$ for Tarone-Ware)

There is no reason to reject $\mathcal{H}_0$ – the length of life of the patients from the three age groups is similar, although the p values are quite close to the significance level of 0.05.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Looking at the hazard values (the ratio of observed to expected events), we notice that they are slightly higher with each successive age group: $\langle 0.68, 0.93, 1.43 \rangle$. Although these differences are not statistically significant, they may suggest the existence of a trend of increasing hazard with age.

Trend among several survival curves

By introducing into the test information about the ordering of the compared categories (using the age variable, in which the age ranges are numbered 1, 2, and 3, respectively), we will check whether there is a trend in the compared curves.
Hypotheses:

$\mathcal{H}_0$: there is no trend in the survival time curves of patients after transplantation (a trend dependent on patients' age at the time of the transplantation),
$\mathcal{H}_1$: the older the patients at the time of the transplantation, the greater/smaller the probability of their survival over a given period of time.

For trend:	
Chi-square statistic	5.1135
Degrees of freedom	1
p-value	0.0237

$p = 0.0237$ for the log-rank test ($p = 0.0317$ for Gehan's, $p = 0.0241$ for Tarone-Ware)

The survival curves are arranged in a statistically significant trend: the older the patient at the time of the transplantation, the smaller the probability of surviving a given period of time.

The lowest curve on the Kaplan-Meier graph is that for people aged $\langle$55 years; 60 years); above it is the curve for patients aged $\langle$50 years; 55 years), and the highest curve is that for patients aged $\langle$45 years; 50 years).

Survival curves for strata

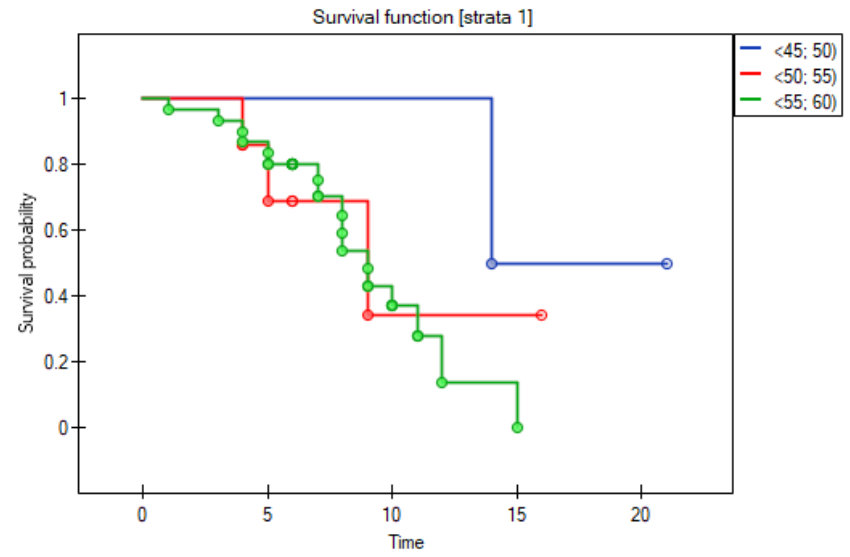
We will now check whether the trend observed earlier is independent of the hospital in which the transplantation was performed. For this purpose we will choose hospital as the stratum variable.

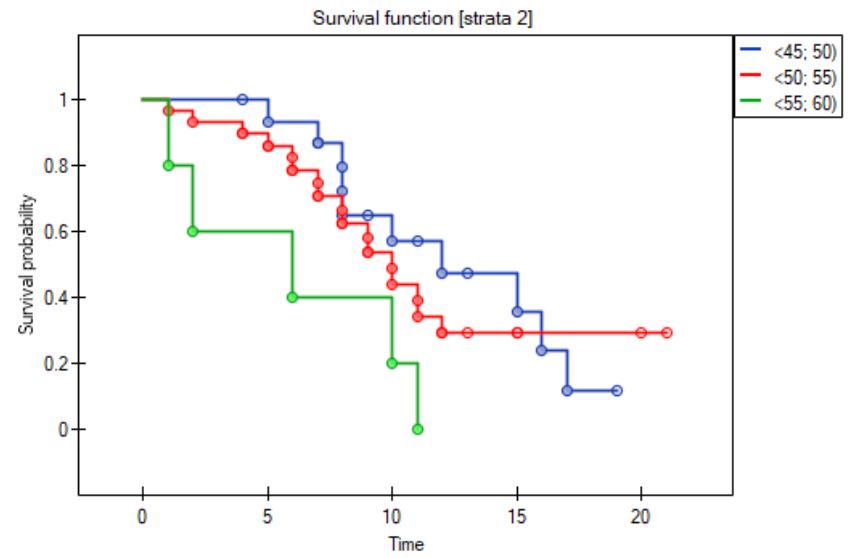
Copyright ©2010-2026 U-smile Stats – All rights reserved

Comparison groups	
Analysed variables	time
	status
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	age (nominal)
Strata variable	hospital
Frequency	89
Failure events	dead
Censored	alive
Test: LogRank	
Strata: 1	
Chi-square statistic	2.4137
Degrees of freedom	2
p-value	0.2991
For trend:	
Chi-square statistic	2.3571
Degrees of freedom	1
p-value	0.1247
Strata: 2	
Chi-square statistic	5.5427
Degrees of freedom	2
p-value	0.0626
For trend:	
Chi-square statistic	3.0283
Degrees of freedom	1
p-value	0.0818
Common for stratas	
Chi-square statistic	6.2594
Degrees of freedom	2
p-value	0.0437
For trend:	
Chi-square statistic	5.3744
Degrees of freedom	1
p-value	0.0204

Logrank				
Strata	Group	Obs.	Exp.	Obs./Exp.
1	<45; 50)	1	3.1485	0.3176
1	<50; 55)	3	3.3935	0.884
1	<55; 60)	17	14.458	1.1758
2	<45; 50)	10	12.5065	0.7996
2	<50; 55)	17	17.4904	0.972
2	<55; 60)	5	2.0031	2.4961
Common	<45; 50)	11	15.655	0.7027
Common	<50; 55)	20	20.8839	0.9577
Common	<55; 60)	22	16.4611	1.3365

Copyright ©2010-2026 U-smile Stats – All rights reserved





The report first presents an analysis of the individual strata, both the test results and the hazard values. In the first stratum an increasing trend in hazard is visible, though not statistically significant; a trend in the same direction (bordering on significance) is observed in the second stratum. Pooling these trends in a joint stratified analysis allowed us to obtain a significant trend in the survival curves.

Stratified analysis: the trend is significant regardless of hospital

The older the patient at the time of the transplantation, the smaller the probability of surviving a given period of time – regardless of the hospital where the transplantation was performed.

A comparative analysis of the survival curves, adjusted for strata, gives a significant result for the log-rank and Tarone-Ware tests and a non-significant result for Gehan's test, which may indicate that the differences between the curves are not as visible in the early periods of survival time as in the later ones. Looking at the hazard ratio for the curves compared in pairs

Copyright ©2010-2026 U-smile Stats – All rights reserved

Logrank

Strata	Group	Hazard r.	-95%CI	+95%CI
1	<50; 55)/<45; 50)	2.7834	0.6005	12.9012
1	<55; 60)/<45; 50)	3.702	1.0941	12.5261
1	<55; 60)/<50; 55)	1.3301	0.4078	4.3382
2	<50; 55)/<45; 50)	1.2156	0.5883	2.5119
2	<55; 60)/<45; 50)	3.1218	0.7024	13.8742
2	<55; 60)/<50; 55)	2.5682	0.5952	11.0803
Common	<50; 55)/<45; 50)	0.7337	0.381	1.4128
Common	<55; 60)/<45; 50)	0.5257	0.2632	1.0502
Common	<55; 60)/<50; 55)	0.7166	0.3756	1.3671

we can localize the significant differences. The smallest hazard ratio was obtained for the comparison of the curve for the youngest group with the curve for the oldest group: 0.53; the 95% confidence interval $\langle 0.26; 1.05 \rangle$ does contain the value 1, but is close to it, which may suggest significant differences between the corresponding curves. To confirm this supposition, the data filter in the analysis window can be used to compare the curves pairwise.

Data Filter

variable	condition	value
4-age	=	<45; 50)
4-age	=	<50; 55)

OR

Note! For pairwise comparisons, one of the corrections used for multiple comparisons should be applied and the significance level modified. Here, with Bonferroni's correction for three comparisons, the significance level will be 0.017. For simplicity we will use only the log-rank test.

$\langle 45$ years; 50 years) vs $\langle 50$ years; 55 years)

Common for stratas	
Chi-square statistic	0.5884
Degrees of freedom	1
p-value	0.443

$\langle 45$ years; 50 years) vs $\langle 55$ years; 60 years)

Common for stratas	
Chi-square statistic	8.9447
Degrees of freedom	1
p-value	0.0028

$\langle 50$ years; 55 years) vs $\langle 55$ years; 60 years)

Common for stratas	
Chi-square statistic	2.2412
Degrees of freedom	1
p-value	0.1344

As expected, statistically significant differences concern only the survival curves for the youngest and oldest age groups.

Copyright ©2010-2026 U-smile Stats – All rights reserved

FOR THOSE INTERESTED

Differences among survival curves
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad S_1(t) = S_2(t) = ... = S_k(t), \quad \text{for all } t, \\ \mathcal{H}_1: &\quad \text{not all } S_i(t) \text{ are equal.} \end{aligned}$$

Formulas:

Test statistic (chi-square):

$$\chi^2 = U' V^{-1} U$$

where:

$U_i = \sum_{j=1}^{m} w_j(d_{ij} - e_{ij})$,

V – the covariance matrix of dimensions $(k-1) \times (k-1)$:

diagonal: $\sum_{j=1}^{m} w_j^2 \frac{n_{ij}(n_j - n_{ij}) d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$,

off-diagonal: $\sum_{j=1}^{m} w_j^2 \frac{n_{ij} n_{lj} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$,

m – the number of time points at which the event of interest (death) occurred,

$d_j = \sum_{i=1}^{k} d_{ij}$ – the observed number of events in the j-th time point,

d_{ij} – the observed number of events in the i-th group in the j-th time point,

$e_{ij} = \frac{n_{ij} d_j}{n_j}$ – the expected number of events in the i-th group in the j-th time point,

$n_j = \sum_{i=1}^{k} n_{ij}$ – the number of cases at risk in the j-th time point.

This statistic asymptotically (for large expected counts) has the χ^2 distribution with $df = k - 1$ degrees of freedom.

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Trend among survival curves
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \text{there is no trend, in the studied population, in the placement of curves } S_1, S_2, ..., S_k, \\ \mathcal{H}_1: &\quad \text{there is a trend, in the studied population, in the placement of curves } S_1, S_2, ..., S_k. \end{aligned}$$

Formulas:

Test statistic (chi-square):

$$\chi^2 = \frac{(c'U)^2}{c'Vc}$$

where $c = (c_1, c_2, ..., c_k)$ – a vector of weights for the compared groups, informing about their natural order (usually consecutive natural numbers). This statistic asymptotically (for large expected counts) has the χ^2 distribution with one degree of freedom.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! To be able to perform a trend analysis in the survival curves, the grouping variable must be a numerical variable whose values indicate the natural order of the groups – these numbers are treated in the analysis as the weights $c_1, c_2, ..., c_k$.

Survival curves for strata
Hypotheses for differences among survival curves:

$$\begin{aligned} \mathcal{H}_0: &\quad S_1^*(t) = S_2^*(t) = ... = S_k^*(t), \quad \text{for all } t, \\ \mathcal{H}_1: &\quad \text{not all } S_i^*(t) \text{ are equal.} \end{aligned}$$

Hypotheses for the trend analysis in survival curves:

$$\begin{aligned} \mathcal{H}_0: &\quad \text{there is no trend, in the studied population, in the placement of curves } S_1^*, S_2^*, ..., S_k^*, \\ \mathcal{H}_1: &\quad \text{there is a trend, in the studied population, in the placement of curves } S_1^*, S_2^*, ..., S_k^*, \end{aligned}$$

where $S_1^*(t), S_2^*(t), ..., S_k^*(t)$ – the survival curves after adjustment for the variable defining the strata. The calculations for the test statistics are based on the formulas described for the tests without strata, with the difference that matrices U and V are replaced with the sums of matrices $\sum_{l=1}^{L} U$ and $\sum_{l=1}^{L} V$, summed over the strata created by the adjusting variable ($l = 1, 2, ..., L$).

The p value, determined on the basis of the test statistic, is compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Cox proportional hazards regression

WHEN TO USE:

Cox regression, also known as the Cox proportional hazards model (Cox PH model) – Cox D.R. (1972)[49], is the most widely used regression approach in survival analysis. It allows the study of the simultaneous influence of many independent variables (X_1, X_2, ..., X_k) on survival time, e.g. the influence of the treatment method, age and other prognostic factors on lifetime. It is, in a sense, a non-parametric approach – it does not require knowledge of the shape of the hazard or survival function, which is the source of its popularity.

ASSUMPTIONS

- **hazard proportionality** – the effect of each independent variable on the hazard must be constant over time (see *Analysis of model residuals* below);
- no strong collinearity of the independent variables;
- a sufficiently large number of observations relative to the number of estimated parameters.

Note! The number of observations should be at least ten times greater than the ratio of the number of

Copyright ©2010-2026 U-smile Stats – All rights reserved

estimated model parameters (k) to the smaller of the proportions of censored or complete observations (p), i.e. $n \geq 10k/p$ (Peduzzi et al. 1995[135]).

Note! The independent variables should not be collinear – collinearity leads to uncertain estimation and very high values of the errors.

DEFINITIONS OF MEASURES

The function on which the model is based describes the resulting hazard as the product of two quantities, only one of which depends on time (t):

$$h(t, X_1, X_2, ..., X_k) = h_0(t) \cdot \exp\left(\sum_{i=1}^{k} \beta_i X_i\right),$$

where $h(t, X_1, ..., X_k)$ is the resulting hazard (the risk changing over time, depending on other factors, e.g. the treatment method), $h_0(t)$ is the baseline hazard (the hazard assuming that all explanatory variables are equal to zero), and $\sum_{i=1}^{k} \beta_i X_i$ is a (usually linear) combination of the independent variables $X_1, ..., X_k$ and the model parameters $\beta_1, ..., \beta_k$, called **regression coefficients**. The preparation of dummy variables and interactions is discussed in chapter 25.0.2 Preparation of variables for analysis in multidimensional models.

The coefficients are estimated with the so-called **„partial" maximum likelihood method** (the iterative Newton-Raphson algorithm) – „partial" because the search for the maximum of the likelihood function L is performed directly only for the complete data, while the censored data are taken into account indirectly. The correction for tied ranks is based on Breslow's method[28].

Note! The convergence criterion of the iterative Newton-Raphson algorithm can be controlled with two parameters: the limit of the number of iterations and the convergence criterion (the value below which the improvement of the estimation is considered insignificant, which ends the algorithm).

Hazard Ratio (HR) – for each independent variable a unit Hazard Ratio (HR) is calculated:

$$HR_i = e^{\beta_i}.$$

It expresses the change in the risk of the unfavorable event when a given independent variable increases by 1 unit, with the remaining variables of the model held constant. Interpretation:

- $HR > 1$ – a stimulating effect of the variable: by how much the risk of the event increases when the independent variable increases by one level;
- $HR < 1$ – an inhibiting effect of the variable: by how much the risk of the event decreases when the independent variable increases by one level;
- $HR \approx 1$ – the variable has no effect on the occurrence of the event.

Note! If the analysis is performed for a model other than linear, or an interaction is included, then – analogously to the logistic regression model – the appropriate HR is calculated by changing the formula of the combination of independent variables.

Model verification – on the basis of a coefficient and its standard error we infer whether a given independent variable has a significant effect on the dependent variable; for that purpose we use the Wald test, which verifies the hypothesis $\mathcal{H}_0 : \beta_i = 0$ (equivalently $HR_i = 1$) against $\mathcal{H}_1 : \beta_i \neq 0$ ($HR_i \neq 1$).

The quality of the whole model, in turn, is assessed with several general measures:

Copyright ©2010-2026 U-smile Stats – All rights reserved

- **information criteria** (AIC, $AICc$, BIC) – they are based on the information entropy carried by the model, i.e. they estimate the information lost when a given model describes the studied phenomenon; we choose the model with the **minimum** value of a given criterion. They are a kind of compromise between goodness of fit and model complexity – they grow with the number of model parameters, the faster the smaller the number of observations. $AICc$ is a correction of AIC recommended for small samples, and BIC additionally penalizes model complexity more strongly (Volinsky and Raftery 2000[168]). An information criterion is not an absolute measure – if all the compared models describe reality poorly, no criterion will give us that warning.

- **Pseudo R**^{2} (McFadden R^2) – a measure of model fit, the counterpart of the coefficient of multiple determination R^2 from multiple linear regression; it takes values from the range $\langle 0; 1)$, where values close to 1 mean good fit and 0 – a complete lack of fit. Because of its sensitivity to the number of variables in the model, its corrected versions are also determined: $R^2_{Nagelkerke}$ and $R^2_{Cox-Snell}$.

- **the likelihood ratio test** – the basic tool for assessing the significance of all model variables jointly; it verifies the hypothesis $\mathcal{H}_0$: all $\beta_i = 0$ against $\mathcal{H}_1$: there exists $\beta_i \neq 0$.

- **time-dependent AUC** – in survival analysis the program determines the **time-dependent AUC** (Heagerty et al. 2000[83], Blanche et al. 2013[21]): for each time t at which the event of interest occurred, the observations are divided into **cases** (the event up to time t) and **controls** (observed longer than t, without the event), and the marker is the linear combination of the variables and model parameters. $AUC(t)$ is the probability that a randomly chosen case has a higher marker value than a randomly chosen control; censoring is accounted for by weighting the observations with the inverse probability of censoring (IPCW). The interpretation is the same as for the classical AUC: 0.5 corresponds to random discrimination, 1 – to perfect discrimination, and values above approx. 0.7–0.8 are usually considered good discriminatory ability. Because $AUC(t)$ is a function of time, the report gives the mean $AUC(t)$ and its values at the quartiles and at the end of the event times, and the graph of $AUC(t)$ against time (the ROC curve option in the Add graph window) shows whether the model's ability to predict events is stable across different prediction horizons or changes with time. For the latest times, when only few observations remain in the set of controls, the $AUC(t)$ estimator becomes unstable – jumps at the very end of the graph should not be over-interpreted. No cut-off point and no significance test are determined for $AUC(t)$.

Analysis of model residuals allows the verification of the model assumptions – above all the localization of outliers and the assessment of hazard proportionality (a number of graphical methods for assessing the adequacy of the proportional hazards model have been developed – Lee and Wang 2003[103]). Because of censored observations, residuals are not calculated here in the standard way (as the difference between the observed and predicted value), but as one of three kinds: **Martingale**, **Deviance** or **Schoenfeld**, which can be plotted against time or against the independent variables.

Martingale – interpreted as the difference, over the time $[0, t]$, between the observed number of unfavorable events and the number predicted by the model. The expected value is 0, but the distribution is skewed (from $-\infty$ to 1), which makes the graph harder to interpret.

Deviance – like Martingale residuals, they asymptotically have the expected value 0, but for a well-fitted model they are distributed symmetrically around zero with a standard deviation of 1. A positive value means that the subject survived shorter than the model predicted, a negative one – longer. This is above all a tool for identifying outliers (in the report, residuals further than 3 standard deviations from 0 are marked in red).

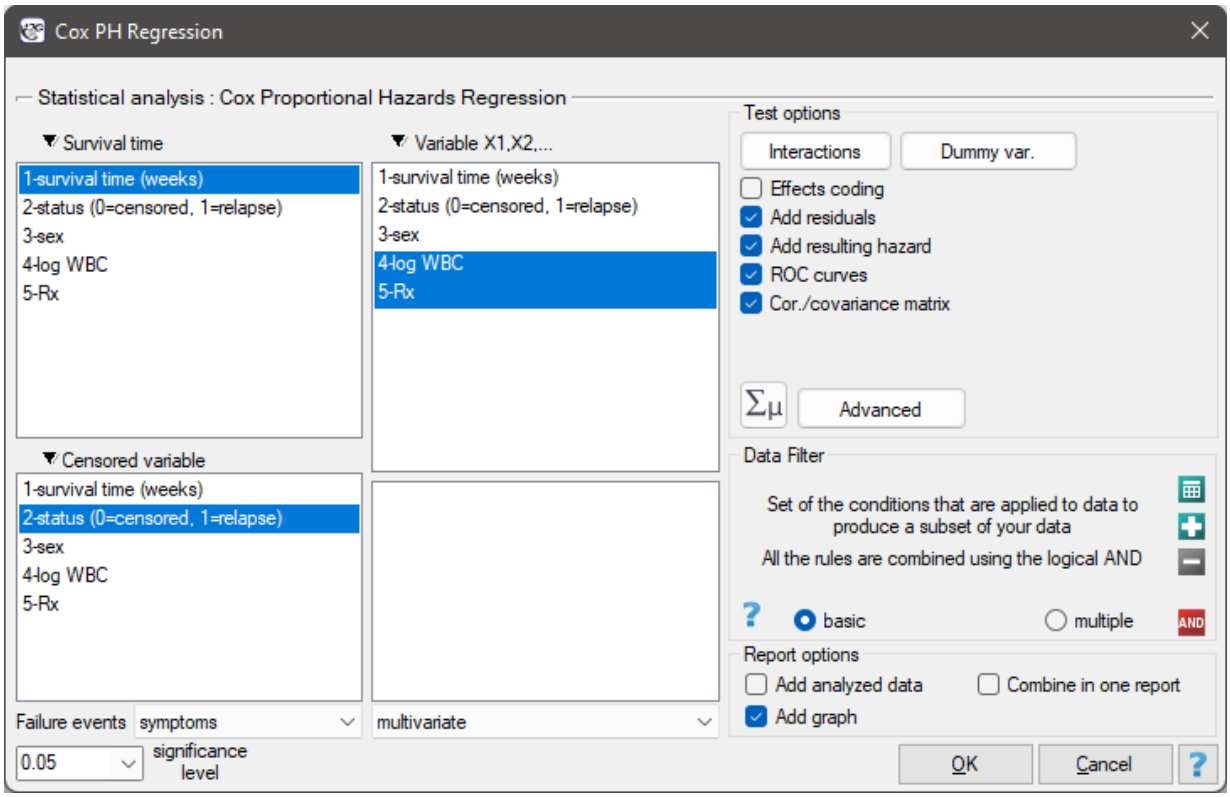
Copyright ©2010-2026 U-smile Stats – All rights reserved

Schoenfeld – calculated separately for each independent variable, only for complete observations; their sum and expected value are 0. They allow the identification of the variable violating the hazard proportionality assumption – it is the one whose residual plot against time forms a systematic pattern (usually the linear dependence of the residuals on time is examined). An even distribution of points around 0 indicates that the hazard proportionality assumption is fulfilled by a given variable.

Note! When the hazard proportionality assumption is not fulfilled for some model variable, a solution may be to perform the Cox analysis separately for each level of that variable (stratification) or to extend the model with an appropriate time-dependent variable.

HOW TO SET UP THE ANALYSIS

The window with settings for Cox regression is accessed via the menu Statistics→Survival analysis→Cox PH Regression.



FOR THOSE INTERESTED

Standard error of a coefficient

$$SE_b = \sqrt{diag(H^{-1})_b},$$

where $diag(H^{-1})$ is the main diagonal of the covariance matrix.

Wald test for a single coefficient

Hypotheses: $\mathcal{H}_0 : \beta_i = 0$ against $\mathcal{H}_1 : \beta_i \neq 0$ (equivalently $\mathcal{H}_0 : HR_i = 1$ against $\mathcal{H}_1 : HR_i \neq 1$).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Formulas: The test statistic:

$$\chi^2 = \left(\frac{b_i}{SE_{b_i}}\right)^2$$

asymptotically (for large sample sizes) has the χ^2 distribution with 1 degree of freedom.

On the basis of the test statistic, the p value is determined and compared with the significance level α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Information criteria (based on L_{FM} – the maximum of the likelihood function of the full model, L_0 – the maximum of the likelihood function of the model with the intercept only, k – the number of parameters, d – the number of unfavorable events):

$$AIC = -2\ln L_{FM} + 2k, \qquad AICc = AIC + \frac{2k(k+1)}{d-k-1}, \qquad BIC = -2\ln L_{FM} + k\ln(d).$$

Pseudo R^2

$$R^2_{Pseudo} = 1 - \frac{\ln L_{FM}}{\ln L_0}, \qquad R^2_{Nagelkerke} = \frac{1 - e^{-(2/d)(\ln L_{FM} - \ln L_0)}}{1 - e^{(2/d)\ln L_0}}, \qquad R^2_{Cox-Snell} = 1 - e^{\frac{(-2\ln L_0) - (-2\ln L_{FM})}{d}}.$$

Likelihood ratio test for all model variables
Hypotheses: $\mathcal{H}_0$: all $\beta_i = 0$ against $\mathcal{H}_1$: there exists $\beta_i \neq 0$. **Formulas:** The test statistic:

$$\chi^2 = -2\ln(L_0/L_{FM}) = -2\ln(L_0) - (-2\ln(L_{FM}))$$

asymptotically (for large sample sizes) has the χ^2 distribution with k degrees of freedom.

On the basis of the test statistic, the p value is determined and compared with α: if $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; if $p > \alpha$, there is no reason to reject $\mathcal{H}_0$.

Time-dependent AUC
For the marker $M = \sum_{i=1}^{k} b_i X_i$ and a time t at which the event of interest occurred, the *cumulative/dynamic* estimator with IPCW weights (Heagerty et al. 2000[83], Blanche et al. 2013[21]) has the form:

$$AUC(t) = \frac{\sum_i \sum_j w_i(t)\, w_j(t) \left[\mathbf{1}(M_i > M_j) + \frac{1}{2}\mathbf{1}(M_i = M_j)\right]}{\sum_i w_i(t) \sum_j w_j(t)},$$

where the summation over i covers the cases ($T_i \leq t$, $\delta_i = 1$) with weight $w_i(t) = 1/\hat{G}(T_i^-)$, and over j – the controls ($T_j > t$) with weight $w_j(t) = 1/\hat{G}(t)$; $\hat{G}$ is the Kaplan-Meier estimator of the censoring distribution (in which censoring of an observation is treated as the „event"). $AUC(t)$ is calculated at all unique times of the event of interest, and the mean $AUC(t)$ given in the report is the arithmetic mean of these values.

Comparison of Cox PH regression models

WHEN TO USE:

Because of the possibility of simultaneous analysis of many independent variables in one Cox regression model, the problem of choosing the optimal model arises. When selecting independent variables, one

Copyright ©2010-2026 U-smile Stats – All rights reserved

should remember that the model should contain variables strongly related to survival time and weakly correlated with one another. Comparing models with different numbers of variables allows us to assess whether adding (or removing) a given variable significantly improves the model.

DEFINITIONS OF MEASURES

When comparing models we pay attention to the information criteria (AIC, $AICc$, BIC) and the model fit coefficients (R^2_{Pseudo}, $R^2_{Nagelkerke}$, $R^2_{Cox-Snell}$), as well as to the result of the **likelihood ratio test**, which verifies whether the maximum of the likelihood function of the full model (L_{FM}) differs significantly from that of the reduced model (L_{RM}).

The decision about the choice of the model is made on the basis of all the above quantities jointly. If the compared models do not differ significantly, we should choose the one with fewer variables – this means that the variables present in the full model and absent in the reduced one do not carry significant information. If the difference is statistically significant, the model with more variables (with the greater R^2) is significantly better.

In PQStat the comparison of models can be performed in two ways:

- **manually** – the researcher builds the **full** model (with more variables) and the **reduced** one (created from the full model by removing the variables superfluous from the point of view of the studied phenomenon), and then, based on the results of the comparison, chooses the better of them;
- **automatically** – the program builds successive models itself and indicates the best one according to the chosen criterion (described below).

Automatic model selection

Model terms. The unit that the automatic methods add to or remove from the model is a term: a single variable, a whole dummy variable (all its 0–1 columns at once) or an interaction. The hierarchy principle applies: an interaction can be in the model only when its main effects are in it, and only the interaction can be removed before them.

The selection criterion is chosen in the analysis window:

- AIC, AICc or BIC – information criteria; the model with the lower value of the criterion is better. A move (adding or removing a term) is accepted only when it lowers the criterion;
- p of the nested test – the likelihood ratio test comparing the model before and after the change, with a given p threshold. A term is removed when p is greater than the threshold, and added when p is smaller than the threshold. The same p value is shown in the report as the comparison of neighbouring models;
- validation error – the value of $-2 \ln PL$ per event determined out of sample (described below: Out-of-sample evaluation of models). A move is accepted only when it reduces this error. The criterion requires the Validation option to be switched on – choosing it while validation is off switches cross-validation on.

Selection methods:

- **backward elimination** – starts from the model with all the variables; in each step the term whose removal lowers the criterion the most is removed; it ends when no removal lowers the criterion any further;

Copyright ©2010-2026 U-smile Stats – All rights reserved

- **forward selection** – starts from the null model (no variables, the baseline hazard only); in each step the term that lowers the criterion the most is added; it ends when no addition lowers the criterion any further;
- **bidirectional stepwise** – starts from the null model; in each step both additions and removals of terms are considered and the best move is chosen; it ends when no move improves the criterion;
- **all subsets** – all models consistent with the hierarchy principle are fitted (a maximum of 15 terms), and the best model of each size is presented in the report; this method uses the information criteria or the validation error (the p criterion is not available in it). With many variables the computation may take longer, because the number of fitted models grows exponentially.

All the models are computed on the same set of observations – rows with missing values in any of the candidate variables are discarded once, before the analysis. Thanks to this, the information criteria and the likelihood ratio tests are comparable between models.

The report contains: the selection method and criterion, the number of terms and of fitted models, the information criteria of the null model, the successive models of the selection path (each with the value of $-2\ln L$, the AIC, $AICc$ and BIC criteria, the R^2_{Pseudo}, $R^2_{Nagelkerke}$ and $R^2_{Cox-Snell}$ coefficients, and the table of coefficients and hazard ratios) and the comparison of each neighbouring pair of models with the likelihood ratio test (the χ^2 statistic, degrees of freedom, p value). The best model according to the chosen criterion is marked in the report. If no term improves the null model, the report states that the null model is the best. With the all subsets method, neighbouring models may not be nested – in that case the test is not performed.

Note! The p values and confidence intervals in an automatically selected model are too optimistic, because the same data set was used both to select the variables and to estimate their effects. Automatic selection should be treated as an aid in searching for a model, and the final choice should be based on subject-matter knowledge. The selected model should be built again in the Cox PH Regression window and the proportional hazards assumption verified.

Out-of-sample evaluation of models

The information criteria, the fit coefficients and the tests are determined on the same observations on which the coefficients of the models were estimated, which makes the assessment of the models too optimistic – the more so, the more variables the model contains. For this reason, as in comparing multiple regression models, the analysis window makes it possible to evaluate the models on observations that did not take part in fitting them (the Validation option): k**-fold cross-validation** (each model is additionally fitted k times, each time without one of the k parts of the data) or a **random test set** (the models are fitted on the training part and evaluated on the test part set aside). The split is random, determined by the Seed, and **stratified by status**: complete and censored observations are split separately, so that the share of events is similar in every part.

For each model the report then gives two additional measures:

- $-2\ln PL$ **per event** – the logarithm of the partial likelihood determined out of sample, multiplied by -2 and divided by the number of events. With a test set it is the partial likelihood of the test observations for the coefficients estimated on the training set, and in cross-validation – the cross-validated partial likelihood of Verweij and van Houwelingen[169]. The lower the value, the better the model. This is the **validation error** – the best model by validation is indicated according to it, and the validation error selection criterion is based on it;

Copyright ©2010-2026 U-smile Stats – All rights reserved

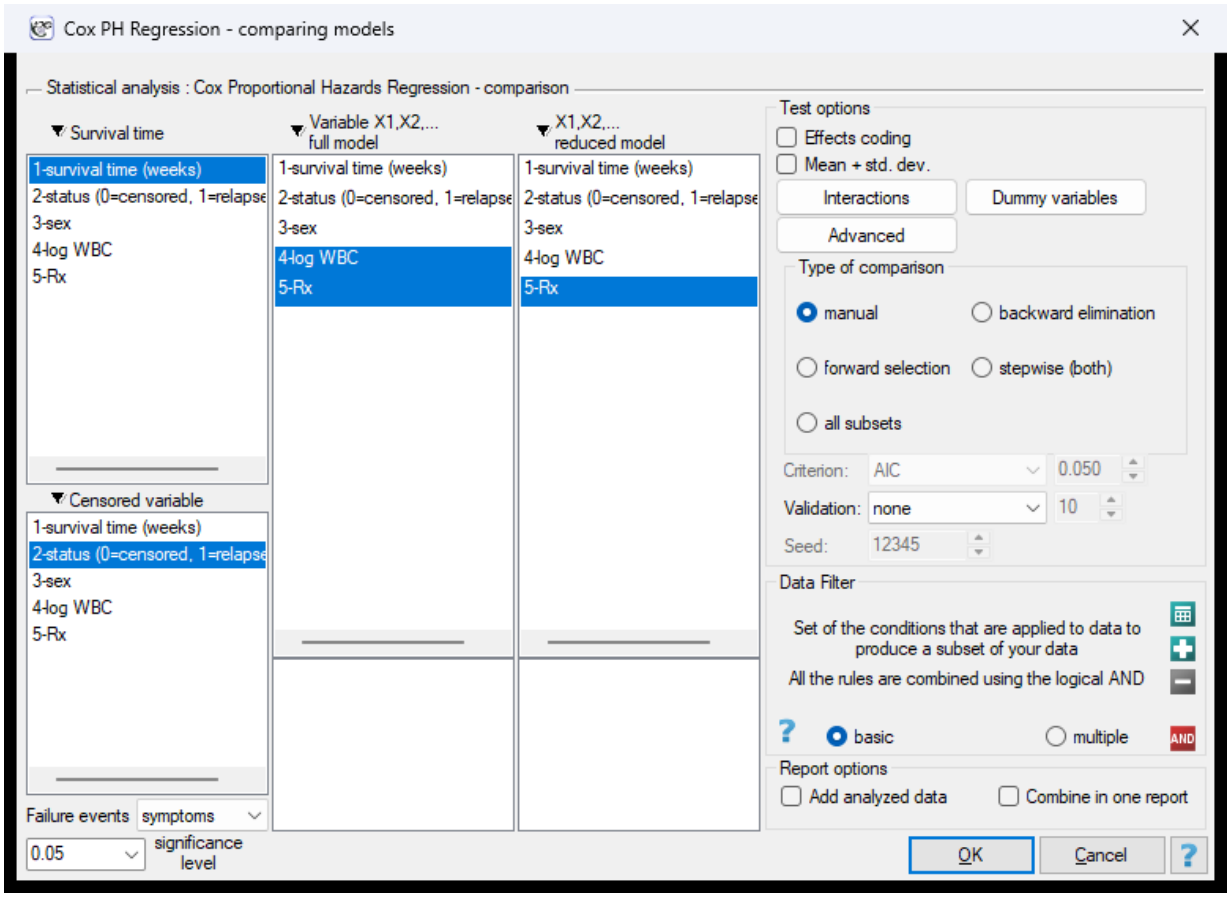
- **Harrell's C** – the concordance index[80]: the percentage of comparable pairs of observations evaluated out of sample in which the observation with the higher predicted risk experienced the event earlier (0.5 – random ordering, 1 – perfect ordering).

The point of reference is the **null model** (without variables – the same risk for all observations); its validation error is given in the model selection block of the report together with the **best model by validation**, indicated regardless of the selection criterion chosen. Out-of-sample evaluation also works in the manual comparison of models.

Note! When the validation error is the selection criterion, the validation error of the selected model is slightly too optimistic – out of many models, the one that performed best on the given split was chosen. With a small number of events, the validation results depend on the random split – it is then worth checking them for several values of the seed.

HOW TO SET UP THE ANALYSIS

The window with settings for model comparison is accessed via the menu Statistics→Survival analysis→Cox PH Regression – comparing models.



In the window we indicate the survival time, the censored variable with the Failure events value, and the independent variables of the full and the reduced model (in automatic selection – the candidate variables). Model selection options:

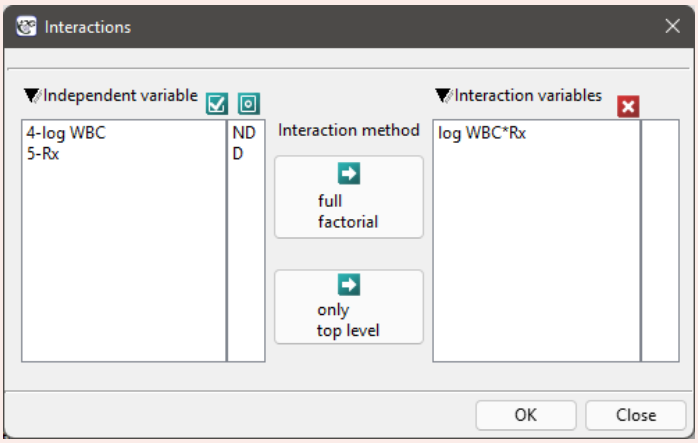
Copyright ©2010-2026 U-smile Stats – All rights reserved

- Type of comparison – manual or one of the automatic selection methods,
- Criterion – AIC, AICc, BIC, p (LR / F test) with the threshold given in the adjacent field, or validation error; in the manual mode the criterion is not used,
- Validation – none (evaluation on the training set only), k-fold CV, k = with the number of folds in the adjacent field (from 2 to 100, 10 by default) or test set, % = with the share of the test set as a percentage of the observations (from 5 to 90, 30 by default),
- Seed – the seed of the random split of the data (12345 by default); the same seed means the same split.

CALCULATION AND INTERPRETATION

Example 28.2. (remissionLeukemia.pqs file)
The analysis is based on the leukemia data described in the work of Freireich et al. 1963[66] and analyzed further by many authors, including Kleinbaum and Klein 2005[97]. The data contain information about the time (in weeks) of remaining in remission, until the patient was withdrawn from the study because of the end of remission (a relapse of symptoms) or because information about the patient was censored. The end of remission is the unfavorable event – it is treated as a **complete** observation. An observation is **censored** if the patient remains in the study to the end and remission does not end, or if the patient leaves the study.
Patients were assigned to two groups: treated traditionally (marked as 1, sometimes called the „placebo group") and treated with a new method (marked as 0). Information was collected about the patients' sex (1=man, 0=woman) and about the values of the white blood cell count indicator, marked as „log WBC", which is a known prognostic factor.

The aim of the study is to determine the influence of the treatment method on the time of remaining in remission, taking into account possible confounders and interactions. In the analysis we focus on the „Rx" variable (1=placebo, 0=new treatment); the „log WBC" variable is placed in the model as a possible confounder. To assess the possible effect of the interaction of „Rx" and „log WBC", we also consider a third variable being their product, added to the model with the Interactions button.



We build three Cox models:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Model A contains only the „Rx" variable

Model									
	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
Rx	1.5092	0.4096	0.7065	2.3119	13.5783	0.0002	4.5231	2.0268	10.0938

Model B contains the „Rx" variable and the potential confounder „log WBC"

Model									
	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
log WBC	1.6043	0.3293	0.9589	2.2498	23.7321	<0.0001	4.9746	2.6088	9.486
Rx	1.2941	0.4221	0.4668	2.1214	9.3989	0.0022	3.6476	1.5948	8.3426

Model C contains the „Rx" variable, the „log WBC" variable and the potential effect of their interaction: „Rx $\times$ log WBC"

Model									
	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
log WBC	1.8028	0.4467	0.9272	2.6783	16.2864	0.0001	6.0665	2.5275	14.5609
Rx	2.3549	1.681	-0.9398	5.6497	1.9625	0.1612	10.5375	0.3907	284.2005
log WBC*Rx	-0.3422	0.5197	-1.3609	0.6765	0.4335	0.5103			

The interaction variable of „Rx" and „log WBC" included in model C is not significant in it ($p = 0.5103$) according to the Wald test. A similar result is obtained by comparing, with the likelihood ratio test, model C with model B in the Cox PH Regression – comparing models module: the non-significant result ($p = 0.5134$) means that model C (with the interaction) is not significantly better than model B (without the interaction).

Chi-square - models comparison	0.4271
Degrees of freedom	1
p-value	0.5134

Comparison of models C and B: $p = 0.5134$ (not significant)

We reject model C – the interaction „Rx" $\times$ „log WBC" does not contribute significant information. We move on to the comparison of models B and A.

The HR for „Rx" in model B is 3.65, which means that the hazard for the „placebo group" is about 3.6 times greater than for the group treated with the new method. Model A, containing only the „Rx" variable, is usually called the „crude" model, because it ignores the effect of potential confounders – in this model the HR for „Rx" is already 4.52, i.e. considerably more than in model B. The 95% confidence interval of these estimators is also important here: for „Rx" in model A its width is 8.06 (10.09 minus 2.03), while in model B it is narrower: 6.74 (8.34 minus 1.60) – so model B gives a more precise estimation of HR than model A. To finally decide which of them estimates the treatment effect better, we compare them once more with the likelihood ratio test.

Comparison of models B and A: $p < 0.0001$ (significant)

Model B is significantly better than model A – it has the lowest values of the information criteria (AIC=148.6, AICc=149, BIC=151.4) and high values of fit (Pseudo $R^2_{McFadden} = 0.2309$, $R^2_{Nagelkerke} = 0.7662$, $R^2_{Cox-Snell} = 0.7647$).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Cox Proportional Hazards Regression - comparison	
Analysed variables	survival time (weeks)
	status (0=censored, 1=relap
	log WBC
	Rx
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	42
Number of variables in the model 1	2
Convergence criterion met	
-2 Log Likelihood	144.5585
AIC - Akaike criterion	148.5585
AICc - corrected Akaike criterion	149.003
BIC - Bayesian criterion	151.3609
Pseudo R2 (McFadden)	0.2309
R2 (Nagelkerke)	0.7662
R2 (Coxa-Snella)	0.7647
Number of variables in the model 2	1
Convergence criterion met	
-2 Log Likelihood	172.7592
AIC - Akaike criterion	174.7592
AICc - corrected Akaike criterion	174.9021
BIC - Bayesian criterion	176.1604
Pseudo R2 (McFadden)	0.0809
R2 (Nagelkerke)	0.3985
R2 (Coxa-Snella)	0.3977
Chi-square - models comparison	28.2007
Degrees of freedom	1
p-value	<0.0001

Out-of-sample evaluation. We will also check whether the advantage of model B holds for patients who did not take part in fitting the models. We repeat the comparison of models B and A, choosing in the Validation field the k-fold CV, k = option with 10 folds (seed 12345). In the report model B (full) is marked as Model 1, and model A (reduced) – as Model 2.

Out-of-sample evaluation	cross-validation
Number of folds (k)	10
Seed of the random split	12345
Null model: -2 ln PL / event	8.1196
Best model by validation (-2 ln PL / event)	Model 1

	training set	cross-validation ($k = 10$)	
model	AIC	$-2 \ln PL$ per event	Harrell's C
null		8.1196	
Model B (Rx, log WBC)	148.5585	6.7100	0.8871
Model A (Rx)	174.7592	7.6269	0.7258

Model A, containing only the treatment method, predicts the time of remaining in remission better than the null model (7.6269 versus 8.1196), but it is only the addition of „log WBC" that clearly reduces the validation error (to 6.7100) and raises the concordance of the predictions with the actual order of

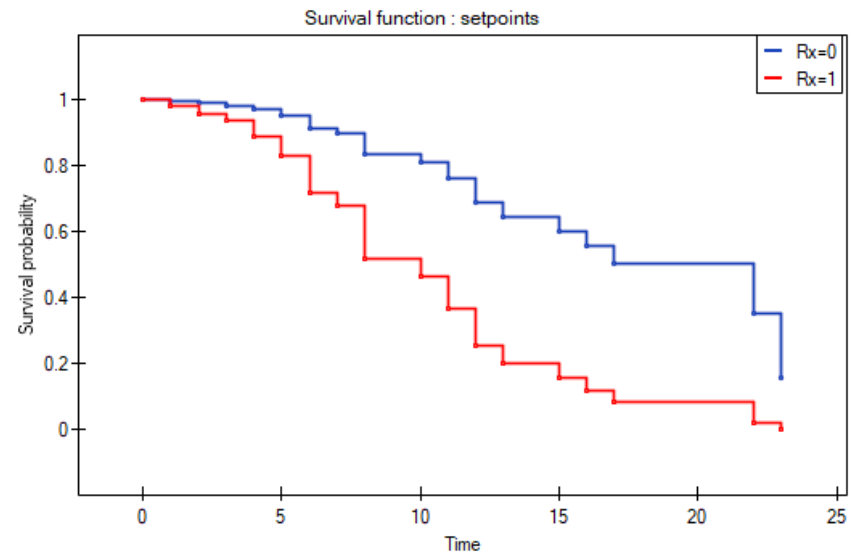
Copyright ©2010-2026 U-smile Stats – All rights reserved

events from $C = 0.7258$ to $C = 0.8871$. The advantage of model B is therefore not a result of fitting to the data on which it was built – it also holds for patients outside the training set.

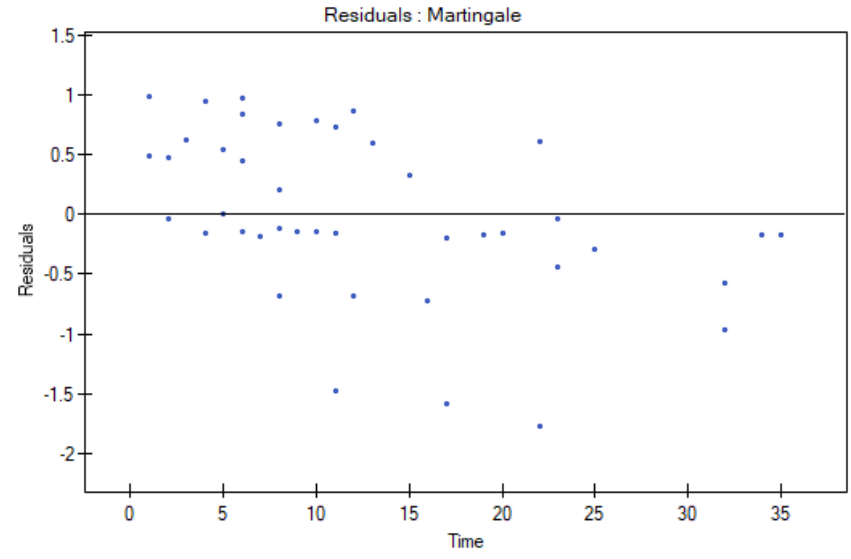
cross-validation ($k = 10$): $-2 \ln PL$ per event $= 6.71$ (model B), 7.63 (model A), 8.12 (null model); Harrell's C: 0.89 and 0.73

The best model by validation is model B – „log WBC" improves prediction out of sample as well.

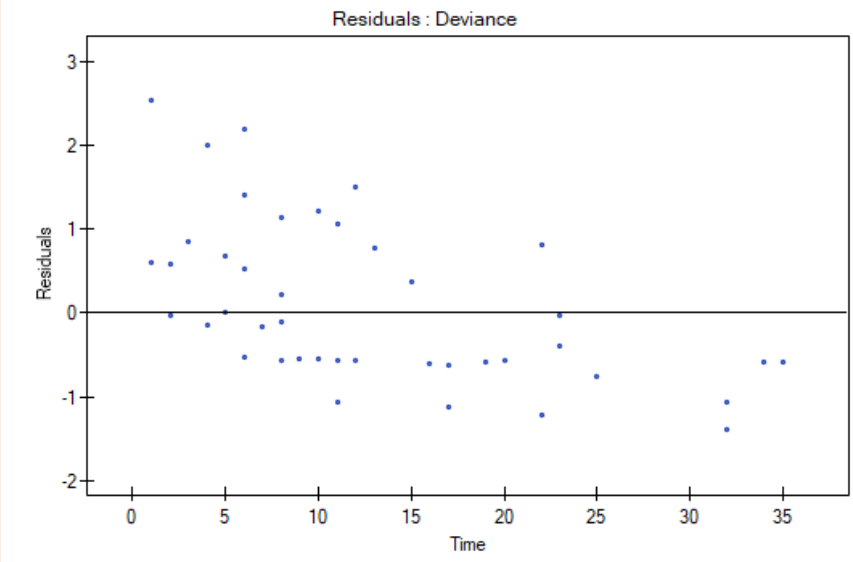
The analysis is complemented with the presentation, for model B, of the survival curves (remaining in remission) of both treatment groups, adjusted for the influence of „log WBC". To plot them, after selecting the Add graph option we check Survival function: in subgroups..., and then – to quickly build a graph of two curves – we choose Quick subgroups and indicate the Rx variable. The Advanced subgroups option allows building any number of arbitrarily defined curves.

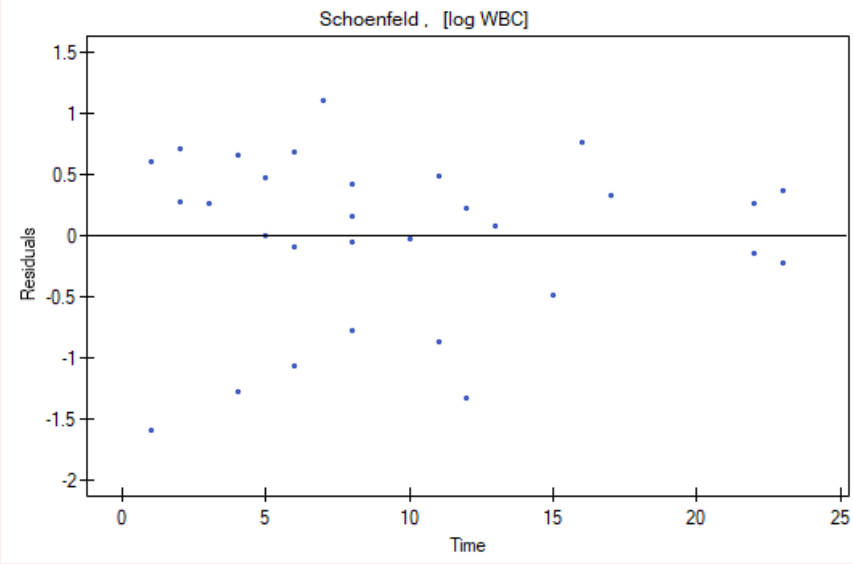


Finally, we assess the assumptions of Cox regression by analyzing the model residuals against time.

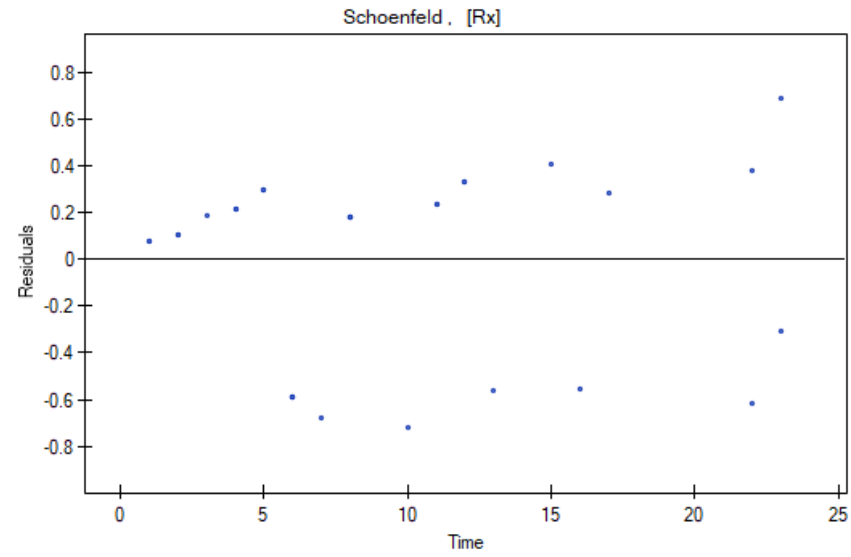


Copyright ©2010-2026 U-smile Stats – All rights reserved





Copyright ©2010-2026 U-smile Stats – All rights reserved



We do not observe outliers, although the Martingale and Deviance residuals become lower for longer times. The Schoenfeld residuals have a symmetrical distribution with respect to time. For the Schoenfeld residuals the analysis of the graph can be supported with tests checking whether the points of the residual plot form some pattern, e.g. a linear relationship – for that purpose we copy the Schoenfeld residuals and time to a datasheet and test the relationship we are looking for. The result of such a test for each variable indicates that the hazard proportionality assumption is fulfilled (when it is not statistically significant) or violated (when it is significant) – such a variable can then be excluded from the model. For the „log WBC" and „Rx" variables the symmetrical distribution of the residuals suggests that the hazard proportionality assumption is fulfilled; this can be confirmed by checking Pearson's linear or Spearman's monotonic correlation of these residuals with time.

Next, we can add the sex variable to the model – with caution, because it is known from the literature that sex may influence the survival function in leukemia remission in a way that is not proportional over time. We therefore build a Cox model for three variables: „Sex", „Rx" and „log WBC", and before interpreting the coefficients we check the Schoenfeld residuals – copying them together with time to a new datasheet and checking Spearman's monotonic correlation. We obtain $p = 0.0259$ (for sex), $p = 0.6192$ (for log WBC) and $p = 0.1490$ (for Rx), which confirms the violation of the hazard proportionality assumption by the sex variable.

Schoenfeld residuals vs time: $p = 0.0259$ for sex (significant)

The „Sex" variable violates the hazard proportionality assumption – we build the Cox model separately for women and men, performing the analysis twice with the data filter switched on (first for the female sex, then for the male sex).

For women

Model									
	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
log WBC	1.1701	0.4986	0.1929	2.1473	5.5083	0.0189	3.2224	1.2128	8.5617
Rx	0.2667	0.5659	-0.8425	1.3759	0.2221	0.6374	1.3057	0.4307	3.9586

For men

Copyright ©2010-2026 U-smile Stats – All rights reserved

Model									
	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
log WBC	1.6389	0.519	0.6216	2.6562	9.9705	0.0016	5.1496	1.8619	14.2422
Rx	1.859	0.7291	0.43	3.2881	6.5014	0.0108	6.4176	1.5373	26.7909

FOR THOSE INTERESTED

Out-of-sample evaluation

Let $\ell(\beta)$ denote the logarithm of the partial likelihood (with the Breslow correction for tied times) determined for the coefficients β. Cross-validation uses the cross-validated partial likelihood of Verweij and van Houwelingen[169]:

$$CVPL = \sum_{j=1}^{k} \left[\ell\left(\hat{\beta}_{(-j)}\right) - \ell_{(-j)}\left(\hat{\beta}_{(-j)}\right) \right],$$

where $\hat{\beta}_{(-j)}$ are the coefficients estimated without the j-th part of the data, ℓ – the logarithm of the partial likelihood of all the observations, and $\ell_{(-j)}$ – of the observations without the j-th part. The validation error is $-2\,CVPL/d$, where d is the number of events. With a test set the validation error is $-2\,\ell_T(\hat{\beta}_U)/d_T$, where ℓ_T is the logarithm of the partial likelihood of the test observations, $\hat{\beta}_U$ – the coefficients estimated on the training set, and d_T – the number of events in the test set. In the null model $\beta = 0$ is assumed.

Harrell's C[80] is the ratio of the number of concordant pairs to the number of comparable pairs; the pairs are formed within each evaluated part of the data (in cross-validation – within each fold, for the predictions of the model fitted without that fold) and counted jointly. A pair is comparable when the observation with the shorter time is complete, and concordant – when that observation has the higher predicted risk (a pair with equal predicted risk counts as half of a concordant pair).

Cox regression with time-dependent variables

WHEN TO USE:

The basic assumption of the Cox model is **hazard proportionality**: the effect of each independent variable on the hazard must be constant over time, i.e. the Hazard Ratio HR_i does not change as time passes. When this assumption is violated for one of the variables (an indication may be, e.g., a non-random, systematic pattern of the Schoenfeld residuals against time – see Analysis of model residuals), one of the solutions – besides stratifying the analysis (separate models built for each level of that variable) – is to extend the model with an additional term describing how the effect of the variable changes over time.

DEFINITIONS OF MEASURES

The extension of the model consists in adding a term being the product of the studied independent variable X_j and some function of time $g(t)$:

$$h(t, X_1, ..., X_k) = h_0(t) \cdot \exp\left(\sum_{i=1}^{k} \beta_i X_i + \delta \cdot X_j \cdot g(t) \right),$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where the product $X_j \cdot g(t)$ is called a **time-dependent variable** (time-dependent covariate). As a result, the partial hazard associated with the variable X_j is no longer constant and depends on time:

$$HR_j(t) = e^{\beta_j + \delta \cdot g(t)}.$$

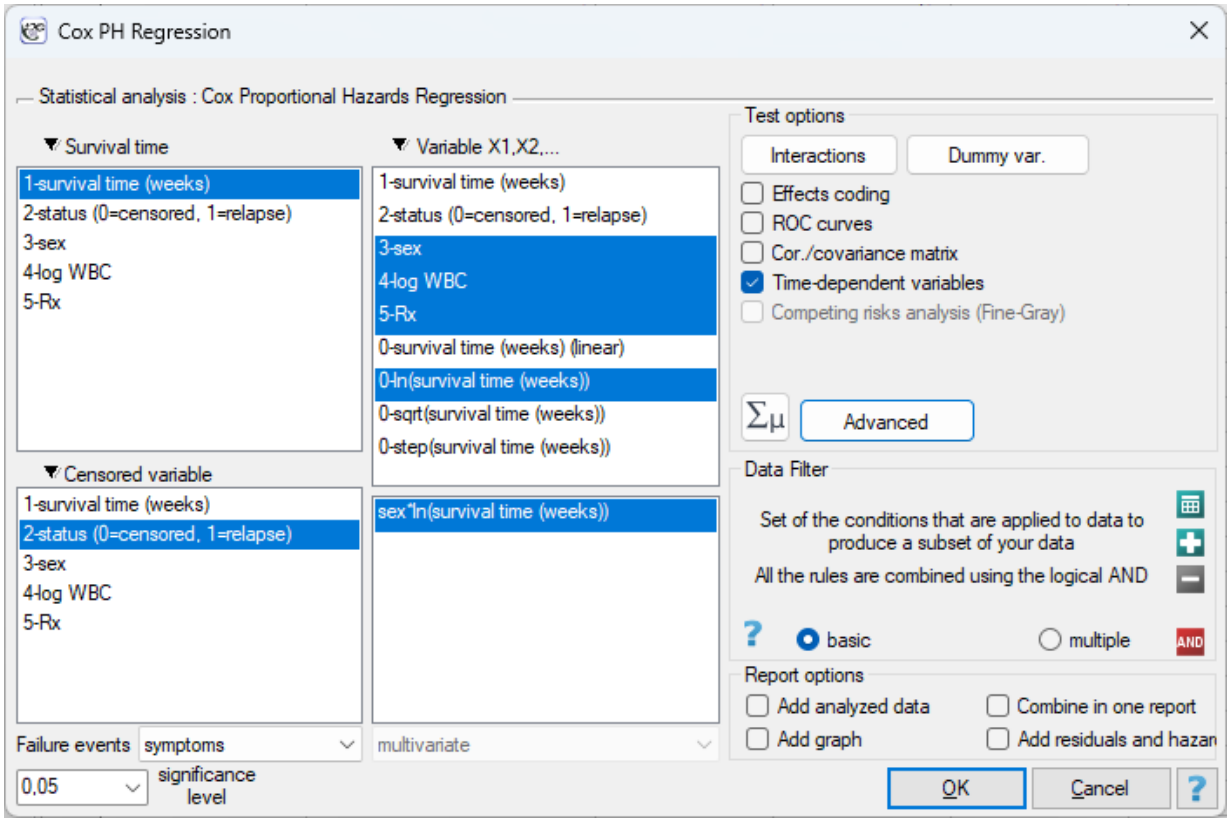
The Wald test for the coefficient δ allows checking whether the effect of the variable X_j really changes over time: $\delta = 0$ means that the effect of X_j is constant (the hazard proportionality assumption is fulfilled), and $\delta \neq 0$ – that the effect changes over time (the assumption is violated).

In PQStat the function $g(t)$ can be chosen from among four additional time pseudo-variables, with which a product is built with any independent variable of the model:

- **time (linear)** – $g(t) = t$,
- **ln(time)** – $g(t) = \ln(t)$,
- **sqrt(time)** – $g(t) = \sqrt{t}$,
- **step(time)** – $g(t)$ is a step (Heaviside) function, taking the value 0 before a fixed time threshold and 1 after it is exceeded.

HOW TO SET UP THE ANALYSIS

The window with settings for Cox regression with time-dependent variables is accessed via the menu Statistics→Survival analysis→Cox PH Regression – the same window as for standard Cox regression, in which we additionally check the Time-dependent variables option. Once it is checked and the variable defining the survival time is selected, the four additional time pseudo-variables listed above appear on the list of variables used for building interactions (the Interactions button) – so the function $g(t)$ is defined by choosing one of them.

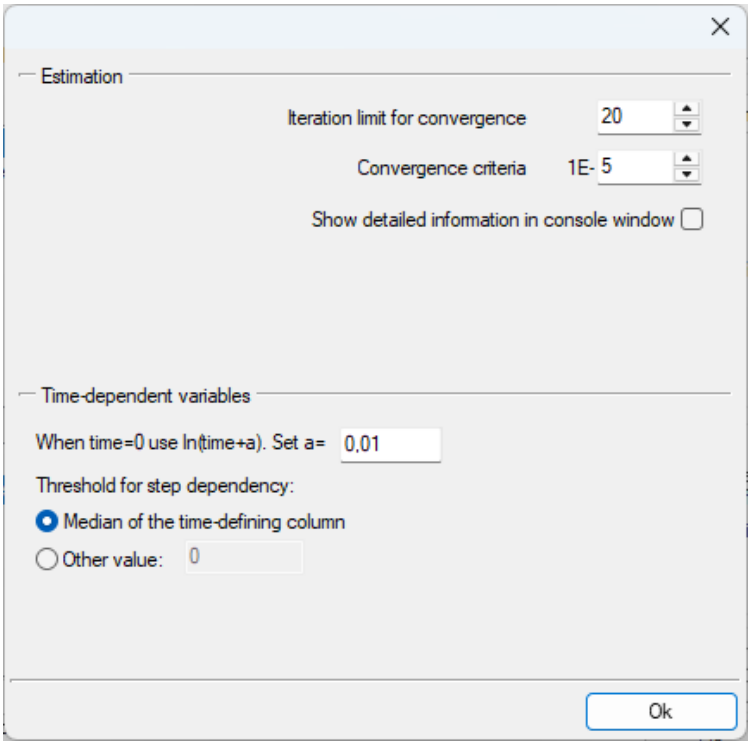


Copyright ©2010-2026 U-smile Stats – All rights reserved

Note! The Time-dependent variables option and the Competing risks analysis (Fine-Gray) option are mutually exclusive.

In the advanced settings window (the gear-icon button), besides the already known convergence parameters of the Newton-Raphson algorithm, two additional parameters related to time-dependent variables are available:

- the value a used to calculate $\ln(t+a)$ for observations whose survival time equals 0 (by default the logarithmic function is not defined in that case),
- the threshold for the step function – it can be the median of the variable defining time (the default option) or any other value indicated by the researcher.



CALCULATION AND INTERPRETATION

Example (28.2) cont. *(remissionLeukemia.pqs file)*

In the leukemia remission example the analysis of Schoenfeld residuals showed a violation of the hazard proportionality assumption by the sex variable (1=man, 0=woman; a significant monotonic relationship between the residuals and time, p=0.0259), while for the Rx and log WBC variables the assumption was fulfilled. Before we build the extended model, let us look at what this violation looks like in the data and what it leads to in an ordinary Cox model.

Step 1. Kaplan-Meier curves for sex

The comparison of the survival curves (remaining in remission) of women and men in the Comparison of groups module gives a non-significant result: log-rank test $\chi^2 = 0.56$, $p = 0.4555$, hazard ratio $HR_{m/f} = 0.78$ (95%CI: $\langle 0.38; 1.60 \rangle$).

Copyright ©2010-2026 U-smile Stats – All rights reserved

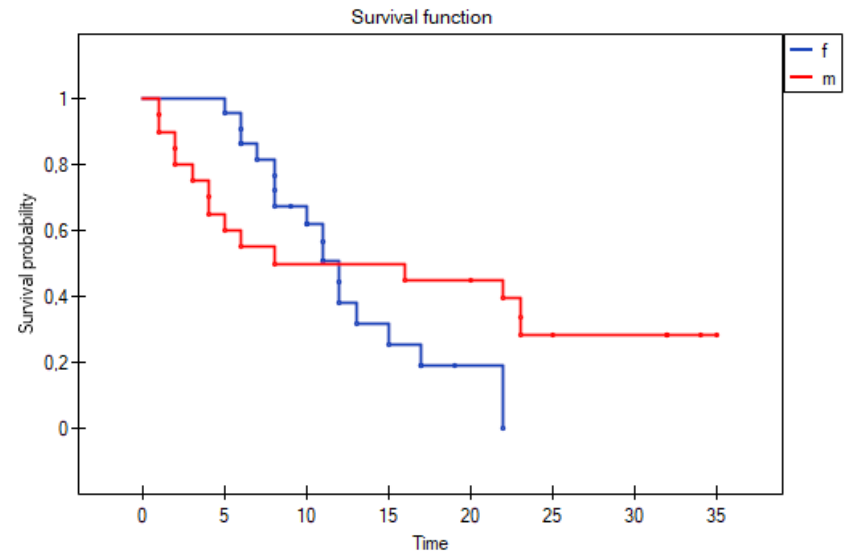
Comparison groups	
Analysed variables	survival time (weeks)
	status (0=censored, 1=relap
Number of unspecified	0
Number of missing data	0
Significance level	0,05
Grouping variable	sex
Frequency	42
Failure events	symptoms
Censored	no symptoms
Test: LogRank	
Chi-square statistic	0,556864
Degrees of freedom	1
p-value	0,455526

Logrank

Group	Obs.	Exp.	Obs./Exp.
f	16	14,156938	1,130188
m	14	15,843062	0,883668

Logrank

Group	Hazard r.	-95%CI	+95%CI
m/f	0,781877	0,381796	1,601199



The graph, however, shows why the test „does not see" the difference: the curves **cross** at about week 11–12. In the first weeks men leave remission much faster (their curve lies lower), but those who get through that period stay in remission longer than women – after week 12 the curve for men lies higher and stabilizes at about 0.28, while the curve for women drops to 0. Crossing curves are the classic picture of a **violation of the hazard proportionality assumption**: the hazard of men relative to women is not constant but changes direction over time. The log-rank test, which assigns the same weight to all time points, sums differences of opposite signs – and loses power.

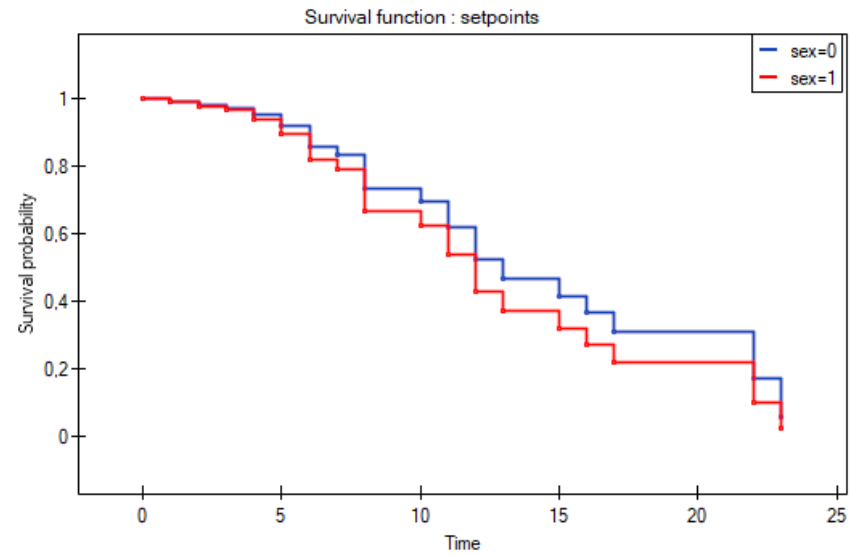
Step 2. A Cox model ignoring the violation of the assumption

The same „cancelling out" effect is observed in an ordinary Cox model with three variables: sex, log WBC and Rx.

Model

	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
sex	0,263171	0,449435	-0,617706	1,144048	0,342879	0,558172	1,301049	0,53918	3,13945
log WBC	1,593619	0,329996	0,946839	2,240399	23,32127	0,000001	4,921527	2,577549	9,397077
Rx	1,390877	0,456646	0,495867	2,285886	9,277232	0,00232	4,018371	1,641922	9,834394

Copyright ©2010-2026 U-smile Stats – All rights reserved



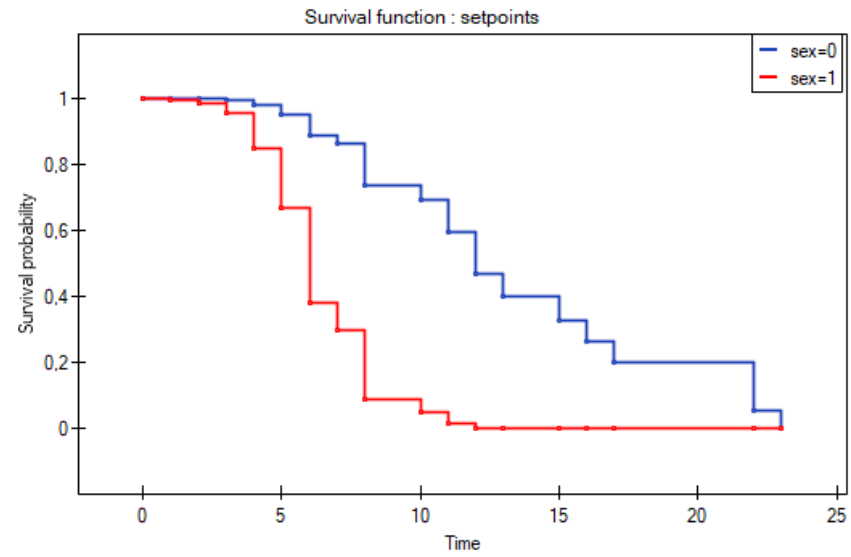
The coefficient of the sex variable is not significant ($HR = 1.30$, 95%CI: $\langle 0.54; 3.14 \rangle$, $p = 0.5582$), and the survival curves for both sexes plotted from the model (at the means of the other variables) lie close to each other and – by the model's assumption – cannot cross. The model averages the early, very unfavorable effect of being a man with the later favorable effect and ends up with „nothing". The conclusion that sex is irrelevant to remission time would therefore be wrong – it is an artifact of the unfulfilled assumption.

Step 3. A Cox model with a time-dependent variable

Instead of splitting the analysis into two separate models (for women and for men), we build one joint model in which sex interacts with a function of time. Because we are interested not in *whether*, but in *how, as time passes*, the effect of sex changes, we choose the logarithmic function $g(t) = \ln(t)$ – the effect of clinical variables in survival analysis usually changes fastest at the beginning of follow-up and stabilizes with time, which is well captured by the logarithm. In the analysis window we check the Time-dependent variables option, and in the Interactions window we define the product sex×ln(time). The model thus contains four terms: sex, log WBC, Rx and sex×ln(time).

Model									
	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
sex	4,057084	1,944145	0,24663	7,867537	4,354826	0,036904	57,805481	1,279706	2611,126514
log WBC	1,409521	0,335513	0,751927	2,067115	17,649162	0,000027	4,093995	2,121084	7,901993
Rx	0,985126	0,478369	0,04754	1,922711	4,240899	0,039461	2,678149	1,048689	6,839476
sex x ln(surviva	-1,911986	0,921656	-3,718398	-0,105574	4,303601	0,038032			

Copyright ©2010-2026 U-smile Stats – All rights reserved



The coefficient of the sex×ln(time) interaction is $\delta = -1.912$ ($SE = 0.922$, $\chi^2_{Wald} = 4.30$, $p = 0.0380$) and is statistically significant – this formally confirms that the effect of sex changes over time, i.e. what the Schoenfeld residuals and the crossing Kaplan-Meier curves suggested earlier. The coefficient of the sex variable itself has also become significant: $\beta_{sex} = 4.057$ ($p = 0.0369$).

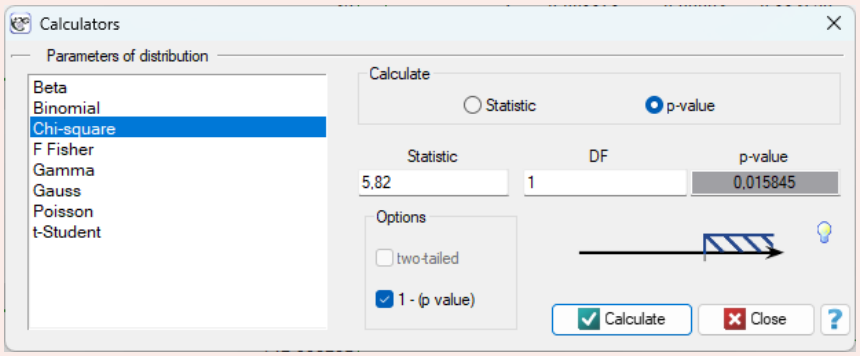
When interpreting these two coefficients, one has to remember that β_{sex} describes the effect of sex only at $t = 1$ (because $\ln(1) = 0$), and not the effect averaged over the whole follow-up period. The hazard of men relative to women in any week t is calculated from the formula $HR_{sex}(t) = e^{\beta_{sex}+\delta\cdot\ln(t)} = e^{4.057-1.912\cdot\ln(t)}$:

t (weeks)	1	2	4	8	12	16	23
$HR_{sex}(t)$	57.8	15.4	4.1	1.1	0.50	0.29	0.14

At the beginning of follow-up the risk of leaving remission is many times higher for men than for women, but this excess decreases quickly; after about 8 weeks ($t^* = e^{4.057/1.912} \approx 8.4$) the hazards of both sexes become equal, and later the direction of the relationship reverses – men who have remained in remission now have a risk of relapse several times *lower* than women. This is exactly the picture shown by the Kaplan-Meier curves in step 1 – the model with a time-dependent variable not only confirms it, but allows it to be measured, with simultaneous adjustment for log WBC and Rx.

The extended model is also simply better than the model from step 2: it has a higher likelihood ($-2\ln L = 138.4$ versus 144.2), lower information criteria ($AIC = 146.4$ versus 150.2; $BIC = 152.0$ versus 154.4) and higher fit coefficients (Pseudo $R^2_{McFadden} = 0.264$ versus 0.233). The likelihood ratio test comparing both models (a difference in $-2\ln L$ of 5.82 with 1 degree of freedom) gives $p = 0.0158$, i.e. adding the time-dependent term significantly improves the model.

Copyright ©2010-2026 U-smile Stats – All rights reserved



$\delta_{sex\times\ln(t)} = -1.91$, $p = 0.0380$; $HR_{sex}(t)$ from 57.8 ($t = 1$) through 1.0 ($t \approx 8.4$) to 0.14 ($t = 23$)

The effect of sex is not constant over time: men have a much higher risk of leaving remission in the first weeks, and a much lower one later. The ordinary Cox model (and the log-rank test) average these opposite effects into a non-significant „no difference"; the model with a time-dependent variable reveals and measures this change.

FOR THOSE INTERESTED

Wald test for the coefficient of the interaction with the time variable
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \delta = 0 \quad (\text{the effect of } X_j \text{ is constant over time} - \text{the hazard proportionality assumption is fulfilled}), \\ \mathcal{H}_1: &\quad \delta \neq 0 \quad (\text{the effect of } X_j \text{ changes over time} - \text{the hazard proportionality assumption is violated}). \end{aligned}$$

The test statistic and the way of inference are analogous to the Wald test for the other coefficients of the Cox model.

Competing risks analysis – the Fine-Gray model

WHEN TO USE:

In many studies the observed subjects may experience one of several, mutually exclusive types of unfavorable events. For example, an oncological patient may die because of the underlying disease, but may also die of an entirely different cause (e.g. cardiovascular), before the event of interest occurs. The occurrence of such an alternative event is called a **competing risk** – it prevents the event of interest from being observed, and at the same time it is not ordinary censoring, because it provides additional information (we know why the observation ended).

Applying the standard Cox model, in which competing events would be treated as ordinary censoring, leads to biased estimates – the standard $1 - S(t)$ function then overestimates the true risk of the event of interest, because some subjects would never have experienced it (they died earlier of a different cause). The appropriate tool for analysis in such a situation is the **cumulative incidence function** (CIF)

Copyright ©2010-2026 U-smile Stats – All rights reserved

together with the **Fine-Gray** regression model (Fine and Gray 1999)[64], based on the so-called **subdistribution hazard**.

DEFINITIONS OF MEASURES

The **cumulative incidence function (CIF)** expresses, for the event of interest, the probability of its occurrence up to time t, taking into account the fact that competing events may prevent it. The sum of the CIF for the event of interest, the CIF for the competing events and the survival function $S(t)$ equals 1 at every time point.

The **Fine-Gray subdistribution hazard model** defines the subdistribution hazard $h^*(t)$ in a way analogous to how the Cox model describes the standard hazard, but it is based on a different risk set: subjects who experienced the competing event are **not removed** from the risk set – they remain in it (with an appropriate weight) until the end of follow-up, simulating a situation in which they could never experience the event of interest, but are still being observed. These weights (inverse probability of censoring weighting, IPCW) are calculated from the inverse of the Kaplan-Meier estimator of the censoring distribution. The model parameters are estimated – just like in the standard Cox model – by the maximum likelihood method (the iterative Newton-Raphson algorithm), with the difference that the likelihood function takes into account the IPCW weights described above.

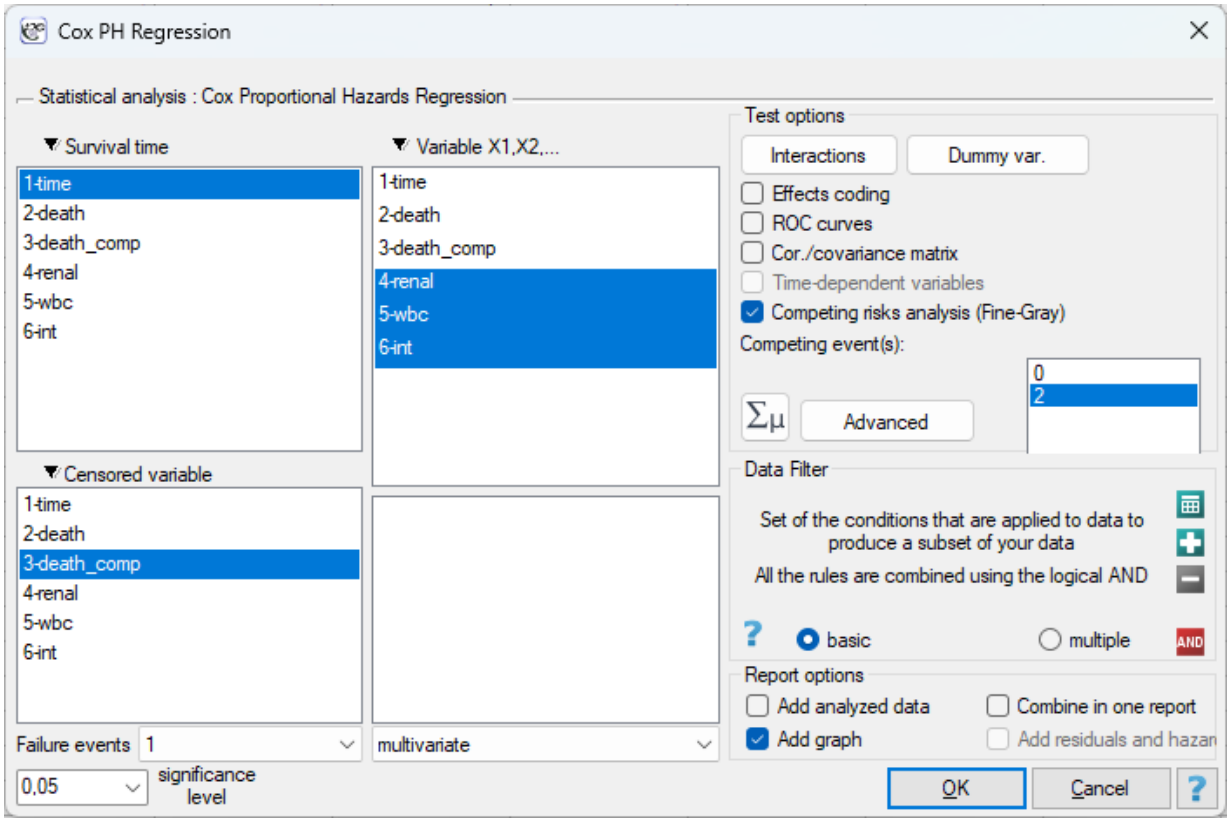
Subdistribution Hazard Ratio (sHR) – for each independent variable a subdistribution hazard ratio (sHR) is calculated. Unlike the standard HR of the Cox model (which describes the effect of a variable on the rate of occurrence of the event of interest only among subjects still „at risk" of that event), sHR describes the effect of the independent variable directly on the cumulative incidence function $CIF(t)$, which makes it easier to interpret clinically (e.g. for prognosis) in the context of competing risks. The direction of interpretation ($sHR > 1$, $sHR < 1$, $sHR \approx 1$) is analogous to that of the standard HR described earlier, and the significance of individual coefficients is likewise verified with the Wald test.

Note! The Fine-Gray model is a separate model from the standard Cox model computed on the same data in which competing events are simply excluded or ignored (treated as no event) – the two models answer different research questions and can lead to different, sometimes even opposite, conclusions (the so-called *sign reversal* effect described in the literature) for the same independent variable.

HOW TO SET UP THE ANALYSIS

The window with settings for the Fine-Gray model is accessed via the menu Statistics→Survival analysis→Cox PH Regression – the same window as for standard Cox regression, in which we additionally check the Competing risks analysis (Fine-Gray) option. Once it is checked, based on the values of the variable selected as the censoring variable, a Competing event(s): list appears, on which we indicate – among the values other than the code of the event of interest – one or more values denoting competing events. Status values selected neither as the event of interest nor as a competing event remain true censoring.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Note! Because model residuals are calculated differently (Martingale, Deviance and Schoenfeld residuals are not defined for the CIF), the Add residuals and hazard option is unavailable for the Fine-Gray model, as is checking the Time-dependent variables option at the same time. The ROC curve option, i.e. the time-dependent AUC, is available – in a version accounting for competing events (see below).

In the report from a Fine-Gray analysis the counts of complete and censored observations are additionally broken down by the code (value of the status variable) they correspond to, and the number of censored observations is further split into the individual competing events. The standard HR column is replaced with an sHR (*Fine-Gray*) column.

Analogously to the standard Cox model (the Add graph option), for the Fine-Gray model the program allows plotting **cumulative incidence function (CIF) graphs**:

- **the baseline CIF** – the cumulative incidence at the mean values of the covariates,
- **CIF for selected variable values** – analogously to the survival function in the standard Cox model, for values of the independent variables indicated by the researcher,
- **CIF for subgroups** – separate CIF curves built for each category of the indicated grouping variable (the Quick subgroups / Advanced subgroups options).

Note! Unlike the Kaplan-Meier survival function, the CIF curve usually does not tend to 0 or to 1 – its limiting value is the estimated probability that a given subject will ever experience the event of interest before experiencing a competing event.

Time-dependent AUC for competing risks. The ability of the Fine-Gray model to discriminate patients who will experience the event of interest is assessed – as for the standard Cox model – with the time-dependent AUC, with one important difference in the definition of controls. The cases at time t are only

Copyright ©2010-2026 U-smile Stats – All rights reserved

the subjects who experienced the **event of interest** up to time t; the controls are both the subjects observed longer than t without any event and the subjects who experienced a **competing event** up to time t – they will certainly no longer experience the event of interest, and a good model should assign them a lower risk. The interpretation of $AUC(t)$ remains unchanged: 0.5 – random discrimination, 1 – perfect discrimination; the graph of $AUC(t)$ against time (the ROC curve option in the Add graph window) shows how the quality of the prediction of the event of interest changes for successive time horizons.

CALCULATION AND INTERPRETATION

Example 28.3. (multipleMyeloma_competingRisks.pqs file)
In a study of 445 patients diagnosed with multiple myeloma, the efficacy of a new drug (the int variable: 0=standard treatment, 1=new drug) in prolonging survival time (the time variable, in days) was compared. For the patients the baseline white blood cell count (wbc) and the presence of renal failure (renal) were also recorded – a complication frequently accompanying multiple myeloma and considered an unfavorable prognostic factor.

The death of a patient could occur because of the progression of the cancer (the event of interest) or from an entirely different cause, e.g. cardiovascular or infectious, not directly related to the myeloma (the competing event). This information is coded by the death_comp variable: 0 – the patient is alive at the end of follow-up (censoring), 1 – death from myeloma, 2 – death from another cause. The data also contain a simplified death variable (death from any cause, without distinguishing the reason), usually used for the standard Cox model.

The standard model (not appropriate for these data)

For comparison we first build a standard Cox model based on the death variable, in which death from a cause other than myeloma (the competing event) is incorrectly treated the same way as death from myeloma (the event of interest):

Model

	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
renal	2,432497	0,242623	1,956965	2,908028	100,51764	<0,000001	11,387278	7,077816	18,320641
wbc	0,053077	0,015474	0,022749	0,083405	11,765627	0,000603	1,054511	1,02301	1,086982
int	0,779894	0,294228	0,203217	1,356571	7,025902	0,008034	2,181242	1,225338	3,882858

The Fine-Gray model

Next we build the model appropriate for these data – the Fine-Gray model – indicating death_comp as the censoring variable, code 1 (death from myeloma) as the event of interest, and checking code 2 (death from another cause) on the Competing event(s): list.

Frequency	445
Failure events (1)	43
competing:	
(2)	45
Censored (0)	357

Model

	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	sHR (Fine-Gr	-95% CI	+95% CI
renal	3,654627	0,349788	2,969055	4,340198	109,16315	<0,000001	38,653086	19,473498	76,722769
wbc	0,057522	0,019021	0,020241	0,094804	9,145211	0,002494	1,059209	1,020448	1,099443
int	-0,406406	0,50775	-1,401578	0,588765	0,64065	0,423475	0,66604	0,246208	1,801763

Copyright ©2010-2026 U-smile Stats – All rights reserved

The obtained sHR for the int variable is $sHR = 0.67$ (95%CI: $\langle 0.25; 1.80 \rangle$), $p = 0.4235$, which means there is no basis to claim that the new drug significantly affects the risk of death from multiple myeloma once the competing risk of death from another cause is properly accounted for. A different picture is given by the standard model (incorrectly ignoring the distinction between causes of death), in which the HR for int is $HR = 2.18$, $p = 0.008034$ – suggesting a significant, unfavorable effect of the new drug on the risk of death from any cause. This is a good illustration of the sign reversal effect described earlier: the drug may slightly increase the risk of death from other causes (e.g. as a result of adverse effects), which in the standard model is mixed with the risk of death from the cancer itself, suggesting an apparent unfavorable effect of the drug on the course of the myeloma – while the Fine-Gray model shows that the drug has no significant effect on the risk of death specifically from myeloma.

$HR_{int} = 2.18$, $p = 0.008034$ (standard model) versus $sHR_{int} = 0.67$, $p = 0.4235$ (Fine-Gray model)

The new drug apparently increases the overall risk of death, but does not significantly increase the risk of death specifically from myeloma – the effect visible in the standard model disappears once competing events are properly accounted for.

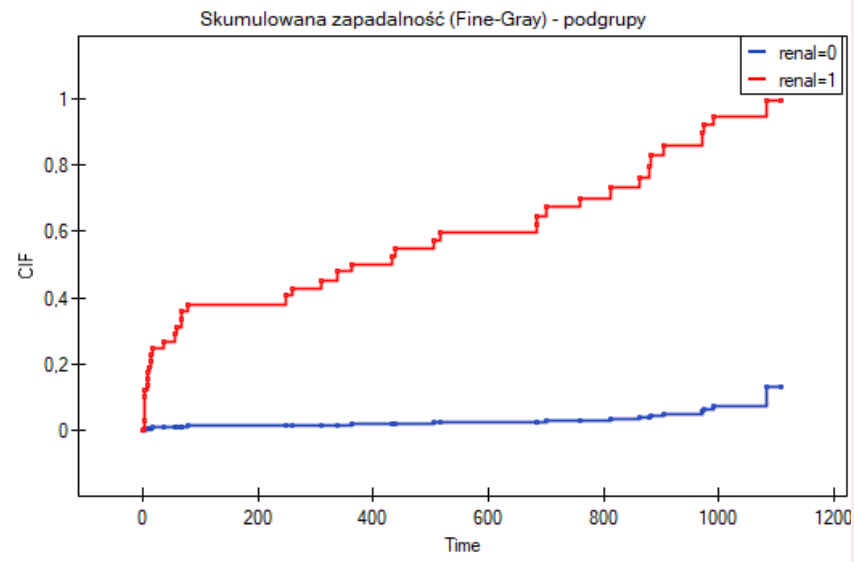
The opposite situation is observed for renal failure (renal) – here both models consistently indicate a strong, statistically significant relationship with the risk of death, and in the Fine-Gray model the effect is even more pronounced: $sHR = 38.65$ (95%CI: $\langle 19.47; 76.72 \rangle$), $p < 0.000001$, versus $HR = 11.39$ (95%CI: $\langle 7.08; 18.32 \rangle$), $p < 0.000001$ in the standard model. Renal failure is therefore not only a significant risk factor for death from any cause, but above all strongly – and even more strongly than the standard model would suggest – related to the risk of death from the progression of the myeloma itself, which is consistent with its well-known role as an unfavorable prognostic factor in this disease (kidney damage usually accompanies an advanced form of myeloma).

$sHR_{renal} = 38.65$, 95%CI: $\langle 19.47; 76.72 \rangle$, $p < 0.000001$ (Fine-Gray model) versus $HR_{renal} = 11.39$, $p < 0.000001$ (standard model)

Renal failure is strongly and consistently related to the risk of death from myeloma – the effect is even more pronounced once the competing risk of death from another cause is properly accounted for.

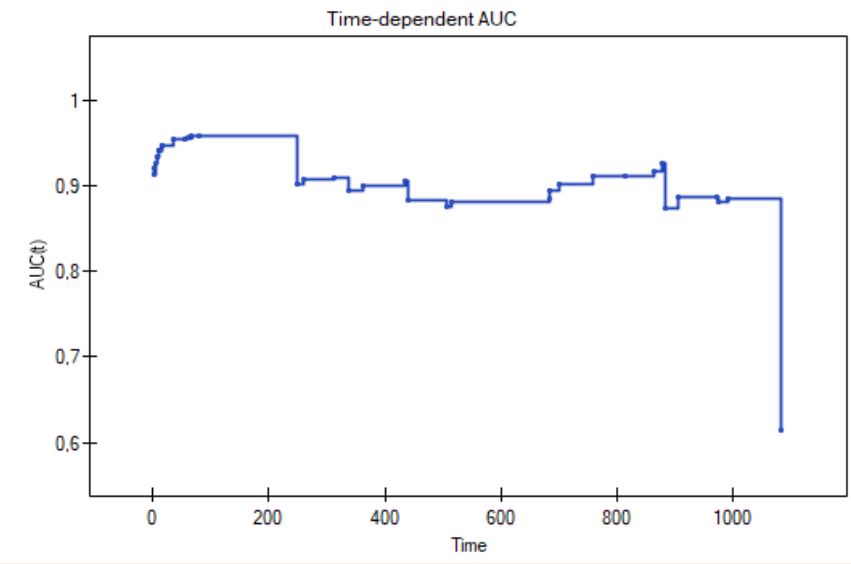
Finally, we compare the CIF curves for death from myeloma in the groups defined by the renal variable:

Copyright ©2010-2026 U-smile Stats – All rights reserved



The CIF curve for patients with renal failure (renal=1) rises much faster and reaches much higher values than the curve for patients without this complication (renal=0) – already after about 100 days of follow-up the cumulative probability of death from myeloma in this group exceeds 0.4, and towards the end of follow-up it approaches 1. In patients without renal failure the cumulative probability of death from myeloma remains low (below 0.15) throughout the follow-up period. The graph thus confirms the numerical sHR result: in these data renal failure is associated with a significantly higher risk of death from multiple myeloma.

Finally, we assess the ability of the built Fine-Gray model to discriminate the patients who will die from myeloma, by selecting the ROC curve option in the Add graph window, i.e. the graph of the time-dependent AUC:



The values of $AUC(t)$ are high throughout the follow-up period. In the first weeks $AUC(t)$ rises from about 0.91 to about 0.96 and stays at this level for the first approx. 250 days, and then – as the controls lose event-free patients and gain deaths from other causes – it decreases to the level of 0.88–0.92, at

Copyright ©2010-2026 U-smile Stats – All rights reserved

which it remains almost until the end of follow-up. This means that the model (and in practice mainly the renal variable) discriminates very well between patients who will die from myeloma by a given time t and those who will not die from this cause by that time – regardless of the prediction horizon. The sharp drop of $AUC(t)$ at the last point (about 1080 days) is an estimation artifact: at such a late time only a few observations remain in the risk set, and an estimator based on a handful of case–control pairs becomes unstable; the values at the very end of the time axis should therefore not be interpreted.

$AUC(t) \approx 0.88$–0.96 for t from a few to about 1000 days

The Fine-Gray model discriminates patients at risk of death from myeloma very well and stably over time – the quality of the prediction does not depend significantly on the adopted time horizon.

FOR THOSE INTERESTED

Cumulative incidence function (CIF)

$$CIF(t) = P(T \leq t,\, \varepsilon = 1) = \sum_{t_i \leq t} S(t_{i-1}) \cdot h_1(t_i),$$

where $S(t_{i-1})$ is the overall (all-cause) survival function just before t_i, and $h_1(t_i)$ is the cause-specific hazard for the event of interest at t_i.

The Fine-Gray subdistribution hazard model

$$h^*(t, X_1, ..., X_k) = h_0^*(t) \cdot \exp\left(\sum_{i=1}^{k} \beta_i X_i\right)$$

The IPCW weights used to keep subjects in the risk set after the occurrence of a competing event are calculated from the inverse of the Kaplan-Meier estimator of the censoring distribution $\hat{G}(t)$:

$$w_i(t) = \frac{\hat{G}(t)}{\hat{G}(\min(T_i, t))},$$

where the estimator $\hat{G}$ is calculated the same way as the Kaplan-Meier estimator of the survival function, but with the roles reversed – true censoring is treated here as the „event", while both the event of interest and the competing events are treated as censoring for this auxiliary estimator.

Subdistribution Hazard Ratio (sHR)

$$sHR_i = e^{\beta_i}.$$

Time-dependent AUC for competing risks

The formula is the same as for the standard Cox model (see Cox proportional hazards regression); only the set of controls and their weights change: besides the subjects with $T_j > t$ (weight $1/\hat{G}(t)$), the controls also include the subjects who experienced a competing event up to time t ($T_j \leq t, \varepsilon_j = 2$), with weight $1/\hat{G}(T_j^-)$. The estimator $\hat{G}$ is the same Kaplan-Meier estimator of the censoring distribution that is used to determine the IPCW weights of the model (Blanche et al. 2013[21]).

Copyright ©2010-2026 U-smile Stats – All rights reserved

29 META-ANALYSIS

The number of scientific papers appearing has grown very significantly over the last decade. This makes access to knowledge easier, but makes it harder to keep up with the constant flow of new information. For example, if a doctor were to apply a new treatment for their patients based on a single scientific paper they had read, they could make a mistake – other, equally credible papers may contradict the effectiveness of that treatment. For a doctor's decision to be as well justified as possible, they should know most of the scientific papers published on the topic, and constantly reviewing an ever-growing body of literature would take so much time that there would be none left for treating patients. **Meta-analysis** allows such a review to be done quickly, because it is a statistical summary of the results described in many independent papers on the same topic.

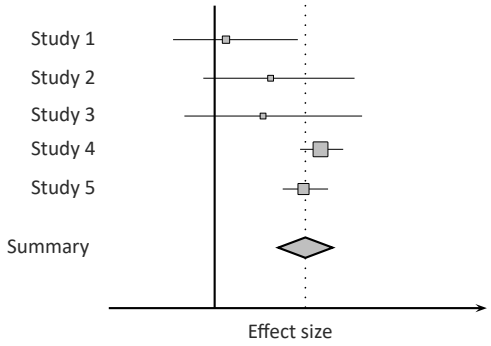
Meta-analysis in PQStat is performed based on the following measures (**final effects**):

- **Difference of means**,
- **Cohen's d**,
- **Hedges' g**,
- **Ratio of means**,
- **Odds ratio (OR)**,
- **Relative risk (RR)**,
- **Risk difference (RD)**,
- **Pearson coefficient**,
- **AUC for the ROC curve**,
- **Proportion**.

Introduction to meta-analysis and data preparation

WHEN TO USE:

Meta-analysis is used when we want to statistically summarize the results of many independent scientific papers concerning the same issue. The image most commonly associated with meta-analysis is the **forest plot**, showing the results of individual studies together with their combined summary.



Copyright ©2010-2026 U-smile Stats – All rights reserved

For the selected literature to be summarized jointly, it must be consistent in its description, and the measures reported in it must be equivalent.

ASSUMPTIONS

For a given scientific paper to be usable in a meta-analysis, it should make it possible to determine:

- the **final effect** – i.e. the statistical measure indicating the result obtained in the paper. In practice this can be various kinds of quantities, e.g. difference of means, odds ratio, relative risk, etc.
- the **effect error**, i.e. **SE** *(Standard Error)* – describing the precision of the study performed. The smaller the error (SE), the greater the precision of a given study, and consequently the greater its **weight** and influence on the meta-analysis result.
- the **sample size**, i.e. the number of objects on which the study was performed – required in some situations.

Note! It often happens that a scientific paper does not directly contain all of the elements listed above. In such a case, it is necessary to look in the paper for the data on the basis of which these measures can be calculated.

Note! PQStat performs meta-analysis calculations on data containing: the final effect, the effect error and, in some situations, the sample size. It is recommended that, before starting a meta-analysis, the data for each publication be entered in the data preparation window – this is particularly useful when a given paper does not directly report these three measures.

DEFINITIONS OF MEASURES

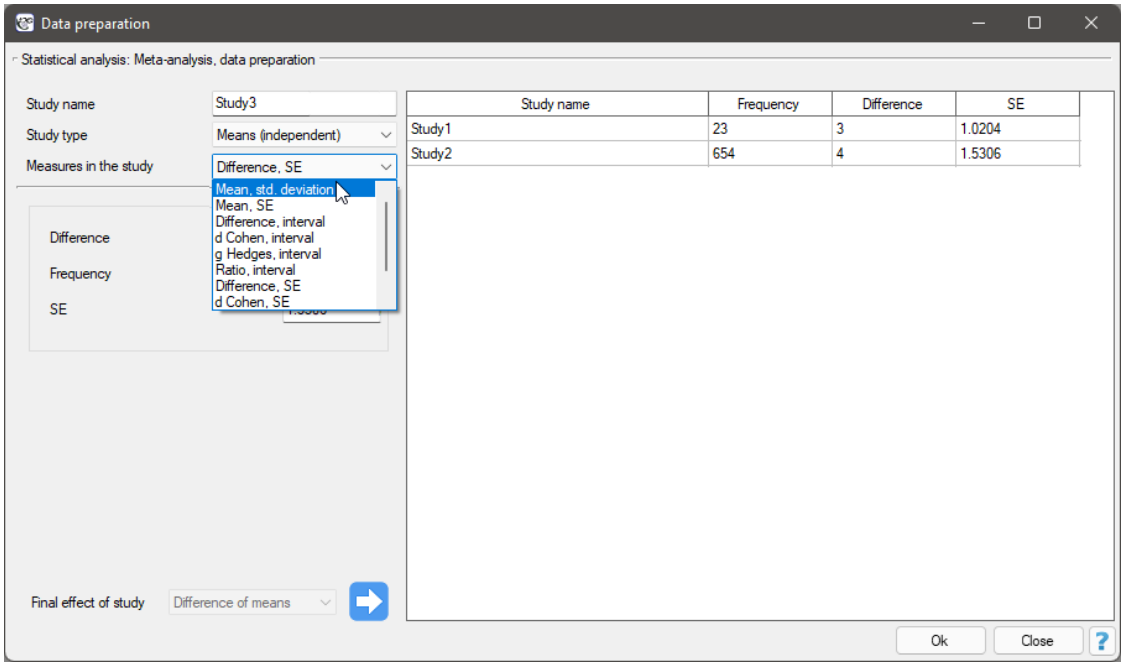
The **p-value** (statistical significance) is not used directly in meta-analysis. The same effect size may be statistically significant in a large study and non-significant in a study based on a small sample – and vice versa: a small effect size may be statistically significant in a large study, and a large effect size may be non-significant in a small study. This results from the power of statistical tests. When testing statistical significance we only check whether an effect exists at all (whether it differs from zero), not whether it is large enough to translate into desired outcomes – e.g. the fact that a drug statistically significantly lowers blood pressure by 1 mmHg will not lead to it being used, since from a clinical point of view this is too small a value. Meta-analysis focuses primarily on the magnitude of individual effects, not on their statistical significance, so it does not matter much whether the papers used in the meta-analysis themselves indicate statistical significance of a given effect or not.

In PQStat, however, statistical significance is calculated for each study, determined from the ratio of the effect to its error. This is an asymptotic approach, based on the normal distribution, dedicated to large groups. If a different test was used in the cited study to check statistical significance, the results obtained may differ slightly.

HOW TO SET UP THE ANALYSIS

The data preparation window for meta-analysis is opened via the menu:
Advanced statistics→Meta-analysis→Data preparation.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In the data preparation window for meta-analysis, the researcher first provides the name of the study being entered. This name should uniquely identify the study, since it will describe it throughout all meta-analysis results, including on charts. The desired Final effect of the study, its error and sample size are calculated based on the measures obtained from the relevant scientific paper. The measures contained in studies, on the basis of which final effects can be calculated, are presented in the table below:

Copyright ©2010-2026 U-smile Stats – All rights reserved

Study type	Measures in the study	Final effect
Independent means	Means, deviations	**a**, **b**, **c**, **e**
	Means, SE	**a**, **b**, **c**, **e**
	Difference, interval	**a**
	Cohen's d, interval	**b**
	Hedges' g, interval	**c**
	Ratio, interval	**e**
	Difference, SE	**a**
	Cohen's d, SE	**b**
	Hedges' g, SE	**c**
	Ratio, SE	e
Dependent means	Means, deviations	**a**, **b**, **c**
	Means, SE	**a**, **b**, **c**
	Difference, interval	**a**
	Cohen's d, interval	**b**
	Hedges' g, interval	**c**
	Difference, deviation	**a**, **b**, **c**
	Difference, SE	**a**, **b**, **c**
	Cohen's d, SE	**b**
	Hedges' g, SE	**c**
Mean vs. set value	Mean, deviation	**a**
	Mean, SE	**a**
Mean – summary	Mean, deviation	**d**
	Mean, SE	**d**
Tables	2x2 table	**b**, **c**, **f**, **g**, **h**
	OR, interval	**b**, **c**, **f**
	RR, interval	**g**
	RD, interval	**h**
	Cohen's d, interval	**b**, **c**, **f**
	Hedges' g, interval	**b**, **c**, **f**
	OR, SE	**b**, **c**, **f**
	RR, SE	**g**
	RD, SE	**h**
	Cohen's d, SE	**b**, **c**, **f**
	Hedges' g, SE	**b**, **c**, **f**
Correlation	Coefficient	**i**
	Coefficient, interval	**i**
	Coefficient, SE	**i**
ROC curve	AUC, SE	**j**
	AUC, interval	**j**
One proportion	Given group size	**k**
	Given proportion	**k**

where the individual final effects are:

a - Difference of means

b - Cohen's d

c - Hedges' g

Copyright ©2010-2026 U-smile Stats – All rights reserved

d - Mean

e - Ratio of means

f - Odds ratio (OR)

g - Relative risk (RR)

h - Risk difference (RD)

i - Pearson coefficient

j - AUC (ROC curve)

k - Proportion

Note! When determining the error of coefficients such as OR or RR and other table-based coefficients, in the presence of zero values in the tables, or when determining the error of a proportion, when the proportion is 0 or 1, a continuity correction is applied using an increase factor of 0.5. The confidence interval for a proportion is determined according to the exact Clopper-Pearson method[40].

PERFORMING AND INTERPRETING THE ANALYSIS

Example 29.1. We are interested in the effect of cigarette smoking on the risk of disease X. We want to perform a meta-analysis whose final effect will be the relative risk (RR). Under this assumption, the papers selected for the meta-analysis must make it possible to calculate the **RR** and its **error**.

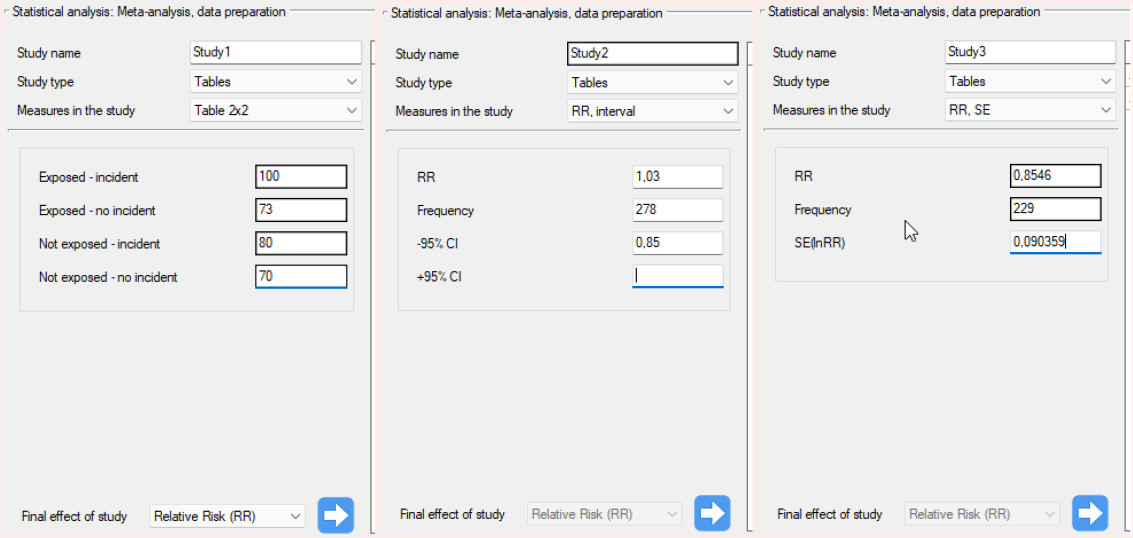
Step 1. Based on the description of the final effect (see the table above), it was established that the relative risk (described as result **g**) can be determined in PQStat in three situations, i.e. if a given scientific paper reports: the RR and its confidence interval, or the RR together with its error, or the appropriate sample sizes in four categories, i.e. a 2x2 table.

Step 2. 10 papers meeting the inclusion criteria and allowing the relative risk to be determined (see step 1) were selected for the meta-analysis. The data needed, contained in the selected papers, is:

Study 1: group sizes: (smokers and ill)=100, (smokers and not ill)=73, (non-smokers and ill)=80, (non-smokers and not ill)=70,
Study 2: group sizes: (smokers and ill)=182, (smokers and not ill)=172, (non-smokers and ill)=180, (non-smokers and not ill)=172,
Study 3: group sizes: (smokers and ill)=157, (smokers and not ill)=132, (non-smokers and ill)=125, (non-smokers and not ill)=201,
Study 4: group sizes: (smokers and ill)=19, (smokers and not ill)=15, (non-smokers and ill)=35, (non-smokers and not ill)=20,
Study 5: group size: 278, RR[95%CI]=1.03[0.85-1.25],
Study 6: group size: 560, RR[95%CI]=1.21[1.05-1.40],
Study 7: group size: 1207, RR[95%CI]=1.04[0.93-1.15],
Study 8: group size: 214, RR[95%CI]=1.15[0.95-1.40],
Study 9: group size: 285, RR[95%CI]=1.36[1.03-1.79],
Study 10: group size: 1968, RR=1.17, SE(lnRR)=0.0437,

Copyright ©2010-2026 U-smile Stats – All rights reserved

Step 3. Using the study preparation window for meta-analysis, the data was entered into the data sheet. The first four studies are entered by selecting tables; studies five through nine are entered by selecting RR and interval; the last study provides all the necessary data, i.e. RR and SE. As the Final effect of the study we set Relative risk (RR):



Each entered study is moved to the window on the right using . Using the OK button, we transfer the prepared studies to the data sheet. Based on the information about individual studies contained in the data sheet, one can proceed to perform the meta-analysis.

Summary effect – fixed effect and random effect

WHEN TO USE:

The most desirable element of a meta-analysis is a summary of the collected studies, i.e. determining a common effect M. Such a summary can be performed in two ways – by determining a fixed effect or a random effect.

DEFINITIONS OF MEASURES

Fixed effect – when calculating the fixed effect, we assume that all studies in the meta-analysis share one common true effect. So if every study concerned the same population (e.g. the same country), then when summarizing the meta-analysis with a fixed effect we assume that the true (population) effect is the same in each of these studies, and consequently – that all factors that could disturb the size of this effect are also the same (e.g. the age or sex of the subjects is similar in every study). Differences in the effects obtained for individual studies then result only from sampling error (the internal error of each study), i.e. from the size of SE.

The fixed effect estimates the population effect – the true effect common to every study. The confidence interval around the fixed effect (the width of the diamond on the forest plot) depends only on the individual SE values.

Random effect – when calculating the random effect, we assume that each study represents a slightly different population, so the true (population) effect is different for each of them. So if every study

Copyright ©2010-2026 U-smile Stats – All rights reserved

concerned a different country, then when summarizing the meta-analysis with a random effect we assume that some factors that could disturb the size of this effect (e.g. level of education or wealth) differ between countries, so the true effect (e.g. the mean increase in fertility) is also somewhat different in each of them. Differences in the effects obtained for individual studies then result from sampling error (the error within each study), i.e. from the size of SE, and from differences between the studied populations, i.e. from the heterogeneity of the studies (T^2). This heterogeneity must not be too large – too great a diversity of the studied populations indicates that there are no grounds for a joint summary.

The random effect estimates the weighted mean of the true (population) effects of individual studies. The confidence interval around the random effect (the width of the diamond on the forest plot) depends on the individual SE values and on T^2.

Confidence interval vs. prediction interval

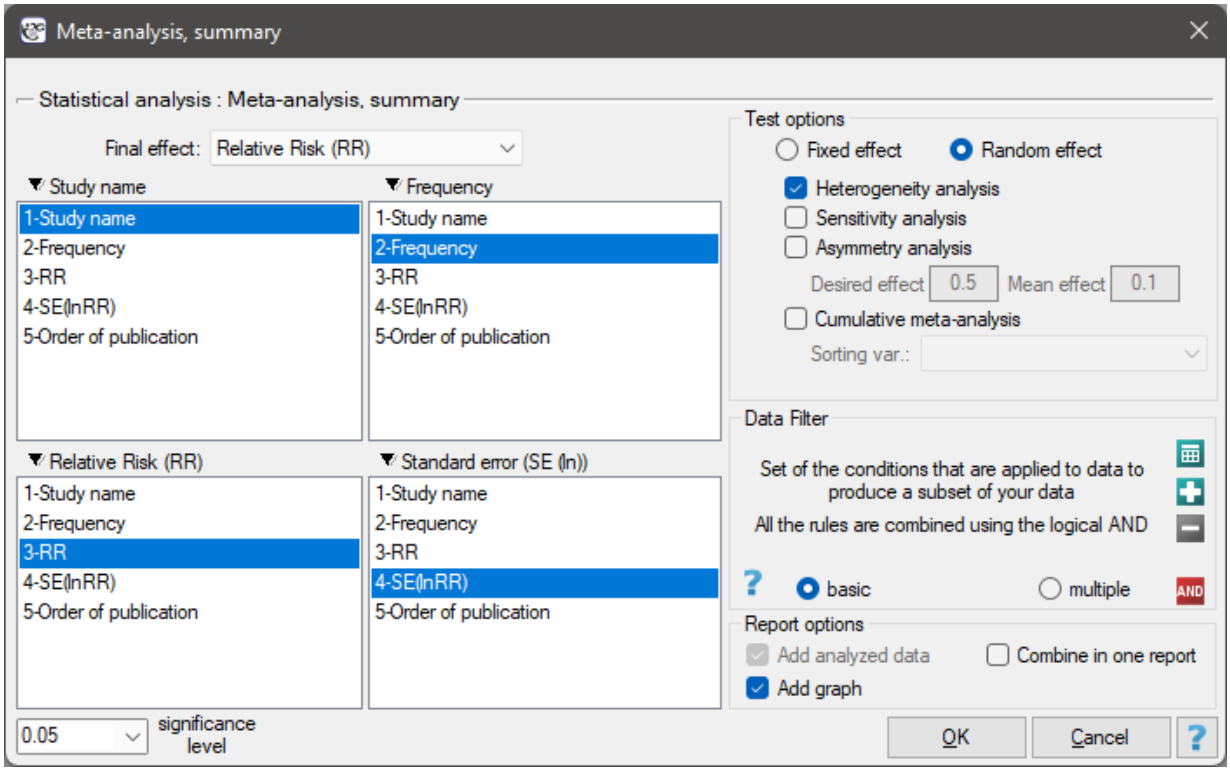
> **95% confidence interval** (the width of the diamond on the forest plot) – means that in 95% of cases of such meta-analyses, the summary random effect will fall within the interval determined by the diamond.
>
> **95% prediction interval** – means that in 95% of cases the true (population) effect of a new study will fall within the determined interval.

Study weight and its contribution – the weight w_i of a study depends on the observed variability: for the fixed effect it results only from the sampling error (SE) of a given study, and for the random effect – additionally from differences between studies, i.e. from the observed variance T^2 (the smaller the error, the greater the weight and the greater the influence of the study on the meta-analysis result). Based on the weights assigned to individual studies, the **contribution** of a given study to the whole analysis is determined, i.e. the percentage that the weight of a given study represents in relation to the total weight of all included studies.

HOW TO SET UP THE ANALYSIS

The window with the meta-analysis settings is opened via the menu:
Advanced statistics→Meta-analysis→Summary.

Copyright ©2010-2026 U-smile Stats – All rights reserved



In this window, depending on the selected final effect, a meta-analysis summary can be performed, along with basic analyses allowing its assumptions to be checked, such as heterogeneity or publication bias (sensitivity testing, asymmetry testing), as well as a cumulative meta-analysis.

PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 29.2. (file MetaanalizaRR.pqs)
The risk of disease X was studied for smokers and non-smokers. Some scientific papers indicated that the risk of disease X was higher for smokers, while others did not prove such a relationship. To establish whether cigarette smoking has an effect on the occurrence of disease X, a meta-analysis was planned. A thorough literature review was performed, and on this basis 10 scientific papers were selected that had data allowing the relative risk (i.e. the risk of illness for smokers relative to the risk of illness for non-smokers) and its error (the precision of a given study) to be calculated. The data was prepared for meta-analysis and saved in a file.

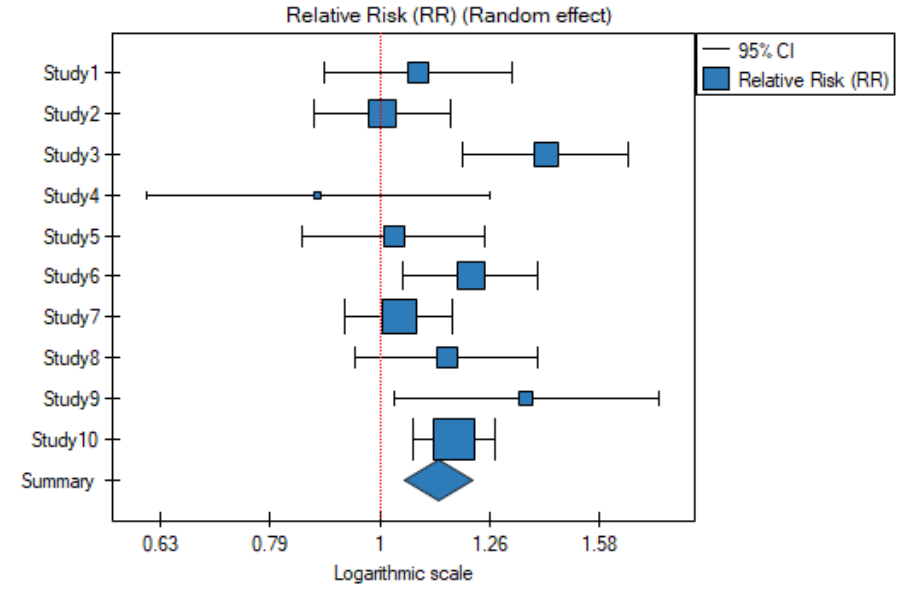
Because the papers included in the meta-analysis came from different centers and covered slightly different populations, the summary was performed by selecting the random effect. The relative risk was selected as the final effect, and the results were presented on a forest plot.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Meta-analysis, summary	
Analysed variables	Study name
	Frequency
	RR
	SE(lnRR)
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Effect	Random effect
Final effect	Relative Risk (RR)
Number of studies	10
Prediction interval of random effect	
-95%CI	0.9307
+95%CI	1.3754

Data :

Name	Frequency	Relative Risk	SE (ln)	-95%CI	+95%CI	Z statistic	p-value	Variance	Weight	Contribution
Study1	323	1.0838	0.1003	0.8905	1.3192	0.8027	0.4221	0.0101	62.9309	8.409%
Study2	706	1.0054	0.0734	0.8707	1.1609	0.0734	0.9415	0.0054	89.1127	11.907%
Study3	615	1.4168	0.0886	1.1911	1.6853	3.9344	0.0001	0.0078	73.1038	9.768%
Study4	89	0.8782	0.1833	0.6131	1.2578	-0.7088	0.4785	0.0336	25.3506	3.387%
Study5	278	1.03	0.098	0.85	1.2481	0.3016	0.7629	0.0096	64.7602	8.653%
Study6	560	1.21	0.0724	1.05	1.3944	2.6342	0.0084	0.0052	90.302	12.066%
Study7	1207	1.04	0.057	0.93	1.163	0.6876	0.4917	0.0033	110.0022	14.699%
Study8	214	1.15	0.0975	0.95	1.3921	1.4338	0.1516	0.0095	65.1908	8.711%
Study9	285	1.36	0.1418	1.03	1.7957	2.1684	0.0301	0.0201	38.5429	5.15%
Study10	1968	1.17	0.0437	1.074	1.2746	3.5928	0.0003	0.0019	129.0799	17.248%
Summary	6245	1.1314	0.0366	1.0532	1.2154	3.3773	0.0007			



The results of four studies (studies 3, 6, 9 and 10) indicate a significantly higher risk of illness for smokers. The result summarizing the performed meta-analysis is also statistically significant and confirms the same effect.

Copyright ©2010-2026 U-smile Stats – All rights reserved

RR[95%CI]=1.13[1.05-1.22]

Cigarette smoking increases the risk of disease X (the summary effect is statistically significant).

Unfortunately, the prediction interval of the random effect is wider: [0.93-1.38], which means that in 95% of cases the true, population relative risk obtained in subsequent studies could be either greater or smaller than one.

Note! Before interpreting the results, the assumptions of the meta-analysis should be checked. In this case, one should consider excluding the third study (see the sensitivity analysis results, the asymmetry analysis, the cumulative meta-analysis, and the heterogeneity assumption).

FOR THOSE INTERESTED

Study weight

Formulas:

For the fixed effect, the variability results only from the sampling error (the error within each study), i.e. from the size of SE:

$$w_i = \frac{1}{SE^2}$$

For the random effect, the variability results from the sampling error (the error within each study), i.e. from the size of SE, and from the differences between studies, i.e. from the observed variance T^2:

$$w_i = \frac{1}{SE^2 + T^2}$$

Heterogeneity testing

WHEN TO USE:

It is hard to expect every study to end up with exactly the same effect size – results obtained in different papers will naturally be somewhat different. Heterogeneity testing is intended to establish to what extent the differences appearing between the effects obtained in different papers affect the summary effect constructed in the meta-analysis. **The summary effect summarizes the results obtained in individual papers well if the differences between the individual effects are natural, i.e. small.** Large differences in the observed effects may indicate that the studies are not homogeneous, and that it is necessary to separate more homogeneous subgroups, e.g. dividing the collected papers into several subgroups with respect to an additional factor. For example: a given drug acts differently on younger people than on older people, so in papers based mainly on data from young people the obtained effect may differ considerably from papers conducted on older people. Dividing the collected papers into more homogeneous subgroups allows for a better estimation of the summary effect for each of these subgroups separately.

Heterogeneity testing checks whether the variability between studies is zero.

DEFINITIONS OF MEASURES

Copyright ©2010-2026 U-smile Stats – All rights reserved

Additional measures describing heterogeneity are the I^2 and H^2 coefficients. **The I^2 coefficient** determines the percentage of the observed variance that results from a true difference in the size of the studied effects (graphically, it reflects the degree of overlap of the confidence intervals of individual studies). Because it lies between 0% and 100%, it is easy to interpret and is widely used: if $I^2 = 0$, the entire observed variance of the effect sizes is "false", so if the confidence interval determined around the I^2 coefficient contains the value 0, the obtained variance can be considered statistically non-significant. The closer the value of I^2 is to 100%, the more one should consider abandoning a joint summary of the studies. It is assumed that $I^2 \approx 25\%$ indicates weak, $I^2 \approx 50\%$ moderate, and $I^2 \approx 75\%$ strong heterogeneity of the studies. **The H^2 coefficient**, on the other hand, is considered relative to the value 1 – if the confidence interval for H^2 contains the value 1, the obtained variance can be considered statistically non-significant, and the higher the value of H^2, the greater the heterogeneity of the studies.

Note! If the heterogeneity test gives a statistically significant result, this is a strong suggestion to abandon a joint summary of all collected studies. If the result is not statistically significant, the studies can be summarized with a common effect – although it is generally suggested that the summary should usually be performed with the random effect. This is because a random-effect summary takes into account the variability between studies (T^2), while a fixed-effect summary does not take it into account. However, if T^2 is small, the result of the fixed-effect model will be close to the results of the random-effect model, and when $T^2 = 0$, both models will give exactly the same result.

HOW TO SET UP THE ANALYSIS

The Heterogeneity testing option is available in the meta-analysis Summary window (see above).

PERFORMING AND INTERPRETING THE ANALYSIS

Example (29.2) cont. *(file MetaanalizaRR.pqs)*
When studying the effect of cigarette smoking on the occurrence of disease X, the heterogeneity assumption of the studies was checked. For this purpose, the Heterogeneity testing option was selected in the analysis window.

Heterogeneity analysis	
Q-statistic	17.398
Degrees of freedom	9
p-value	0.0428

Heterogeneity analysis			
	Values	-95%CI	+95%CI
T2	0.0058	0	0.0188
H2	1.9331	0.9339	4.0014
I2	48.27%	0%	75.009%

$p = 0.04284$ for the Q statistic (statistically significant, $\alpha = 0.05$)

The variance of the observed effects is not zero ($T^2 = 0.0058$), and the coefficient $I^2 = 48.27\%$ indicates moderate heterogeneity of the studies.

Only the confidence interval for the H^2 coefficient considers the variability between studies to be non-significant (the interval for this coefficient is [0.93-4.00]). Bearing the above results in mind, it should be

Copyright ©2010-2026 U-smile Stats – All rights reserved

considered whether the collected papers can be summarized with one common final effect (a common relative risk), or whether it is worth determining a more homogeneous group of papers and performing the analysis again.

FOR THOSE INTERESTED

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \tau^2 = 0, \\ \mathcal{H}_1: &\quad \tau^2 \neq 0, \end{aligned}$$

where:
τ^2 – the variance of the true (population) effects of individual studies.

Formulas:
The test statistic is of the form:

$$Q = T^2 W + k - 1$$

where:
T^2 – the variance of the observed effects,
W – a coefficient calculated based on the weights assigned to individual studies,
k – number of studies.

This statistic asymptotically (for large sample sizes) has a χ^2 distribution with degrees of freedom calculated by the formula $df = k - 1$.

The p value determined based on the test statistic is compared with the significance level α: when $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; when $p > \alpha$, there are no grounds to reject $\mathcal{H}_0$.

Formulas (coefficients I^2 and H^2):
The I^2 and H^2 coefficients are given by the formulas:

$$I^2 = \frac{H^2 - 1}{H^2}, \qquad H^2 = \frac{Q}{k - 1}.$$

Sensitivity testing

WHEN TO USE:

The summary effect of a study may change depending on which studies we include in the analysis and which we exclude from it. It is the researcher's responsibility to check how sensitive the analysis is to changes in the study selection criteria. Checking sensitivity allows the **changes in the summary effect** resulting from removing a given study to be established. The studies should be similar enough that removing one of them does not completely change the interpretation of the summary effect.

DEFINITIONS OF MEASURES

The value assigned to a given study called **contribution of the remaining studies** defines the percentage

Copyright ©2010-2026 U-smile Stats – All rights reserved

represented by the combined weight of the studies remaining in the analysis after a given study has been excluded. **Precision change**, on the other hand, indicates how the precision of the summary effect (the width of the confidence interval) will change when a given study is excluded from the analysis. A good illustration of a sensitivity analysis is a forest plot of the effect size and a plot of the precision changes when individual studies are excluded.

HOW TO SET UP THE ANALYSIS

The Sensitivity testing option is available in the meta-analysis Summary window (see above).

PERFORMING AND INTERPRETING THE ANALYSIS

Example (29.2) cont. *(file MetaanalizaRR.pqs)*
When studying the effect of cigarette smoking on the occurrence of disease X, the sensitivity of the analysis to excluding individual studies was checked. For this purpose, the Sensitivity testing option was selected in the analysis window, along with the forest plot (sensitivity) and bar plot (sensitivity).

Sensitivity

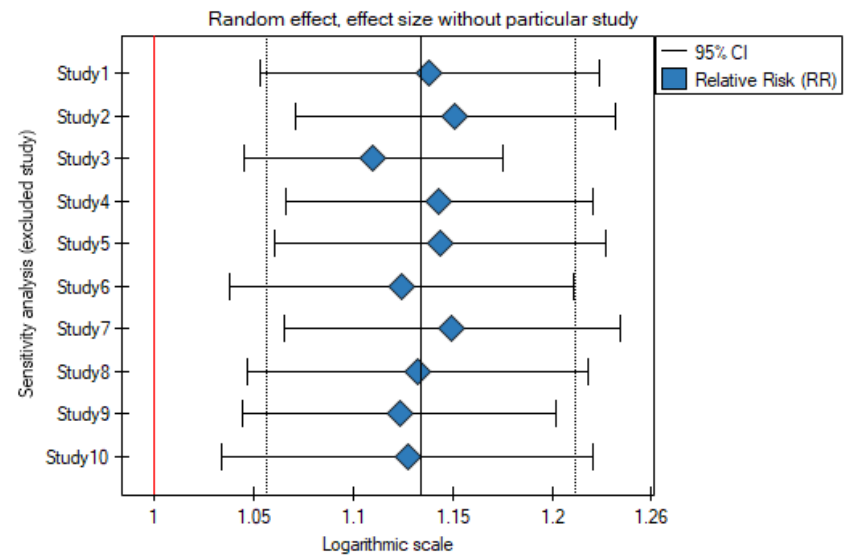
Excluded studies	Frequency	Relative Risk	SE (ln)	-95%CI	+95%CI	Z statistic	p-value	Remaining c	Precision cha
Study1	5922	1.136	0.04	1.0504	1.2286	3.1896	0.0014	91.591%	9.851%
Study2	5539	1.1496	0.0379	1.0673	1.2382	3.6788	0.0002	88.093%	5.338%
Study3	5630	1.1068	0.0306	1.0424	1.1752	3.3168	0.0009	90.232%	-18.14%
Study4	6156	1.1411	0.0363	1.0628	1.2253	3.6369	0.0003	96.613%	0.145%
Study5	5967	1.1417	0.0391	1.0574	1.2326	3.3865	0.0007	91.347%	7.998%
Study6	5685	1.1211	0.0406	1.0354	1.2139	2.8163	0.0049	87.934%	10.056%
Study7	5038	1.1479	0.0397	1.062	1.2408	3.4737	0.0005	85.301%	10.236%
Study8	6031	1.1296	0.0403	1.0438	1.2225	3.0247	0.0025	91.289%	10.099%
Study9	5960	1.1203	0.037	1.0418	1.2047	3.0661	0.0022	94.85%	0.36%
Study10	4277	1.1244	0.044	1.0316	1.2255	2.6667	0.0077	82.752%	19.549%

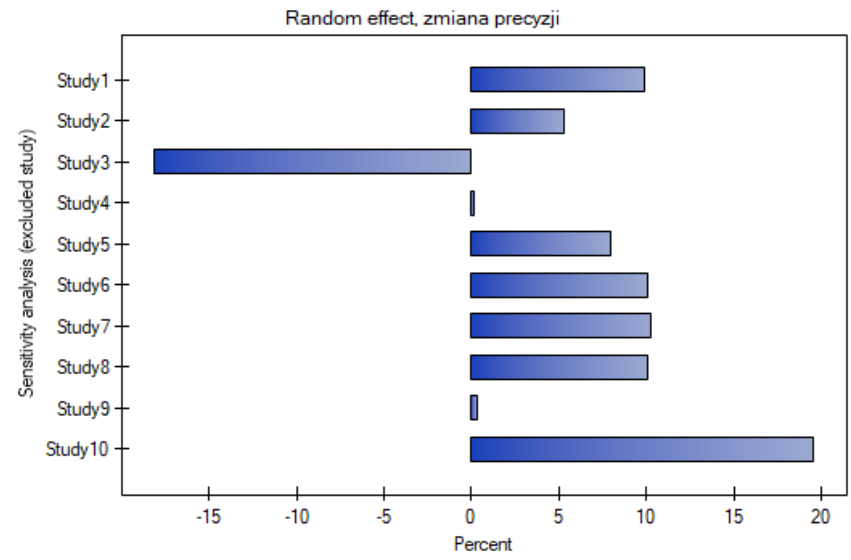
The results for the common relative risk, even when individual studies are not included, remain statistically significant.

Excluding Study 3

The precision of the obtained summary increases – the confidence interval for the final effect is then about 18% narrower.

Copyright ©2010-2026 U-smile Stats – All rights reserved





The analysis of the charts leads to the same conclusions – the narrowest interval and the most favorable change in precision are obtained after excluding Study 3.

Asymmetry testing (publication bias)

WHEN TO USE:

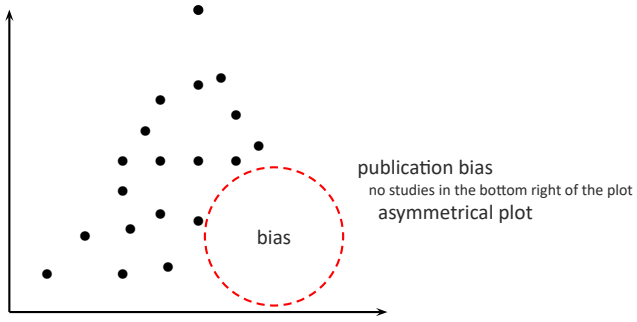
Symmetry of the obtained effects usually indicates the absence of **publication bias**. However, it should be remembered that many objective factors can disturb symmetry – e.g. studies with statistically non-significant effects, or small studies, are often not published, which makes it much harder to find such results. At the same time, there are no sufficiently comprehensive and universal statistical tools for detecting asymmetry, and as a consequence a considerable proportion of meta-analyses are published

Copyright ©2010-2026 U-smile Stats – All rights reserved

despite a diagnosed asymmetry; such studies, however, require good justification for proceeding this way.

DEFINITIONS OF MEASURES

Funnel plot



The standard way of checking for publication bias in the form of asymmetry is a funnel plot, showing the relationship between the size of a study (Y axis) and the size of the summary effect (X axis). It is assumed that large studies (placed in the upper part of the plot), in a properly selected set, are located close to each other and determine the center of the funnel, while smaller studies are located lower and are more varied and symmetrically distributed. Instead of the study size on the Y axis, the effect error of a given study can be shown, which is a better solution – the effect error is a measure that indicates the quality of the study and also carries information about its sample size.

Egger's test – since the interpretation of a funnel plot is always subjective, the Egger coefficient (Egger, 1997[55]) – the intercept of a fitted regression line – can be helpful. This coefficient is based on the correlation between the reciprocal of the standard error and the ratio of the effect size to its error. The further the value of the coefficient is from 0, the greater the asymmetry. The direction of the coefficient determines the type of asymmetry: a positive value together with a positive confidence interval indicates too large an effect size in small studies, and a negative value together with a negative confidence interval – too small an effect size in small studies.

Note! Egger's test should be used only when there is a large diversity in study sizes and there is a medium-sized study present. With few studies (a small number k), it is difficult to achieve a significant result despite a visible asymmetry.

Testing the "fail-safe" number – the number of papers that would be needed to undermine the obtained summary effect. **Rosenthal's Nfs** (Rosenthal, 1979[143]) determines the number of papers not indicating the presence of an effect (e.g. a mean difference of 0, an odds ratio of 1, etc.) that is needed to reduce the summary effect from statistically significant to non-significant. Rosenthal (1984[132]) defined the number of papers constituting the **cutoff point** as $5k + 10$. By determining the ratio of Nfs to the cutoff point, we obtain the **coefficient(fs)** – if it is greater than 1, the probability of publication bias is minimal. **Orwin's Nfs** (Orwin, 1983[132]), on the other hand, determines the number of papers with the average effect indicated by the researcher that is needed to reduce the summary effect to the desired size indicated by the researcher.

HOW TO SET UP THE ANALYSIS

Copyright ©2010-2026 U-smile Stats – All rights reserved

The Asymmetry testing option is available in the meta-analysis Summary window (see above).

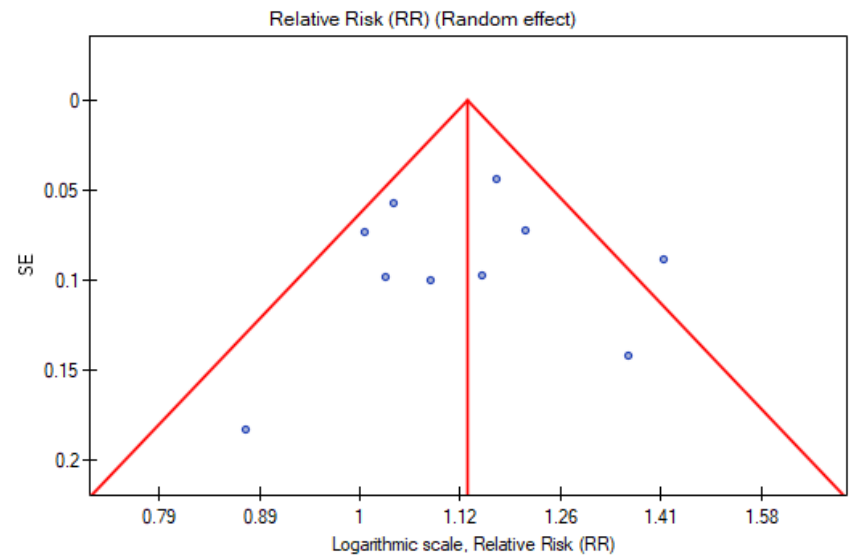
PERFORMING AND INTERPRETING THE ANALYSIS

Example (29.2) cont. *(file MetaanalizaRR.pqs)*
When studying the effect of cigarette smoking on the occurrence of disease X, the asymmetry assumption of the studies, i.e. publication bias, was checked. For this purpose, the Asymmetry testing option was selected in the analysis window, along with the funnel plot.

Asymmetry analysis	
b Egger coefficient	-0.1419
SE(b)	1.268
-95% CI for b coefficient	-3.0659
+95% CI for b coefficient	2.7822
t-test statistic for b	-0.1119
Degrees of freedom	8
p-value	0.9137
Number of studies 'Fail-safe'	
Rosenthal's method	
N(fs)	72
Cut-off point N(fs)	60
Index (fs)	1.2
Orwin's method	
Desired effect	1.12
Mean effect	1.11
N(fs)	11

Egger's test: $p = 0.9137$ **(not statistically significant,** $\alpha = 0.05$**)**

No publication bias – the distribution of the studies on the funnel plot is practically symmetrical.



Copyright ©2010-2026 U-smile Stats – All rights reserved

The points representing the individual studies are symmetrically distributed on the funnel plot. Admittedly, one study lies outside the triangle (Study 3), but it is close to its edges, so based on the plot too we have no fundamental objections regarding the selection of studies – the only doubt concerns the third study.

The "fail-safe" number of publications determined using Rosenthal's method is large, at 72. Therefore, for the summary effect (the relative risk common to all studies) to become statistically non-significant (i.e. for cigarette smoking to have no effect on the risk of disease X), 72 more papers, in which the relative risk equals one, would need to be added to the collected papers. The obtained effect can therefore be considered stable, since it will not be easy (with a small number of papers) to undermine the obtained effect.

The obtained common relative risk is RR=1.13. Using Orwin's method, it was checked how many papers with a relative risk of 1.11 would need to be included for the common relative risk to fall to 1.12. The result obtained was 11 papers. When reducing the relative risk from 1.11 to 1.10, however, only 5 papers are needed for the common relative risk to be 1.12.

FOR THOSE INTERESTED

Egger's test
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad b = 0, \\ \mathcal{H}_1: &\quad b \neq 0, \end{aligned}$$

where:
b – the intercept in Egger's regression equation.

Formulas:
The test statistic is of the form:

$$t = \frac{b}{SE(b)}$$

where:
$SE(b)$ – the standard error of the intercept.

The test statistic has a t-Student distribution with $k-2$ degrees of freedom.

The p value determined based on the test statistic is compared with the significance level α: when $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; when $p > \alpha$, there are no grounds to reject $\mathcal{H}_0$.

Rosenthal's Nfs

$$Nfs = \frac{\left(\sum_{i=1}^{k} Z_i\right)^2}{Z_c^2} - k$$

where:
Z_i – the value of the test statistic (with normal distribution) for a given study,
Z_c – the critical value of the normal distribution for the given significance level,
k – number of studies in the meta-analysis.

Orwin's Nfs

$$Nfs = k\frac{M - M_d}{M_d - \overline{M_{fs}}}$$

Copyright ©2010-2026 U-smile Stats – All rights reserved

where:
M – the summary effect obtained in the meta-analysis,
M_d – the desired effect size indicated by the researcher,
$\overline{M_{fs}}$ – the average effect of the papers added to the analysis, indicated by the researcher.

Cumulative meta-analysis

WHEN TO USE:

The typical purpose of performing a cumulative meta-analysis is to show how the effect has changed since the last meta-analysis on a given topic was performed/published, or how it has changed over recent years. In that case, subsequent studies are added chronologically (according to the timeline), and each time a summary effect is calculated. Equally important is a cumulative analysis examining the change of the summary effect depending on the magnitude of the influence of a selected additional factor – then the studies are sorted according to the magnitude of this factor, and **for successively added studies a common summary effect is calculated**.

DEFINITIONS OF MEASURES

Depending on the purpose of the cumulation, a variable according to which individual studies will be sorted should be chosen, i.e. the order in which studies are added to the meta-analysis summary should be established – this can be any numeric variable. The value assigned to a given study called **cumulative contribution** defines the percentage represented by the combined weight of the studies included in the analysis, i.e. the given study and the studies preceding it. **Precision change**, on the other hand, indicates how the precision of the summary effect (the width of the confidence interval) will change when a given study is added to the studies preceding it. A good illustration of a cumulative analysis is a forest plot of the effect size and a plot of the precision changes as individual studies are added.

HOW TO SET UP THE ANALYSIS

The Cumulative meta-analysis option is available in the meta-analysis Summary window (see above).

PERFORMING AND INTERPRETING THE ANALYSIS

Example (29.2) cont. *(file MetaanalizaRR.pqs)*
When studying the effect of cigarette smoking on the occurrence of disease X, it was checked how the obtained results evolved over time. For this purpose, the Cumulative meta-analysis option was selected in the analysis window, along with the variable according to which successive papers would be included in the meta-analysis, and the forest plot (cumulative) and bar plot (cumulative) were selected.

Copyright ©2010-2026 U-smile Stats – All rights reserved

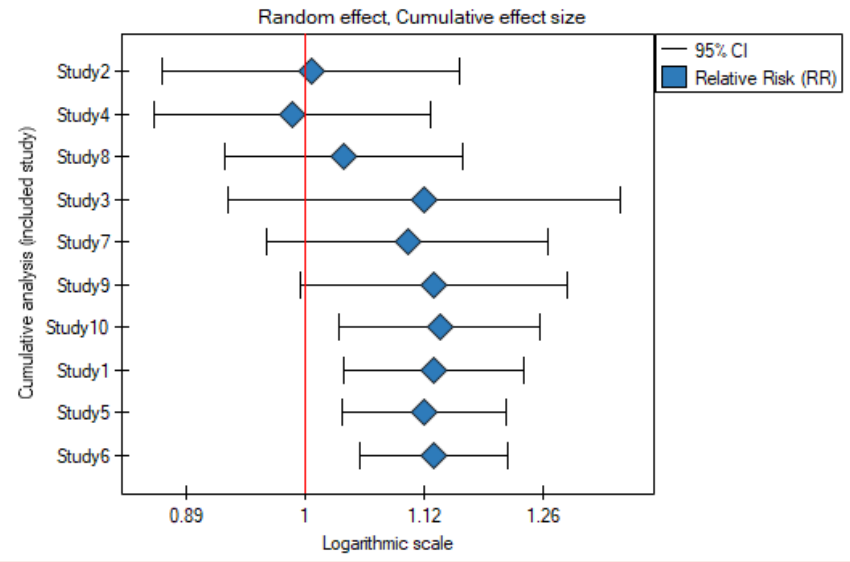
Cumulative										
Name	Frequency	Relative Risk	SE (ln)	-95%CI	+95%CI	Z statistic	p-value	Cumulative	Precision cha	Sorting valu
Study2	706	1.0054	0.0734	0.8707	1.1609	0.0734	0.9415	11.907%	NA	1
Study4	795	0.9868	0.0681	0.8634	1.1277	-0.1952	0.8452	15.295%	-8.922%	2
Study8	1009	1.0377	0.0586	0.9251	1.1641	0.6317	0.5276	24.006%	-9.557%	3
Study3	1624	1.1219	0.0968	0.928	1.3562	1.1878	0.2349	33.774%	79.154%	4
Study7	2831	1.104	0.0695	0.9634	1.2652	1.4237	0.1545	48.473%	-29.526%	5
Study9	3116	1.1322	0.0659	0.9951	1.2882	1.8847	0.0595	53.623%	-2.869%	6
Study10	5084	1.1387	0.0498	1.0327	1.2555	2.6064	0.0091	70.871%	-24.022%	7
Study1	5407	1.132	0.0443	1.0379	1.2346	2.8011	0.0051	79.28%	-11.714%	8
Study5	5685	1.1211	0.0406	1.0354	1.2139	2.8163	0.0049	87.934%	-9.184%	9
Study6	6245	1.1314	0.0366	1.0532	1.2154	3.3773	0.0007	100%	-9.137%	10

As new papers appeared, the obtained summary effect gained in strength, and its significance was achieved after Study 10 was added to the earlier papers, and then also with subsequent studies.

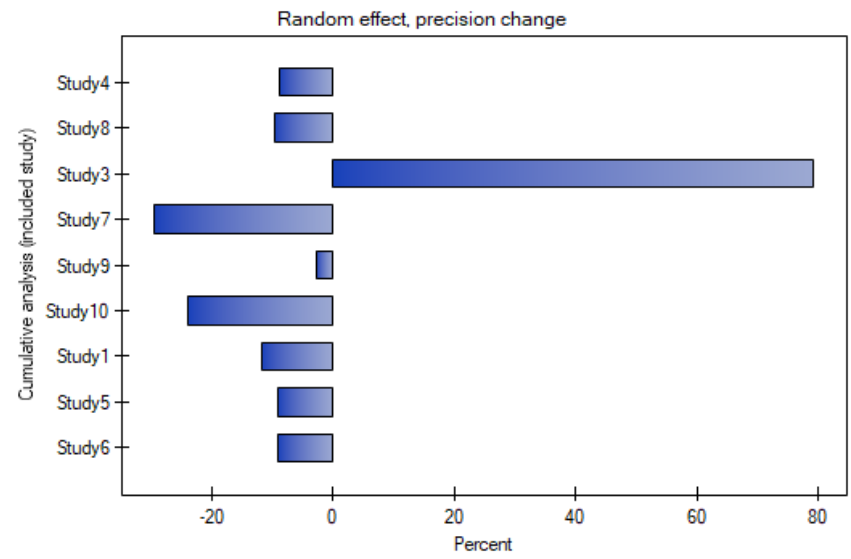
Adding Study 3

Exceptionally weakens the precision of the summary – the confidence interval of the summary relative risk widened by about 79.15% at that point.

In general, adding subsequent papers increased the precision of the obtained relative risk, with the exception of the situation in which Study 3 was added. This effect can be seen in the table and in the accompanying charts.



Copyright ©2010-2026 U-smile Stats – All rights reserved



As a result, one should consider excluding Study 3 from the meta-analysis performed.

Group comparison

WHEN TO USE:

There are situations in which the collected data concern the same effect, obtained on the same population, but under somewhat different conditions – e.g. some studies were performed under condition A, and some under condition B. It may then be interesting to compare the summary effects obtained for each group. Demonstrating differences between the summary effects can be the main goal of a meta-analysis – in that case it is not advisable to summarize both subgroups with a single common effect. If, however, the researcher is aware that the studies were performed under different conditions, but a joint summary of all the studies still seems purposeful, then after demonstrating the absence of statistically (or clinically) significant differences, they can perform a joint summary that takes into account the division into subgroups A and B, i.e. determine a common summary corrected for the different experimental conditions.

For example: country A has a somewhat different climate than country B. Having a series of studies concerning both countries and examining the vegetation of these countries, we can check whether climatic conditions affect the obtained study effects. A comparative analysis of the subgroups determined in this way will allow us to assess whether climate has a fundamental influence on the obtained results, and whether the results of the studies covering both countries can be summarized with one common effect, or whether separate summaries should be determined for each country. Another example could be dividing studies into those with full randomization and those without it, to check whether the studies without randomization give effects close enough to the studies with randomization to be included in a further, joint analysis.

DEFINITIONS OF MEASURES

A comparison of groups can be performed by choosing as the summary effect: the fixed effect, the

Copyright ©2010-2026 U-smile Stats – All rights reserved

random effect (separate T^2), or the random effect (pooled T^2), where T^2 is the variance of the observed effects.

- **Fixed effect** is chosen when we assume that the studies within each group share one common true (i.e. population) effect.
- **Random effect (separate T^2)** is chosen when we assume that the studies within each group represent slightly different populations, and that the groups differ in the variance between studies.
- **Random effect (pooled T^2)** is chosen when we assume that the studies within each group represent slightly different populations, but the variance between studies is the same regardless of the group to which they belong.

As a result, each group is described separately with the chosen summary effect. The main goal is to compare the groups, i.e. to establish whether the compared groups differ in their true (population) summary effect – in practice this means checking whether the variance of the groups' summary effects is zero, i.e. **testing the heterogeneity of the groups**. The description and interpretation of the results of this analysis correspond to the chapter Heterogeneity testing, except that in the case of group comparison the heterogeneity concerns the summary effects of the compared groups, not individual studies, and the result depends on the chosen summary effect. If the result is statistically significant (the Q statistic, the I^2 coefficient, or the H^2 coefficient), this is a strong suggestion to abandon a joint summary of the compared groups.

Additional options for group comparison analysis:

Heterogeneity testing within groups The ability to analyze each group separately for heterogeneity (as described in the chapter Heterogeneity testing). The results obtained, in particular the determined variance T^2, make it easier to choose between the random effect (separate T^2) and the random effect (pooled T^2) for the group comparison.

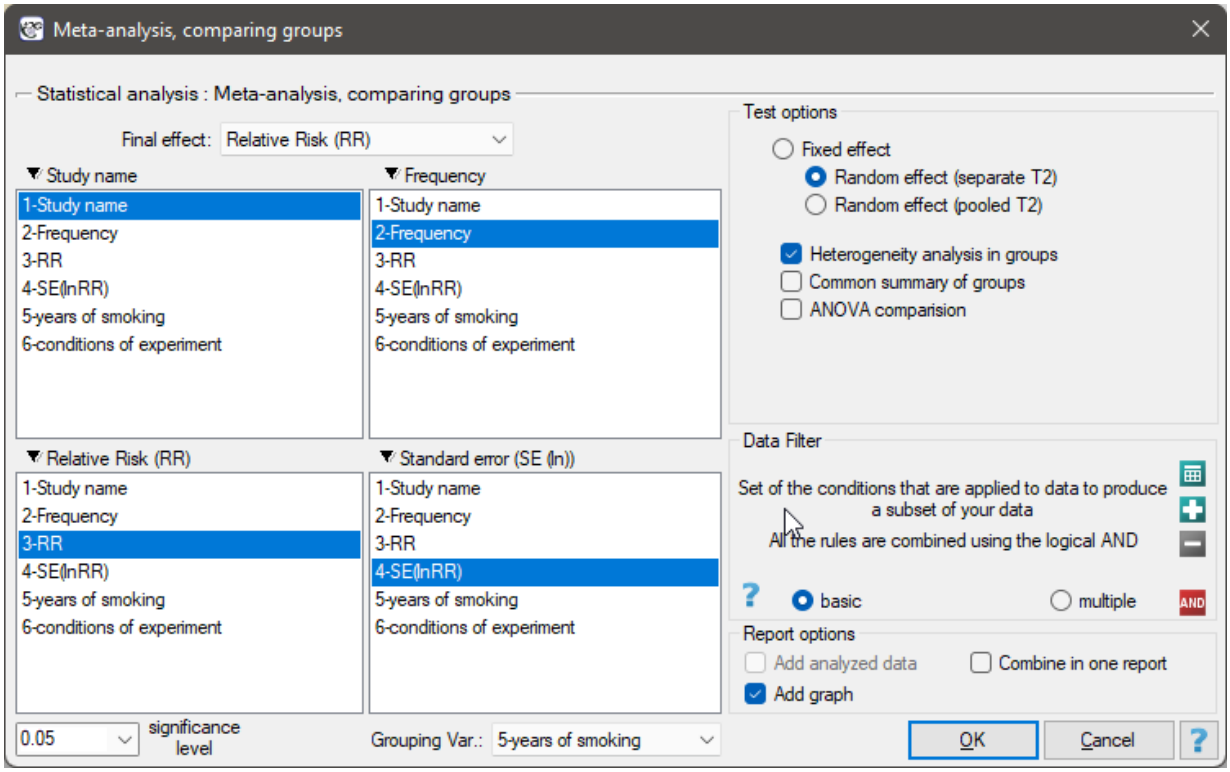
Joint summary of the groups If the differences obtained between the summary effects of the individual groups are small and non-significant, a joint summary of these groups, corrected for the division into the indicated groups, can be performed. The result of the joint summary depends on the observed differences (variability between studies and between groups), i.e. on the choice of summary effect. A good illustration of a joint summary of groups is a forest plot showing the results of individual studies together with the summary of each group and the joint summary of the groups.

ANOVA comparison An additional option for comparing groups, based on a different mathematical model than the comparison by testing the heterogeneity of groups, although both methods give converging results. The observed variance is broken down into between-group variance and within-group variance, and the latter – into the variance of each group separately.

HOW TO SET UP THE ANALYSIS

The window with the group comparison settings for meta-analysis is opened via the menu: Advanced statistics→Meta-analysis→Group comparison.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 29.3. (file MetaanalizaRR.pqs)
The risk of disease X was studied for smokers and non-smokers. To establish whether the length of time spent smoking has an effect on the occurrence of disease X, a meta-analysis was planned. A thorough literature review was performed, and on this basis 17 scientific papers were selected with a description of the relative risk and its error (the precision of a given study). Because the studies concerned different smoking durations, 3 groups of studies were distinguished:

(1) studies concerning people smoking for more than 10 years,

(2) studies concerning people smoking for 5 to 10 years,

(3) studies concerning people smoking for less than 5 years.

In addition, a division was made into two different conditions under which the studies were performed (different inclusion/exclusion criteria for subjects). The data was prepared for meta-analysis and saved in a file. The goal of performing the meta-analysis was to compare the age groups. In addition, it was checked whether the different experimental conditions translated into differences in the obtained relative risk.

Because the papers included in the meta-analysis came from different centers and covered slightly different populations, the summary was performed by selecting the random effect (separate T2). The relative risk was selected as the final effect, and the results were presented on a forest plot.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Meta-analysis, comparing groups	
Analysed variables	Study name
	Frequency
	RR
	SE(lnRR)
	years of smoking
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	years of smoking
Effect	Random effect separate T2
Final effect	Relative Risk (RR)
Number of groups	3
Heterogeneity analysis	
Comparison groups	
Q-statistic	9.375
Degrees of freedom	2
p-value	0.0092

Heterogeneity analysis			
Comparison groups	Values	-95%CI	+95%CI
T2	0.0046	0.0006	0.0174
H2	4.6875	1.4615	15.0346
I2	78.667%	31.576%	93.349%

$p = 0.0092$ for group heterogeneity (statistically significant, $\alpha = 0.05$)

The groups differ statistically significantly – this can be seen not only in the heterogeneity test, but also in the H^2 coefficient (together with a confidence interval above the value one) and the I^2 coefficient (78%, i.e. high heterogeneity).

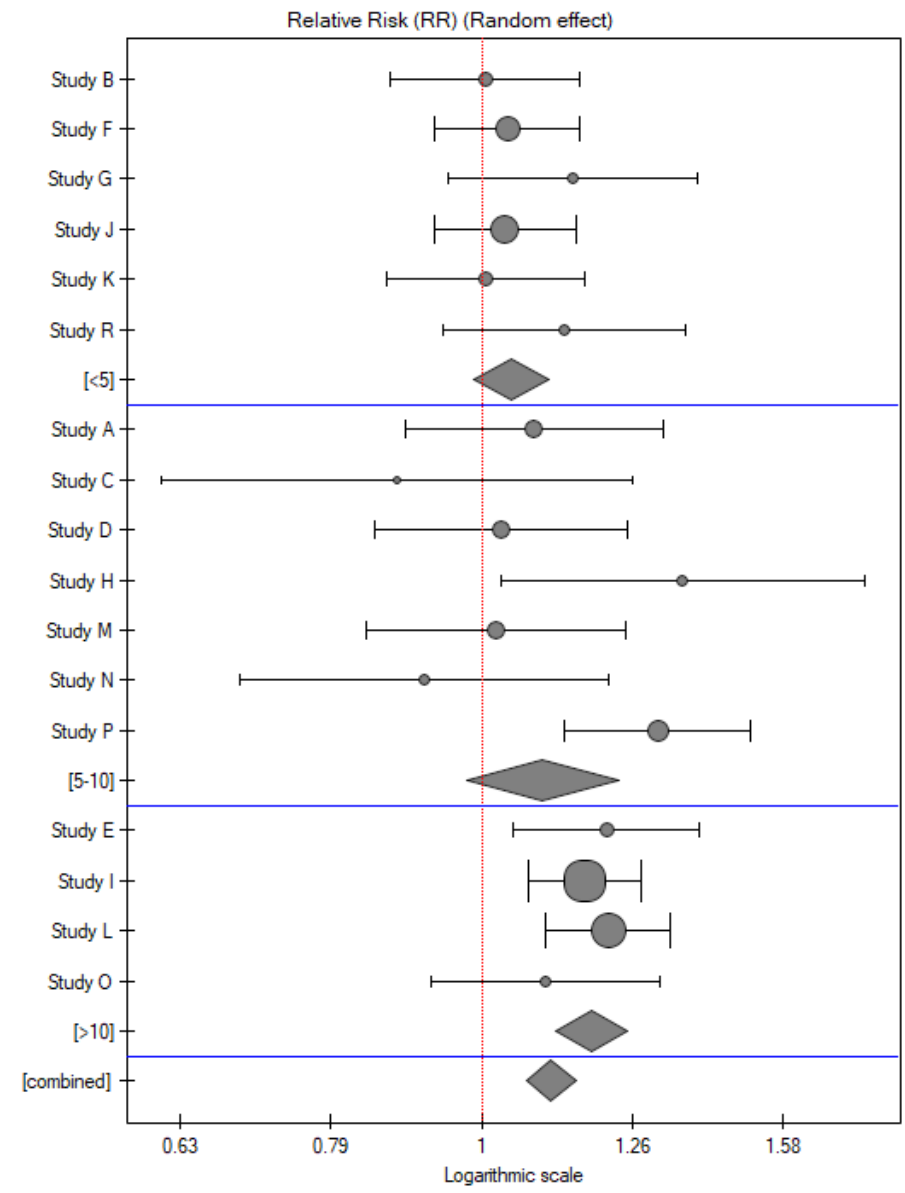
Therefore, the collected papers will not be summarized with a common effect, but only through a separate summary of each group.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Data :										
Name	Frequency	Relative Risk	SE (ln)	-95%CI	+95%CI	Z statistic	p-value	Variance	Weight	Contribution
Group:<5										
Study B	706	1.0054	0.0734	0.8707	1.1609	0.0734	0.9415	0.0054	185.7254	15.47%
Study F	1207	1.04	0.057	0.93	1.163	0.6876	0.4917	0.0033	307.3826	25.603%
Study G	214	1.15	0.0975	0.95	1.3921	1.4338	0.1516	0.0095	105.2394	8.766%
Study J	1155	1.0365	0.0556	0.9295	1.1558	0.6447	0.5191	0.0031	323.6753	26.96%
Study K	626	1.0061	0.0777	0.864	1.1715	0.0777	0.9381	0.006	165.7551	13.807%
Study R	270	1.1347	0.0942	0.9435	1.3647	1.3425	0.1794	0.0089	112.7798	9.394%
Summary	4178	1.0465	0.0289	0.989	1.1075	1.5764	0.1149			
Group:5-10										
Study A	323	1.0838	0.1003	0.8905	1.3192	0.8027	0.4221	0.0101	46.5426	16.338%
Study C	89	0.8782	0.1833	0.6131	1.2578	-0.7088	0.4785	0.0336	22.2014	7.793%
Study D	278	1.03	0.098	0.85	1.2481	0.3016	0.7629	0.0096	47.5356	16.687%
Study H	285	1.36	0.1418	1.03	1.7957	2.1684	0.0301	0.0201	31.7054	11.13%
Study M	250	1.0232	0.1011	0.8392	1.2476	0.2269	0.8205	0.0102	46.1598	16.204%
Study N	150	0.9164	0.1439	0.6911	1.2151	-0.6068	0.544	0.0207	31.1001	10.917%
Study P	481	1.3081	0.0731	1.1336	1.5095	3.6762	0.0002	0.0053	59.629	20.932%
Summary	1856	1.097	0.0592	0.9767	1.2321	1.5626	0.1182			
Group:>10										
Study E	560	1.21	0.0724	1.05	1.3944	2.6342	0.0084	0.0052	190.9672	15.096%
Study I	1968	1.17	0.0437	1.074	1.2746	3.5928	0.0003	0.0019	523.6452	41.395%
Study L	1583	1.2134	0.0485	1.1034	1.3345	3.9872	0.0001	0.0024	424.8383	33.584%
Study O	414	1.1027	0.0892	0.9257	1.3135	1.0952	0.2734	0.008	125.5514	9.925%
Summary	4525	1.1835	0.0281	1.12	1.2505	5.9907	<0.0001			

The forest plot likewise presents the summary of each group and does not contain a joint summary of the groups.

Copyright ©2010-2026 U-smile Stats – All rights reserved



Additionally, the homogeneity within each group was checked, to make sure that they could be summarized separately.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Heterogeneity analysis	
<5	
Q-statistic	2.2729
Degrees of freedom	5
p-value	0.8102
5-10	
Q-statistic	11.7292
Degrees of freedom	6
p-value	0.0683
>10	
Q-statistic	1.0555
Degrees of freedom	3
p-value	0.7878

Heterogeneity analysis			
<5	Values	-95%CI	+95%CI
T2	0	0	0.0041
H2	0.4546	0.1153	1.7914
I2	0%	0%	44.177%
5-10	Values	-95%CI	+95%CI
T2	0.0114	0	0.0434
H2	1.9549	0.8268	4.6221
I2	48.846%	0%	78.365%
>10	Values	-95%CI	+95%CI
T2	0	0	0.0045
H2	0.3518	0.0539	2.2978
I2	0%	0%	56.48%

The results of the comparison concerning the different study conditions, on the other hand, indicate that these conditions have no significant effect on the summary effect. In this case, it is possible to calculate a common final effect corrected for the different study conditions, i.e. a joint summary of both groups.

Copyright ©2010-2026 U-smile Stats – All rights reserved

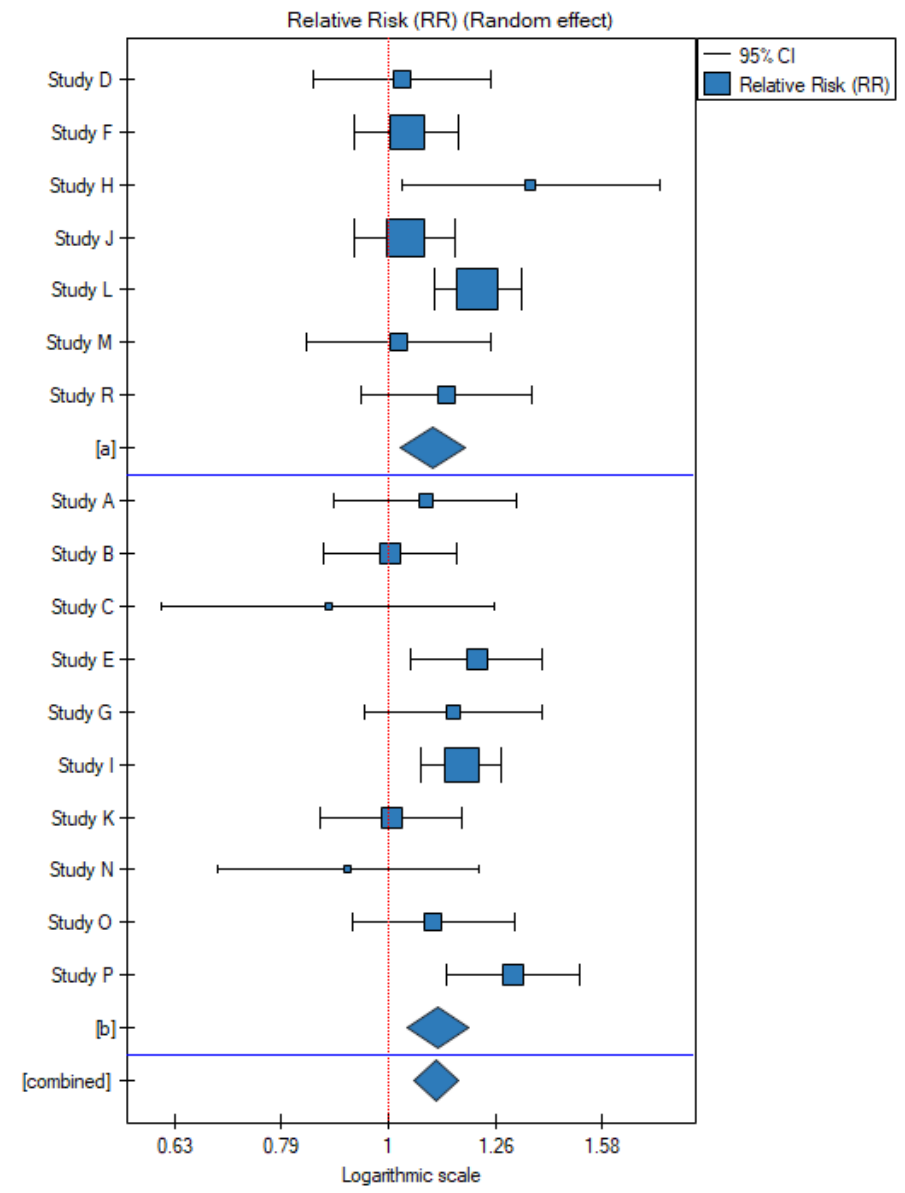
Meta-analysis, comparing groups	
Analysed variables	Study name
	Frequency
	RR
	SE(lnRR)
	conditions of experiment
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	conditions of experiment
Effect	Random effect separate T2
Final effect	Relative Risk (RR)
Number of groups	2
Heterogeneity analysis	
a	
Q-statistic	9.4928
Degrees of freedom	6
p-value	0.1477
b	
Q-statistic	14.5886
Degrees of freedom	9
p-value	0.1029
Comparison groups	
Q-statistic	0.0436
Degrees of freedom	1
p-value	0.8345

Heterogeneity analysis			
a	Values	-95%CI	+95%CI
T2	0.0031	0	0.0147
H2	1.5821	0.6668	3.7541
I2	36.794%	0%	73.363%
b	Values	-95%CI	+95%CI
T2	0.004	0	0.0155
H2	1.621	0.773	3.399
I2	38.308%	0%	70.58%
Comparison groups	Values	-95%CI	+95%CI
T2	0	NA	NA
H2	0.0436	NA	NA
I2	0%	NA	NA

Copyright ©2010-2026 U-smile Stats – All rights reserved

Data :										
Name	Frequency	Relative Risk	SE (ln)	-95%CI	+95%CI	Z statistic	p-value	Variance	Weight	Contribution
Group:a										
Study D	278	1.03	0.098	0.85	1.2481	0.3016	0.7629	0.0096	78.7338	10.055%
Study F	1207	1.04	0.057	0.93	1.163	0.6876	0.4917	0.0033	157.4759	20.111%
Study H	285	1.36	0.1418	1.03	1.7957	2.1684	0.0301	0.0201	43.0949	5.504%
Study J	1155	1.0365	0.0556	0.9295	1.1558	0.6447	0.5191	0.0031	161.6444	20.644%
Study L	1583	1.2134	0.0485	1.1034	1.3345	3.9872	0.0001	0.0024	183.4613	23.43%
Study M	250	1.0232	0.1011	0.8392	1.2476	0.2269	0.8205	0.0102	75.0299	9.582%
Study R	270	1.1347	0.0942	0.9435	1.3647	1.3425	0.1794	0.0089	83.5859	10.675%
Summary	5028	1.101	0.0357	1.0265	1.1809	2.6918	0.0071			
Group:b										
Study A	323	1.0838	0.1003	0.8905	1.3192	0.8027	0.4221	0.0101	71.0889	8.029%
Study B	706	1.0054	0.0734	0.8707	1.1609	0.0734	0.9415	0.0054	106.4033	12.018%
Study C	89	0.8782	0.1833	0.6131	1.2578	-0.7088	0.4785	0.0336	26.5792	3.002%
Study E	560	1.21	0.0724	1.05	1.3944	2.6342	0.0084	0.0052	108.1033	12.21%
Study G	214	1.15	0.0975	0.95	1.3921	1.4338	0.1516	0.0095	73.9861	8.357%
Study I	1968	1.17	0.0437	1.074	1.2746	3.5928	0.0003	0.0019	168.8162	19.068%
Study K	626	1.0061	0.0777	0.864	1.1715	0.0777	0.9381	0.006	99.5331	11.242%
Study N	150	0.9164	0.1439	0.6911	1.2151	-0.6068	0.544	0.0207	40.4279	4.566%
Study O	414	1.1027	0.0892	0.9257	1.3135	1.0952	0.2734	0.008	83.481	9.429%
Study P	481	1.3081	0.0731	1.1336	1.5095	3.6762	0.0002	0.0053	106.9341	12.078%
Summary	5531	1.1123	0.0336	1.0414	1.1881	3.1673	0.0015			
Summary										
a	5028	1.101	0.0357	1.0265	1.1809	2.6918	0.0071	0.0013	783.026	46.933%
b	5531	1.1123	0.0336	1.0414	1.1881	3.1673	0.0015	0.0011	885.353	53.067%
Summary	10559	1.107	0.0245	1.0551	1.1614	4.1514	<0.0001			

Copyright ©2010-2026 U-smile Stats – All rights reserved



FOR THOSE INTERESTED

Hypotheses for testing group heterogeneity:

$$\begin{aligned} \mathcal{H}_0: &\quad \tau^2 = 0, \\ \mathcal{H}_1: &\quad \tau^2 \neq 0, \end{aligned}$$

where:
τ^2 – the variance of the true (population) summary effects of the compared groups.

The p value determined based on the test statistic is compared with the significance level α: when $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; when $p > \alpha$, there are no grounds to reject $\mathcal{H}_0$.

ANOVA comparison

Copyright ©2010-2026 U-smile Stats – All rights reserved

Procedure:
Using the ANOVA approach, the total observed variance is broken down into between-group variance and within-group variance, and the latter into the variance of each group separately. As a result, the following Q statistics are determined:

- Q statistic (group 1) – examines the portion of the total variance relating to the first group, i.e. the variance between studies within the first group,
- Q statistic (group 2) – similarly for the second group,
- ...
- Q statistic (group g) – similarly for the last group,
- Q statistic (within groups) $= Q$ (group 1) $+ Q$ (group 2) $+ \ldots +$ Q (group g) – examines the portion of the total variance relating to the inside of the individual groups,
- Q **statistic (between groups)** – examines the portion of the total variance relating to the differences between groups (a result identical to the group heterogeneity test),
- Q statistic (total) – examines the variance between all studies.

Each of the above Q statistics has a χ^2 distribution with the appropriate number of degrees of freedom.

Meta-regression

WHEN TO USE:

Meta-regression is performed in a manner analogous to the regression analysis described in the chapter Multiple linear regression. In the case of meta-regression, the studied objects are the individual studies: their results (e.g. odds ratios, relative risks, mean differences) constitute the dependent variable Y, i.e. the explained variable, while the additional conditions under which these studies were conducted constitute the independent variables (X_1, X_2, ..., X_k), i.e. the explanatory variables. Just as in traditional regression models, the independent variables can interact, and those described on a nominal scale can be subject to special coding (more information in the chapter Preparing variables for analysis in multivariate models).

Meta-regression can be performed by choosing the fixed effect or the random effect.

- **Fixed effect** is chosen when we assume that the studies represent one common true effect, such that all factors that could disturb its magnitude are the same – except for the factors examined as independent variables in the model. This is a rare situation in practice, since it requires fully controlled conditions (difficult to satisfy when the studies are conducted by different centers and different researchers). It would be justified, for example, if all studies were performed by one center, on the same population, with only those conditions described by the examined characteristic being changed.
- **Random effect** is chosen when we assume that the studies may represent somewhat different populations, i.e. not all factors that could disturb the magnitude of the studied effect are descri-

Copyright ©2010-2026 U-smile Stats – All rights reserved

bed in every paper (they can be assumed to be similar, but do not have to be identical). The use of the random effect is more common, since individual studies are usually conducted by different centers under somewhat different conditions.

ASSUMPTIONS

– the number of independent variables should be small – smaller than the number of papers on which the study is based ($n \geq k + 1$).

DEFINITIONS OF MEASURES

Statistical significance of individual variables in the model – based on the coefficient and its error, we infer whether the independent variable for which this coefficient was estimated has a significant effect on the final effect.

Quality of the constructed meta-regression **model** is assessed with several measures:

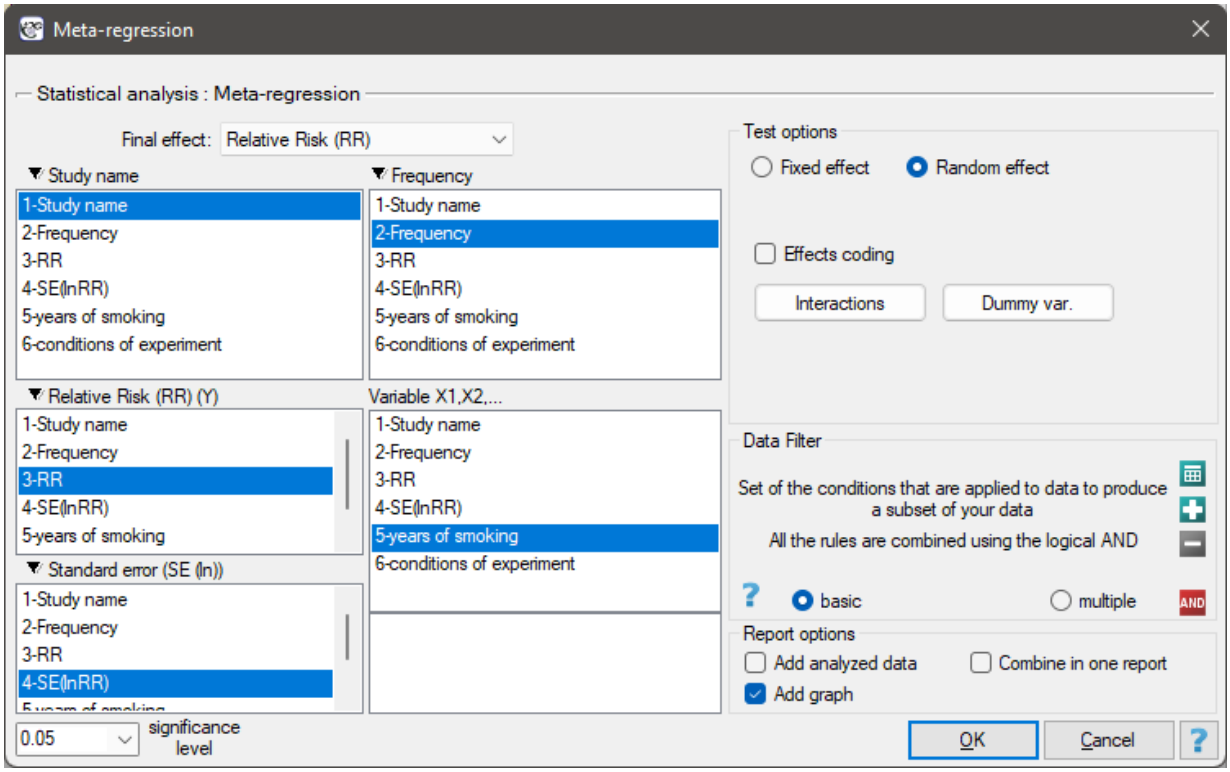
- **Coefficient** R^2 – a measure of model fit, expressing the percentage of variability between study effects explained by the model. The value of this coefficient lies in the range $< 0; 1 >$, where 1 means a perfect fit of the model, and 0 – a complete lack of fit.
- **Coefficient** I^2 – determines the percentage of the observed variance that results from a true difference in the size of the studied effects (cf. Heterogeneity testing).

Statistical significance of all variables in the model is checked using the ANOVA approach, determining the Q(model) statistic – the observed variance between studies is then broken down into the variance explained by the model and the residual variance (not explained by the model).

HOW TO SET UP THE ANALYSIS

The window with the meta-regression settings is opened via the menu:
Advanced statistics→Meta-analysis→Meta-regression.

Copyright ©2010-2026 U-smile Stats – All rights reserved



PERFORMING AND INTERPRETING THE ANALYSIS

Example (29.3) cont. *(file MetaanalizaRR.pqs)*
The risk of disease X was studied for smokers and non-smokers. To establish whether the length of time spent smoking has an effect on the occurrence of disease X, and whether the different experimental conditions translated into differences in the obtained relative risk, a meta-analysis comparing the identified groups of studies was performed. Based on the comparison of study groups, it was established that the last group (those smoking the longest, i.e. more than 10 years) shows a relationship between smoking and the occurrence of disease X, while for the groups with a shorter smoking duration no significant effect could be obtained. It was noted, however, that the effect systematically increases with the number of years of smoking. To test the hypothesis of a significant increase in the risk of disease X with increasing years of smoking, two regression models were built: in the first, the grouping variable Years of smoking was treated as a continuous variable, and in the second as a categorical (dummy) variable with a reference group of people smoking for less than 5 years. The data was prepared for meta-analysis and saved in a file.

Because the papers included in the meta-analysis came from different centers and covered slightly different populations, the meta-regression was performed by selecting the random effect. The relative risk was selected as the final effect, and the results were presented on a chart.

Model						
	b coeff.	b error	-95% CI	+95% CI	Z statistic	p-value
intercept	-0.0142	0.0447	-0.1019	0.0734	-0.3181	0.7504
years of smokin	0.0614	0.0203	0.0217	0.1011	3.0302	0.0024

Copyright ©2010-2026 U-smile Stats – All rights reserved

Model						
	b coeff.	b error	-95% CI	+95% CI	Z statistic	p-value
intercept	0.0462	0.0304	-0.0133	0.1057	1.5222	0.1279
years of smokin	0.0666	0.0513	-0.0338	0.1671	1.2999	0.1937
years of smokin	0.1218	0.043	0.0375	0.2062	2.8301	0.0047

A significant relationship between smoking duration and the size of the relative risk in both models

With the passage of smoking time (moving to the next group of years of smoking), the logarithm of the relative risk of disease X increases by 0.06139. In the second model, relative to those smoking for less than 5 years, the logarithm of the relative risk increases by 0.06663 for those smoking 5 to 10 years, and by as much as 0.12182 for those smoking more than 10 years.

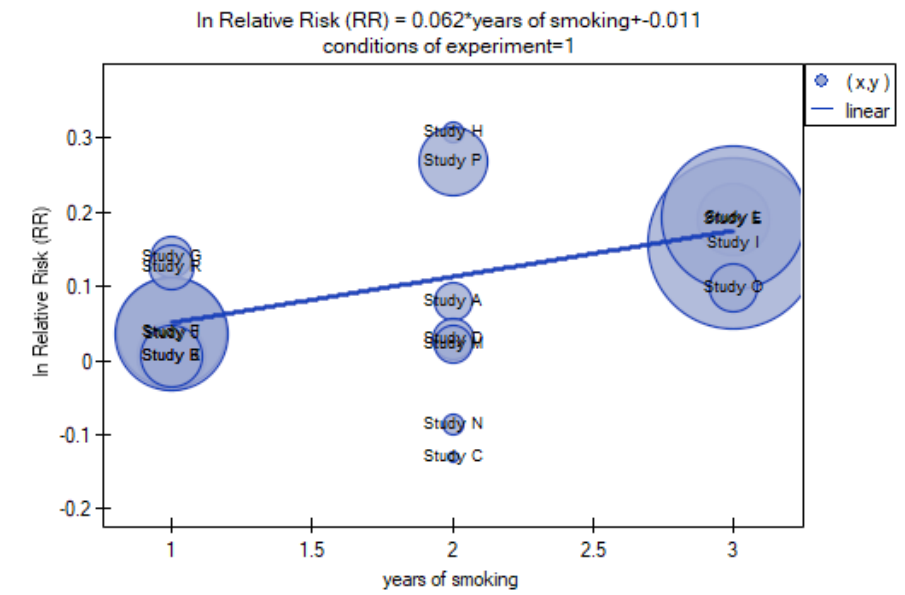
Because part of the studies was conducted according to other criteria (under different conditions), the results obtained for both models were corrected for the different study conditions.

Model						
	b coeff.	b error	-95% CI	+95% CI	Z statistic	p-value
intercept	-0.0027	0.0688	-0.1375	0.1322	-0.0386	0.9692
years of smokin	0.0621	0.0222	0.0185	0.1057	2.7908	0.0053
conditions of ex	-0.0085	0.0396	-0.0861	0.0691	-0.2155	0.8294

Model						
	b coeff.	b error	-95% CI	+95% CI	Z statistic	p-value
intercept	0.0595	0.0667	-0.0714	0.1903	0.8907	0.3731
years of smokin	0.0654	0.0536	-0.0397	0.1705	1.2196	0.2226
years of smokin	0.123	0.0473	0.0304	0.2157	2.6041	0.0092
conditions of ex	-0.0089	0.0418	-0.0908	0.0729	-0.2135	0.831

The correction performed did not change the underlying trend, so it can be concluded that the risk of disease X increases with the number of years of smoking, regardless of the methodology adopted (inclusion/exclusion criteria for subjects). The relationship obtained for the first model, assuming that the studies were conducted under condition “a” (indicated as the first condition), is shown in the chart.

Copyright ©2010-2026 U-smile Stats – All rights reserved



FOR THOSE INTERESTED

Statistical significance of an individual coefficient
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \beta_i = 0, \\ \mathcal{H}_1: &\quad \beta_i \neq 0. \end{aligned}$$

Formulas:
The test statistic is of the form:

$$Z = \frac{b_i}{SE_{b_i}}$$

The test statistic has a normal distribution.

The p value determined based on the test statistic is compared with the significance level α: when $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; when $p > \alpha$, there are no grounds to reject $\mathcal{H}_0$.

Coefficient R^2

$$R^2 = T^2_{(model)} + T^2_{(total)},$$

where:
$T^2_{(model)}$ – the between-study variance explained by the model,
$T^2_{(total)}$ – the total between-study variance.

Statistical significance of all variables in the model
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \text{all } \beta_i = 0, \\ \mathcal{H}_1: &\quad \text{there exists } \beta_i \neq 0. \end{aligned}$$

Procedure:
Using the ANOVA approach, the observed variance between studies is broken down into the variance

Copyright ©2010-2026 U-smile Stats – All rights reserved

explained by the model and the residual variance (not explained by the model). As a result, the following Q statistics are determined:

- Q statistic (residuals) – examines the portion of the total variance that is not explained by the model,
- Q statistic (model) – examines the portion of the total variance that is explained by the model,
- Q statistic (total) – examines the variance between all studies.

Each of the above Q statistics has a χ^2 distribution with the appropriate number of degrees of freedom.

The p value determined based on the test statistic is compared with the significance level α: when $p \leq \alpha$, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$; when $p > \alpha$, there are no grounds to reject $\mathcal{H}_0$.

Copyright ©2010-2026 U-smile Stats – All rights reserved

30 RELIABILITY ANALYSIS

Scale reliability

WHEN TO USE:

Reliability analysis is most often performed when we are building a composite scale, i.e. a questionnaire or test made up of many individual items (questions) which we sum up into a single score. We then want to check whether the individual questions really measure the same thing – in other words, to what extent we can trust that the total score obtained from the scale accurately reflects the studied trait (e.g. competence, level of anxiety, job satisfaction). This is what is meant by the **reliability** of a scale, understood as its internal consistency.

ASSUMPTIONS

- a scale made up of at least several items (questions), which are summed to a single overall score,
- the individual scale items should measure the same construct, i.e. the correlation between them should be high – this assumption can be checked by calculating the matrix of Pearson correlation coefficients between the scale items.

DEFINITIONS OF MEASURES

Although various measures of agreement can be used for the issue of reliability, the most commonly used technique here is Cronbach's α coefficient and so-called split-half reliability.

Cronbach's α coefficient (Cronbach, 1951[51]) was given this name for the first time by Cronbach in 1951. In simplified terms, it measures the ratio of the variability of the individual scale items to the variability of the whole scale (i.e. the sum of those items) – in other words, it shows how strongly the individual questions "hold together". The α coefficient can take any value less than or equal to 1, including negative values, although only positive values are meaningful. If all scale items are perfectly reliable, the reliability coefficient equals 1. Besides the plain coefficient α_C, a standardised version ($\alpha_{standard}$) is also calculated, based on the mean correlation between the scale items.

The usefulness of particular scale items can be assessed with the help of:

- the value of the α_C coefficient calculated after removing a given scale item,
- the value of the standard deviation of the scale calculated after removing a given scale item,
- the mean value of the scale calculated after removing a given scale item,
- the value of the correlation between a given item and the sum of the remaining items.

If, after removing an item, the α_C coefficient clearly increases, this is a signal that the item fits poorly with the rest of the scale, and it is worth considering removing or rewording it.

Copyright ©2010-2026 U-smile Stats – All rights reserved

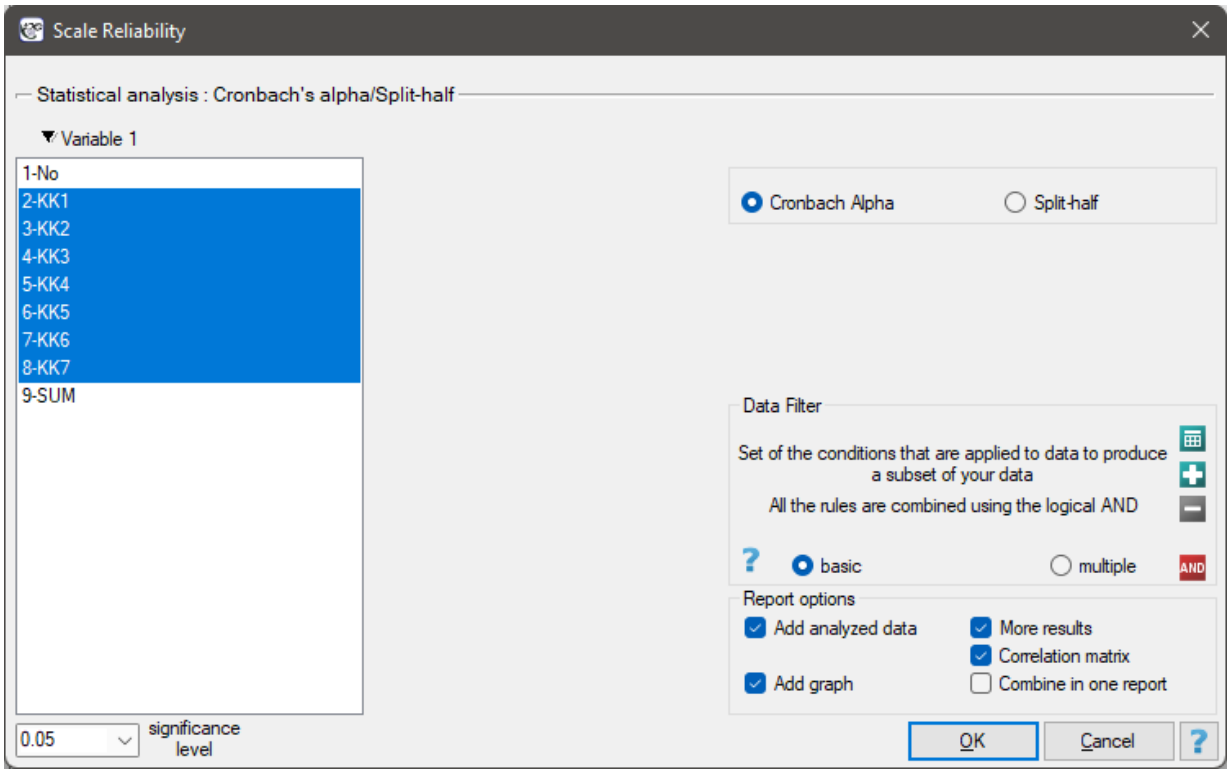
Split-half reliability is a different approach to the same issue. It consists of randomly dividing the scale items into two halves and checking how far the two halves agree with (are correlated with) each other. For this purpose, the Spearman-Brown split-half reliability coefficient is used, published independently by Spearman (1910)[160] and Brown (1910)[35], as well as an alternative formula proposed by Guttman (1945)[75]. If two randomly chosen halves of the scale are perfectly correlated, the split-half reliability coefficient equals 1.

Note! It is generally assumed that, for a scale to be regarded as reliable, the scale reliability coefficients (α_C, $\alpha_{standard}$, r_{SH}, r_{SHG}) should be greater than 0.6 and less than 1.

Additionally, for a scale found to be reliable, the **standard error of measurement** (SEM) is determined, which indicates how large a spread of the measurement result can be expected due to the imperfect reliability of the scale – the lower the reliability, the larger the measurement error.

HOW TO SET UP THE ANALYSIS

The window with the scale reliability options is opened through the Statistics→Scale reliability menu.



PERFORMING AND INTERPRETING THE ANALYSIS

EXAMPLE 30.1. (scale.pqs file)
A "competence scale" developed in-house is used, among other tools, to help assess the suitability of prospective employees of a certain consulting company. In addition to a job interview, candidates fill in

Copyright ©2010-2026 U-smile Stats – All rights reserved

a questionnaire that includes, among others, questions making up the ”competence scale”. The scale contains 7 questions. For each question, from 1 to 5 points can be scored, where 1 denotes the lowest and 5 the highest possible mark. A total of 35 points can be obtained. The scores on this scale obtained by 24 candidates are shown in the table below:

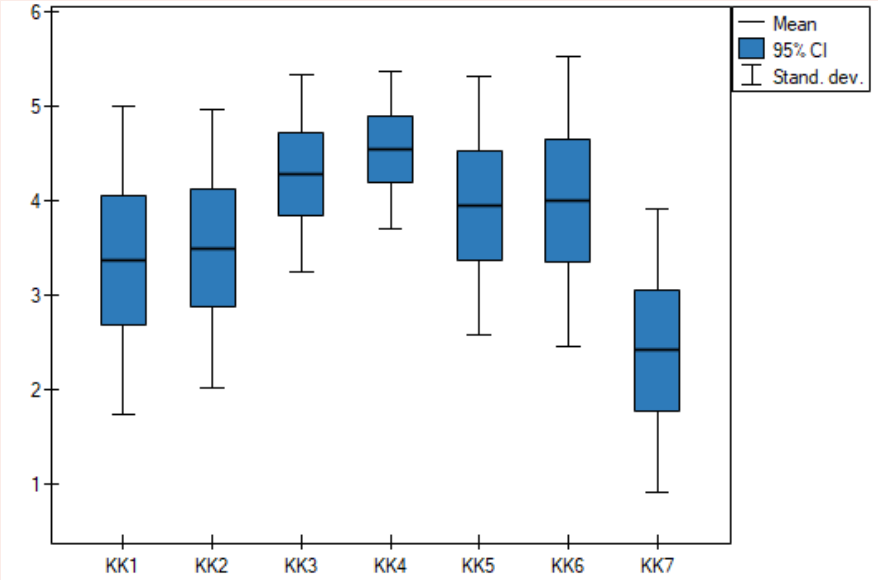
We want to check the accuracy of the tool used to assess competence (the scale applied). For this purpose, a reliability analysis will be performed.

The correlation matrix shows that the last item is the most weakly correlated with the others. We therefore suspect that this item does not measure the same construct as the remaining items.

Correlations						
	KK2	KK3	KK4	KK5	KK6	KK7
KK1	0.712	0.265	0.355	0.338	0.225	0.164
KK2	.	0.326	0.443	0.378	0.269	0.196
KK3	.	.	0.512	0.498	0.735	0.058
KK4	.	.	.	0.67	0.409	-0.049
KK5	.	.	.	.	0.478	-0.14
KK6	.	.	.	.	.	-0.151

Copyright ©2010-2026 U-smile Stats – All rights reserved

Deleted item	KK7
Scale mean if item deleted	23.6667
Scale standard deviation if item deleted	5.7307
Correlation between deleted item and sum of remaining	0.027
Cronbach Alpha if item deleted	0.8036
Group size	24
Number of items	7
Mean of scale	26.0833
Standard deviation of scale	5.9631
Cronbach Alpha for scale	0.7368
-95% CI dla Cronbach Alpha for scale	0.5371
+95% CI dla Cronbach Alpha for scale	0.8705
Standard error of measurement	3.0592
Average correlation between pairs of items	0.3185
Standardized Cronbach alpha	0.7659



Cronbach's alpha coefficient $= 0.736805$

The competence scale turned out to be a reliable scale. The mean correlation coefficient between the scale items is 0.31847.

A more detailed analysis of each item shows that all of them, with the exception of the last one, contribute to the reliability of the whole scale to a similar degree. The correlation of item KK7 with the remaining scale items is the weakest, at 0.026954. If item KK7 were removed from the scale, the value of Cronbach's coefficient would rise to 0.803619.

Similar conclusions can be drawn on the basis of the split-half reliability analysis performed on the items randomly divided into 2 parts (KK1, KK3, KK5) (KK2, KK4, KK6, KK7).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Cronbach's alpha/Split-half	
Analysed variables	KK1
	KK3
	KK5
	KK2
	KK4
	KK6
	KK7
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Group size	24
Mean of scale	26.0833
Standard deviation of scale	5.9631
Correlation between two halves of scale	0.7509
Split-half reliability	0.8577
Standard error of measurement	2.2494
Guttman split-half reliability	0.8565
First half	
Number of items	3
Names of items	KK1;KK3;KK5
Mean	11.625
Standard deviation	3.076
Cronbach Alpha	0.6071
Second half	
Number of items	4
Names of items	KK2;KK4;KK6;KK7
Mean	14.4583
Standard deviation	3.2966
Cronbach Alpha	0.417

Split-half reliability: Spearman-Brown $= 0.857705$**, Guttman** $= 0.856531$

The halves are fairly well correlated – the correlation coefficient between the halves is 0.750862. The value of Cronbach's alpha coefficient is, however, too low for the second half (0.416958), because it contains item KK7, which is weakly correlated with the remaining scale items.

After removing this item and repeating the analysis, all the reliability coefficients are indeed higher.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Cronbach's alpha/Split-half	
Analysed variables	KK1
	KK3
	KK5
	KK2
	KK4
	KK6
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Group size	24
Mean of scale	23.6667
Standard deviation of scale	5.7307
Correlation between two halves of scale	0.8229
Split-half reliability	0.9029
Standard error of measurement	1.786
Guttman split-half reliability	0.9023
First half	
Number of items	3
Names of items	KK1;KK3;KK5
Mean	11.625
Standard deviation	3.076
Cronbach Alpha	0.6071
Second half	
Number of items	3
Names of items	KK2;KK4;KK6
Mean	12.0417
Standard deviation	2.9263
Cronbach Alpha	0.5864

FOR THOSE INTERESTED

Formulas

Cronbach's α coefficient:

$$\alpha_C = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} sd_i^2}{sd_t^2}\right),$$

where:
k – the number of items in the scale,
sd_i^2 – the variance of item i,
sd_t^2 – the variance of the sum of all items.

Standardised reliability coefficient:

$$\alpha_{standard} = \frac{k\overline{r_p}}{1 + (k-1)\overline{r_p}},$$

where:
$\overline{r_p}$ – the mean of the Pearson correlation coefficients of all $(k(k-1)/2)$ scale items.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Spearman-Brown split-half reliability:

$$r_{SH} = \frac{2r_p^*}{1 + r_p^*},$$

where:
r_p^* – the Pearson correlation coefficient between the halves of the scale.

Guttman split-half reliability:

$$r_{SHG} = 2\left(1 - \frac{sd_{t1}^2 + sd_{t2}^2}{sd_t^2}\right),$$

where:
sd_{t1}^2, sd_{t2}^2 – the variance of the first and second half of the scale,
sd_t^2 – the variance of the sum of all items.

Standard error of measurement (SEM), determined for a reliable scale:

$$SEM = sd_t\sqrt{1 - \alpha_C}$$ – for Cronbach's reliability coefficient

or

$$SEM = sd_t\sqrt{1 - r_{SH}}$$ – for the split-half reliability coefficient

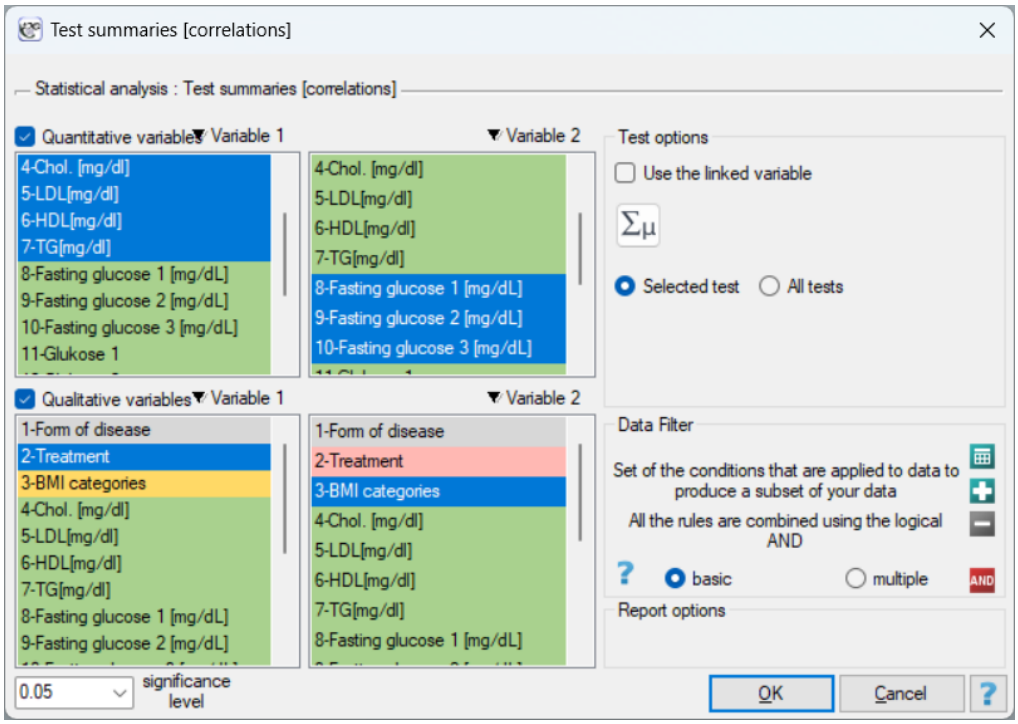
Copyright ©2010-2026 U-smile Stats – All rights reserved

31 TEST SUMMARIES

Instead of running successive tests one at a time, the program lets you run a whole set of several analyses at once, matched to the kind of groups being compared. Quantitative data will also be described with means, medians, etc., and qualitative data with counts and percentages.

HOW TO RUN A SUMMARY

The Test summaries options window is opened via the menu Statistics→Summaries→Test summaries, after which you select the appropriate group of analyses.



There are four groups of summaries to choose from:

1. Comparison of two dependent groups:
   - The Student's t-test for dependent groups
   - The Wilcoxon matched-pairs test
   - The Bowker-McNemar test
   - Normality of distribution test Kolmogorov-Smirnov (or another one selected by the user)
   - and others ...
2. Comparison of two independent groups:
   - The Student's t-test for independent groups
   - The Mann-Whitney U test
   - Chi-square, Fisher, OR/RR
   - Normality of distribution test Kolmogorov-Smirnov (or another one selected by the user)

Copyright ©2010-2026 U-smile Stats – All rights reserved

- and others ...

3. Comparison of multiple independent groups (see the description below):
   - ANOVA for independent groups and Welch's ANOVA
   - The Kruskal-Wallis ANOVA
   - Chi-square and Fisher's test for large tables
   - Levene's test of equality of variances, normality of distribution test (as above)
   - post-hoc tests (Tukey HSD, Tamhane T2, Dunn, pairwise table comparisons) and homogeneous groups
   - and others ...
4. Correlation:
   - Pearson's linear correlation
   - Spearman's monotonic correlation
   - Chi-square, Fisher, correlation coefficient
   - Normality of distribution test Kolmogorov-Smirnov (or another one selected by the user)
   - and others ...

WHAT THE PROGRAM RETURNS

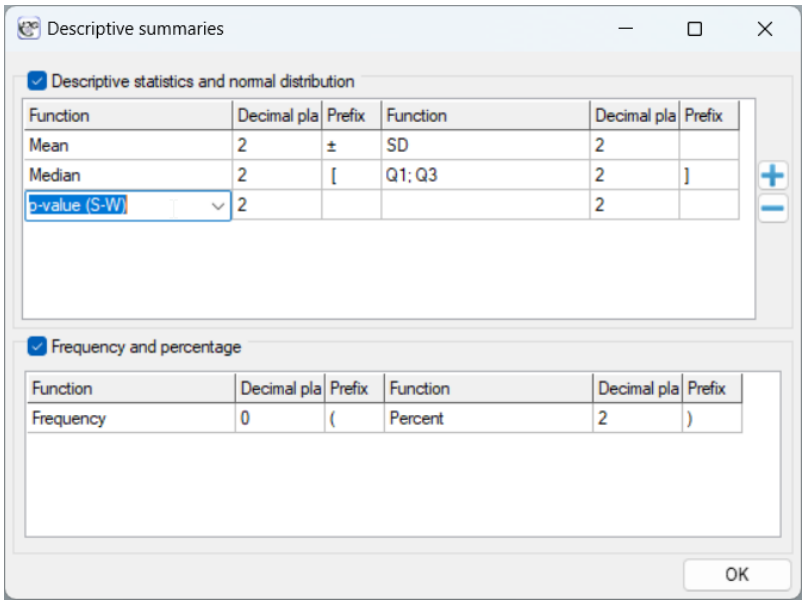
For each analyzed variable, depending on whether it is quantitative or qualitative, the program can return the following results in the report:

- **the selected test** - chosen automatically according to the rule described below;
- **all tests** together with their accompanying coefficients and measures, regardless of whether the minimum conditions for their use are met.

HOW THE PROGRAM SELECTS THE TEST AUTOMATICALLY

- If, before the analysis, the user does not describe the individual variables with the appropriate scales, quantitative data are treated as an interval scale, and qualitative data as a nominal scale.
- Testing for normality of distribution is based on the result of the normality test selected by the user when setting up the descriptive statistics.

Copyright ©2010-2026 U-smile Stats – All rights reserved



– If the user does not indicate any test for assessing normality of distribution in the descriptive-statistics window, normality will be checked using the Kolmogorov-Smirnov test.

Note! The analyses we propose are robust to small deviations from the normal distribution, and the Kolmogorov-Smirnov test is the most conservative of the available normality tests – so a lack of normality is only detected when the tested distribution differs greatly from the normal distribution. Normality of distribution is tested: (1) for the comparison of two independent groups – based on the data in each group, (2) for the comparison of two dependent groups – based on the difference between measurements, (3) for correlation – based on the model residuals, (4) for the comparison of multiple independent groups – based on the data in each group (the condition must hold in all groups).

RUNNING THE ANALYSIS AND INTERPRETING THE RESULTS

Example 31.1. (Summaries.pqs file)
We want to make an automatic comparison of two independent groups: patients with a chronic condition versus patients with an acute form of the disease. The comparison is based on data described on the interval scale: Chol Tot., LDL, HDL, TG, on the ordinal feature BMI, and on nominal data: Treatment, Estradiol 1, Estradiol 2, Estradiol 3. To do this, we select the menu Summaries → Test summaries [two independent groups] and select the grouping variable: Treatment, then select the interval data (in the section for quantitative variables) and the nominal and ordinal data (in the section for qualitative variables). We select the Selected test option and run the analysis.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note

Test summaries [two independent groups]					
Variable	Measures	Form of disease[acute]	Form of disease[chronic]	p-value	test
Chol. [mg/dl]				0.5433	(M-W)
	Mean ± SD	160.12±23.77	160.82±30.31		
	Median [Q1; Q3]	165.2 [135.38; 178.53]	158.5 [141.25; 170.33]		
	p-value (S-W)	0.14	<0.01		
LDL[mg/dl]				0.2241	(M-W)
	Mean ± SD	86.83±20	82.91±27.81		
	Median [Q1; Q3]	91.35 [69.55; 102.35]	79.75 [64.83; 94.33]		
	p-value (S-W)	0.31	0.01		
HDL[mg/dl]				0.0339	(t-st)
	Mean ± SD	53.27±9.7	58.99±9.97		
	Median [Q1; Q3]	52.5 [48.28; 56.7]	58 [52.88; 67.55]		
	p-value (S-W)	0.52	0.8		
TG[mg/dl]				0.3557	(M-W)
	Mean ± SD	103.25±40.09	93.36±35.59		
	Median [Q1; Q3]	95.45 [73.48; 119.35]	86.8 [64.65; 116.18]		
	p-value (S-W)	0.08	0.05		

Variable	Categories	Form of disease[acute]	Form of disease[chronic]	p-value	test
Treatment				0.1583	(chi2)
	Drug A	19(79.17%)	21(61.76%)		
	Drug B	5(20.83%)	13(38.24%)		
BMI categories				<0.0001	(chi2 trend)
	norm	6(25%)	32(94.12%)		
	overweight	8(33.33%)	0(0%)		
	obesity	10(41.67%)	2(5.88%)		
Estradiol 1				0.0077	(chi2)
	above the norm	17(70.83%)	12(35.29%)		
	norm	7(29.17%)	22(64.71%)		
Estradiol 2				0.6822	(chi2 Yates)
	above the norm	2(8.33%)	3(8.82%)		
	norm	22(91.67%)	31(91.18%)		
Estradiol 3				0.988	(chi2 Yates)
	above the norm	3(12.5%)	3(8.82%)		
	norm	21(87.5%)	31(91.18%)		

As a result, we obtain both a description of each group and the statistical test chosen for the comparison.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Rule for selecting the test – from the report's Note

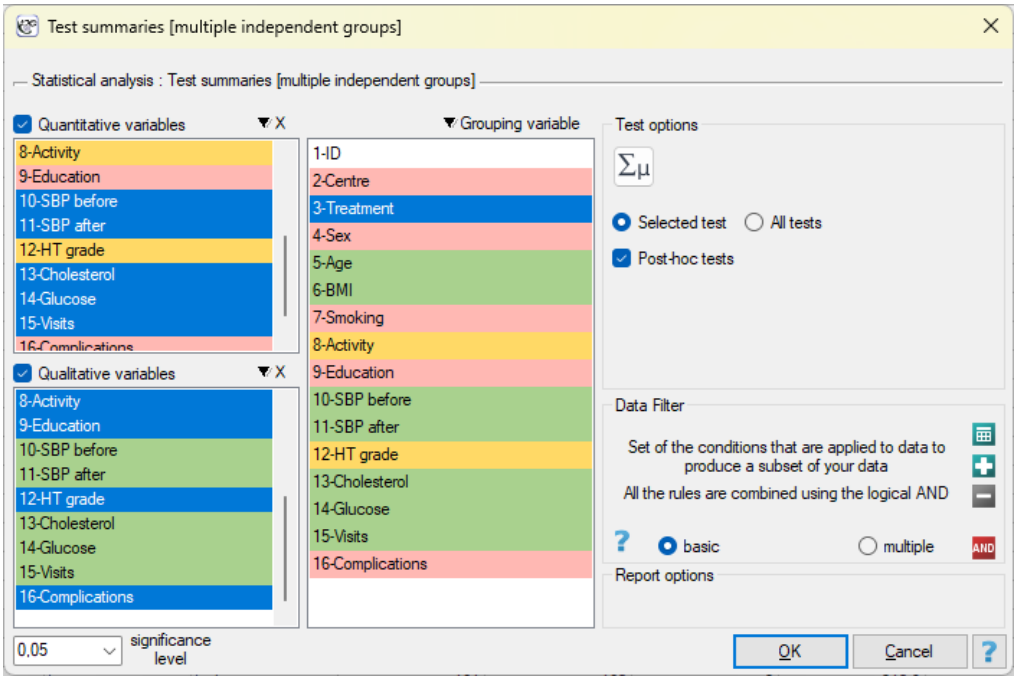
At the top of the report, the Note contains a description explaining the rule used for selecting the statistical test appropriate for the given analysis:

"**Quantitative variables**:
For the interval scale with normality of distribution, the Student's t-test (t-st) was used, or its Cochran-Cox correction (C-C) when the group variances differed. For the interval scale, when the normality-of-distribution condition was not met, as well as for the ordinal scale, the Mann-Whitney test (M-W) was used. Normality of the data was tested with the Shapiro-Wilk test, and equality of variances with the Fisher-Snedecor test (F-S). If no scale was assigned to the analyzed variables, the data were assumed to come from an interval scale.

Qualitative variables:
For the nominal scale, the chi-square test (chi2) was used, and when Cochran's condition was not met, the Fisher exact test (Fisher exact) or, for 2x2 tables with a sample size greater than 40, the Yates correction (chi2-Yates). For the ordinal scale, the chi-square test for trend was used. If no scale was assigned to the analyzed variables, the data were assumed to come from a nominal scale."

SUMMARY FOR MULTIPLE INDEPENDENT GROUPS

The summary Statistics→Summaries→Test summaries [multiple independent groups] compares at the same time two or more groups defined by a grouping variable (e.g. several centres, several treatment schemes, age categories). For two groups it gives results consistent with the summary for two independent groups, but with tests intended for multiple groups (ANOVA instead of the Student's t-test, etc.), so with exactly two groups it is more convenient to use the two-group summary.



In the summary window we indicate the grouping variable and – as in the other summaries – the qu-

Copyright ©2010-2026 U-smile Stats – All rights reserved

antitative and qualitative variables, and then choose:

- Selected test or All tests – in the latter case, for quantitative variables the report contains columns with the result of the normality test in the groups, the Levene and Brown-Forsythe tests, ANOVA, Welch's ANOVA, Brown-Forsythe F^* ANOVA, the Kruskal-Wallis ANOVA, the Conover test of equality of rank variances and the effect-size measures η^2 and ϵ^2; for qualitative variables – the results of the chi-square and Fisher exact tests for $r \times c$ tables and, for the ordinal scale, the Kruskal-Wallis ANOVA for category codes;
- Post-hoc tests – when this option is checked, for variables whose main test turned out significant and there are at least three groups, multiple comparisons are performed: the Tukey HSD test after ANOVA, the Tamhane T2 test after Welch's ANOVA, the Dunn test with the Benjamini-Hochberg correction after the Kruskal-Wallis ANOVA, and for qualitative variables pairwise group comparisons with the chi-square or Fisher exact test with the Benjamini-Hochberg correction. In the post-hoc column the report lists the pairs of groups that differ significantly, and in the homogeneous groups (HG) columns the groups that do not differ significantly share a common letter.

The descriptive measures for each group by default also include the size N; as in the other summaries, their scope is set in the descriptive statistics window.

Example 31.2. (summaries_groups.pqs file)
The file contains data of 1000 hypertensive patients treated in four centres with three schemes: Standard, New drug and Placebo. We want to check whether the treatment schemes differ in patient characteristics and treatment outcomes. The comparison is based on quantitative data: Age, BMI, SBP before, SBP after (systolic pressure before and after treatment), Cholesterol, Glucose and Visits (number of visits), on the ordinal features Activity and HT grade, and on nominal data: Sex, Smoking and Complications. The scales of the ordinal and nominal variables are already set in the file.
To do this, we select the menu Statistics→Summaries→Test summaries [multiple independent groups], indicate Treatment as the grouping variable, select the listed interval variables in the section for quantitative variables and the ordinal and nominal variables in the section for qualitative variables. We check the Selected test and Post-hoc tests options and run the analysis.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Note

Test summaries [multiple independent groups]

Variable	Measures	Treatment[New drug]	Treatment[Placebo]	Treatment[Standa	p-value	test	post-hoc	HG New drug	HG Placebo	HG Standard
Age					0,421784	(K-W)	none	A	A	A
	Średnia ± SD	56,38±10,77	56,41±10,44	55,37±11,19						
	Mediana [Q1; Q3]	56 [49; 64]	57 [49; 64]	56 [47; 63]						
	Wartość p (S-W)	0,41	0,07	0,01						
	N	333	295	372						
BMI					0,427722	(K-W)	none	A	A	A
	Średnia ± SD	27,91±3,94	28,26±4,34	28,09±4,11						
	Mediana [Q1; Q3]	28,1 [25,4; 30,6]	28,7 [25,1; 31,4]	27,9 [25,3; 31,13]						
	Wartość p (S-W)	0,42	0,03	0,12						
	N	329	287	364						
SBP before					0,516973	(K-W)	none	A	A	A
	Średnia ± SD	152,28±13,38	152,44±13,95	153,05±13,2						
	Mediana [Q1; Q3]	152 [143; 161]	153 [143; 162]	154 [144; 162]						
	Wartość p (S-W)	0,05	0,01	0,02						
	N	333	295	372						
SBP after					<0,000001	(K-W)	New drug-Pla	A	B	C
	Średnia ± SD	135,61±16,93	149,46±16,86	142,83±15,8						
	Mediana [Q1; Q3]	134 [123; 146]	149 [138; 161]	144 [132; 153]						
	Wartość p (S-W)	0,02	0,57	0,28						
	N	333	295	372						
Cholesterol					0,076908	(ANOVA)	none	A	A	A
	Średnia ± SD	206,64±33,75	202,16±31,3	207,75±31,96						
	Mediana [Q1; Q3]	207,6 [181,45; 229,68	201,3 [182,8; 223,9]	209 [183,15; 230,						
	Wartość p (S-W)	0,21	0,81	0,17						
	N	320	287	359						
Glucose					0,13009	(K-W)	none	A	A	A
	Średnia ± SD	100,34±16,97	100,97±16,66	103,42±19,79						
	Mediana [Q1; Q3]	98,8 [87,9; 109,48]	99,1 [88,6; 112,75]	101,1 [89,4; 116]						
	Wartość p (S-W)	<0,01	<0,01	<0,01						
	N	326	291	365						
Visits					0,081599	(K-W)	none	A	A	A
	Średnia ± SD	3,06±1,74	2,98±1,55	3,24±1,72						
	Mediana [Q1; Q3]	3 [2; 4]	3 [2; 4]	3 [2; 4]						
	Wartość p (S-W)	<0,01	<0,01	<0,01						
	N	333	295	372						

Copyright ©2010-2026 U-smile Stats – All rights reserved

Variable	Categories	Treatment[New drug]	Treatment[Placebo]	Treatment[Standa	p-value	test	post-hoc	HG New drug	HG Placebo	HG Standard
Centre					0,30118	(chi2)	none	A	A	A
	West	68(20,42%)	67(22,71%)	93(25%)						
	South	76(22,82%)	77(26,1%)	75(20,16%)						
	North	98(29,43%)	69(23,39%)	109(29,3%)						
	East	91(27,33%)	82(27,8%)	95(25,54%)						
Sex					0,509218	(chi2)	none	A	A	A
	M	164(49,25%)	132(44,75%)	172(46,24%)						
	F	169(50,75%)	163(55,25%)	200(53,76%)						
Smoking					0,712116	(chi2)	none	A	A	A
	yes	96(28,83%)	82(27,8%)	114(30,65%)						
	no	237(71,17%)	213(72,2%)	258(69,35%)						
Activity					0,929372	(K-W)	none	A	A	A
	moderate	140(43,21%)	121(41,58%)	159(43,44%)						
	low	115(35,49%)	109(37,46%)	118(32,24%)						
	high	69(21,3%)	61(20,96%)	89(24,32%)						
Education					0,44659	(chi2)	none	A	A	A
	secondary	150(45,05%)	142(48,14%)	186(50%)						
	primary	92(27,63%)	67(22,71%)	82(22,04%)						
	higher	91(27,33%)	86(29,15%)	104(27,96%)						
HT grade					0,629838	(K-W)	none	A	A	A
	3	8(2,4%)	9(3,05%)	9(2,42%)						
	2	91(27,33%)	77(26,1%)	112(30,11%)						
	1	234(70,27%)	209(70,85%)	251(67,47%)						
Complication:					0,03746	(chi2)	Placebo-Stan	A+B	A	B
	yes	32(9,61%)	48(16,27%)	44(11,83%)						
	no	301(90,39%)	247(83,73%)	328(88,17%)						

For each variable the report gives a description of the three groups (size, mean with standard deviation or median with quartiles, and for qualitative variables counts and percentages), the name of the selected test and its p value. For variables with a distribution close to normal in every group (e.g. Age, BMI, SBP before) the ANOVA is selected (or Welch's ANOVA when Levene's test shows a difference in variances), for variables with a skewed distribution (Glucose, Visits) and for ordinal variables – the Kruskal-Wallis ANOVA, and for nominal variables – the chi-square test. Since the patients were randomly assigned to the schemes, the baseline characteristics (Age, BMI, SBP before, Sex, Smoking) should not differ significantly between the groups; differences are expected in the pressure after treatment (SBP after) and in the frequency of complications. For variables with a significant result the post-hoc column indicates which pairs of schemes differ significantly (for SBP after – all three pairs, with the lowest pressure after the new drug), and the homogeneous groups (HG) columns mark with a common letter the schemes that cannot be distinguished.

The same file allows a summary for two independent groups (grouping variable Sex or Smoking) and a summary for multiple groups by centre (Centre, four groups) or by hypertension grade (HT grade, three groups).

Copyright ©2010-2026 U-smile Stats – All rights reserved

Rule for selecting the test for multiple groups – from the report's Note

In the Note at the top of the report the program describes the rule used for selecting the test:

"**Quantitative variables**:
For the interval scale with normality of distribution in all groups and equality of variances, the one-way ANOVA for independent groups (ANOVA) was determined; when the variances of the groups differed, the Welch ANOVA (Welch) was determined. For the interval scale, when the condition of normality of distribution was not met, as for the ordinal scale, the Kruskal-Wallis ANOVA (K-W) was determined. Normality of the distribution was tested with the Shapiro-Wilk test and equality of variances with the Levene test. If the scale was not marked for the analyzed variables, it was assumed that the data came from the interval scale.

Qualitative variables:
For the nominal scale, the chi-square test (chi2) was determined, and when Cochran's condition was not met the Fisher exact test (Fisher exact) or, for 2x2 tables with a sample size greater than 40, the Yates correction (chi2-Yates) was determined. For the ordinal scale, the Kruskal-Wallis ANOVA (K-W) for category codes and the Conover test of equality of rank variances were determined. If the scale was not marked for the variables analyzed, it was assumed that the data came from the nominal scale.

Post-hoc tests:
Determined only for at least 3 groups and a significant main test: Tukey HSD (ANOVA), Tamhane T2 (Welch), Dunn with the Benjamini-Hochberg correction (K-W), pairwise group comparisons with the chi-square or Fisher exact test with the Benjamini-Hochberg correction (RxC tables). The post-hoc column lists the pairs of groups that differ significantly, and in the homogeneous groups (HG) columns the groups that do not differ significantly share a common letter."

Copyright ©2010-2026 U-smile Stats – All rights reserved

32 THE WIZARD

The Wizard is a tool that helps novice users navigate the basic statistics available in the program. It offers hints about the assumptions that should be checked before selecting the appropriate statistical test. The final step of the Wizard is choosing the appropriate statistical test and opening the options window for that test.

The Wizard can be launched via:
- Statistics→Wizard,
- the  button on the toolbar.

The Wizard window that opens contains a field for choosing the kind of analysis the user wants to perform. The user can choose:

Comparison - 1 group - to compare measurement values for 1 population with a value specified by the user. This population is represented by data gathered in 1 column, either in raw form or cumulated into a frequency table.

Comparison - 2 groups - to compare measurement values between 2 populations. These populations are represented by data gathered in raw form in 2 columns, or cumulated into a contingency table.

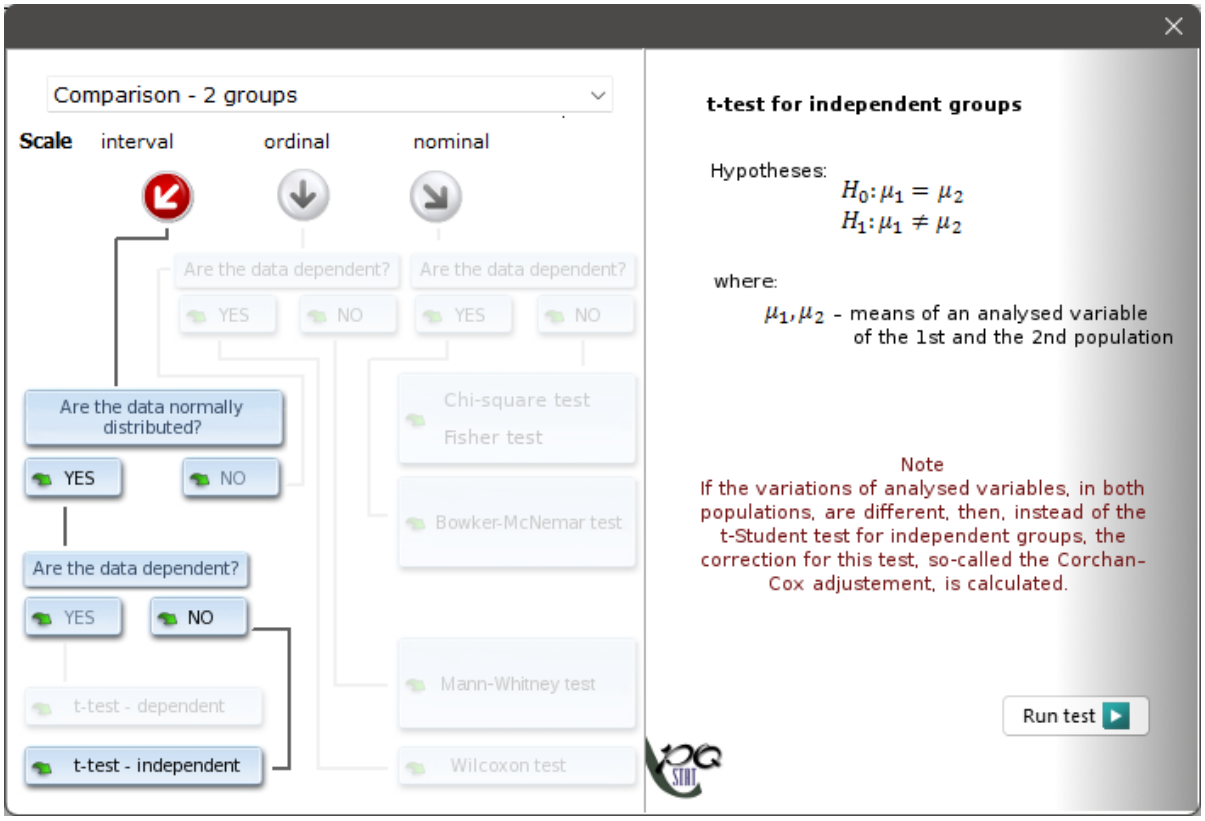
Comparison - more than 2 groups - to compare measurement values between several populations. These populations are represented by data gathered in raw form in several columns.

Correlation - to check for a relationship between 2 population parameters. These features are represented by data gathered in raw form in 2 columns, or cumulated into a contingency table.

Agreement - to examine the agreement between several obtained results. These features are represented by data gathered in raw form in several columns, or cumulated into a contingency table.

Once the kind of analysis has been chosen, a graph appears, divided according to the scale on which the studied population features were measured (interval scale, ordinal scale, nominal scale).

Copyright ©2010-2026 U-smile Stats – All rights reserved



The user moves through the graph by choosing the correct answer to each question asked. Having followed the chosen path through the graph, the user can then run the test that, based on the answers given, is appropriate for solving the statistical problem at hand.

33 OTHER NOTES

33.1 FILE FORMATS

PQS - the default data file format for ℙℚStat files; it represents all objects created with ℙℚStat.

PQX - the XML-standard ℙℚStat file format (UTF-8 encoding); it can be used to represent all objects in the ℙℚStat program; recommended on systems with a small amount of RAM.

Copyright ©2010-2026 U-smile Stats – All rights reserved

Literatura

[1] Arthur D., Vassilvitskii S. (2007). *k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. 1027–1035*

[2] Abdi H. (2007), *Bonferroni and Sidak corrections for multiple comparisons", in N.J. Salkind (ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage*

[3] D'Agostino R.B. and Pearson E.S. (1973), *Tests of departure from normality. Empirical results for the distribution of b2 and sqrt(b1). Biometrika, 60, 613-622*

[4] D'Agostino R.B., Belanger A., D'Agostino Jr.R B. (1990), *A suggestion for using powerful and informative tests of normality. American Statistician, 44, 3 16-321*

[5] Agresti A., Coull B.A. (1998), *Approximate is better than "exact" for interval estimation of binomial proportions. American Statistics 52: 119-126*

[6] Altman D.G., (1998), *Confidence intervals for the number needed to treat. BMJ. 317(7168): 1309–1312*

[7] Aroian, L. A. (1947), *The probability function of the product of two normally distributed variables. Annals of Mathematical Statistics, 18, 265-271.*

[8] Bland J.M., Altman D.G. (1999), *Measuring agreement in method comparison studies. Statistical Methods in Medical Research 8:135-160.*

[9] Anscombe F.J. (1981), *Computing in Statistical Science through APL. Springer-Verlag, New York*

[10] Armitage P. (1955), *Tests for Linear Trends in Proportions and Frequencies. Biometrics. 11 (3): 375–386*

[11] Armitage P., Berry G., (1994), *Statistical Methods in Medical Research (3rd edition); Blackwell*

[12] Armitage P., Colton T., (2009), *Encyclopedia of Biostatistics. John Wiley and Sons.*

[13] Austin P.C., (2009), *The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making; 29(6):661-77*

[14] Austin P.C., (2011), *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research 46, 3: 399–424*

[15] Barnard G.A. (1989), *On alleged gains in power from lower p-values. Statistics in Medicine 8:1469-1477*

[16] Baron R. M., Kenny D. A. (1986), *The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. Journal of Personality and Social Psychology, 51, 1173-1182.*

[17] Beal S.L. (1987), *Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. Biometrics 43: 941-950*

[18] Bender R. (2001), *Calculating confidence intervals for the number needed to treat. Controlled Clinical Trials 22:102–110*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[19] Benjamini Y. and Hochberg Y. (1995), *Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57, 289–300*

[20] Betty R. Kirkwood and Jonathan A. C. Sterne (2003), *Medical Statistics (2nd ed.). Meassachusetts: Blackwell Science, 177—188, 240—248*

[21] Blanche P., Dartigues J.F., Jacqmin-Gadda H. (2013), *Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Statistics in Medicine, 32(30):5381–5397*

[22] Bland J.M., Altman D.G. (1986), *Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327 (8476): 307–10*

[23] Bland J.M., Altman D.G. (1999), *Measuring agreement in method comparison studies. Statistical Methods in Medical Research 8 (2): 135–60*

[24] Bowker A.H. (1948), *Test for symmetry in contingency tables. Journal of the American Statistical Association, 43, 572-574*

[25] Box G. E. , Cox D. R. (1964), *An analysis of transformations. Journal of the Royal Statistical Society, Series B 26: 211–252*

[26] Breslow N.E., Day N.E. (1980), *Statistical Methods in Cancer Research: Vol. I - The Analysis of Case-Control Studies. Lyon: International Agency for Research on Cancer*

[27] Breslow N.E. (1996), *Statistics in epidemiology: the case-control study', Journal of the American Statistical Association, 91, 14—28*

[28] Breslow N.E. (1974), *Covariance analysis of censored survival data. Biometrics, 30(1):89–99*

[29] Brier G.W. (1950), *Verification of forecasts expressed in terms of probability. Monthly Weather Review 78:1-3*

[30] Brookmeyer R. and Crowley J. (1982a), *A confidence interval for the median survival time. Biometrics 38, 29-41*

[31] Brown L.D., Cai T.T., DasGupta A. (2001), *Interval Estimation for a Binomial Proportion. Statistical Science, Vol. 16, no. 2, 101-133*

[32] Brown M.B., Forsythe A. B. (1974a), *Robust tests for equality of variances. Journal of the American Statistical Association, 69,364-367*

[33] Brown M. B., Forsythe A. B. (1974), *The ANOVA and multiple comparisons for data with heterogeneous variances. Biometrics, 30, 719-724*

[34] Brown M. B., Forsythe A. B. (1974), *The small sample behavior of some statistics which test the equality of several means. Technometrics, 16, 385-389*

[35] Brown W. (1910), *Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296-322*

[36] Hochberg Y. (1988), *A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–803*

[37] Chow S.C., Shao J., and Wang H. (2008). *Sample Size Calculations in Clinical Research, Second Edition. Chapman and Hall/CRC. Boca Raton, Florida.*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[38] Cicchetti D. and Allison T. (1971),*A new procedure for assessing reliability of scoring eeg sleep recordings. American Journal EEG Technology, 11, 101-109*

[39] Cleveland, W. S. (1979),*Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association. 74*

[40] Clopper C. and Pearson S. (1934), *The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26: 404-413*

[41] Cochran W.G. (1950), *The comparison ofpercentages in matched samples. Biometrika, 37, 256-266*

[42] Cochran W.G. (1952), *The chi-square goodness-of-fit test. Annals of Mathematical Statistics, 23, 315-345*,

[43] Cochran W. G. (1954), *Some methods for strengthening the common chi-square tests. Biometrics, 10(4) 17-45 1*

[44] Cochran W.G. and Cox G.M. (1957), *Experimental designs (2nd 4.). New York: John Wiley and Sons.*

[45] Cohen J. (1960), *A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 10,3746*

[46] Cohen J. (1968), *Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220*

[47] Cohen J. (1988), *Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, Hillsdale, New Jersey*

[48] Conover W. J. (1999), *Practical nonparametric statistics (3rd ed). John Wiley and Sons, New York*

[49] Cox D.R. (1972), *Regression models and life tables. Journal of the Royal Statistical Society, B34:187-220*

[50] Cramkr H. (1946), *Mathematical models of statistics. Princeton, NJ: Princeton University Press.*

[51] Cronbach L.J. (1951), *Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334*

[52] DeLong E.R., DeLong D.M., Clarke-Pearson D.L., (1988), *Comparing the areas under two or more correlated receiver operating curves: A nonparametric approach. Biometrics 44:837-845*

[53] Dunn O. J. (1964), *Multiple comparisons using rank sums. Technometrics, 6: 241–252*

[54] Durbin J. (1951), *Incomplete blocks in ranking experiments. British Journal of Statistical Psychology, 4: 85–90*

[55] Egger M., Smith G. D., Schneider M., Minder C. (1997), *Bias in meta-analysis detected by a simple, graphical test. BMJ, 315(7109):629-634*

[56] Epps T.W., Pulley L.B. (1983), *A test for normality based on the empirical characteristic function. Biometrika. 1983;70:723–726*

[57] Fagerland M. W., Lydersen S., and Laake P. (2013), *The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional, BMC Med Res Methodol; 13: 91.*

[58] Fisher R.A. (1934), *Statistical methods for research workers (5th ed.). Edinburgh: Oliver and Boyd*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[59] Fisher R.A. (1935), *The logic of inductive inference. Journal of the Royal Statistical Society, Series A, 98,39-54*

[60] Fisher R.A. (1936), *The use of multiple measurements in taxonomic problems. Annals of Eugenics 7 (2): 179–188*

[61] Fleiss J.L. (1971), *Measuring nominal scale agreement among many raters. Psychological Bulletin, 76 (5): 378–382*

[62] Fleiss J.L., Cohen J. (1973), *The equivalence of weighted kappa and the intraclass correlation coeffcient as measure of reliability. Educational and Psychological Measurement, 33, 613-619*

[63] Fleiss J.L., Levin B., Paik M.C. (2003), *Statistical methods for rates and proportions. 3rd ed. (New York: John Wiley) 598-626*

[64] Fine J.P., Gray R.J. (1999), *A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association, 94(446):496–509*

[65] Freeman G.H. and Halton J.H. (1951), *Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika 38:141-149*

[66] Freireich E.O., Gehan E., Frei E., Schroeder L.R., Wolman I.J., et al. (1963), *The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia. Blood, 21: 699–716*

[67] Friedman M. (1937), *The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32,675-701*

[68] Fritz C.O., Morris P.E., Richler J.J.(2012),*Effect size estimates: Current use, calculations, and interpretation. Journal of Experimental Psychology: General., 141(1):2–18.*

[69] Games P. A., Howell J. F. (1976), *Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. Journal of Educational Statistics, 1, 113-125*

[70] Gehan E. A. (1965a), *A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. Biometrika, 52:203—223*

[71] Gehan E. A. (1965b), *A Generalized Two-Sample Wilcoxon Test for Doubly-Censored Data. Biometrika, 52:650—653*

[72] Goodman L. A. (1960), *On the exact variance of products. Journal of the American Statistical Association, 55, 708-713*

[73] Greenhouse S. W., Geisser S. (1959), *On methods in the analysis of profile data. Psychometrika, 24, 95–112*

[74] Green S.B. (1991), *How many subjects does it take to do a regression analysis? Multivariate Behavioral Research, 26, 499-510*

[75] Guttman L. (1945), *A basic for analyzing test-retest reliabilit. Psychometrika, 10, 255-282*

[76] Hanley J.A. i Hajian-Tilaki K.O. (1997), *Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. Academic radiology 4(1):49-58*

[77] Hanley J.A. i McNeil M.D. (1982), *The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29-36*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[78] Hanley J.A. i McNeil M.D. (1983), *A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148: 839-843*

[79] Hanusz Z., Tarasińska J. (2014), *On multivariate normality tests using skewness and kurtosis, Colloquium Biometricum 44, 139-148*

[80] Harrell F.E., Lee K.L., Mark D.B. (1996), *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 15: 361-387*

[81] Henderson, R. (1916), *Note on graduation by adjusted average.Transactionsof the Actuarial Society of America, 17:43–48*

[82] Henze N., Zirkler B. (1990), *A class of invariant consistent tests for multivariate normality. Comm. Statist. Theory Methods. 1990;19:3595–3617*

[83] Heagerty P.J., Lumley T., Pepe M.S. (2000), *Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics, 56(2):337–344*

[84] Hochberg Y. (1988), *A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika 75 (4): 800–802*

[85] Holm S. (1979), *A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70*

[86] Hotelling H. (1931), *The generalization of Student's ratio. Annals of Mathematical Statistics 2 (3): 360–378*

[87] Hotelling, H. (1947), *Multivariate Quality Control. In C. Eisenhart, M. W. Hastay, and W. A. Wallis, eds. Techniques of Statistical Analysis. New York: McGraw-Hill*

[88] Hotelling H. (1951), *A generalized t 2 test and measurement of multivariate dispersion. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability 1: 23–41*

[89] Huynh H., Feldt L. S. (1976), *Estimation of the Box correction for degrees of freedom from sample data in randomized block and split=plot designs. Journal of Educational Statistics, 1, 69–82*

[90] Iman R. L., Davenport J. M. (1980), *Approximations of the critical region of the friedman statistic, Communications in Statistics 9, 571–595*

[91] Jarque C. M., Bera A. K., (1987)., *A test for normality of Observations and Regression Residuals, International Statistical Review 55, 163-172*

[92] Jones M. C., Marron J. S., Sheather S. J., (1996)., *A brief survey of bandwidth selection for density estimation. J. Amer. Statist. Assoc. 91 401–407*

[93] Jonckheere A. R. (1954), *A distribution-free k-sample test against ordered alternatives. Biometrika, 41: 133–145*

[94] Kaplan E.L., Meier P. (1958), *Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53:457-481*

[95] Kendall M.G. (1938), *A new measure of rank correlation. Biometrika, 30, 81-93.*

[96] Kendall M.G., Babington-Smith B. (1939), *The problem of m rankings. Annals of Mathematical Statistics, 10, 275-287*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[97] Kleinbaum D. G., Klein M. (2005), *Survival Analysis: A Self-Learning Text, Second Edition (Statistics for Biology and Health)*

[98] Kolmogorov A.N. (1933), *Sulla deterrninazione empirica di una legge di distribuzione. Giornde1l'Inst. Ital. degli. Art., 4, 89-91*

[99] Kruskal W.H. (1952), *A nonparametric test for the several sample problem. Annals of Mathematical Statistics, 23, 525-540*

[100] Kruskal W.H., Wallis W.A. (1952), *Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, 583-621*

[101] Lancaster H.O. (1961), *Significance tests in discrete distributions. Journal of the American Statistical Association 56:223-234*

[102] Lawley D. N. (1938), *A generalization of Fisher's z-test. Biometrika 30: 180–187*

[103] Lee E. T., Wang J. W. (2003), *Statistical Methods for Survival Data Analysis, ed. third, Wiley*

[104] Lenth, R. V., (2001), *Some Practical Guidelines for Effective Sample Size Determination. The American Statistician, 55(3), 187-193*

[105] Levene H. (1960), *Robust tests for the equality of variance. In I. Olkin (Ed.) Contributions to probability and statistics (278-292). Palo Alto, CA: Stanford University Press*

[106] Liddell F.D.K. (1983) *Simplified exact analysis of case-referent studies; matched pairs; dichotomous exposure. Journal of Epidemiology and Community Health; 37:82-84.*

[107] Lilliefors H.W. (1967), *On the Kolmogorov-Smimov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62,399-402*

[108] Lilliefors H.W. (1969), *On the Kolmogorov-Smimov test for the exponential distribution with mean unknown. Journal of the American Statistical Association, 64,387-389*

[109] Lilliefors H.W. (1973), *The Kolmogorov-Smimov and other distance tests for the gamma distribution and for the extreme-value distribution when parameters must be estimated. Department of Statistics, George Washington University, unpublished manuscript*

[110] Lloyd S. P. (1982), *Least squares quantization in PCM. IEEE Transactions on Information Theory 28 (2): 129–137*

[111] Lund R.E., Lund J.R. (1983), *Algorithm AS 190, Probabilities and Upper Quantiles for the Studentized Range. Applied Statistics; 34*

[112] Mahalanobis P. C. (1930), *On tests and measures of group divergence. Journal of the Asiatic Society of Bengal 26: 541–588*

[113] Mahalanobis P. C. (1936), *On the generalized distance in statistics. National Institute of Science of India 12: 49–55*

[114] Mann H. and Whitney D. (1947), *On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 1 8 , 5 0 4*

[115] Mantel N. and Haenszel W. (1959), *Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22,719-748*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[116] Mantel N. (1963), *Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. J. Am. Statist. Assoc., 58, 690-700*

[117] Mantel N. (1966), *Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. Cancer Chemotherapy Reports, 50:163—170*

[118] Marascuilo L.A. and McSweeney M. (1977), *Nonparametric and distribution-free method for the social sciences. Monterey, CA: Brooks/Cole Publishing Company*

[119] Mardia K. V. (1970), *Measures of multivariate skewness and kurtosis with applications, Biometrica 57, 519-530*

[120] Mardia K. V. (1974), *Applications of some measuresof multivariate skewness and kurtosis for testing normality and robustness studies, Sankhay B 36, 115-128*

[121] Mauchly J. W. (1940), *Significance test for sphericity of n-variate normal population. Annals of Mathematical Statistics, 11, 204-209.*

[122] McNemar Q. (1947), *Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157*

[123] Mehta C.R. and Patel N.R. (1986), *Algorithm 643. FEXACT: A Fortran subroutine for Fisher's exact test on unordered r*c contingency tables. ACM Transactions on Mathematical Software, 12, 154–161*

[124] Miettinen O.S. (1985), *Theoretical Epidemiology: Principles of Occurrence Research in Medicine. John Wiley and Sons, New York*

[125] Miettinen O.S. and Nurminen M. (1985), *Comparative analysis of two rates. Statistics in Medicine 4: 213-226*

[126] Mimar S.F. (2017), *The Mediation Analysis With the Sobel Test and the Percentile Bootstrap, International Journal of Management and Applied Science, Volume-3, Issue-2*

[127] Nadaraya, E. A. (1964), *On Estimating Regression. Theory of Probability and Its Applications. 9 (1): 141–2*

[128] Newcombe R.G. (1998), *Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods. Statistics in Medicine 17: 873-890*

[129] Newman S.C.(2001), *Biostatistical Methods in Epidemiology. 2nd ed. New York: John Wiley*

[130] Normand S.L. T., Landrum M.B., Guadagnoli E., Ayanian J.Z., Ryan T.J., Cleary P.D., McNeil B.J. (2001), *Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. Journal of Clinical Epidemiology; 54:387–398.*

[131] Ogilvie J. C. (1965), *Paired comparison models with tests for interaction. Biometrics 21(3): 651-64*

[132] Orwin R. G. (1983), *A Fail-SafeN for Effect Size in Meta-Analysis. J Educ Behav Stat, 8(2):157-159*

[133] Oyeyemi G.M. , Adewara A.A., Adebola F.B. and Salau S.I. (2010), *On the Estimation of Power and Sample Size in Test of Independence, Asian Journal of Mathematics and Statistics, 3(3): 139-146*

[134] Page E. B. (1963), *Ordered hypotheses for multiple treatments: A significance test for linear ranks. Journal of the American Statistical Association 58 (301): 216–30*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[135] Peduzzi P., Concato J., Feinstein A.R., Holford T.R. (1995), *Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of Clinical Epidemiology, 48:1503-1510*

[136] Peduzzi P., Concato J., Kemper E., Holford T.R., Feinstein A.R. (1996), *A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology; 49(12):1373-9*

[137] Pillai K. C. (1955), *Some new test criteria in multivariate analysis. Annals of Mathematical Statistics 26: 117–121*

[138] Plackett R.L. (1984), *Discussion of Yates' "Tests of significance for 2x2 contingency tables". Journal of Royal Statistical Society Series A 147:426-463*

[139] Pratt J.W. and Gibbons J.D. (1981), *Concepts of Nonparametric Theory. Springer-Verlag, New York*

[140] Robins, J., Breslow, N., and Greenland S. (1986), *Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. Biometrics 42, 311–323*

[141] Robins, J., Greenland S. and Breslow, N.E. (1986), *A general estimator for the variance of the Mantel–Haenszel odds ratio. American Journal of Epidemiology 124, 719–723*

[142] Rosenbaum P.R., Rubin D.B. (1983a), *The central role of the propensity score in observational studies for causal effects. Biometrika; 70:41–55*

[143] Rosenthal R. (1979), *The "file drawer problem" and tolerance for null results. Psychological Bulletin, 5, 638-641*

[144] Rothman K.J., Greenland S., Lash T.L. (2008), *Modern Epidemiology, 3rd ed. (Lippincott Williams and Wilkins) 221−225*

[145] Roy S. N. (1939), *p-statistics or some generalizations in analysis of variance appropriate to multivariate problems. Sankhya 4: 381–396*

[146] Royston P. (1992), *Approximating the Shapiro–Wilk W-test for non-normality". Statistics and Computing 2 (3): 117–119*

[147] Royston P. (1993b), *A toolkit for testing for non-normality in complete and censored samples. Statistician 42: 37–43*

[148] Rufibach K. (2010), *Assessment of paired binary data; Skeletal Radiology volume 40, pages1–4*

[149] Satterthwaite F.E. (1946), *An approximate distribution of estimates of variance components. Biometrics Bulletin, 2, 1 10-1 14*

[150] Savin N.E. and White K.J. (1977), *The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors. Econometrica 45, 1989-1996*

[151] Schiaparelli, G. V.(1866), *Sul modo di ricavare la vera espressione delle leggidelta natura dalle curve empiricae.Effemeridi Astronomiche di Milano perl'Arno, 857:3–56*

[152] Scott D. W., (1992), *Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley.*

[153] Shapiro S.S. and Wilk M.B. (1965), *An analysis of variance test for normality (complete samples). Biometrika 52 (3–4): 591–611*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[154] Sheather S.J. (2009), *A modern approach to regression with R. New York, NY: Springer*

[155] Shrout P.E., and Fleiss J.L (1979), *Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86, 420-428*

[156] Šidák Z. K. (1967), *Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association, 62 (318): 626–633*

[157] Silverman B. W., (1986), *Density estimation for statistics and data analysis, London: Chapman and Hall*

[158] Skillings J.H., Mack G.A. (1981) *On the use of a Friedman-type statistic in balanced and unbalanced block designs. Technometrics, 23:171–177*

[159] Sobel M. E. (1982). *Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology 13: 290–312*

[160] Spearman C. (1910), *Correlation calculated from faulty data. British Journal of Psychology, 3, 271-295*

[161] Tamhane A. C. (1977), *Multiple comparisons in model I One-Way ANOVA with unequal variances. Communications in Statistics, A6 (1), 15-32*

[162] Tarone R. E., Ware J. (1977), *On distribution-free tests for equality of survival distributions. Biometrica, 64(1):156-160*

[163] Tarone R.E. (1985), *On heterogeneity tests based on efficient scores. Biometrika 72, 91–95*

[164] Terpstra T. J. (1952), *The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. Indagationes Mathematicae, 14: 327–333*

[165] Terrell G. R. (1990), *The maximal smoothing principle in density estimation. Journal of the American Statistical Association 85, 470–477*

[166] Terrell G.R., Scott D. W. (1985), *Oversmoothed nonparametric density estimates. Journal of the American Statistical Association 80, 209-214*

[167] Thode H. C. (2002), *Testing For Normality. CRC Press; 2002. 506 s.*

[168] Volinsky C.T., Raftery A.E. (2000) , *Bayesian information criterion for censored survival models. Biometrics, 56(1):256–262*

[169] Verweij P.J.M., van Houwelingen H.C. (1993), *Cross-validation in survival analysis. Statistics in Medicine 12(24): 2305-2314*

[170] Wallenstein S. (1997), *A non-iterative accurate asymptotic confidence interval for the difference between two Proportions. Statistics in Medicine 16: 1329-1336*

[171] Wallis W.A. (1939), *The correlation ratio for ranked data. Journal of the American Statistical Association, 34,533-538*

[172] Watson, G. S. (1964), *Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A. 26 (4): 359–372*

[173] Welch B. L. (1951), *On the comparison of several mean values: an alternative approach. Biometrika 38: 330–336*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[174] Wilcoxon F. (1945), *Individual comparisons by ranking methods. Biometries, 1,80-83*

[175] Wilcoxon F. (1945), *Individual comparisons by ranking methods. Biometries, 1, 80-83*

[176] Wilcoxon F. (1949), *Some rapid approximate statistical procedures. Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation*

[177] Wilcoxon F. (1949), *Some rapid approximate statistical procedures. Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation*

[178] Wilcoxon F. (1949), *Some rapid approximate statistical procedures. Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation*

[179] Wilson E.B. (1927), *Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association: 22(158):209-212*

[180] Wilks S.S. (1932), *Certain generalizations in the analysis of variance. Biometrika 24: 471–494*

[181] Yates F. (1934), *Contingency tables involving small numbers and the chi-square test. Journal of the Royal Statistical Society, 1,2 17-235*

[182] Youden W.J. (1950), *Index for rating diagnostic tests. Cancer. 3: 32–35*

[183] Kubiak K.B., Więckowska B., Jodłowska-Siewert E., Guzik P. (2024), *Visualising and quantifying the usefulness of new predictors stratified by outcome class: The U-smile method. PLOS ONE 19(5): e0303276*

[184] Więckowska B., Kubiak K.B., Guzik P. (2025), *Evaluating the three-level approach of the U-smile method for imbalanced binary classification. PLOS ONE 20(4): e0321661*

[185] Więckowska B., Guzik P. (2026), *Usmile likelihood evaluation provides robust threshold free assessment of binary classification models for balanced and imbalanced datasets. Scientific Reports 16: 10000*

[186] Yule G. (1900), *On the association of the attributes in statistics: With illustrations from the material ofthe childhood society, and c. Philosophical Transactions of the Royal Society, Series A, 194,257-3 19*

[187] Zar J. H., (2010), *Biostatistical Analysis (Fifth Edition). Pearson Educational*

[188] Zweig M.H., Campbell G. (1993), *Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry 39:561-577*

[189] Bates D., Mächler M., Bolker B., Walker S. (2015), *Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1):1-48*

[190] Belenky G., Wesensten N.J., Thorne D.R., Thomas M.L., Sing H.C., Redmond D.P., Russo M.B., Balkin T.J. (2003), *Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. Journal of Sleep Research 12:1-12*

[191] Klein J.P. (1992), *Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics 48:795-806*

[192] Kuznetsova A., Brockhoff P.B., Christensen R.H.B. (2017), *lmerTest package: tests in linear mixed effects models. Journal of Statistical Software 82(13):1-26*

Copyright ©2010-2026 U-smile Stats – All rights reserved

[193] Laird N.M., Ware J.H. (1982), *Random-effects models for longitudinal data. Biometrics 38:963-974*

[194] Lesnoff M., Laval G., Bonnet P., Abdicho S., Workalemahu A., Kifle D., Peyraud A., Lancelot R., Thiaucourt F. (2004), *Within-herd spread of contagious bovine pleuropneumonia in Ethiopian highlands. Preventive Veterinary Medicine 64:27-40*

[195] McGilchrist C.A., Aisbett C.W. (1991), *Regression with frailty in survival analysis. Biometrics 47:461-466*

[196] Nakagawa S., Schielzeth H. (2013), *A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4:133-142*

[197] Potthoff R.F., Roy S.N. (1964), *A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika 51:313-326*

[198] Self S.G., Liang K.Y. (1987), *Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82:605-610*

[199] Therneau T.M., Grambsch P.M. (2000), *Modeling Survival Data: Extending the Cox Model. Springer, New York*

[200] Zeger S.L., Liang K.Y., Albert P.S. (1988), *Models for longitudinal data: a generalized estimating equation approach. Biometrics 44:1049-1060*

Copyright ©2010-2026 U-smile Stats – All rights reserved