PQStat Software
Statistical Compution Software

User Manual - PQStat

Barbara Więckowska

To the version 1.8.6
P7909040423

www.pqstat.pl

Spis treści

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
- OS: Windows® 2000/XP/Vista/7/8

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.

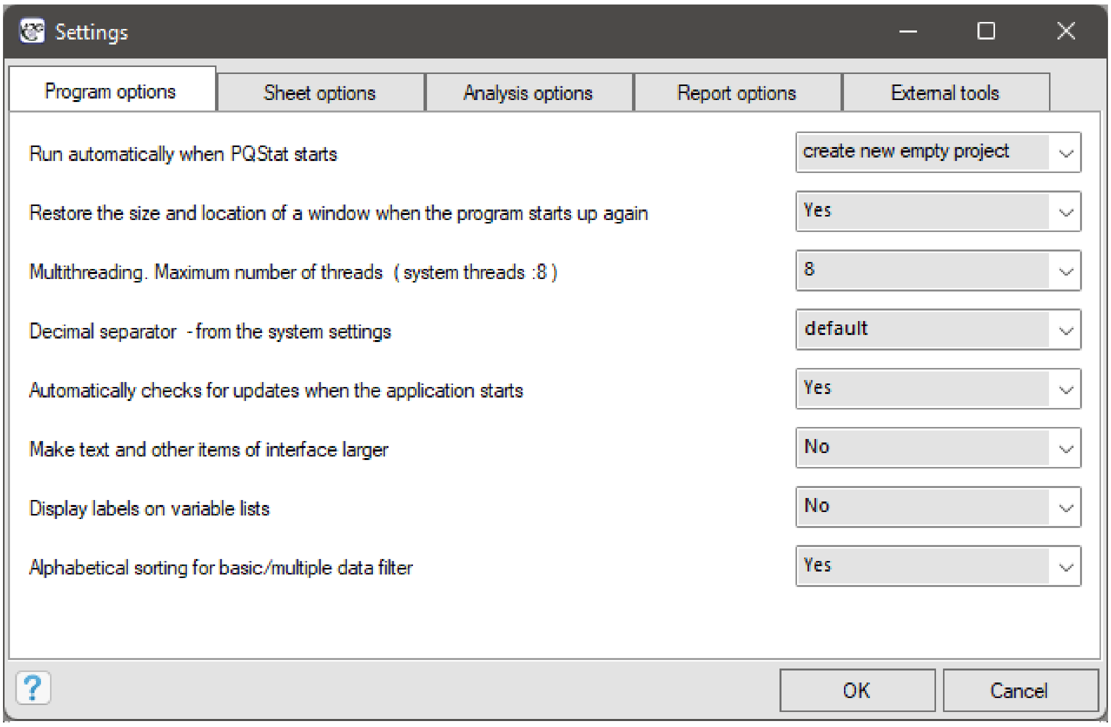
Pressing “Next” will give you a choice of either a full installation or a version not including exemplary data sets. The data sets are used in the user guide.

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.

3 SETTINGS



Run automatically when PQStat starts: create new empty project

Restore the size and location of a window when the program starts up again: Yes

Multithreading. Maximum number of threads (system threads :8): 8

Decimal separator - from the system settings: default

Automatically checks for updates when the application starts: Yes

Make text and other items of interface larger: No

Display labels on variable lists: No

Alphabetical sorting for basic/multiple data filter: Yes

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.

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.

- 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 http://manuals.pqstat.pl/statpqpl:installpl

4 PROGRAM OPERATION

Documents management is based on projects. Each project is a separate file.

A project is similar to a worksheet. It consists of 3 basic elements:

1. Datasheets (including map sheets and matrixs) - the number of sheets in a given project is limited to 1000,
2. Results sheets (reports) - the number of reports in a given datasheet is limited to 2000,
3. Project manager – allows you to change the name of datasheets and result reports, create your own descriptions and notes as well as export.

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) New project

- File→New datasheet (Ctrl+D) Add datasheet

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

- File→Open project (Ctrl+O) Open project

- File→Open examples 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) Save project

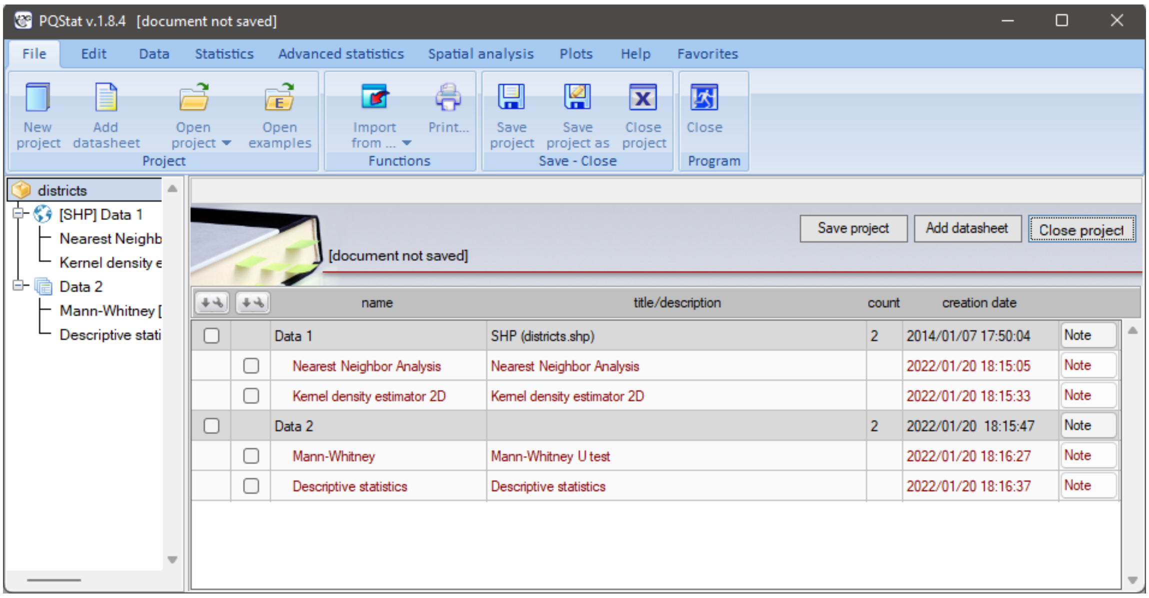
- File→Save project as... Save project as
- Save project button in the Project Manager.

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
- 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).



		name	title/description	count	creation date	
□		Data 1	SHP (districts.shp)	2	2014/01/07 17:50:04	Note
	□	Nearest Neighbor Analysis	Nearest Neighbor Analysis		2022/01/20 18:15:05	Note
	□	Kernel density estimator 2D	Kernel density estimator 2D		2022/01/20 18:15:33	Note
□		Data 2		2	2022/01/20 18:15:47	Note
	□	Mann-Whitney	Mann-Whitney U test		2022/01/20 18:16:27	Note
	□	Descriptive statistics	Descriptive statistics		2022/01/20 18:16:37	Note

4.1 WORKING WITH DATASHEETS

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

4.1.1 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) Add datasheet
- 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, however, that if you have reports or a map attached to a datasheet, deleting the datasheet also deletes any reports/maps 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.

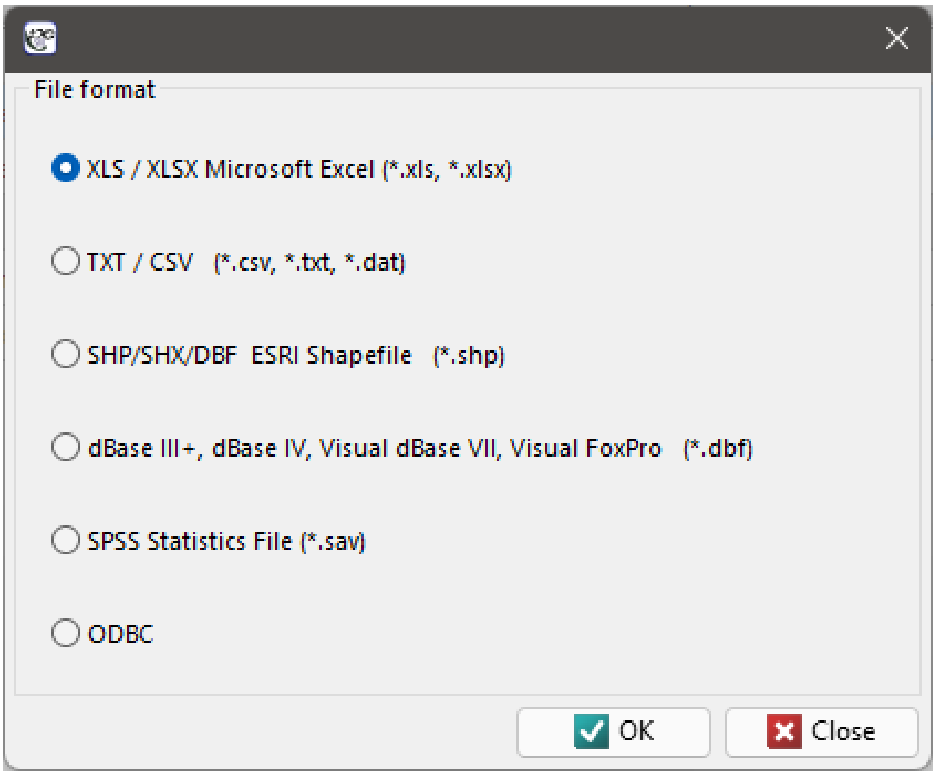
4.1.2 INPUTTING DATA INTO A DATASHEET

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 ... Import from ...



File format

● XLS / XLSX Microsoft Excel (*.xls, *.xlsx)

○ TXT / CSV (*.csv, *.txt, *.dat)

○ SHP/SHX/DBF ESRI Shapefile (*.shp)

○ dBase III+, dBase IV, Visual dBase VII, Visual FoxPro (*.dbf)

○ SPSS Statistics File (*.sav)

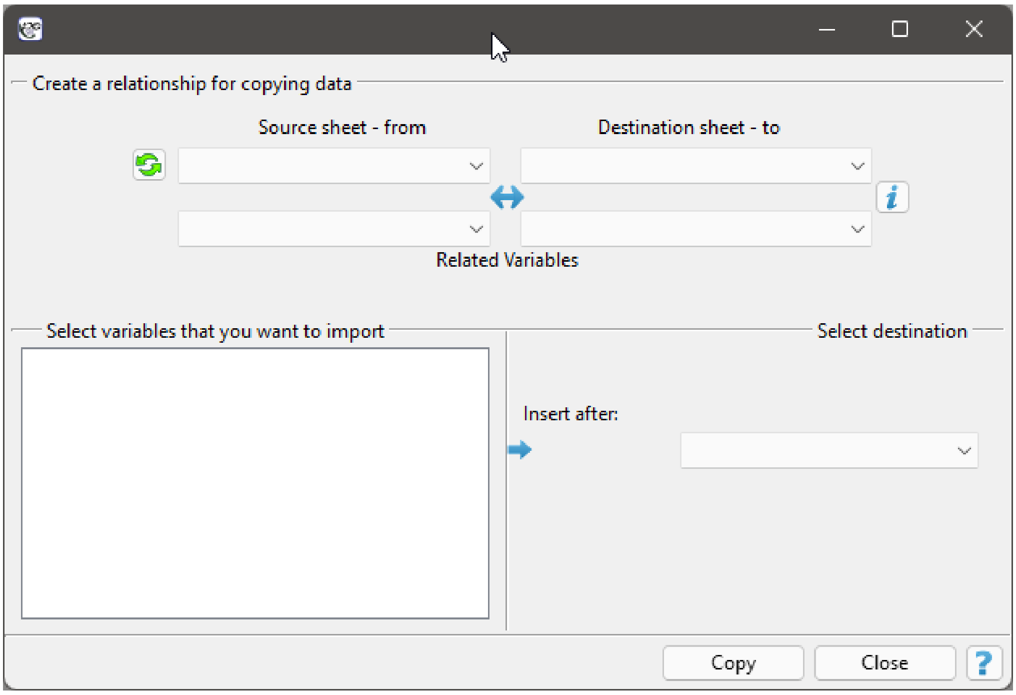
○ ODBC

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. To avoid misinterpretation of special characters, pay attention to these characters in the preview window. 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 Microsft Excel documents applies only to cell values; formatting and formulas cannot be imported.

Copying data with relation

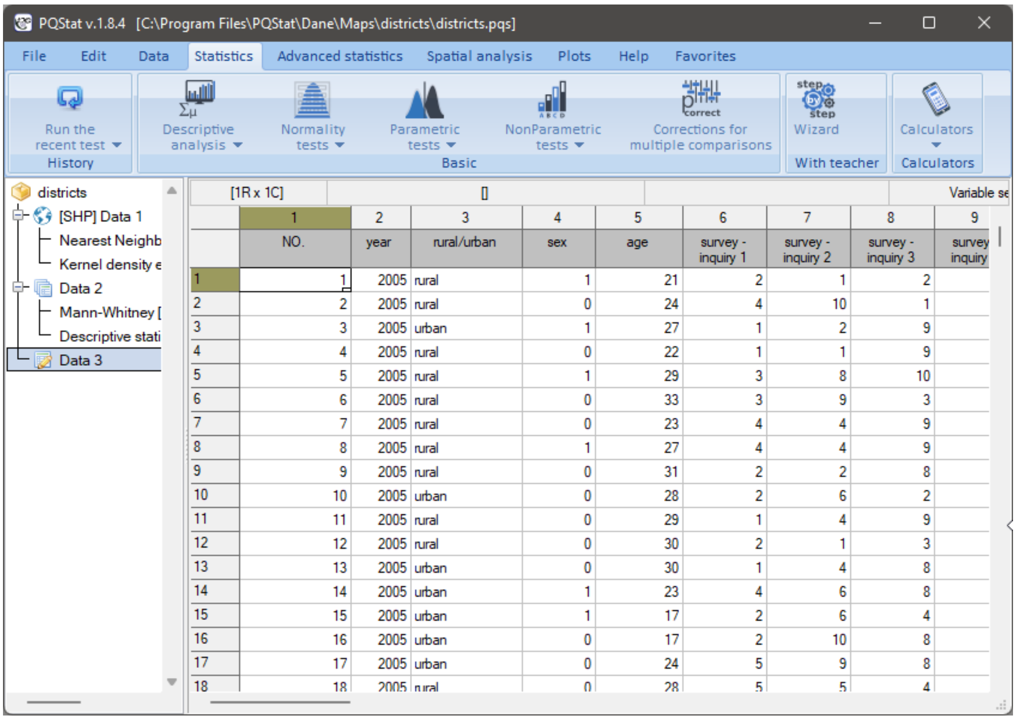
Data from another worksheet can be copied into the selected data sheet based on relations. This type of data copying is performed by selecting menu: Data→Copying with relation... Copying with relation



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.

4.1.3 DATASHEET WINDOW

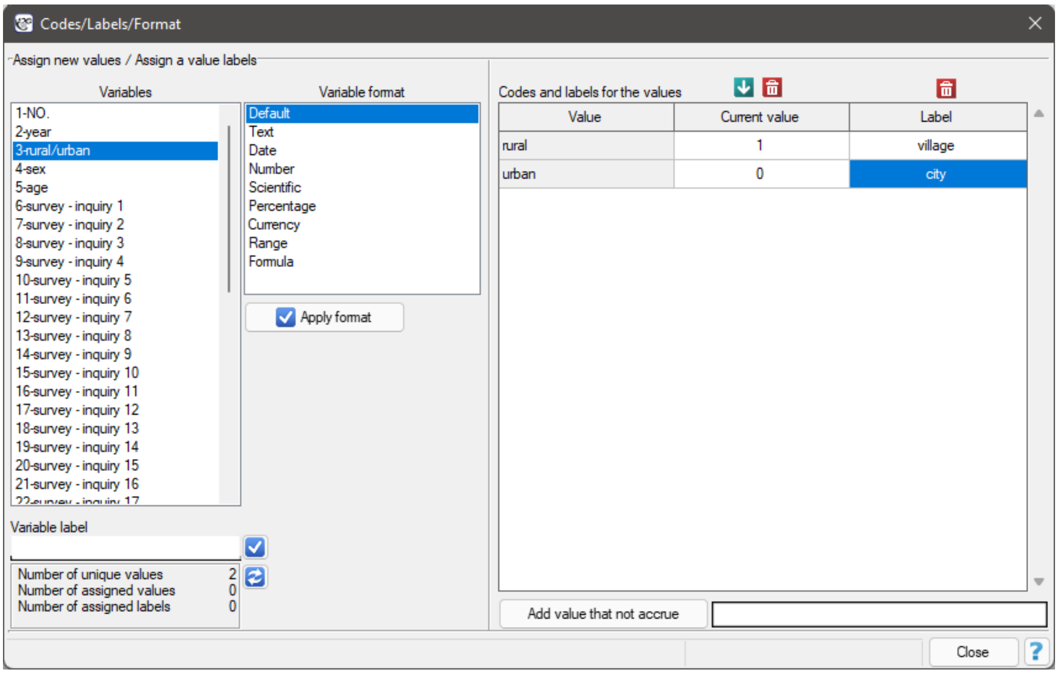
The rows and columns of the datasheet are indicated by consecutive natural numbers. Each column in the space marked in gray can be given its own heading. At the top of the datasheet is the Message bar. This is where the current user information is displayed. The left part of the bar informs about the 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.



	1	2	3	4	5	6	7	8	9
	NO.	year	rural/urban	sex	age	survey - inquiry 1	survey - inquiry 2	survey - inquiry 3	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	

4.1.4 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).


Codes/Labels/Format
Assign new values / Assign a value labels
Variables
Variable format
Codes and labels for the values
Apply format
Variable label
Add value that not accrue
Close
?

Variables:
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 4
10-survey - inquiry 5
11-survey - inquiry 6
12-survey - inquiry 7
13-survey - inquiry 8
14-survey - inquiry 9
15-survey - inquiry 10
16-survey - inquiry 11
17-survey - inquiry 12
18-survey - inquiry 13
19-survey - inquiry 14
20-survey - inquiry 15
21-survey - inquiry 16
22-survey - inquiry 17

Variable format:
Default
Text
Date
Number
Scientific
Percentage
Currency
Range
Formula

☑ Apply format

Value	Current value	Label
rural	1	village
urban	0	city

Number of unique values	2
Number of assigned values	0
Number of assigned labels	0

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

4.1.5 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,
- 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

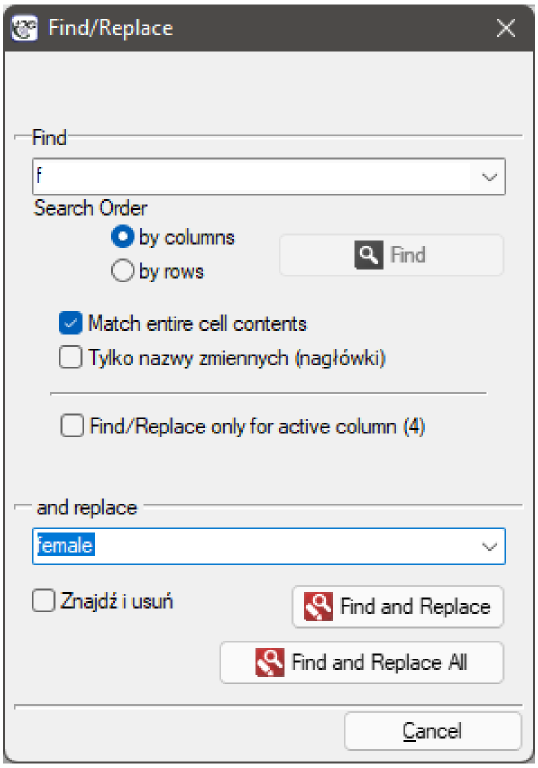
You can insert blank rows or columns above or to the left of an existing row or column. This will move the cells down or to the right. To insert a row(s), select the row(s) above which you want to insert a new row(s), and then choose Insert row from the context menu on the selected row number. Inserting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

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



Find: f

Search Order: ● by columns ○ by rows

☑ Match entire cell contents

□ Tylko nazwy zmiennych (nagłówki)

□ Find/Replace only for active column (4)

and replace: female

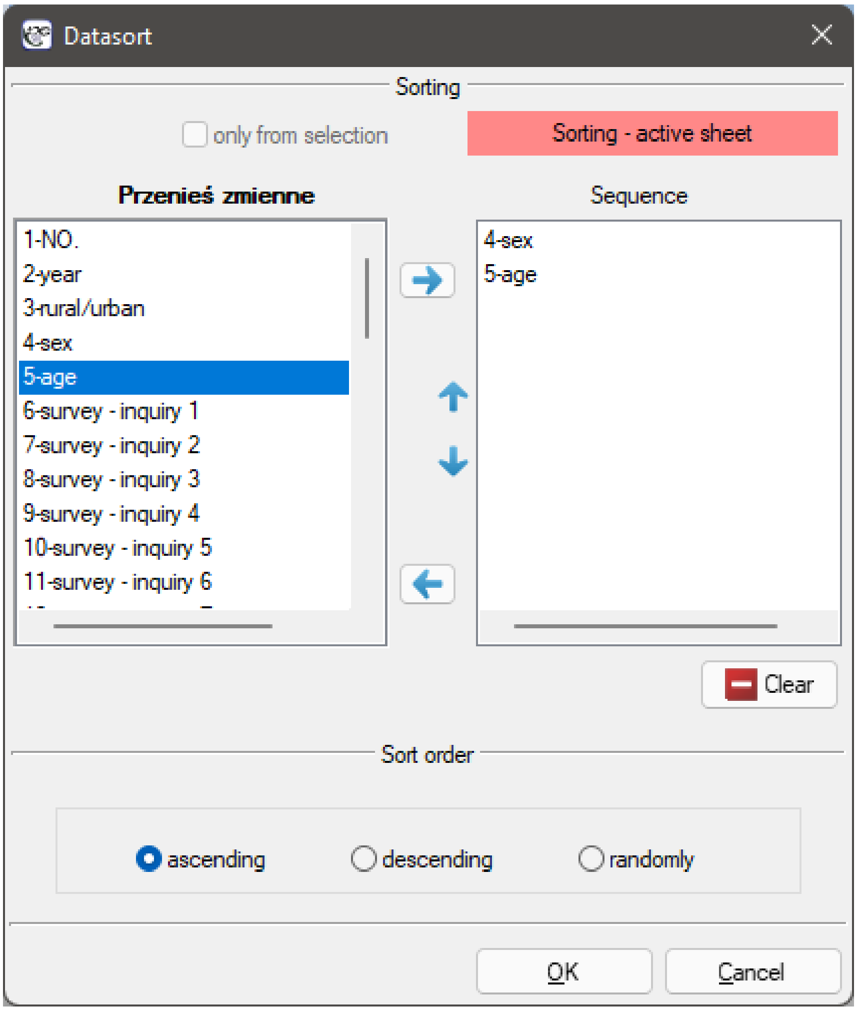
□ Znajdź i usuń

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.

4.1.6 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.



Datasort

Sorting

☐ only from selection

Sorting - active sheet

Przenieś zmienne

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 4
10-survey - inquiry 5
11-survey - inquiry 6

Sequence

4-sex
5-age

Sort order

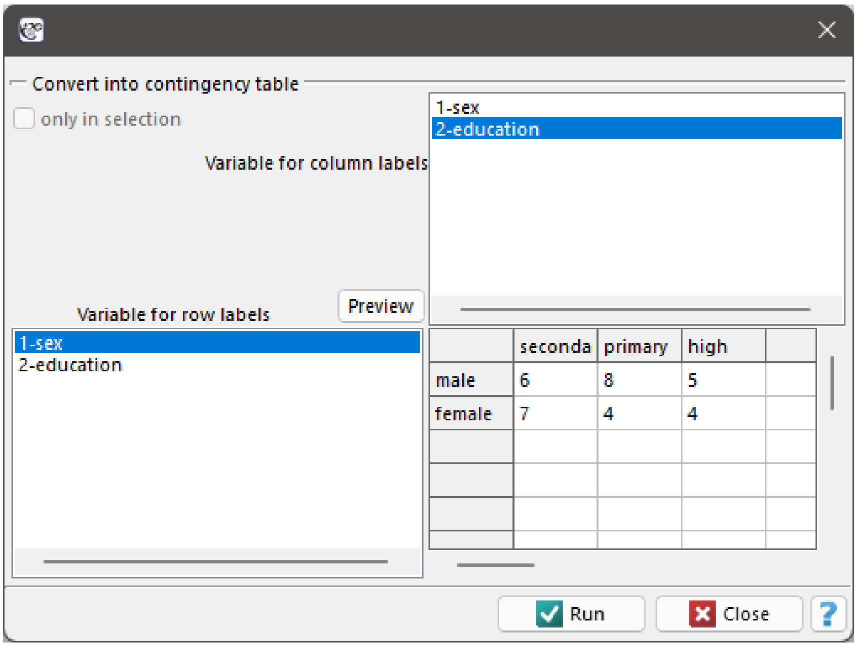
● ascending ○ descending ○ randomly

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

4.1.7 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.



Convert into contingency table

□ only in selection

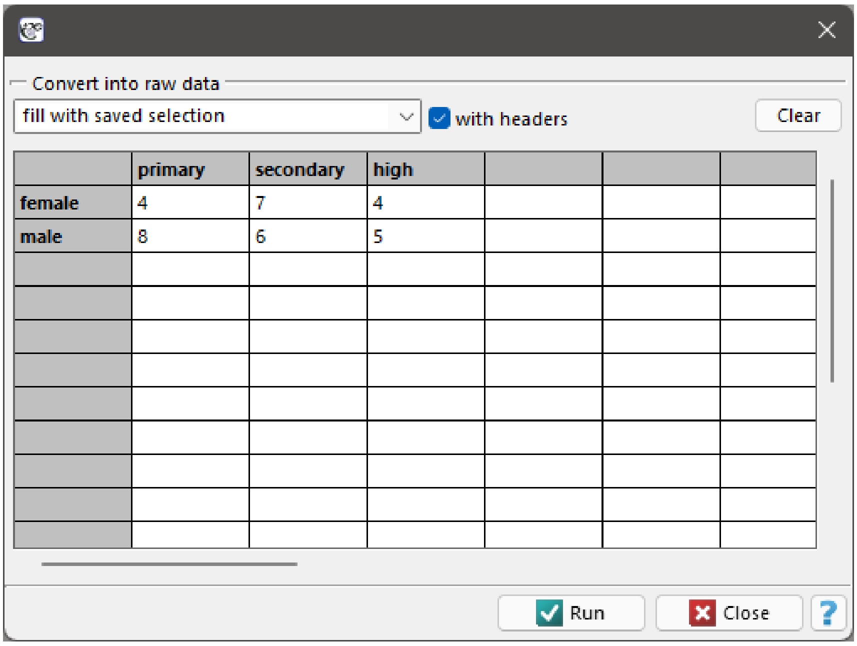
Variable for column labels: 1-sex, 2-education (selected)

Variable for row labels: 1-sex (selected), 2-education

	seconda	primary	high
male	6	8	5
female	7	4	4

4.1.8 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.



Convert into raw data

fill with saved selection

☑ with headers

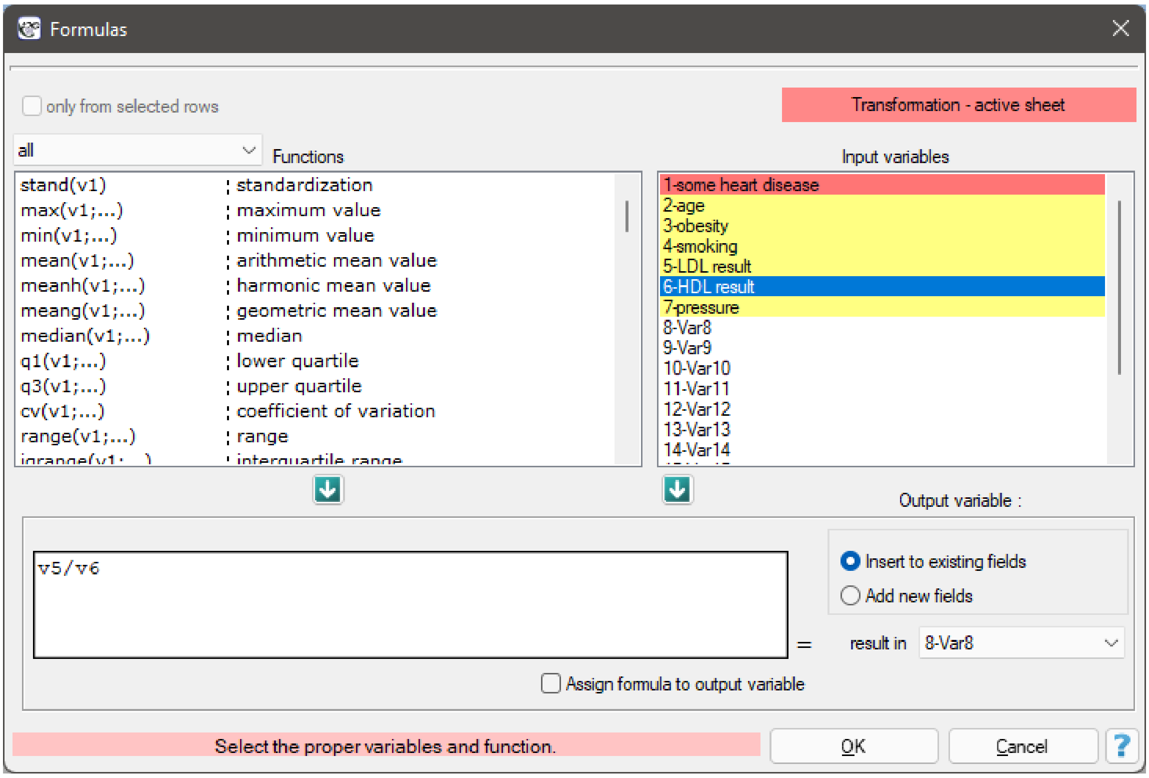
	primary	secondary	high
female	4	7	4
male	8	6	5

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.

4.1.9 FORMULAS

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...



Formulas

□ only from selected rows

Transformation - active sheet

all

Functions

stand(v1)	; standardization
max(v1;...)	; maximum value
min(v1;...)	; minimum value
mean(v1;...)	; arithmetic mean value
meanh(v1;...)	; harmonic mean value
meang(v1;...)	; geometric mean value
median(v1;...)	; median
q1(v1;...)	; lower quartile
q3(v1;...)	; upper quartile
cv(v1;...)	; coefficient of variation
range(v1;...)	; range
iqrange(v1;...)	; interquartile range

Input variables

1-some heart disease
2-age
3-obesity
4-smoking
5-LDL result
6-HDL result
7-pressure
8-Var8
9-Var9
10-Var10
11-Var11
12-Var12
13-Var13
14-Var14

Output variable :

v5/v6

● Insert to existing fields
○ Add new fields

= result in 8-Var8

□ Assign formula to output variable

Select the proper variables and function.

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.

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,
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 4.1.4). 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

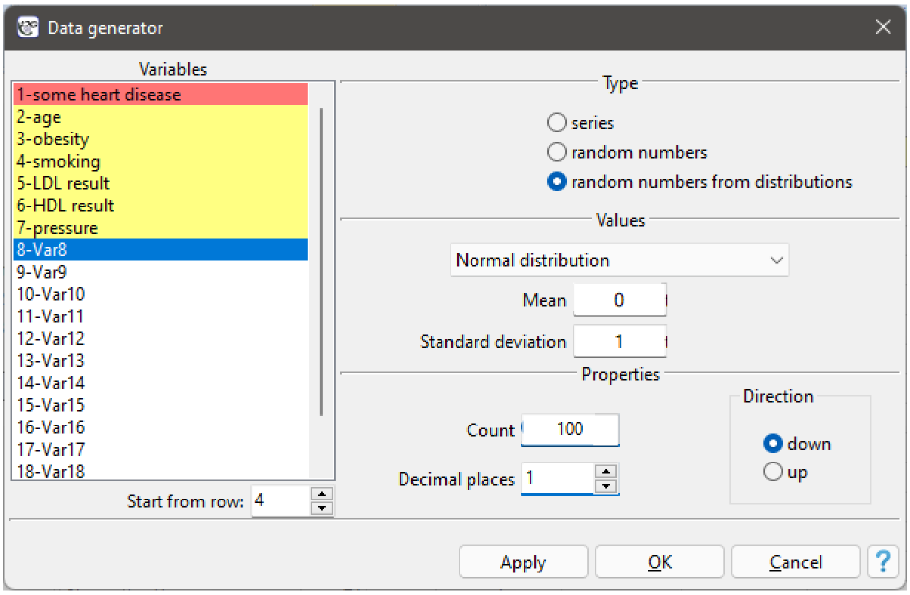
if(question;'yes - answer';'no - answer') - a question is formulated as an expression that can be true or false; the function outputs one value if the expression is true and the other if it is false,
and - conjunction operator - returns true (1) when all conditions it combines are true, false (0) otherwise,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

4.1.10 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...



Data generator

Variables
1-some heart disease
2-age
3-obesity
4-smoking
5-LDL result
6-HDL result
7-pressure
8-Var8
9-Var9
10-Var10
11-Var11
12-Var12
13-Var13
14-Var14
15-Var15
16-Var16
17-Var17
18-Var18

Start from row: 4

Type
○ series
○ random numbers
● random numbers from distributions

Values
Normal distribution
Mean 0
Standard deviation 1

Properties
Count 100
Decimal places 1

Direction
● down
○ up

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.
- 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.

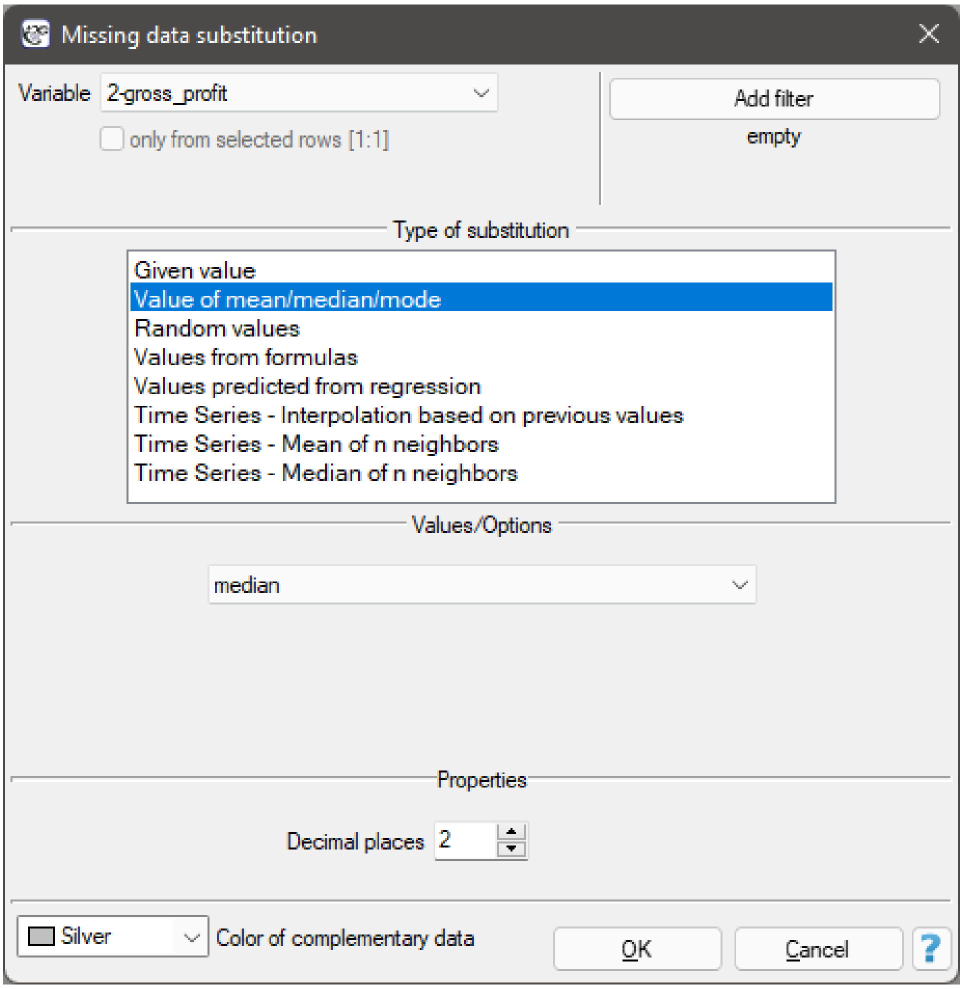
The amount of data generated depends on the value the user enters in the Count box, and the precision depends on the setting of the Decimal places box. The data will be filled in starting from the active cell in either down or up - depending on the selected option. Finally, confirm your selection with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

4.1.11 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...



Missing data substitution

Variable 2-gross_profit

□ only from selected rows [1:1]

Add filter

empty

Type of substitution

Given value
Value of mean/median/mode
Random values
Values from formulas
Values predicted from regression
Time Series - Interpolation based on previous values
Time Series - Mean of n neighbors
Time Series - Median of n neighbors

Values/Options

median

Properties

Decimal places 2

Silver Color of complementary data

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;

- 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 24.2 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 between imputed and real data, the replaced spaces are marked with a chosen color.

Example 4.1. (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.

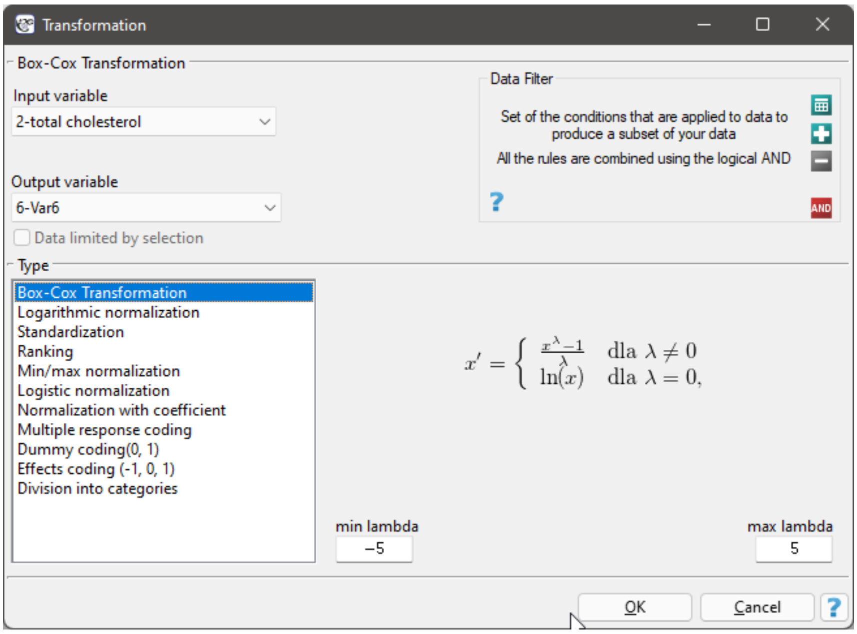
Replacing missing data with the median value is done on datasheet 1 called ”Insert the median”. Set the variable prepared to be inputed as gross profit in the Missing data window and select the method of replacement as the median value. This will result in a value of 46 850 dollars being inputed in place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

4.1.12 TRANSFORMATIONS

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


Transformation
Box-Cox Transformation
Input variable
2-total cholesterol
Output variable
6-Var6
□ Data limited by selection
Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
AND
Type
Box-Cox Transformation
Logarithmic normalization
Standardization
Ranking
Min/max normalization
Logistic normalization
Normalization with coefficient
Multiple response coding
Dummy coding(0, 1)
Effects coding (-1, 0, 1)
Division into categories
$x' = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{dla } \lambda \neq 0 \\ \ln(x) & \text{dla } \lambda = 0, \end{cases}$
min lambda
−5
max lambda
5
OK
Cancel

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 [24] 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:

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

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

Note

If min-max normalization was used before the Box-Cox transformation, then after the Box-Cox transformation, you can return to the previous range by using this transformation again.

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

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

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 24.1 PREPARING VARIABLES FOR ANALYSIS IN MULTI-DIMENSIONAL MODELS.

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 24.1 PREPARING VARIABLES FOR ANALYSIS IN MULTIDIMENSIONAL 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.

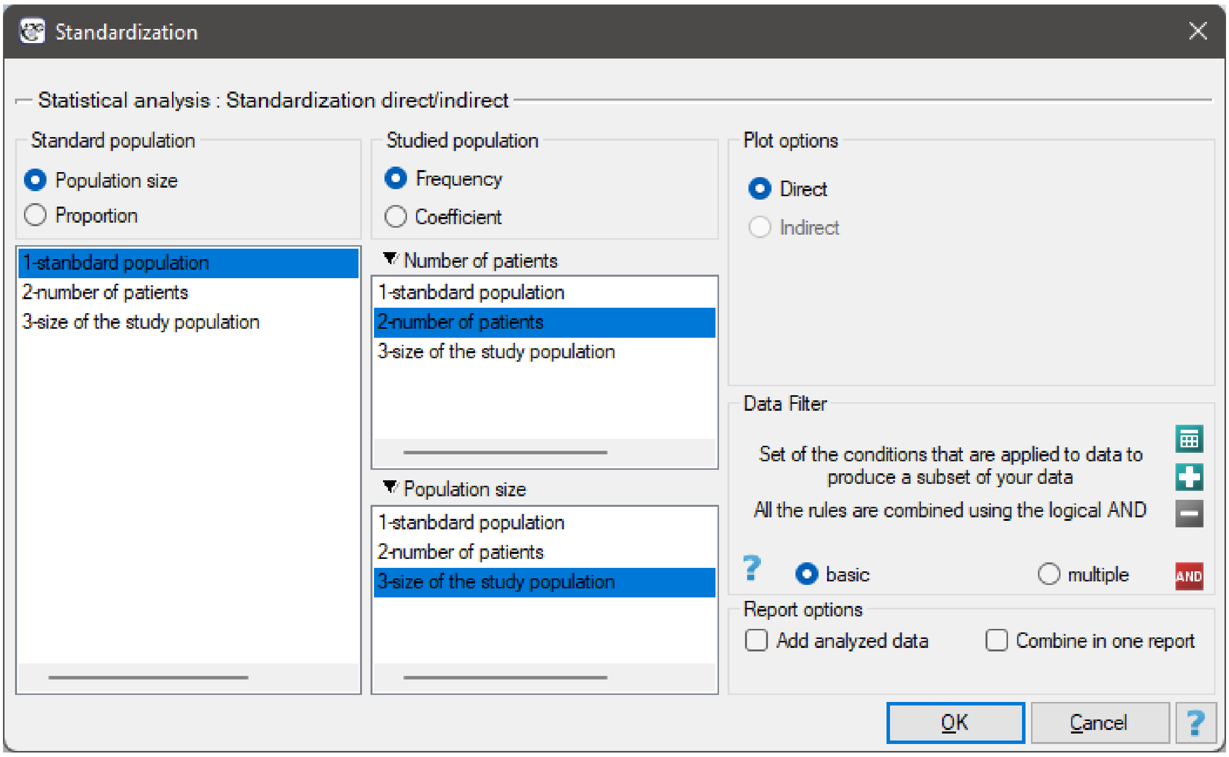
In the division window, it is also possible to select Add color scheme then the column that will store the new data will be color coded according to the indicated scheme.

Example 4.2. (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.

4.1.13 DIRECT AND INDIRECT STANDARDIZATION

Standardization We bring up the direct and indirect standardization window for epidemiological coefficients via Data→Standardization...



Standardization

Statistical analysis : Standardization direct/indirect

Standard population

● Population size
○ Proportion

1-stanbdard population
2-number of patients
3-size of the study population

Studied population

● Frequency
○ Coefficient

Number of patients

1-stanbdard population
2-number of patients
3-size of the study population

Population size

1-stanbdard population
2-number of patients
3-size of the study population

Plot options

● Direct
○ Indirect

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

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).

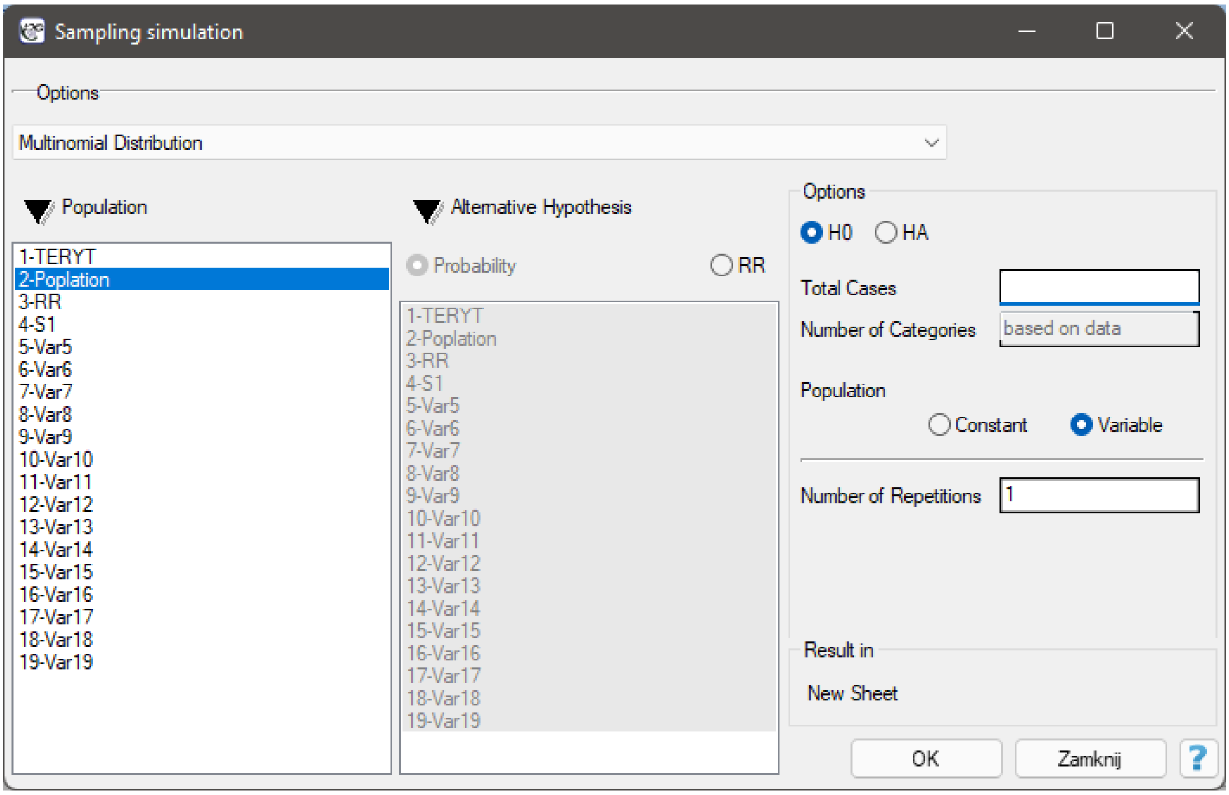
Age and gender are the most commonly used characteristics for standardization, however, the standardization can also be based on other, arbitrary characteristics, which due to their obvious impact on the phenomenon should be ”filtered out” from the study. Such features are called disturbing or confounding 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.

4.1.14 SAMPLING SIMULATION

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



Options: Multinomial Distribution

Population: 2-Poplation

Options: ● H0 ○ HA

Total Cases:

Number of Categories: based on data

Population ○ Constant ● Variable

Number of Repetitions: 1

Result in: New Sheet

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.

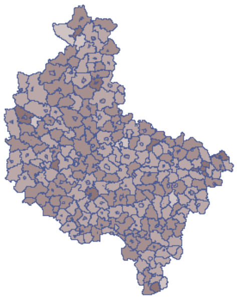
Variable population - assumes that the user is interested in distributing the cases such that the proportion of cases to the population is distributed according to the proposed distribution.

Example 4.3. (simulations.pqs file)
As a basis for the simulation, the population of Wielkopolska in 2013 was used, which according to 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:

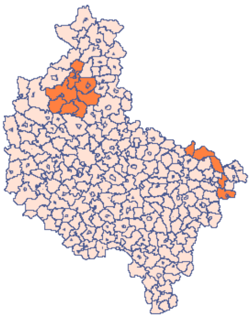
1. Random distribution (based on data from the ”Random” datasheet)
2. Four times higher disease frequency in the indicated municipalities than in the rest of the voivodeship (based on data from the ”Clusters” datasheet)

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 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.

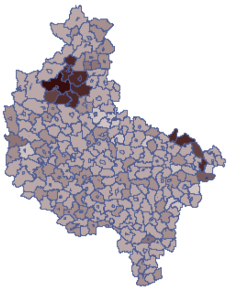


Ref 2. In the ”Clusters” datasheet, as in the previous task, the frequency for the study population is given. This time the higher frequency is expected in some municipalities (indicated on the map), so in addition, in the next column of the datasheet, the value of the relative risk for individual municipalities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



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.

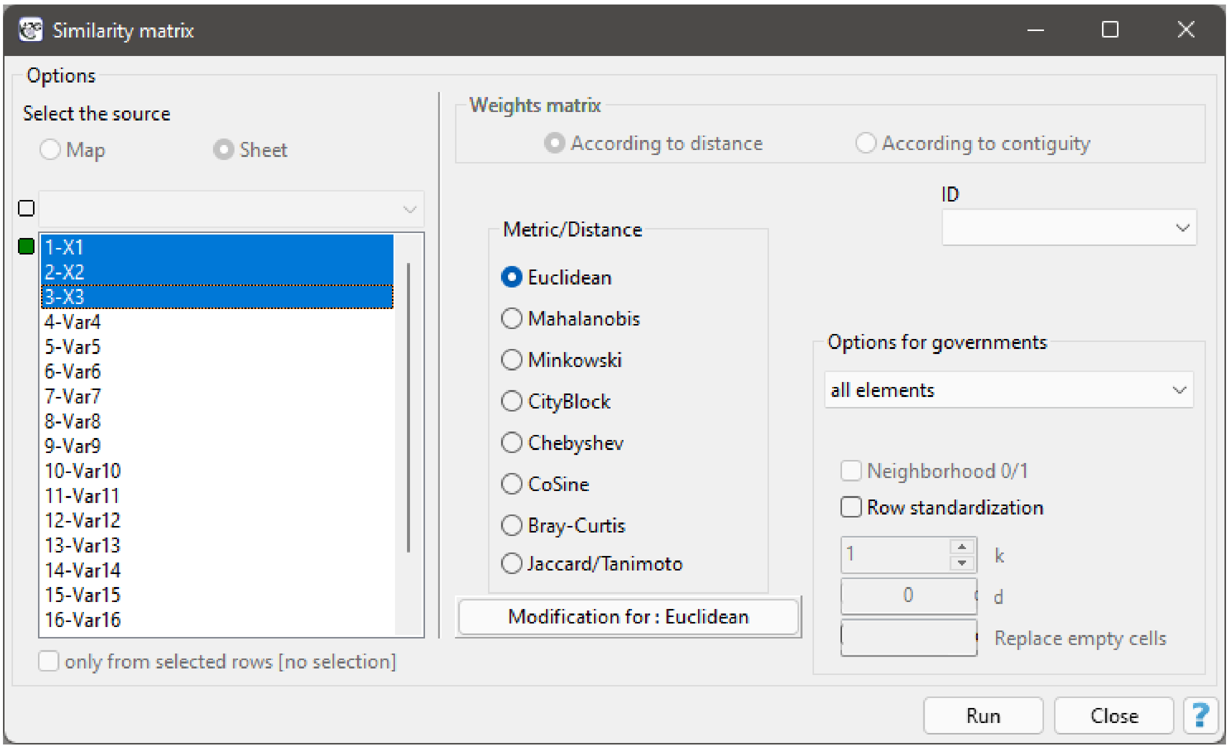
To be able to observe the distribution of the coefficient, assuming greater risk in the indicated municipalities, copy the result obtained to the datasheet ”Clusters” column ”S1”. The formula in column 7 will then be recalculated. On the map, the obtained result is presented using the Map manager Map 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.



4.1.15 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...



Similarity matrix

Options

Select the source

○ Map ● Sheet

Weights matrix

● According to distance ○ According to contiguity

ID

□

🟩 1-X1, 2-X2, 3-X3 (selected), 4-Var4, 5-Var5, 6-Var6, 7-Var7, 8-Var8, 9-Var9, 10-Var10, 11-Var11, 12-Var12, 13-Var13, 14-Var14, 15-Var15, 16-Var16

Metric/Distance

● Euclidean
○ Mahalanobis
○ Minkowski
○ CityBlock
○ Chebyshev
○ CoSine
○ Bray-Curtis
○ Jaccard/Tanimoto

Modification for : Euclidean

Options for governments: all elements

□ Neighborhood 0/1
□ Row standardization

k: 1
d: 0
Replace empty cells

□ only from selected rows [no selection]

Run Close ?

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 the same value ranges. If not, then features with higher ranges would have a greater impact on the similarity score than those with lower ranges. For example, when calculating similarity between people, we can base it on such features as body mass and age. Then body mass in kilograms, in the range of 40 to 150 kg, will have a greater influence on the result than age in years, in the range of 18 to 90 years. To ensure that the effect of each characteristic on the resulting similarity score is balanced, you should normalize/standarize each characteristic before proceeding with the analysis. On the other hand, if you want to decide the magnitude of this influence yourself, after applying the standardization, indicating the type of metric, you should enter the weights you gave yourself.

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:

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

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).

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

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 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

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 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 4.4. (file: flatsSimilarities.pqs)

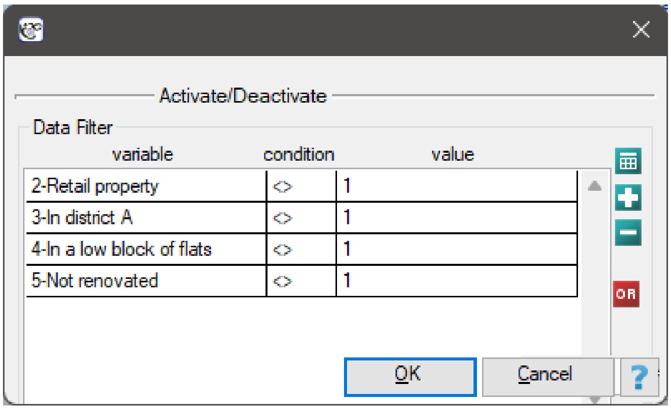
In real estate valuation procedures, for both substantive and legal reasons, the issue of similarity plays an important role. For example, it is an essential prerequisite for grouping objects and assigning them to an appropriate segment.

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:

- 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)....



variable	condition	value
2-Retail property	<>	1
3-In district A	<>	1
4-In a low block of flats	<>	1
5-Not renovated	<>	1

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).

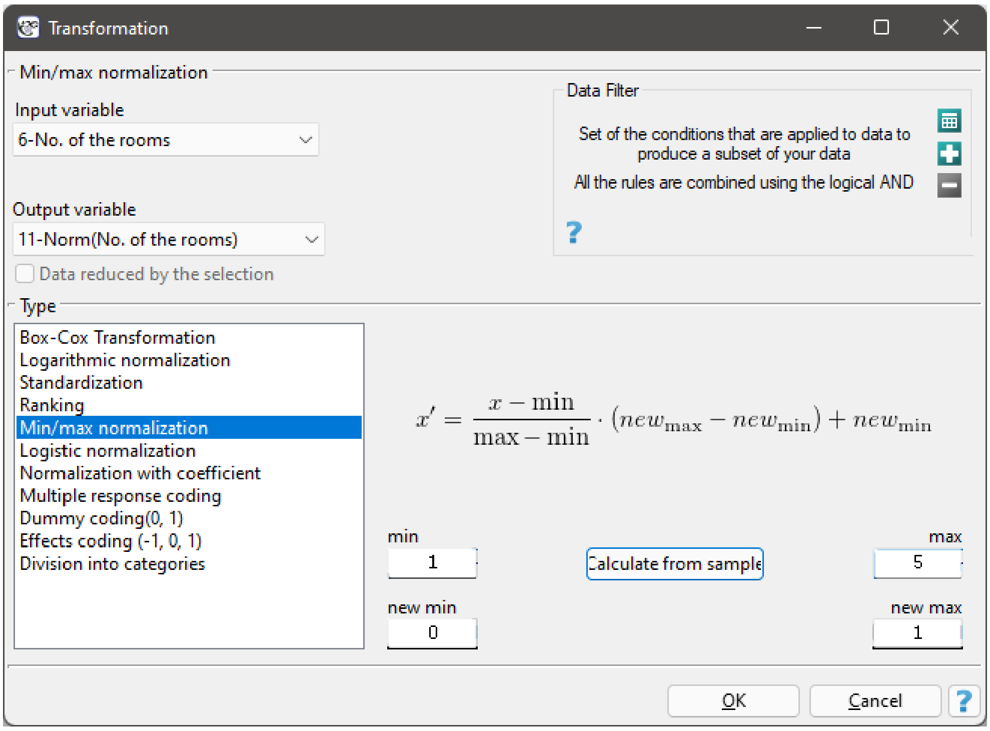
Now we will consider those features that have a strong influence on the customer's decision, but are not decisive:

- 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.

	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".



Input variable: 6-No. of the rooms

Output variable: 11-Norm(No. of the rooms)

□ Data reduced by the selection

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

min: 1

max: 5

new min: 0

new max: 1

The normalized data is shown in the table below.

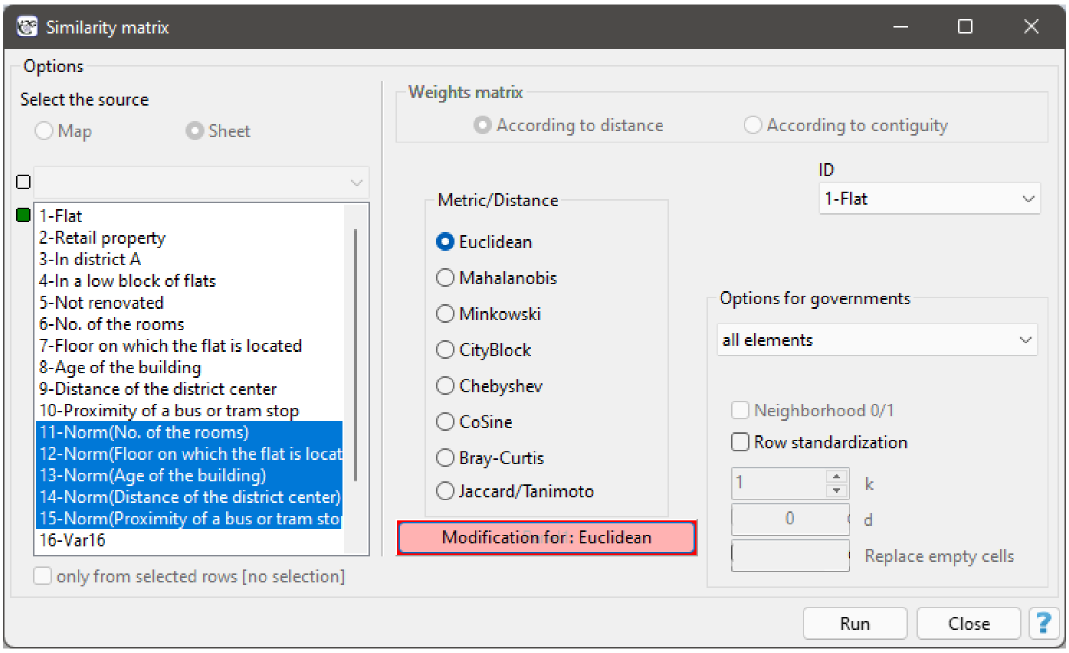
	Norm(Number of rooms)	Norm(Apartment floor)	Norm(Age of the building)	Norm(Proximity to the city center)	Norm(Proximity to the public transport station)
Poszukiwany	0,666666667	0	0,222222222	1	0,142857143
Lokal 10	0.333333333	0.2	0	0	0.285714286
Lokal 12	0	0.4	0	0	0.387755102
Lokal 17	0.666666667	0.2	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".



Similarity matrix

Options

Select the source
○ Map ● Sheet

1-Flat
2-Retail property
3-In district A
4-In a low block of flats
5-Not renovated
6-No. of the rooms
7-Floor on which the flat is located
8-Age of the building
9-Distance of the district center
10-Proximity of a bus or tram stop
11-Norm(No. of the rooms)
12-Norm(Floor on which the flat is locat
13-Norm(Age of the building)
14-Norm(Distance of the district center)
15-Norm(Proximity of a bus or tram sto
16-Var16

□ only from selected rows [no selection]

Weights matrix
● According to distance ○ According to contiguity

Metric/Distance
● Euclidean
○ Mahalanobis
○ Minkowski
○ CityBlock
○ Chebyshev
○ CoSine
○ Bray-Curtis
○ Jaccard/Tanimoto

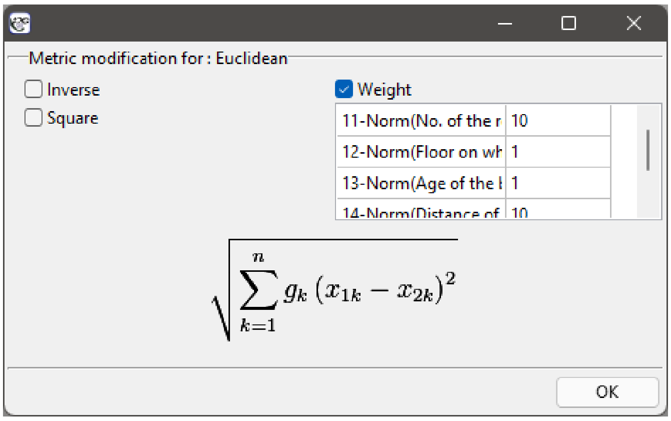
Modification for : Euclidean

ID: 1-Flat

Options for governments: all elements

□ Neighborhood 0/1
□ Row standardization
1 k
0 d
Replace empty cells

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.



Metric modification for : Euclidean

□ Inverse
□ Square
☑ Weight

11-Norm(No. of the r	10
12-Norm(Floor on wh	1
13-Norm(Age of the b	1
14-Norm(Distance of	10

$$\sqrt{\sum_{k=1}^{n} g_k\left(x_{1k}-x_{2k}\right)^2}$$

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	...

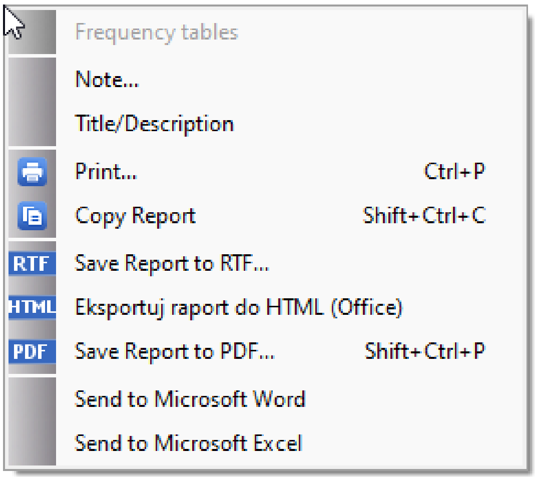
Euclidean with scales	**Wanted**	...
Wanted	0	...
Location 10	3.35	...
Location 12	3.84	...
Location 17	1.44	...
Location 35	2.86	...
Location 88	2.78	...
Location 101	3.45	...
Location 105	3.37	...
Location 122	3.39	...
Location 130	3.43	...
Location 132	1.24	...
Location 135	2.66	...

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.

4.2 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:
  – context menu Delete report (Shift+Del) on the report name in the Navigation tree,
  – Project Manager.

Note, however, that if map layers are attached to the report, deleting the report simultaneously 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) lub 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.

4.3 MENU AND LANGUAGE SETTINGS

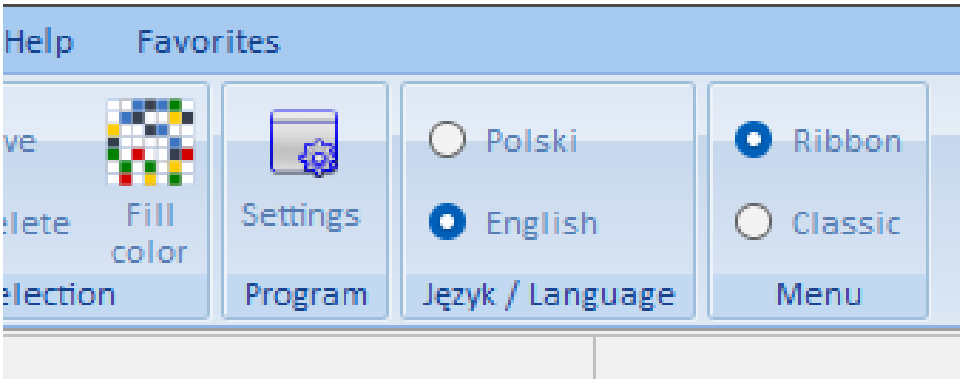
4.3.1 LANGUAGE

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).

4.3.2 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.



5 ORGANIZATION OF WORK WITH THE PROGRAM

All statistical analysis procedures are available in the Statistics menu.

5.1 FORM OF DATA ORGANIZATION

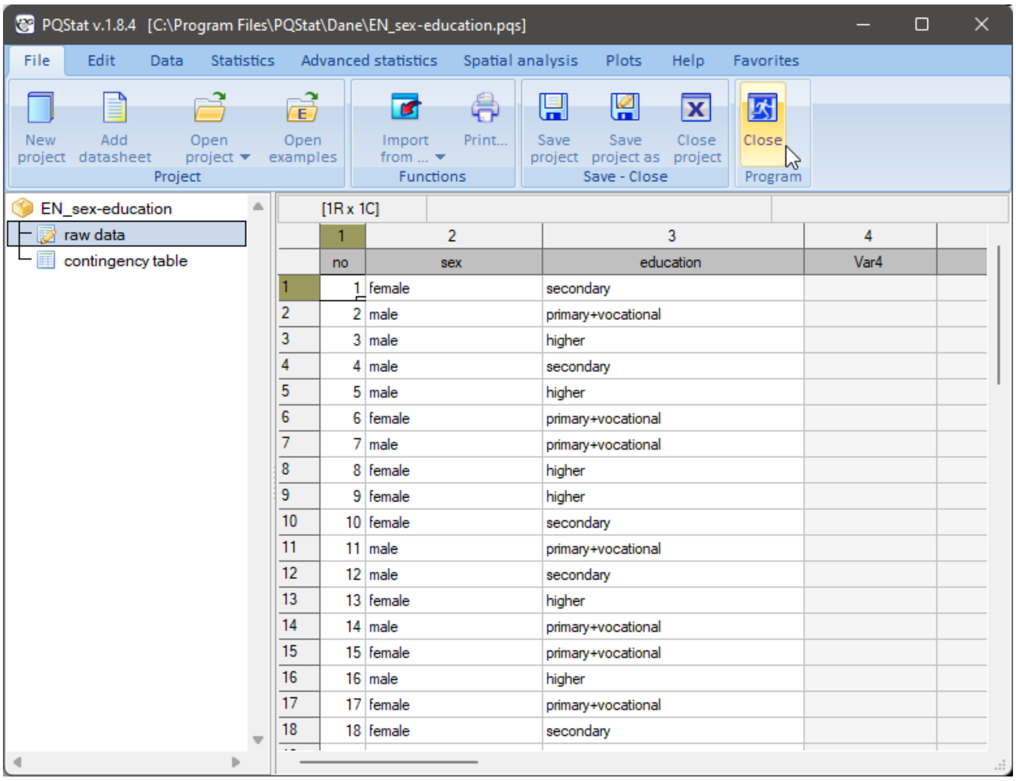
The form of data organization depends on the statistical procedures you plan to perform.

Statistical analysis of the data may involve data collected in the form of a contingency table or raw data. With this, it is possible to convert the data:

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

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

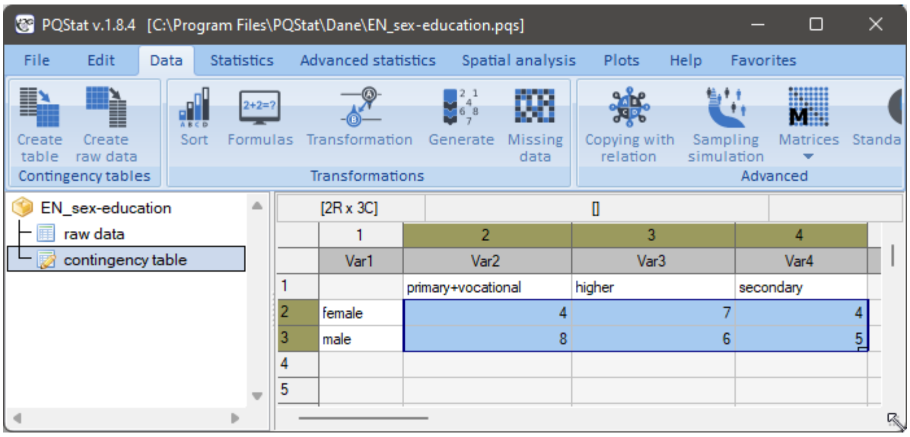
Example 5.1. *Raw data* (sex-education.pqs file)



	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. Contingency table, is a table showing the joint distribution of two variables. The inside of the table are the observed values (natural numbers).

Example 5.2. *Contingency table* (sex-education.pqs file)



	1	2	3	4
	Var1	Var2	Var3	Var4
1		primary+vocational	higher	secondary
2	female	4	7	4
3	male	8	6	5
4				
5				

5.2 DATASHEET WORKSPACE

Normally, when performing a statistical analysis, we have access to the entire workspace of the datasheet. However, the user is free to limit this area by selecting a specific part of the datasheet. Such selection can be done in several ways:

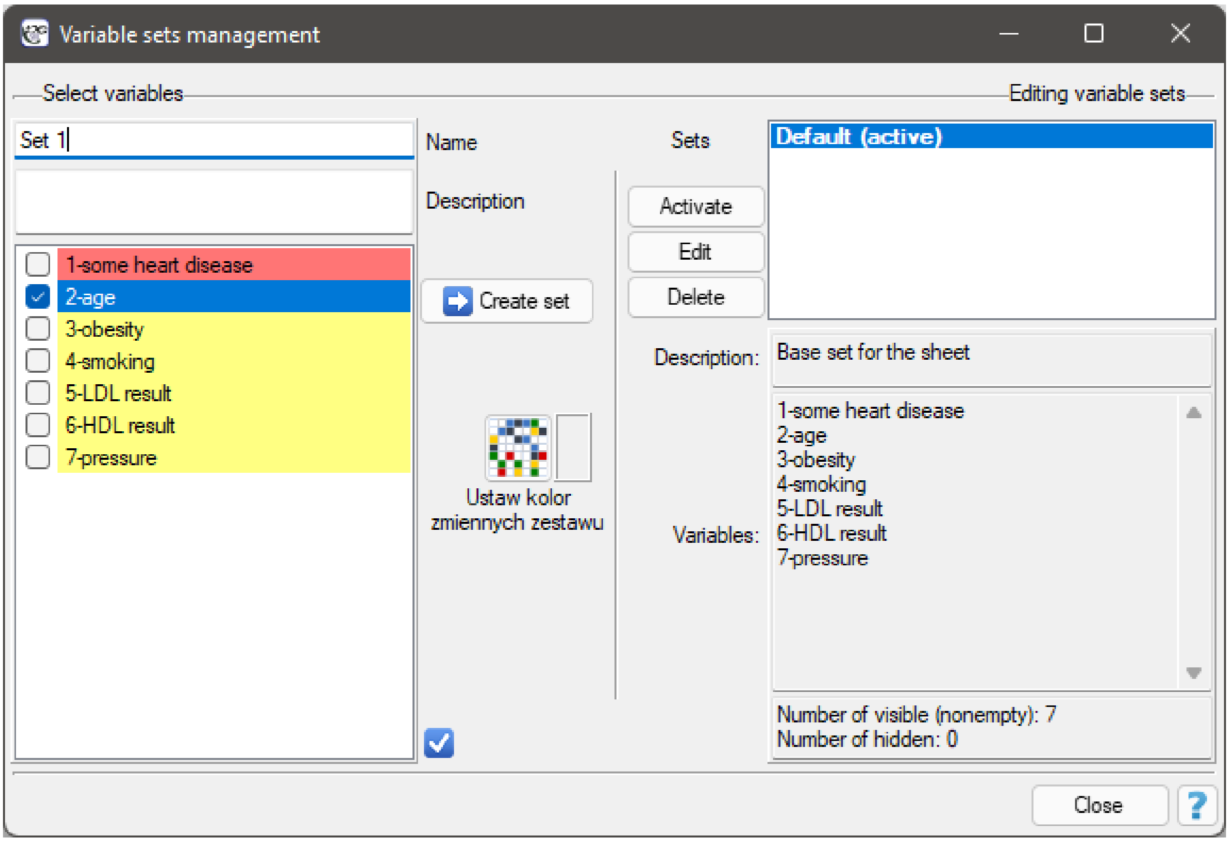
5.2.1 Organizing variables into sets

Variable sets are subsets of variables. Defining different sets is intended to simplify the selection of variables in analysis windows by limiting the displayed list of variables. The selected variable set is displayed in the variable lists in each analysis window and in the datasheet.

By default, the default set of variables is used. The default set is the set consisting of all variables.

Defining sets of variables

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



Variable sets management

Select variables

Name: Set 1

Description:

- ☐ 1-some heart disease
- ☑ 2-age
- ☐ 3-obesity
- ☐ 4-smoking
- ☐ 5-LDL result
- ☐ 6-HDL result
- ☐ 7-pressure

Editing variable sets

Sets: Default (active)

Description: Base set for the sheet

Variables:
1-some heart disease
2-age
3-obesity
4-smoking
5-LDL result
6-HDL result
7-pressure

Number of visible (nonempty): 7
Number of hidden: 0

In the Variable Set Management window, enter a name for the new set (possibly a description) and select the variables in the list that are to make up the set. Confirm your selection with the Create Set button. A variable can belong to multiple sets.

Editing Variable Sets

To edit a set, in the Manage Variable Sets window, I select a set from the list of sets and confirm with the Ok button. The variable selection window will show the selected variables in the set and the unselected variables not belonging to the set.Once the variables are selected/unselected again accordingly, confirm the selection with the Save Changes button.

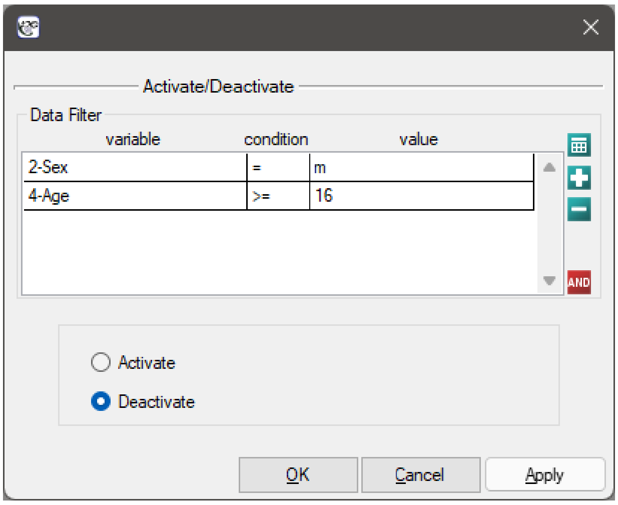
Activating a selected set of variables

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

5.2.2 Case activation/deactivation

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

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



variable	condition	value
2-Sex	=	m
4-Age	>=	16

○ Activate
● Deactivate

Example 5.3. (filter.pqs file)
We are going to do a lot of statistical analyses, on the data from the filtr.pqs file. These analyses are not to apply to boys who are 16 years old or older. To do this, we specify the rows that the analysis will not apply to: select the ⊞ button and set a rule for the gender variable; again, select the ⊞ button and set a rule for the age variable. Remember: to perform this task correctly, all filter conditions should be connected by conjunction (this is indicated by the AND). Leave the Disable option selected and confirm the analysis conditions set this way with OK

When narrowing the datasheet workspace, remember that filter rules can be connected by conjunction or alternative. To change the alternative and conjunction, use the AND OR buttons
More examples of how to use the filter can be found here.

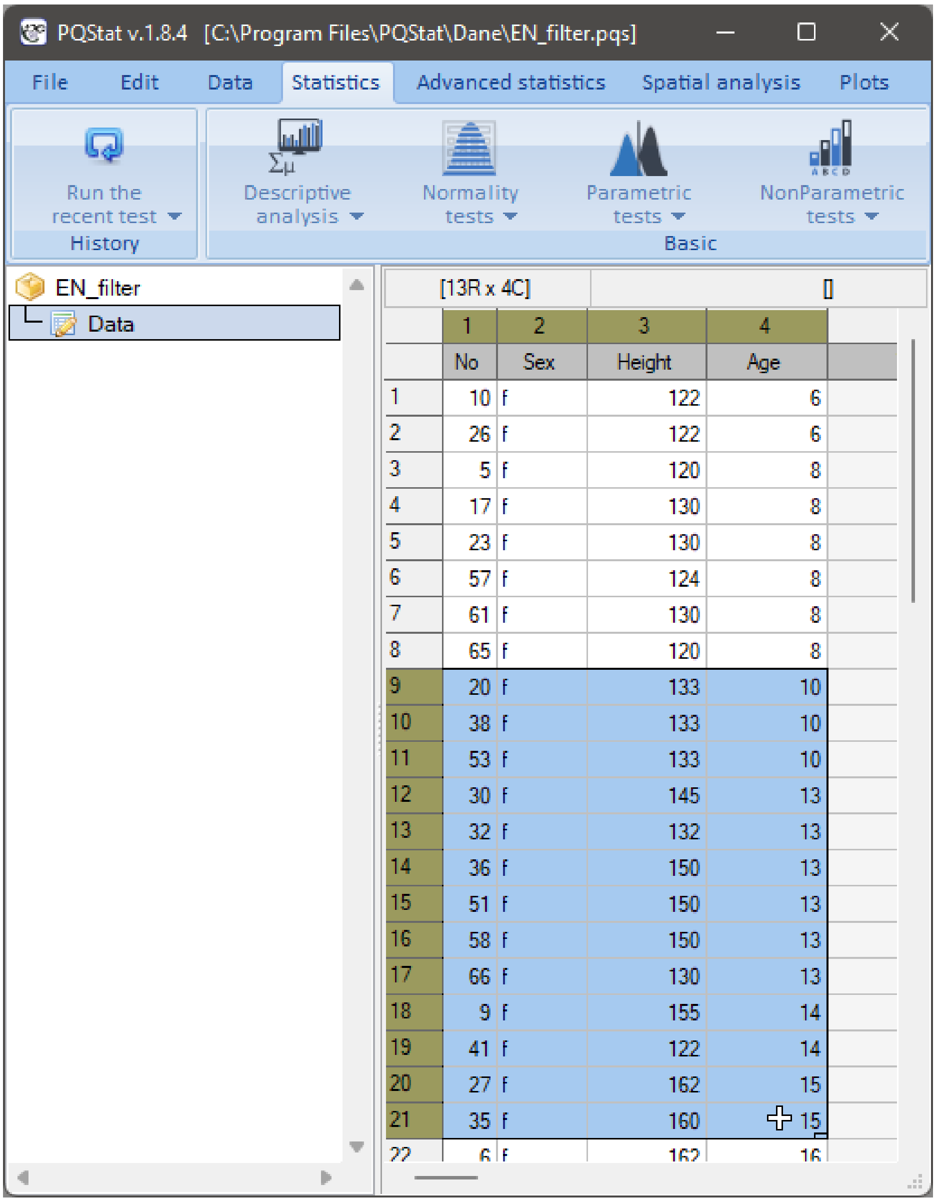
To activate all cases, select the menu Edit →Activate all

5.2.3 Selecting an area

Selecting a consistent area causes the analysis we select next to be performed only on those rows inside the selection and to have available only those columns inside the selection that contain data.

Example 5.4. (filter.pqs file)
We want to determine the descriptive statistics for the height of girls between the ages of 10 and 15.

So we sort the data by the gender column and by the age column, then select the consistent area containing the column with the height of the girls between the ages of 10 and 15, and choose Statistics→Descriptive analysis→Descriptive statistics from the menu.



	1	2	3	4
	No	Sex	Height	Age
1	10	f	122	6
2	26	f	122	6
3	5	f	120	8
4	17	f	130	8
5	23	f	130	8
6	57	f	124	8
7	61	f	130	8
8	65	f	120	8
9	20	f	133	10
10	38	f	133	10
11	53	f	133	10
12	30	f	145	13
13	32	f	132	13
14	36	f	150	13
15	51	f	150	13
16	58	f	150	13
17	66	f	130	13
18	9	f	155	14
19	41	f	122	14
20	27	f	162	15
21	35	f	160	15
22	6	f	162	16

In the Descriptive Statistics Test Options window, select the procedures you want to perform (e.g., select the mean, standard deviation, minimum, and maximum) and the variable to analyse (the height column), and confirm your selection with the OK button.

Narrowing the datasheet workspace by selecting a consistent piece of data causes a message to appear in the analysis window:

Data reduced by selection

5.2.4 Saved Selection

If selected ranges are assigned to the datasheet, they are highlighted by a border. These can be used for those analyses that have the option to enter data directly in the analysis window. Then, using the fill from saved selection button, you can paste the data that is in the selected range.

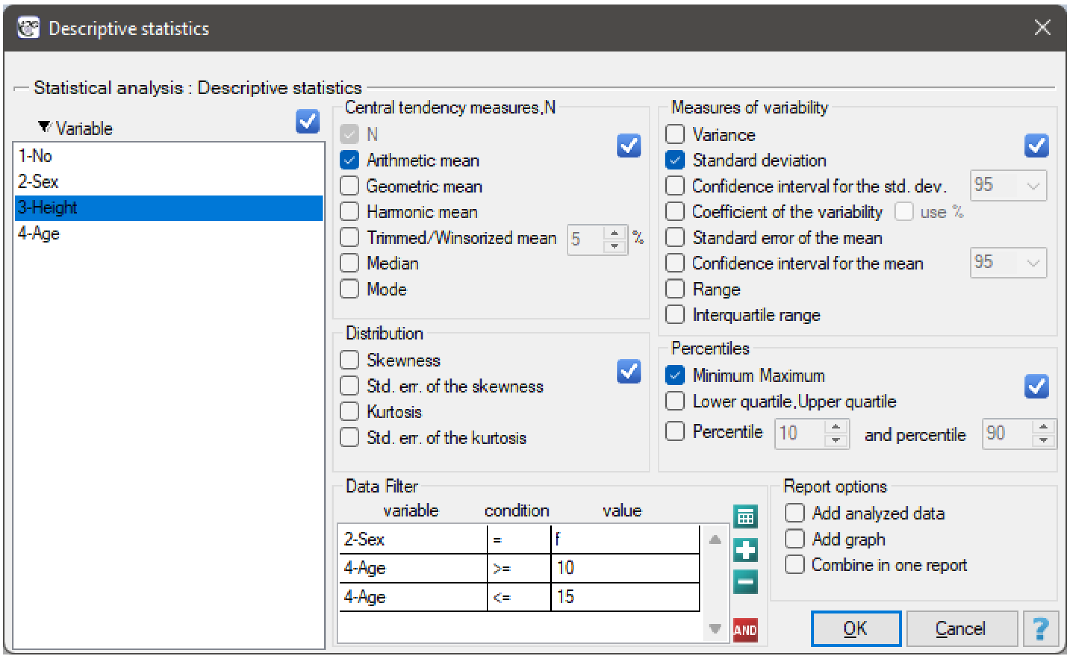
Example 5.5. (layers.pqs file)
We want to determine statistics related to Odds Ratio (OR) analysis for several strata. We will use the data stored as 10 tables – they are highlighted (framed). From the Advanced statistics menu, select Stratified analysis → Mantel-Haenszel OR/RR. In the test options window, we select Contingency table and then set the number of stratas to 10. Each resulting layer can now be filled from the saved selection. When we have filled in all the tables we perform the analysis by selecting the OK button.
Note
To assign another selection to the datasheet select Save Selection (Ctrl+T) from the Edit menu, and to remove assigned selections select Delete selections.

5.2.5 Data filter

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

- **Basic filter** allows you to select one particular subgroup. The selection can be made in two ways:
  - **Automatic method** – this way we can apply one condition to the collected data using the ⊞ button, or multiple conditions by pressing ⊞ again. The conditions will be connected by an alternative or conjunction depending on whether we choose the AND or OR button. Conditions are removed using the ⊟ button.

    Example 5.6. *Automatic basic filter* (filter.pqs file)
    We want to determine the descriptive statistics for the height of girls between the ages of 10 and 15. From the menu, we select Statisticsto Descriptive analysisto Descriptive statistics. In the descriptive statistics test options window, we select the procedures we want to perform (e.g., select the mean, standard deviation, minimum and maximum) and the variable to be analysed (the column with height). We set the filter by adding rules with the ⊞ button. First we set the rule for variable gender , as condition we choose the equals sign and as value we choose the letter g meaning girls. Next, we add another condition and set a filter for the age variable, we choose $>=$ as the condition and enter 10 as the value. In a similar way, we add the age condition $<$=15. Remember: to perform this task correctly, all filter conditions should be connected by conjunction (this is indicated by the AND). Confirm the analysis conditions set in this way with the OK button.



Descriptive statistics

Statistical analysis : Descriptive statistics

Variable ☑
1-No
2-Sex
3-Height
4-Age

Central tendency measures,N ☑
☑ N
☑ Arithmetic mean
□ Geometric mean
□ Harmonic mean
□ Trimmed/Winsorized mean 5 %
□ Median
□ Mode

Measures of variability ☑
□ Variance
☑ Standard deviation
□ Confidence interval for the std. dev. 95
□ Coefficient of the variability □ use %
□ Standard error of the mean
□ Confidence interval for the mean 95
□ Range
□ Interquartile range

Distribution ☑
□ Skewness
□ Std. err. of the skewness
□ Kurtosis
□ Std. err. of the kurtosis

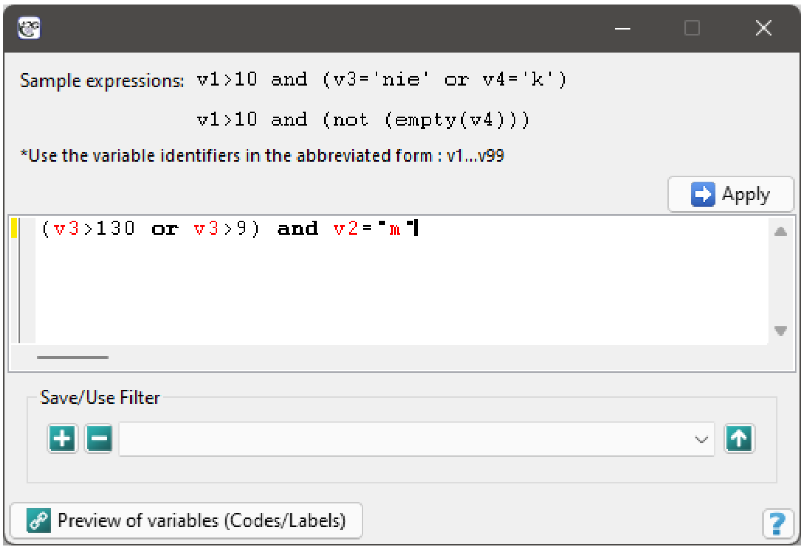
Percentiles ☑
☑ Minimum Maximum
□ Lower quartile,Upper quartile
□ Percentile 10 and percentile 90

Data Filter

variable	condition	value
2-Sex	=	f
4-Age	>=	10
4-Age	<=	15

Report options
□ Add analyzed data
□ Add graph
□ Combine in one report

– **Advanced method** – this way we can write any rule, i.e. both simple and more complex. We start the advanced way by switching the ▦ button to ✎ and then selecting ✚. This brings up a window inside which we can build the filter formula.



Sample expressions: `v1>10 and (v3='nie' or v4='k')`

`v1>10 and (not (empty(v4)))`

*Use the variable identifiers in the abbreviated form : v1...v99

```
(v3>130 or v3>9) and v2="m"
```

When setting up an advanced filter, we use variable numbers preceded by *v*. You can view the variable numbers and their contents using the Preview of varialbes (Codes/Labels) button.

We can use the following logic functions:
and - conjunction operator - checks whether all the conditions it combines are true,
or - alternative operator - checks if at least one of the given conditions is true,
xor - disjunctive alternative operator - checks if exactly one of the given conditions is true,
not - negation operator,
empty(v1) – function checking if there are empty cells,
text(v1) – function checking if there is text in the cells,
numer(v1) – function checking if there is a number in the cells.

The finished filter can be used by selecting the Apply button, or saved for later by selecting the ✚ button. Saved filters can be used by uploading them to the formula window via the ⬆ button.

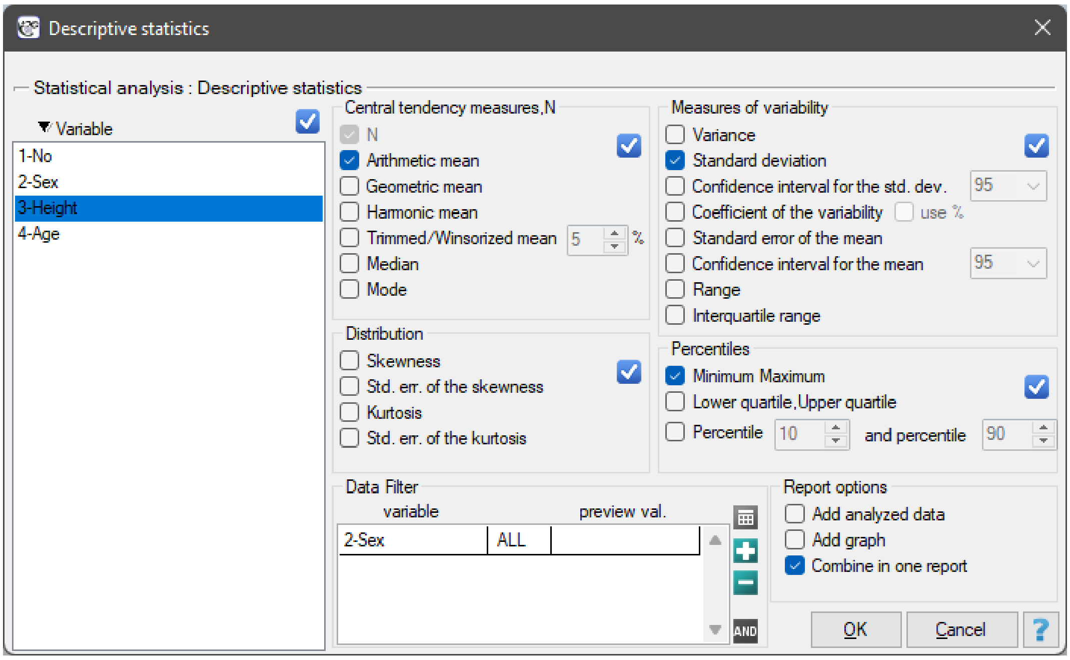
Example 5.7. *Basic advanced filter* (filter.pqs file)

We want to determine the descriptive statistics for the height of boys who are taller than 130 cm or older than 9 years old. From the menu, we select Statistics to Descriptive analysis to Descriptive statistics. In the Descriptive Statistics test options window, select the procedures you want to perform (e.g., select the mean, standard deviation, minimum, and maximum) and the variable to analyse (the height column). Set the filter by entering the formula (v3>130 or v4>9) and v2='m'. Then select Apply and perform the analysis.

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

Example 5.8. *Multiple filter* (filter.pqs file)

We want to determine the descriptive statistics for girls' height and separately for boys' height. From the menu, we select Statistics to Descriptive analysis to Descriptive statistics. In the Descriptive Statistics test options window, select the procedures you want to perform (e.g., select the mean, standard deviation, minimum, and maximum) and the variable to analyse (the height column). We set up the multiple filter by adding a rule with the ✚ button and selecting the Gender variable. Confirm the selected analysis options with the OK button. This will result in 2 reports: a separate one for boys and a separate one for girls.



Descriptive statistics

Statistical analysis : Descriptive statistics

Variable: 1-No, 2-Sex, **3-Height** (selected), 4-Age

Central tendency measures,N
- ☑ N
- ☑ Arithmetic mean
- □ Geometric mean
- □ Harmonic mean
- □ Trimmed/Winsorized mean 5 %
- □ Median
- □ Mode

Measures of variability
- □ Variance
- ☑ Standard deviation
- □ Confidence interval for the std. dev. 95
- □ Coefficient of the variability □ use %
- □ Standard error of the mean
- □ Confidence interval for the mean 95
- □ Range
- □ Interquartile range

Distribution
- □ Skewness
- □ Std. err. of the skewness
- □ Kurtosis
- □ Std. err. of the kurtosis

Percentiles
- ☑ Minimum Maximum
- □ Lower quartile,Upper quartile
- □ Percentile 10 and percentile 90

Data Filter

variable	preview val.
2-Sex	ALL

Report options
- □ Add analyzed data
- □ Add graph
- ☑ Combine in one report

.

5.3 MULTIPLE ANALYSES

To streamline performing the same analysis multiple times, you can:

1. Use the option to memorize the current and previous analysis. PQStat automatically remembers the last analysis performed and the options set in its window. The ability to quickly return to it is provided by the Run the recent test button (ribbon menu ->↺) located on the toolbar.

2. Use the option to memorize the analysis based on the report. PQStat automatically saves the options set in the analysis window together with the analysis report. A button 🕓 located in the context menu of the right mouse button on a report in Navigation tree allows quick return to it.

3. Select multiple variables in the analysis window so that the analysis is run multiple times. Results for each analysis performed will be returned in subsequent reports.

4. Use the multiple filter to run the analysis separately for each subset of the data. Results for each analysis performed will be returned in subsequent reports.

5.4 ORGANIZING REPORTS INTO SETS

When you have a large number of reports, it is convenient to group them into sets. Prepared sets make it easier to navigate through the results and speed up the transfer of results to other programs such as Excel or Word.

Creating sets

Through the context menu Add empty in the Navigation Tree on a datasheet or existing set.

Placing Reports into Sets

Reports can be placed in a set

- at the time they are created - from the analysis window - to do this, in the analysis window, select the ▣ button and choose the set to which the analysis report is to be created, or before the analysis itself, select this set in the Navigation tree;

- after the analysis has been performed - via the drag-and-drop option or the context menu on the name of the report(s) in the Navigation Tree, or via the Project Manager.

Editing sets

Via the context menu in the Navigation Tree on the datasheet or an existing set Edit Package..

Deleting sets

Via the context menu in the Navigation Tree on the set name Delete package (Shift+Del).

5.5 INFORMATION RETURNED TO THE REPORT

In addition to the basic settings for the statistical analysis performed in the test window, you have the option to:

- Include the analysed data in the report.
  The data analysed depending on the test being performed can be returned to the report:
  - in the form of raw records,
  - in the form of contingency tables.

- Include a graph in the report.
  To make sure the appropriate chart is included in the report, in the window of the selected statistical analysis, select Add graph.

- Combine results from multiple analyses into a single report.
  By selecting multiple variables or by applying multiple data filters, multiple analyses of the same type can be performed simultaneously. To make viewing the results easier, all analyses performed in this way can be returned to a single report. To do this, in the analysis window, select Combine in one report.

5.6 IDENTIFICATION OF STATISTICALLY SIGNIFICANT RESULTS

In the report, red is used to denote the p-value of the executed statistical test that is lower than the user-set significance level. 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 program is closed) in the selected test window.

0.05 significance level

6 PLOTS

The PQStat program offers column plots, error plots, box-whisker plots, point plots, and line-point plots.

The window with the plots settings is opened via the plots menu.

Changing the basic plot parameters is possible directly in the plot window. If, however:

- you want to change general plot parameters such as titles, backgrounds, axes, grid lines and the legend – select the Plot General Options tab.;
- you want to change the appearance of the drawn object itself, e.g. shape, style, colors – select the Plot Detailed Options tab.;
- you want to draw additional elements on the plot, e.g., a line – select the Other tab..

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

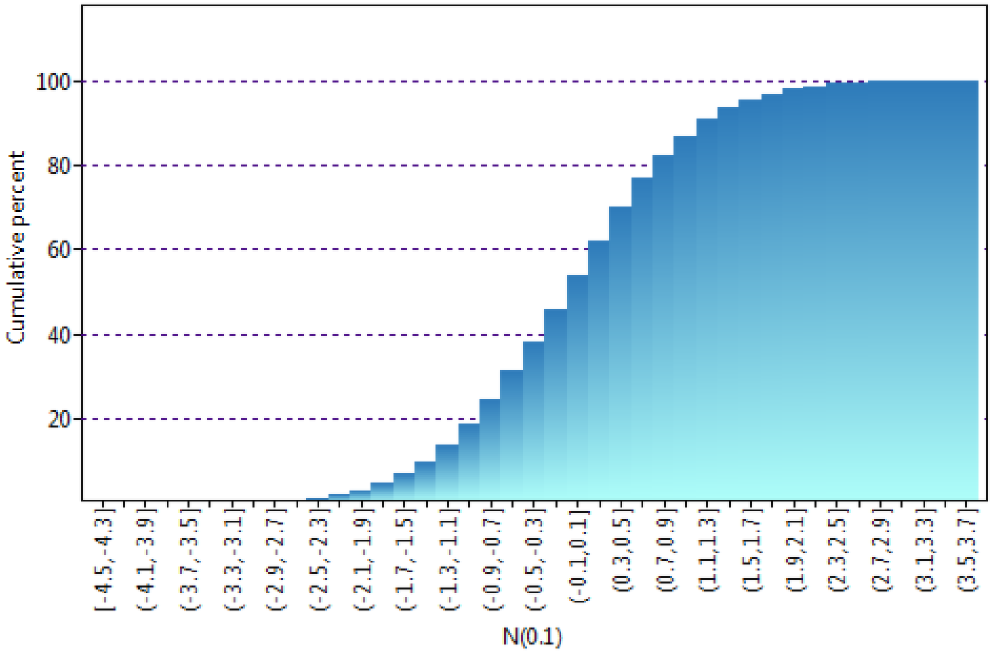
The plot is returned to the report, where it can be:

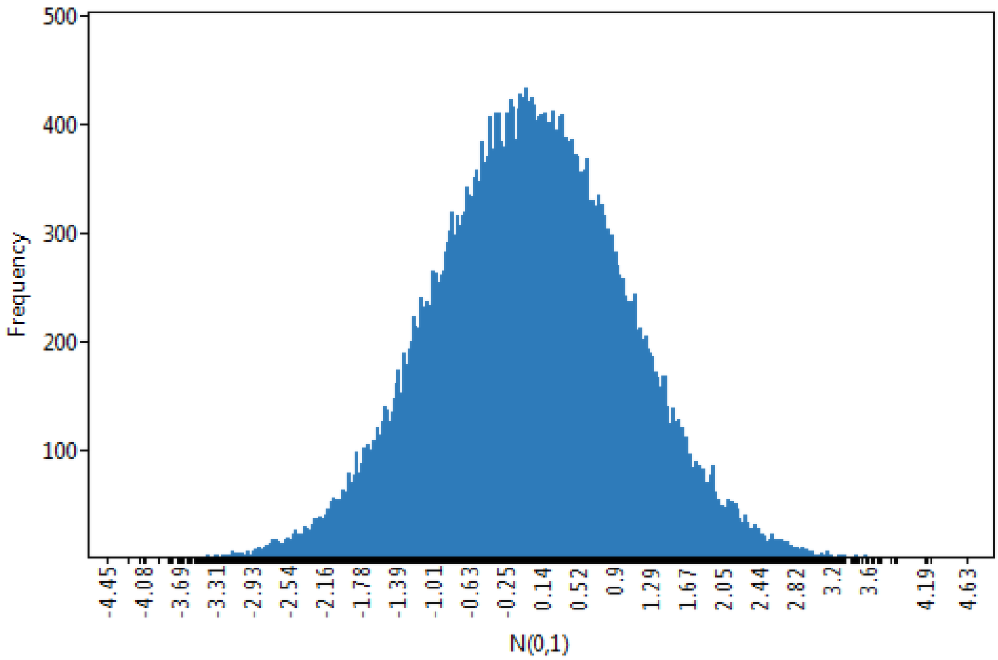
- saved – Save the plot as... in the context menu;
- printed – Print plot in the context menu;
- copied – Copy plot in the context menu;
- edited – this applies to Plot General Options i Plot Detailed Options. To edit a chart, simply double-click on the plot, or select Edit Plot from the context menu. In the plot editing window it is also possible to save the plot in high resolution.

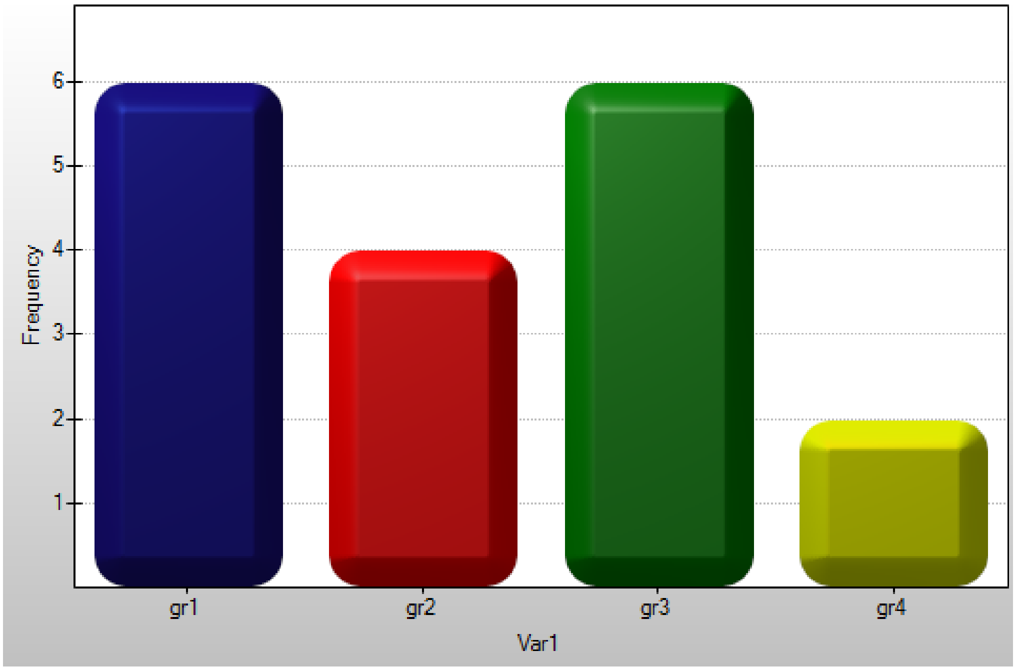
6.1 PLOT GALLERY

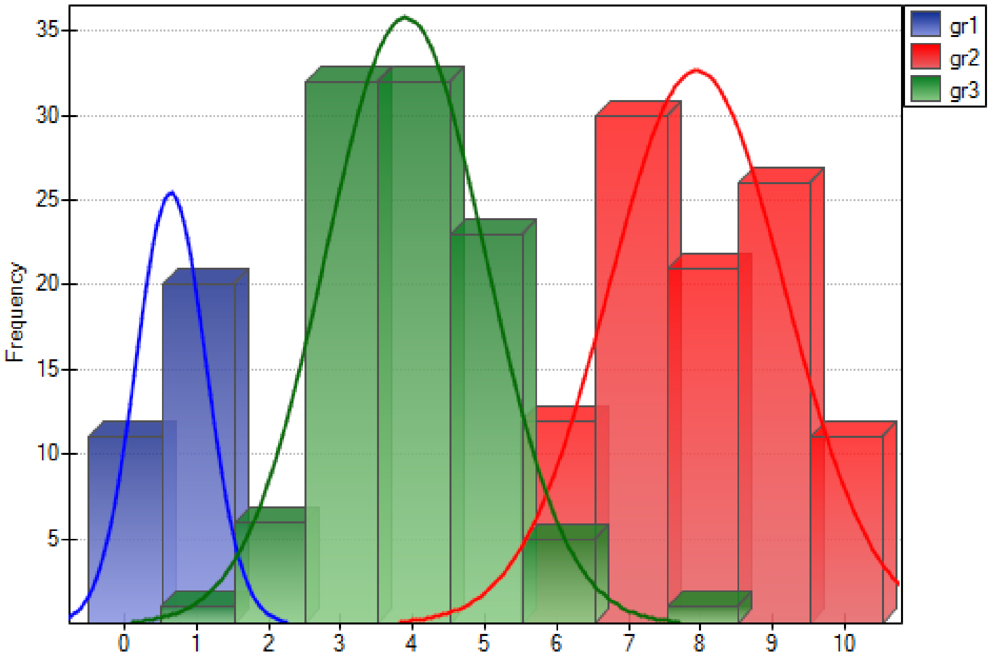
Depending on the type of analysis we can choose from:

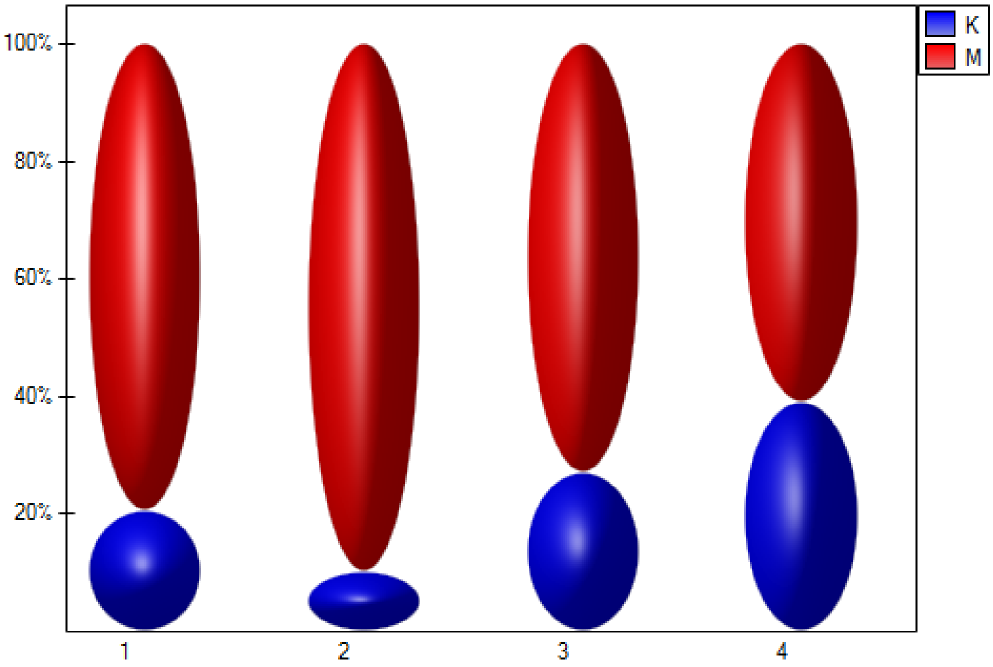
6.1.1 Column plots

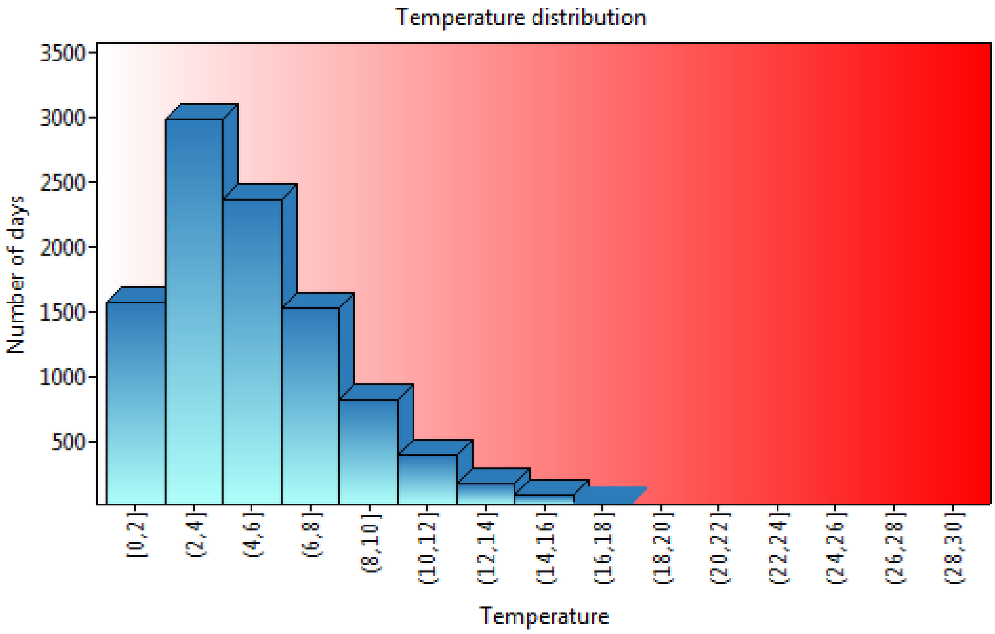


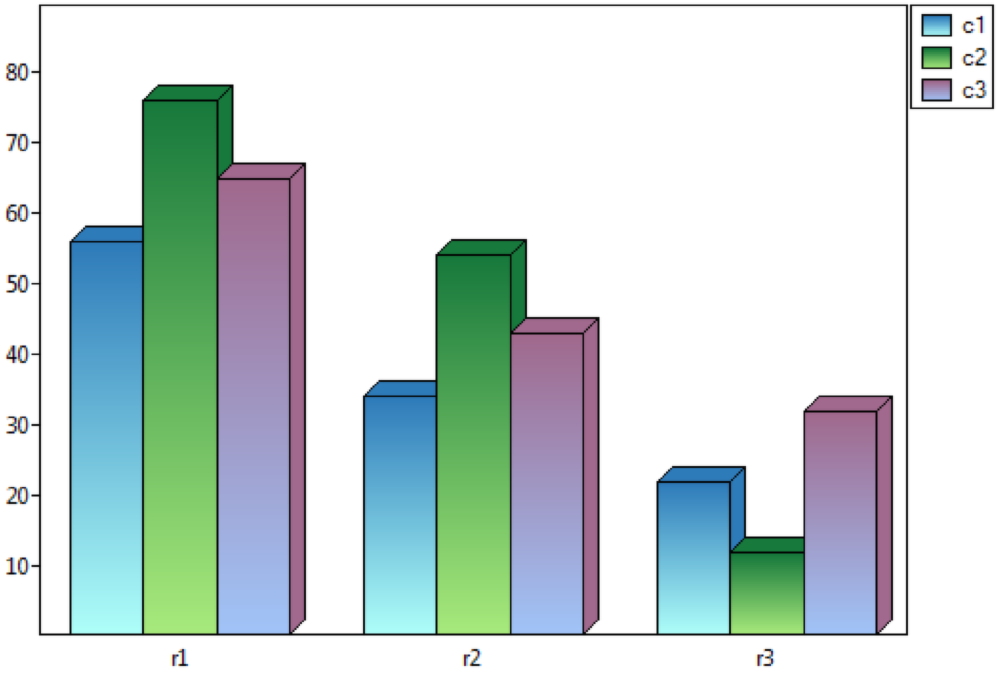

500
400
300
200
100
Frequency
-4.45
-4.08
-3.69
-3.31
-2.93
-2.54
-2.16
-1.78
-1.39
-1.01
-0.63
-0.25
0.14
0.52
0.9
1.29
1.67
2.05
2.44
2.82
3.2
3.6
4.19
4.63
N(0,1)

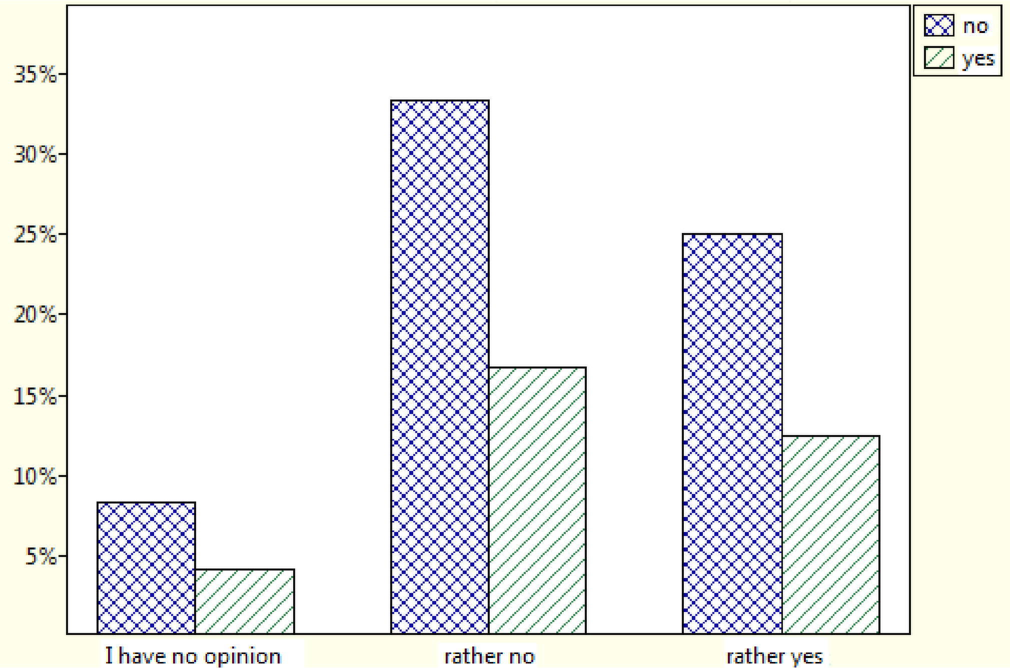

6
5
4
3
2
1
Frequency
gr1
gr2
gr3
gr4
Var1

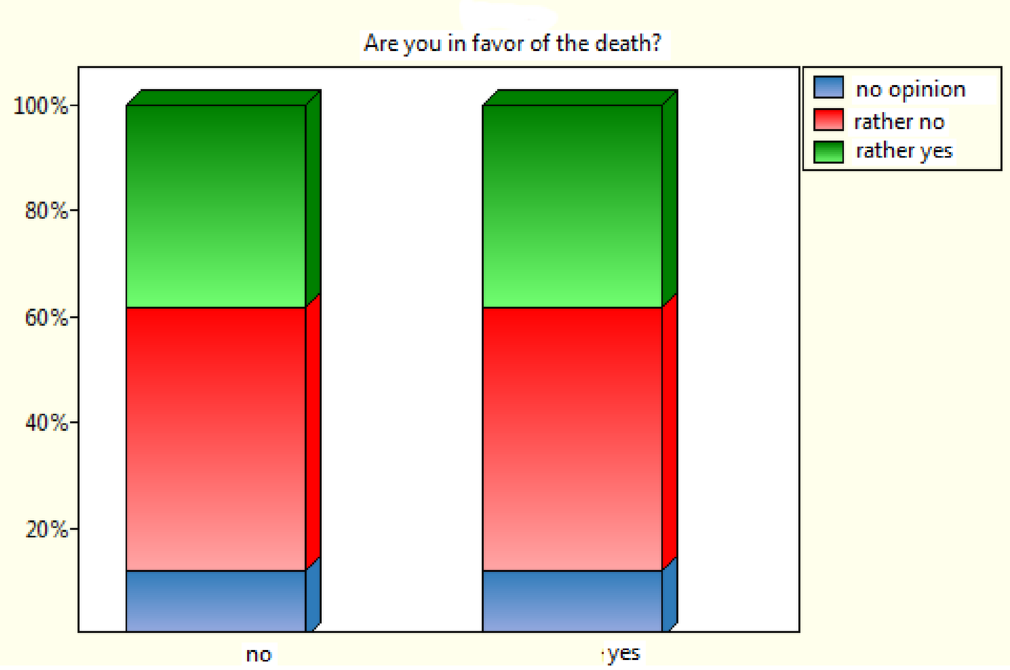

Frequency
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
gr1
gr2
gr3

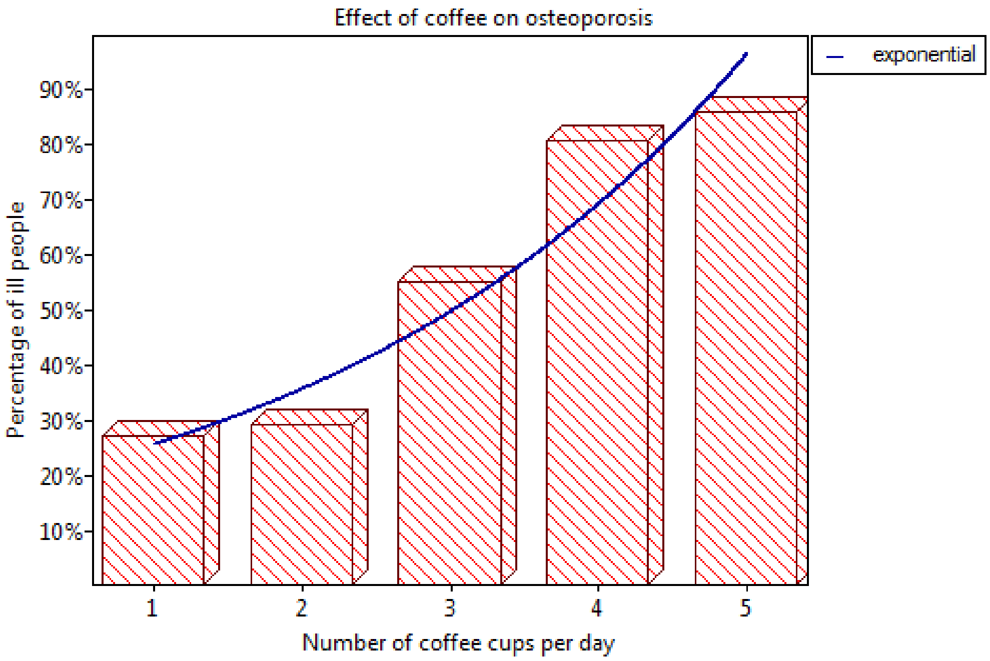

100%
80%
60%
40%
20%
1
2
3
4
K
M


Temperature distribution
Number of days
3500
3000
2500
2000
1500
1000
500
[0,2]
(2,4]
(4,6]
(6,8]
(8,10]
(10,12]
(12,14]
(14,16]
(16,18]
(18,20]
(20,22]
(22,24]
(24,26]
(26,28]
(28,30]
Temperature

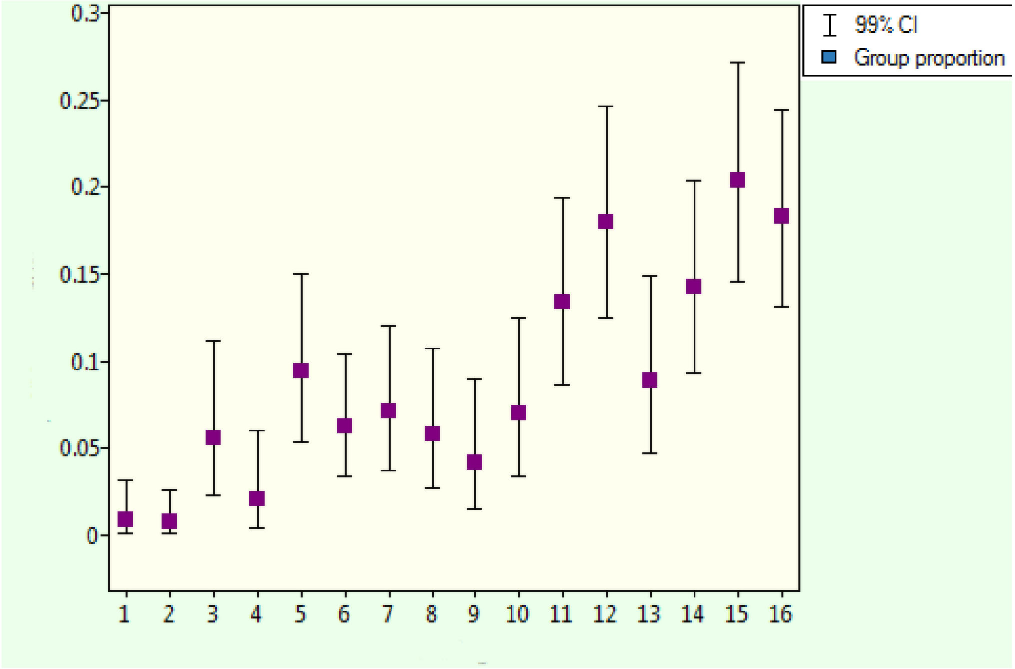

c1
c2
c3
80
70
60
50
40
30
20
10
r1
r2
r3

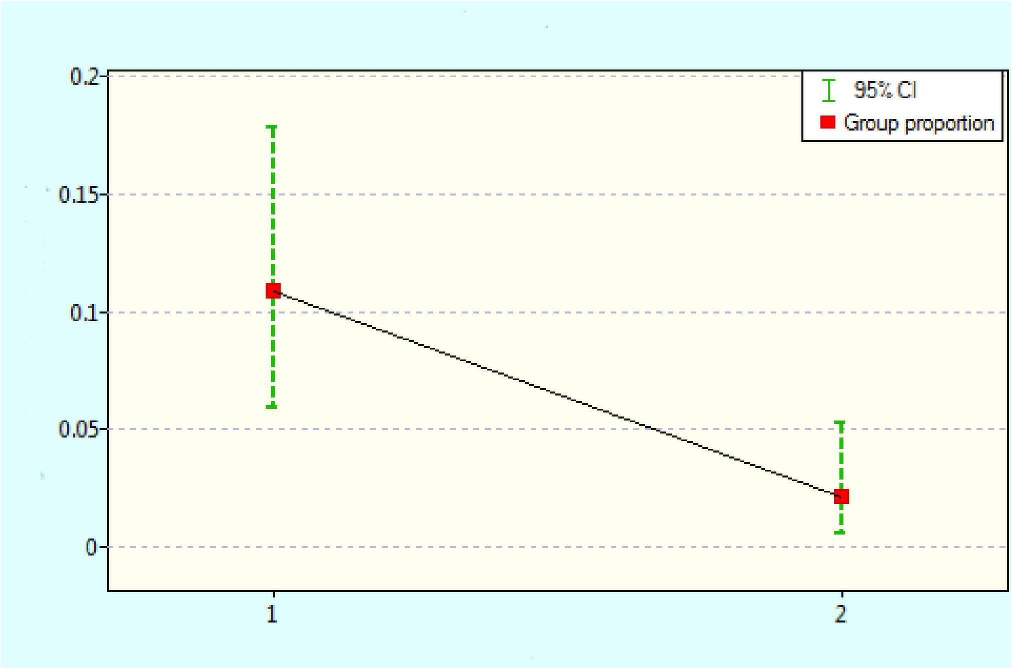

35%
30%
25%
20%
15%
10%
5%
I have no opinion
rather no
rather yes
no
yes

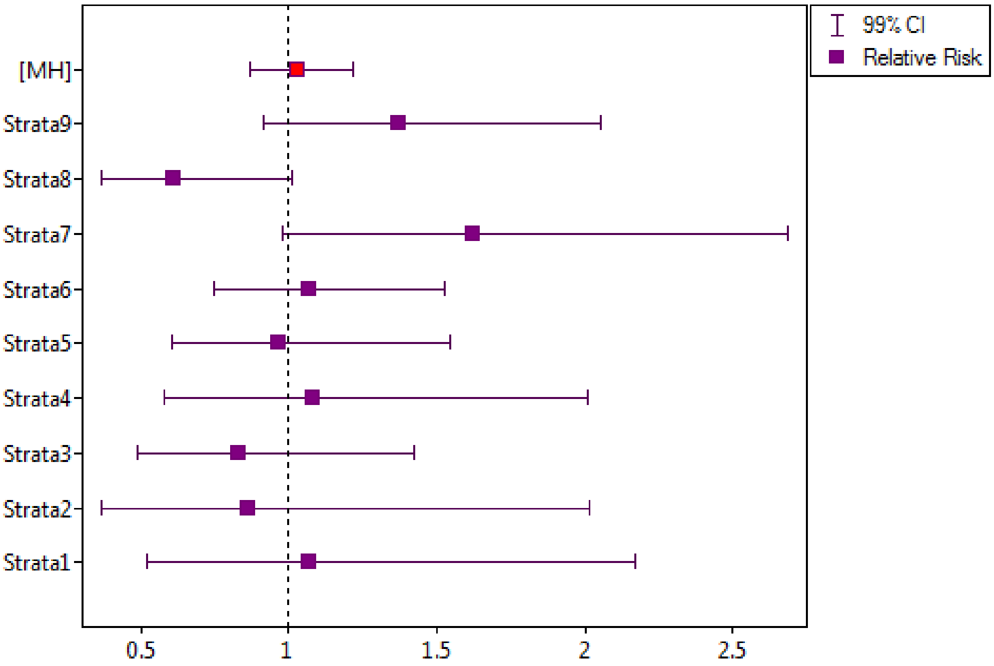

Are you in favor of the death?
100%
80%
60%
40%
20%
no
yes
no opinion
rather no
rather yes



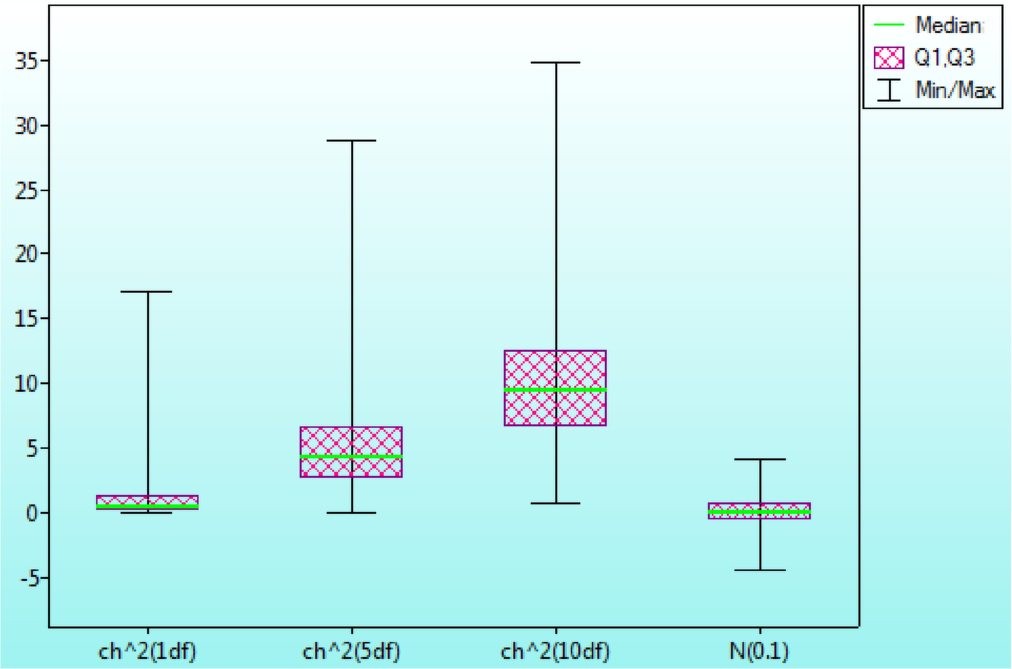
6.1.2 Estimator error plots

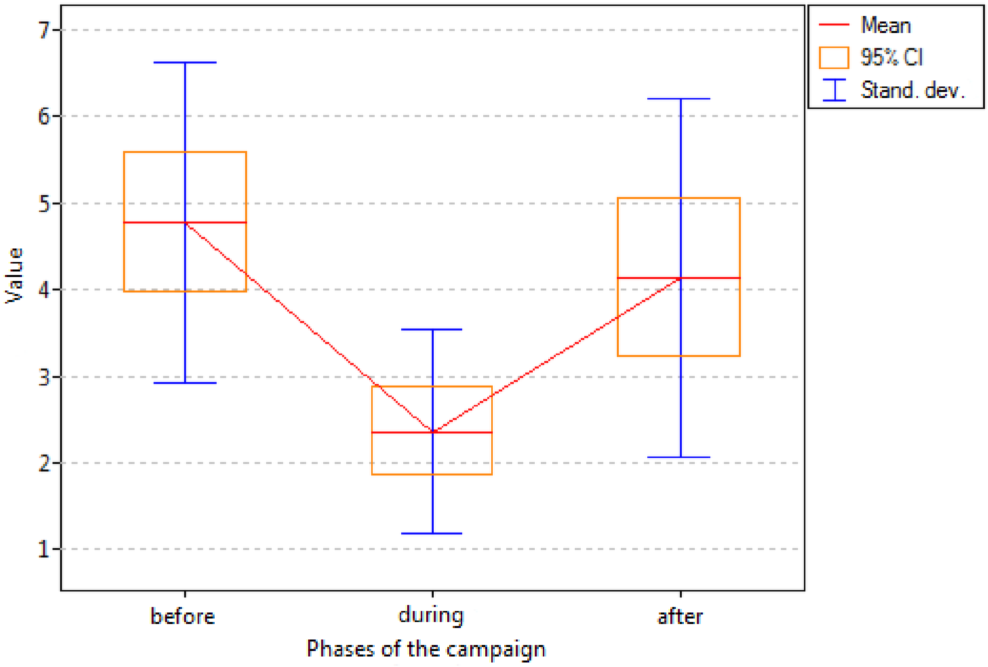


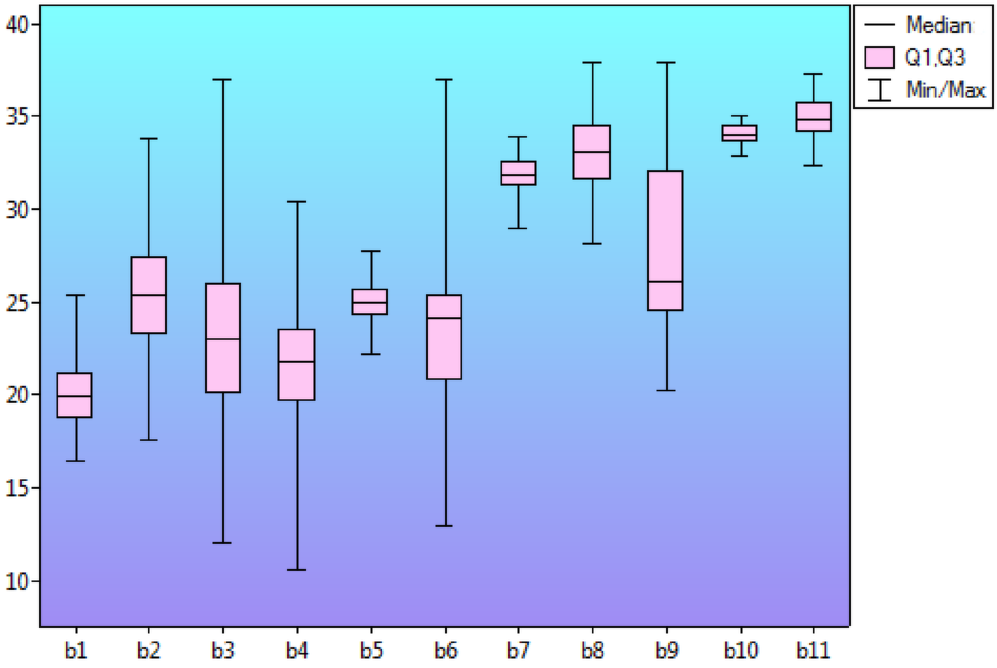

0.2
0.15
0.1
0.05
0
1
2
95% CI
Group proportion


[MH]
Strata9
Strata8
Strata7
Strata6
Strata5
Strata4
Strata3
Strata2
Strata1
0.5
1
1.5
2
2.5
99% CI
Relative Risk

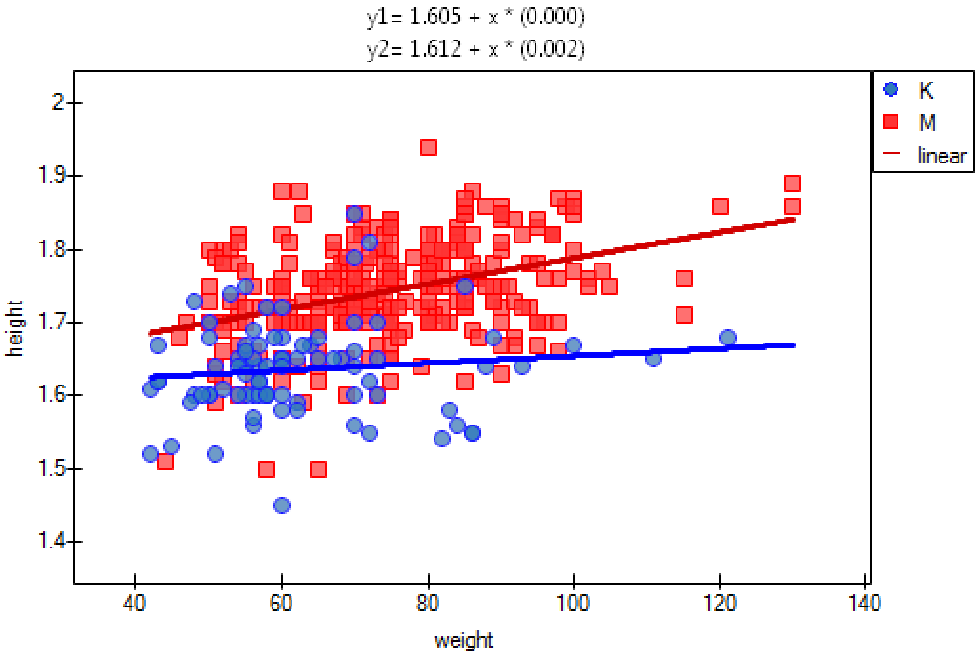
6.1.3 Box-whiskers plots

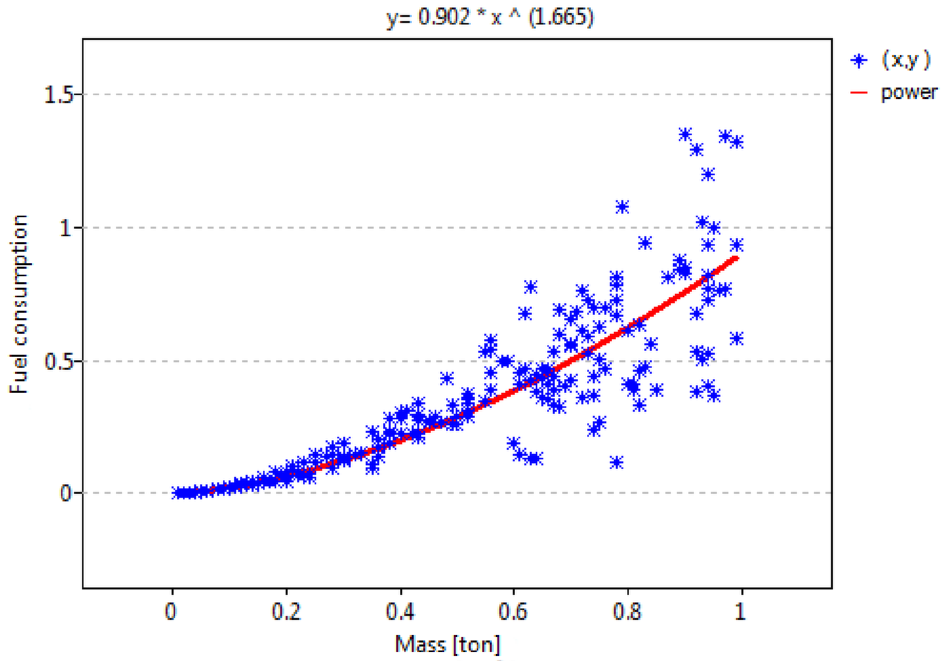


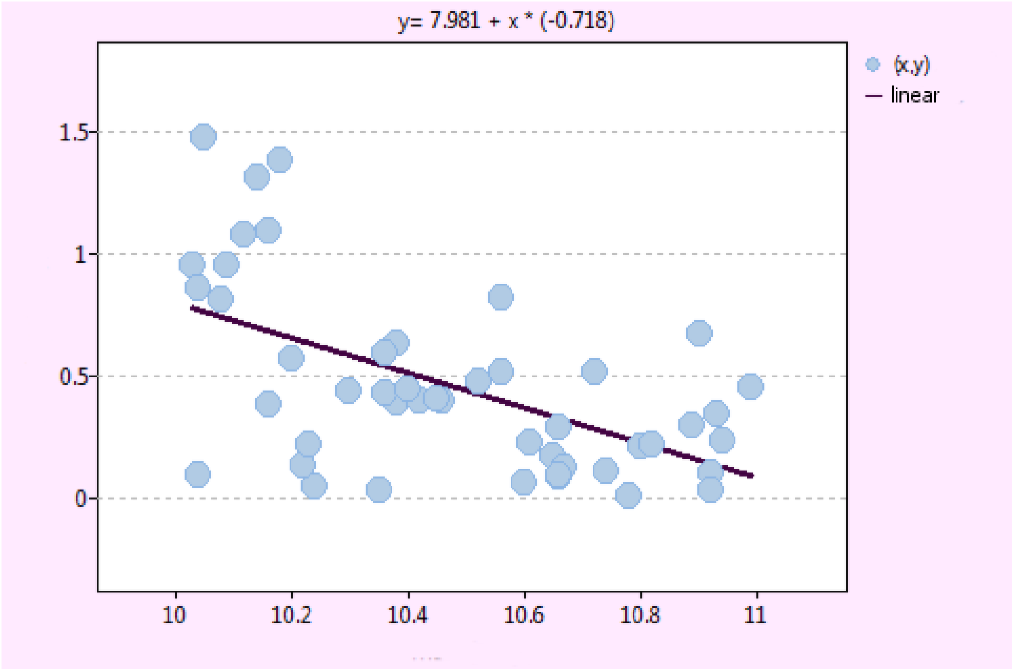


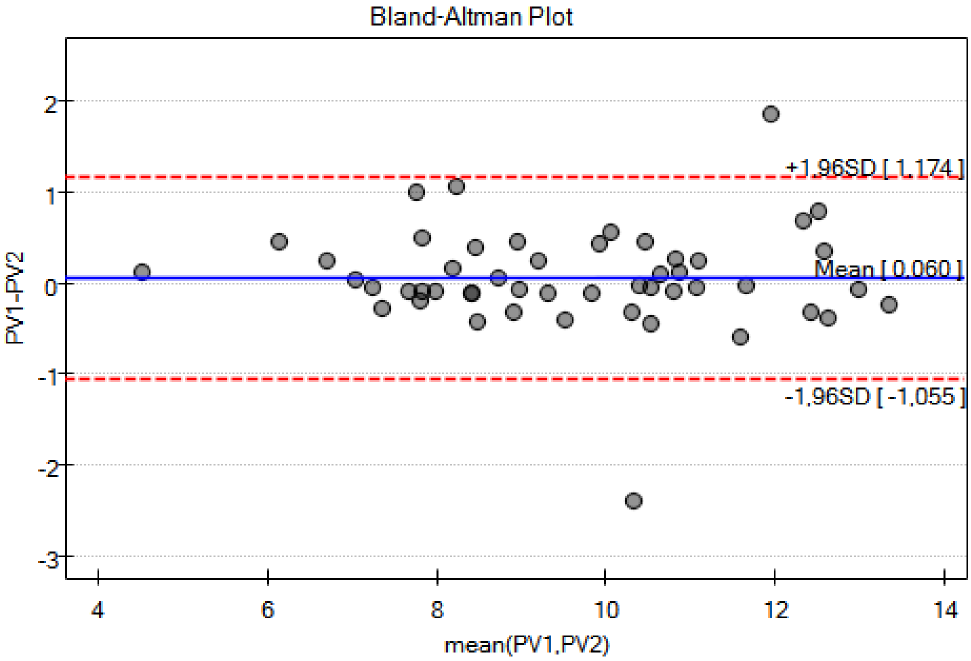


6.1.4 Scatter plots

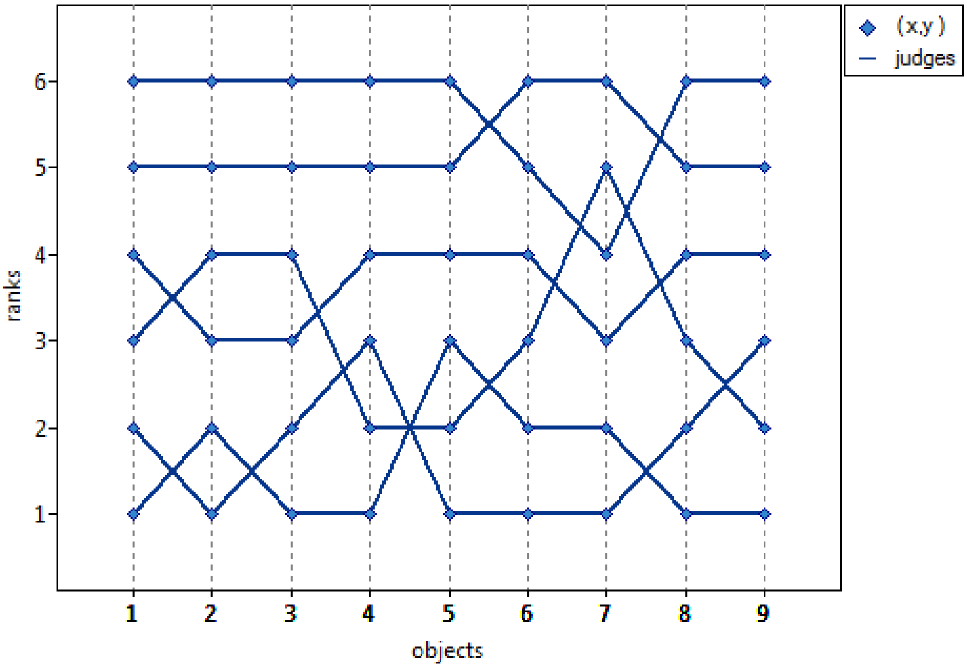


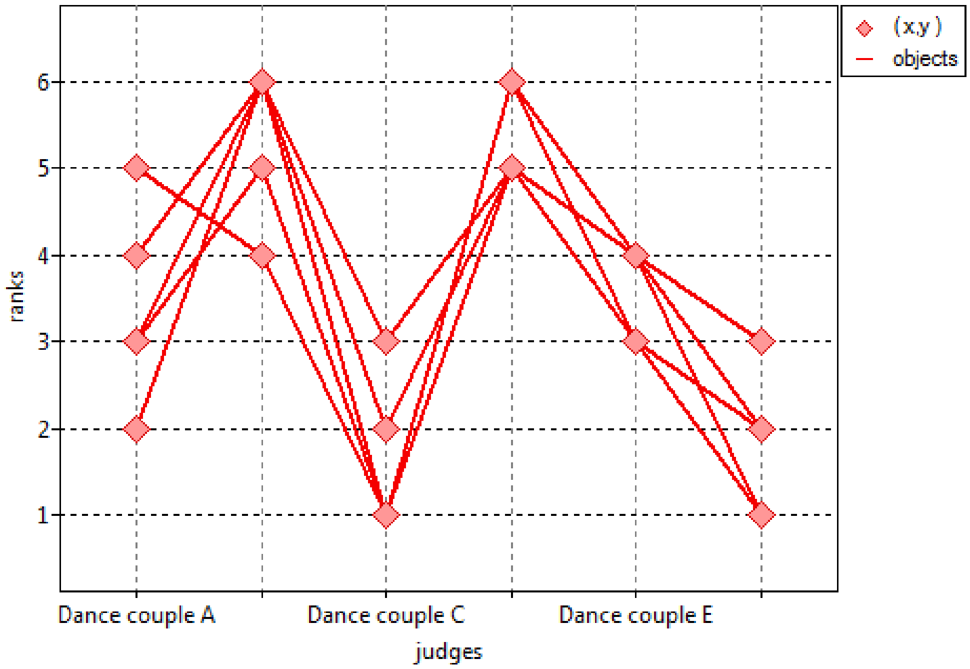

y= 0.902 * x ^ (1.665)
(x,y)
power
Fuel consumption
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1
Mass [ton]


y= 7.981 + x * (-0.718)
(x,y)
linear
1.5
1
0.5
0
10
10.2
10.4
10.6
10.8
11



6.1.5 Line plots




6
5
4
3
2
1
ranks
Dance couple A
Dance couple C
Dance couple E
judges
(x,y)
objects

6.2 LOCAL LINEAR SMOOTHING TECHNIQUES

6.2.1 LOWESS

LOWESS (locally weighted scatterplot smoothing) also known as LOESS (locally estimated scatterplot smoothing) is one of many ”modern” modeling methods based on the least squares method. LOESS combines the simplicity of linear regression with the flexibility of nonlinear regression. Locally weighted regression (LOESS) was independently introduced in several different fields in the late 19th and early 20th centuries (Henderson, 1916[76]; Schiaparelli, 1866[145]).In the statistical literature, the method was introduced independently from different perspectives in the late 1970s by Cleveland, 1979[37], among others. The method used in the PQstat program is based on this particular work. 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 neighborhood of any point x.

Procedure algorithm:

1. For each point in the dataset, we build a window containing adjacent elements. The number of elements in the window is determined by the smoothing parameter q. The higher its value, the smoother the function will be. If this parameter is e.g. 0.2, then about 20% of the data will be in the window and they will be used to build the polynomial (here the unicomial, i.e. the linear function).
   Due to the need to maintain the symmetry of the window, i.e., the data in the window should be above and below the point for which we are building the model, 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 first odd number. This produces a window containing an element x_0 and the corresponding number $k(x_0)$ of elements before and after that element in an ordered collection of data.
   $$k(x_0) = \frac{roundUpNpar(q*n)-1}{2}$$
   For example, if the window is to contain 7 elements, it will contain the element x_0 and the three preceding and three following elements of the sample. The window determined for the first and last sample elements is of the same size but the test element is not symmetrically placed in it i.e. in the middle.

2. At each point in the dataset, a low-degree polynomial (here a linear function) is fitted to a subset of the data located in the window. The fit is performed using a weighted least squares method, which gives more weight to points near the point whose response is estimated and less weight to points further away. In this way, a different polynomial function 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 set must have less weight than points nearby. Here the weights proposed by Cleveland, the so-called tricube, are used
   $$w = (1-d^3)^3,$$
   where the distance d from point x_0 was 0 -for points outside the designated window and was given as the actual distance between points, but normalized to the interval [0,1]-for points located within the window, so that the maximum distances in all windows were the same.

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

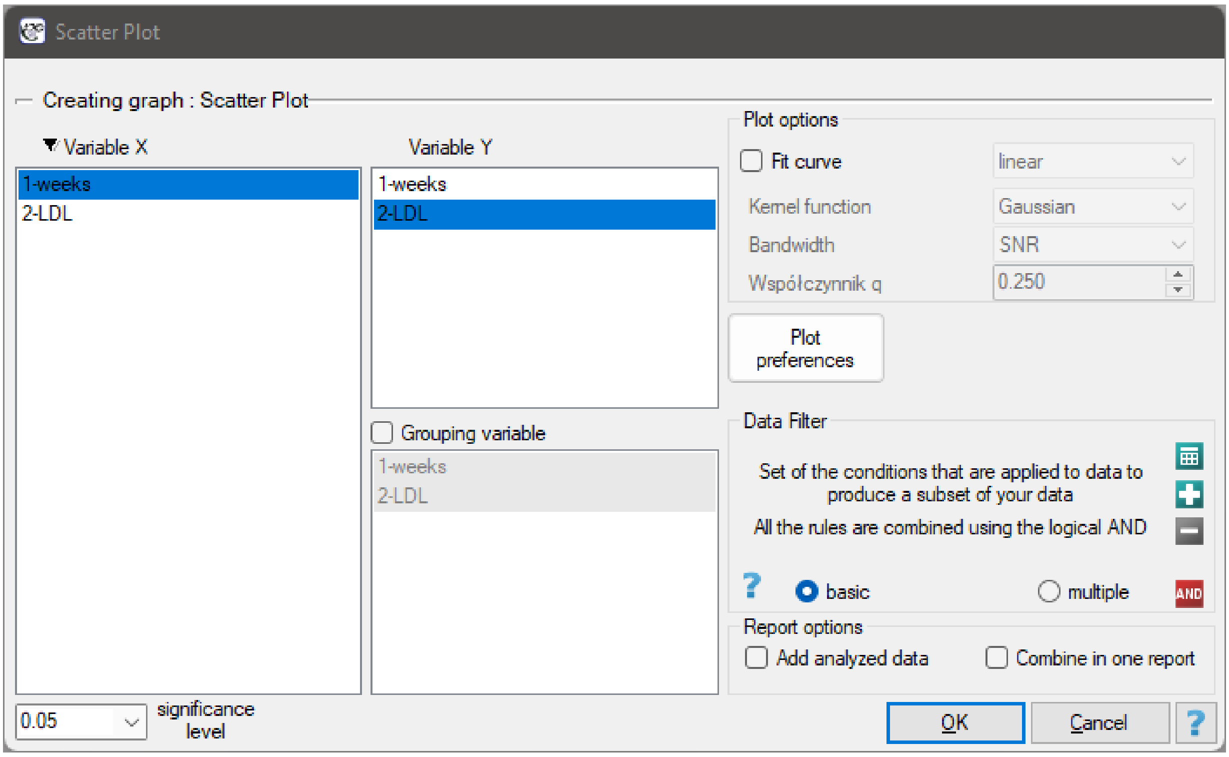
6.2.2 Kernel smoothing

The estimation of the regression function by kernel smoothing is sometimes called the Nadaraya-Watson estimation (Nadaraya, 1964[121]; Watson, 1964[165]). 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 section Kernel estimation
The smoothing parameter h (bandwidth) has a decisive influence on the obtained estimator. The higher the value of the smoothing parameter, the greater the degree of smoothing. It is possible to choose any smoothing parameter by setting a user value. It is also possible to select it automatically by SNR, SROT or OS method. Kernel function has much less influence on the obtained result. We can choose kernels: Gaussian, uniform function (rectangle), triangular, Epanechnikov, quartic or biweight (fourth degree). A description of the different magnitudes of the smoothing parameter and kernel function can be found in the aforementioned chapter.

The window with the settings of the point plot options with the fitted function by the LOWESS method or by kernel smoothing can be found in various analyses. You can also make this graph via the menu Plots→Scatter Plot.



Scatter Plot

Creating graph : Scatter Plot

Variable X: 1-weeks (selected), 2-LDL

Variable Y: 1-weeks, 2-LDL (selected)

□ Grouping variable: 1-weeks, 2-LDL

Plot options

□ Fit curve: linear

Kernel function: Gaussian

Bandwidth: SNR

Współczynnik q: 0.250

Plot preferences

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple AND

Report options

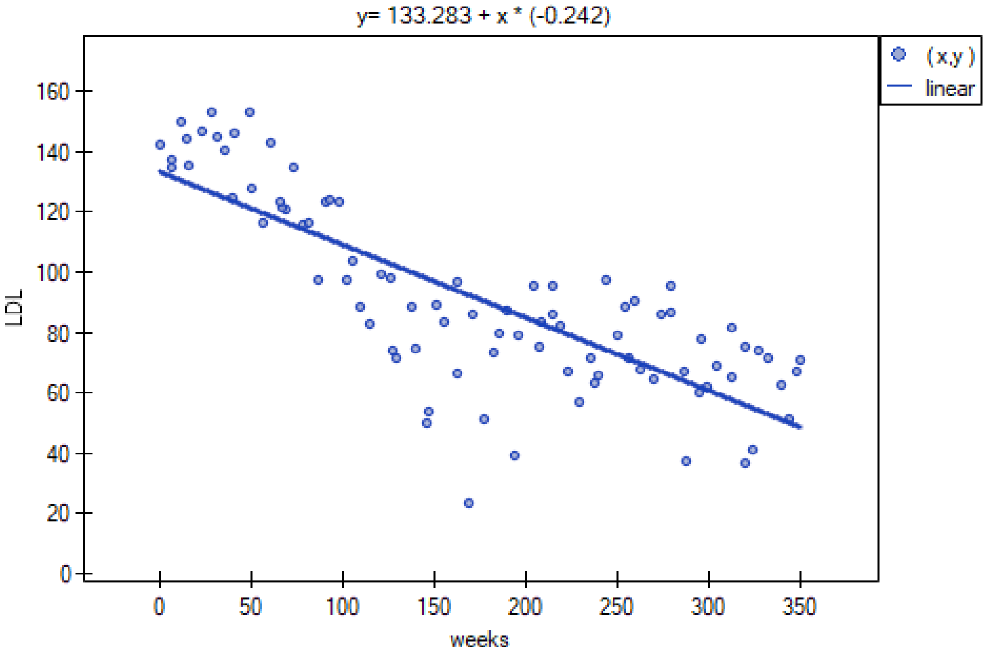
□ Add analyzed data □ Combine in one report

0.05 significance level

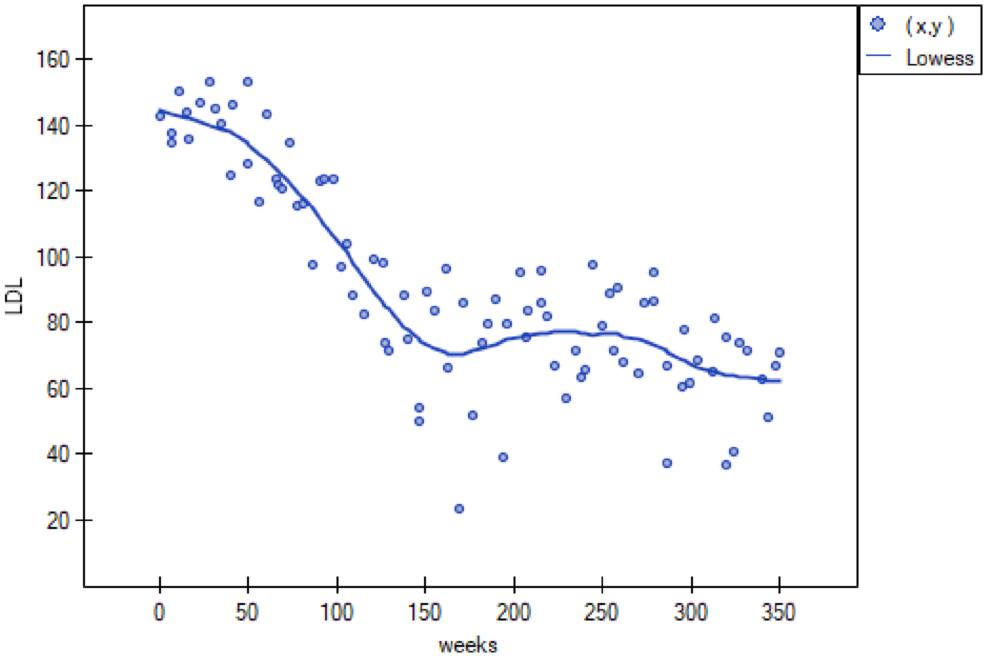
OK Cancel ?

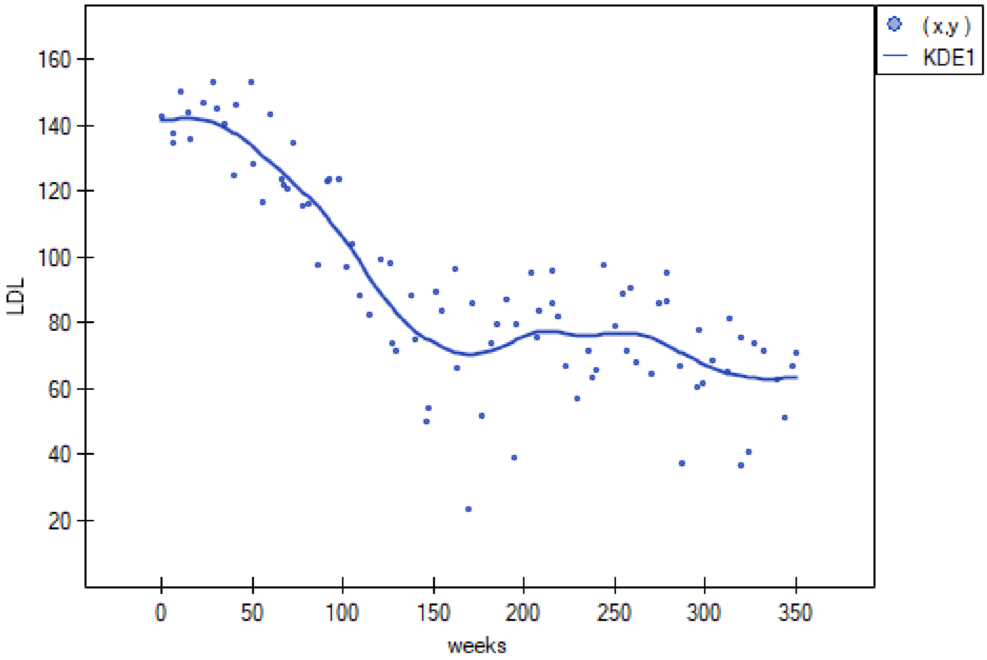
Example (20.2) cont. *(LDL weeks.pqs file)*
The effectiveness of a new therapy designed to lower cholesterol levels in the LDL fraction was tested. 88 people at different stages of the treatment were examined. We will test whether LDL cholesterol levels decrease and stabilize as the treatment is administered over time (time in weeks).

Results are presented by initially fitting a straight line indicating the direction of the relations under study.



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



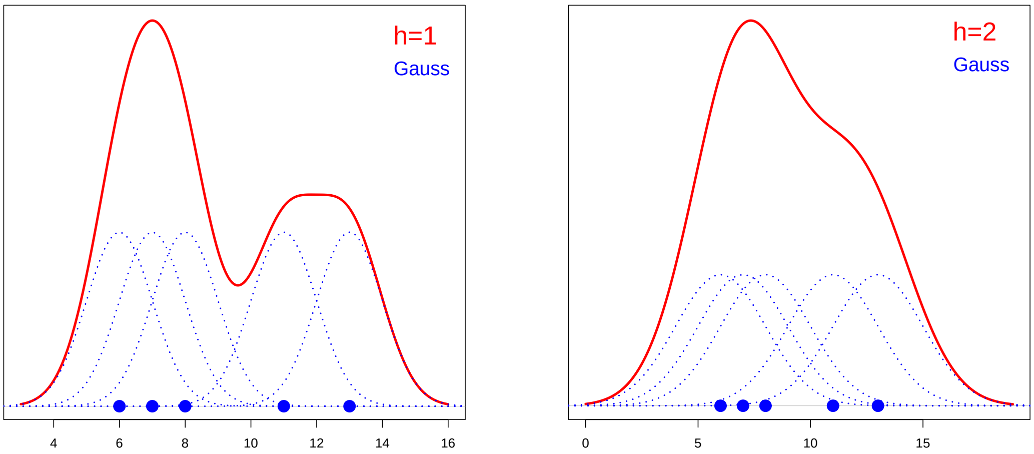


Both methods i.e., both LOWESS and kernel smoothing gave similar results and described the data much better, indicating an initial decline in LDL followed by stabilization near 70 mg/dl.

6.3 KERNEL ESTIMATION

6.3.1 One-dimensional kernel estimator

The one-dimensional kernel density estimator allows you to approximate the density of a data distribution by creating a smoothed density curve in a non-parametric way. It provides a better density estimate than is given by a traditional histogram, which columns form a staircase function.



The kernel estimator is defined based on a properly smoothed kernel $K_h(t_i)$. The smoothing parameter h (*bandwidth*) has a decisive influence on the obtained estimator. The higher the value of the smoothing parameter, the greater the degree of smoothing.

For each point x in the range defined by the data, the density is determined, that is, the value of the kernel estimator at that point is given. This estimator is created by summing the values of the kernel function $K_h(t_i)$ at that point:

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

If we give the individual cases weights w_i, then we can construct a weighted kernel density estimator defined by the formula:

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

Smoothing factors

User - gives you the ability to select any user-specified smoothing factor, but the factor must be positive.

User scaled - is set so that the kernel function can be changed while remaining at the smoothing that was previously chosen for the Gauss kernel. In practice, by choosing a function other than Gauss, the smoothing factor is scaled (Scott, D. W. 1992[146]) so that the smoothing remains at a similar level as it was for the Gauss function. This offers the convenience of switching between different kernels without considering scaling the smoothing parameter. Scaling conversions are made based on the standard deviation:

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

SNR - a smoothing factor constructed from Silverman's method (Silverman B. W. 1986 [151] pp. 45 and 47) and Jones' correction (Jones M. C. et al 1996 [86]) using the standard deviation from the sample rather than the population - as proposed by Silverman:

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

For a non-Gaussian kernel, the smoothing factor is subject to scaling (Scott D. W., 1992[146])

SROT - smoothing factor built on the basis of Silverman's method (Silverman B. W. 1986 [151] pp. 48) with Jones' correction (Jones M. C. i inni 1996[86]):

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

For a non-Gaussian kernel, the smoothing factor is subject to scaling (Scott D. W., 1992[146])

OS - smoothing factor built on the basis of Terrell and Scott's method (Terrell G. R. i Scott D. W. 1985[160], Terrell G. R. 1990[159] pp. 470):

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

For a non-Gaussian kernel, the smoothing factor is subject to scaling (Scott D. W., 1992[146])

Kernel function

The kernel function affects the obtained value of the kernel estimator to a lesser extent than the smoothing parameter. The kernel is a probability density function built around each data point x_i. Typically, it is a symmetric function reaching a maximum at a point x_i, and decreasing its values as one moves away (increasing distance d_i) from that point. The distance from the analysed point is modified by the smoothing parameter h according to the formula: $t_i = \frac{d_i}{h}$.

Depending on your needs, the kernel function can take the form of a function such as:

Gauss

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

uniform (rectanglular)

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

triangular

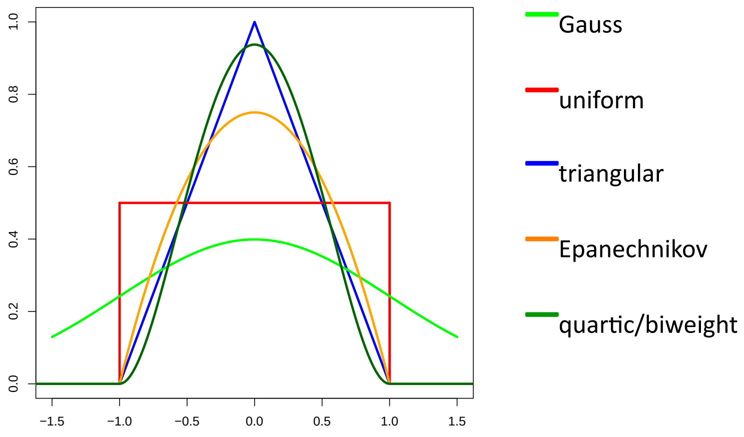
$$K_h(t_i) = \begin{cases} \frac{1-t_i}{h} & \text{if } t_i < 1 \\ 0 & \text{if } t_i \geq 1 \end{cases}$$

Epanechnikov

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

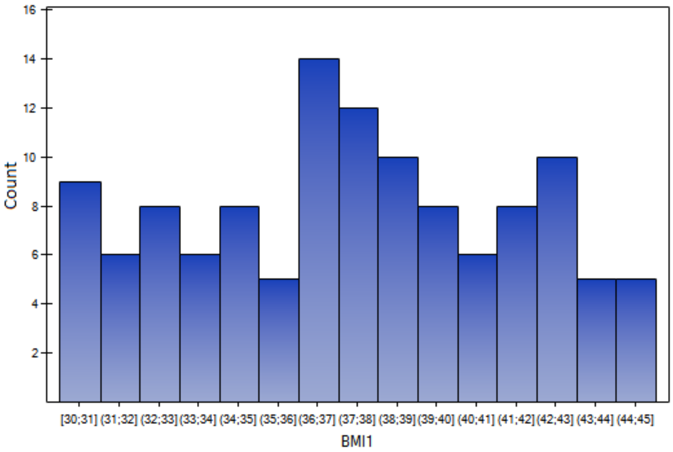
quartic or biweight (fourth degree)

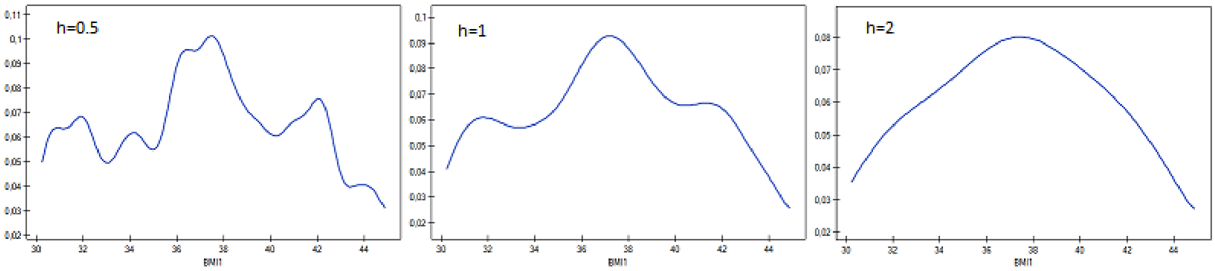
$$K_h(t_i) = \begin{cases} \frac{15}{16}\frac{(1-t_i^2)^2}{h} & \text{ifi } t_i < 1 \\ 0 & \text{if } t_i \geq 1 \end{cases}$$



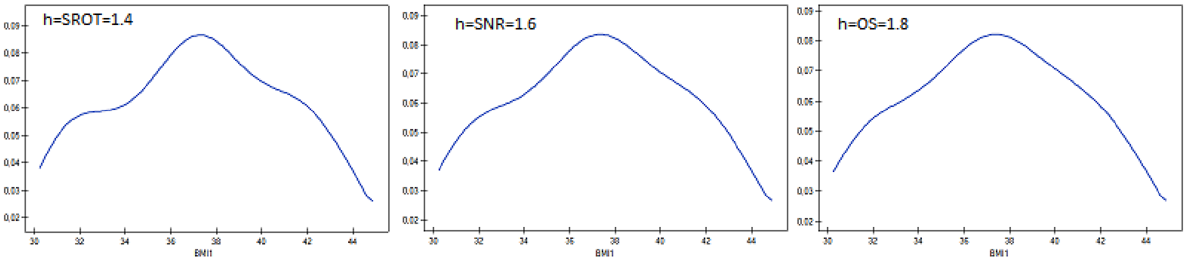
Example 6.1. (BMI.pqs file)

The values of the weight-growth index BMI1 for a certain group of obese subjects were calculated. Their distribution was presented using a histogram with the values divided every 1 BMI unit. The data were also visualized using a kernel density estimator by selecting a Gaussian kernel function and setting the smoothing factors to respectively: 0.5, 1, 2.





The smoothing factors of the kernel estimator suggested by the SROT, SNR, and OS methods reach magnitudes between 1.4 and 2.



6.4 BLAND-ALTMAN PLOT

As noted by Bland and Altman (1986[21], 1999[8]) in clinical medicine, measurements made on the living body are constantly changing and their true value is unknown (e.g., blood pressure), necessitating constant refinement and development of new and better tools to measure them. Usually, when a new method is created, its results are compared with another recognized method, the so-called gold standard. For this purpose, the compatibility of the new method with the previously used method is examined. Of course, the new method cannot be expected to give exactly the same result as the method used so far, but the researcher is interested to see how the results differ. To replace an old method with a new one, the difference between the results of the two methods should be small enough not to pose a problem in clinical interpretation. For example, in a blood pressure measurement, a difference of 20mmHg will be so large that it cannot be considered an acceptable error because it may change the treatment decision. Statistical methods will not answer the question of how large a difference in methods is permissible for the methods to be considered compatible, but appropriate graphical illustration of the differences obtained and the possible limits of variability will help the researcher in making a decision.

A Bland-Altman plot is a point plot, where:

X axis - average of measurements for compared methods;

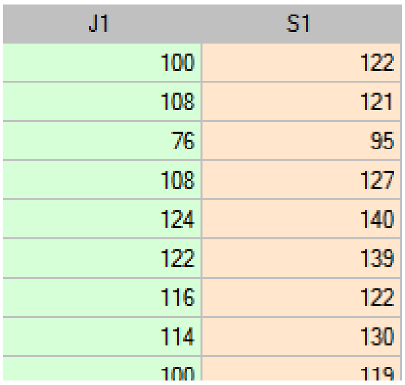
Y axis - difference between measurements for compared methods;

Mean difference- If the results obtained by the new method are consistently greater / less than the old method, then there is a shift, which is called bias, i.e. the line representing the average difference is not at 0, but is shifted significantly up or down from this level.

95% limits of agreement- if the differences have a normal distribution, 95% of the differences will be in the range (Mean difference ± 1.96SD), where SD, is the standard deviation of the differences.
Note1!
The compliance interval defined above is not the same as the limits of agreement.
Note2!
There is no requirement that the data have a normal distribution, only that the distribution of differences does not deviate significantly from the normal distribution. We can check if the differences have a normal distribution by using tests that test for conformity to a normal distribution or a visual interpretation of the normal distribution.

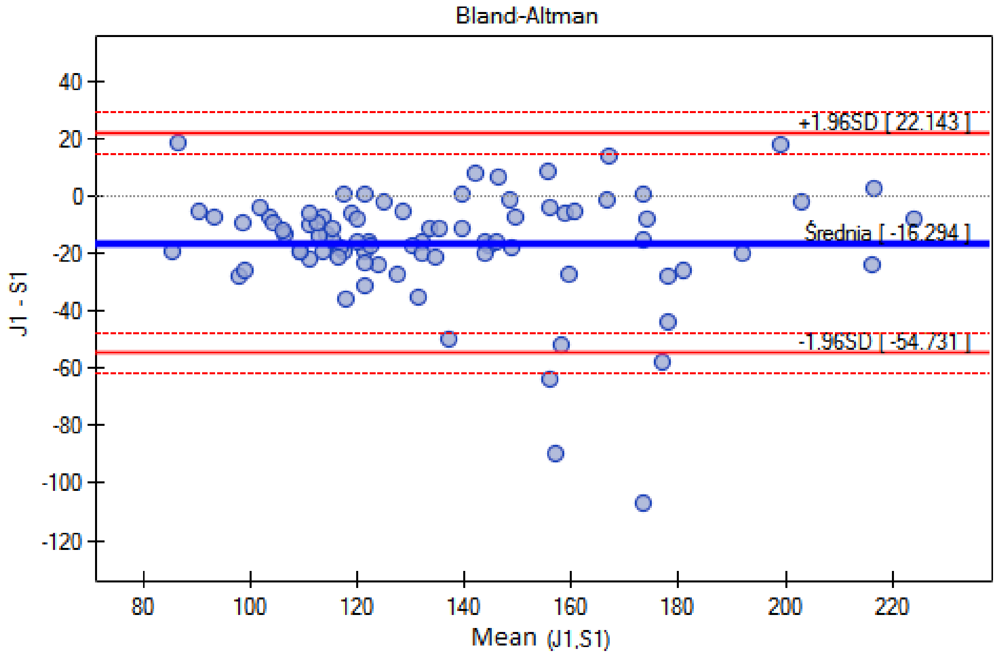
Precision of the limits of agreement- is the interval for the limits, and thus the range of accuracy with which we determine the limits based on a representative sample. The larger the sample and the smaller the variance of the differences, the higher the precision obtained.

Example 6.2. (preassure guage.pqs file)
The example is taken from the work of Bland and Altman (1999[8]). In this study, a semi-automatic blood pressure monitor (S) was compared with the previously traditionally used classic blood pressure monitor (J). For this purpose, systolic blood pressure was measured for 85 patients using both blood pressure 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

A Bland-Altman plot of the collected data indicates that the semi-automatic (S) monitor yields higher results than the classic monitor by an average of 16.3mmHG (the line for the mean difference is 16.3 lower than the absolute agreement shown by the level 0 line). The span of the agreement interval is as high as 76.9mmHG.



For people with hypertension (systolic pressure $\geq$ 140), the changes in pressure can be quite large, so the tested measurement differences can be distorted by actual pressure spikes, so we extracted a subgroup of people with normal pressure and hypertension based on the average pressure value. For each subgroup, we can plot separately by setting a multiple filter for the variable *group* in the test window. The agreement of the methods for people with normal blood pressure will then be much improved (narrower agreement interval).

Bland-Altman plot for repeated measurements

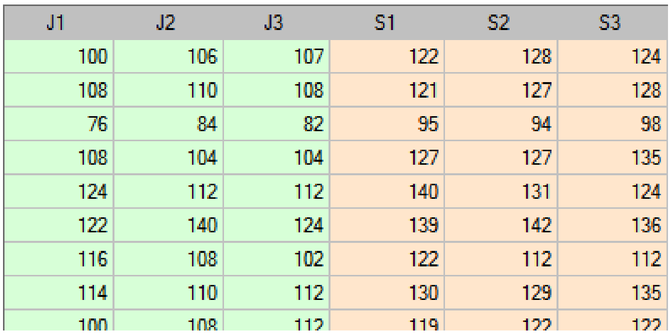
Repeatability of measurements is an important but often overlooked aspect in method agreement testing. A method with higher repeatability is more precise. If the measurements of one of the compared methods are not repeatable (i.e. repeated measurements made on the same objects give rather different results), its agreement with the other method will be low. If the repeatability of both methods is poor, their agreement will be even lower. Consequently, when the repeatability of the old method is poor, the agreement of the new method may be poor, even if the new method has high repeatability. Therefore, although in real research a single measurement is taken for each subject (patient), in research aimed at estimating agreement it is recommended to take measurements several times. This approach provides an opportunity to take into account the reproducibility of the results obtained in studies on agreement of methods.

Note!

By repeated measurements we mean measurements performed independently on the same objects under the same conditions.

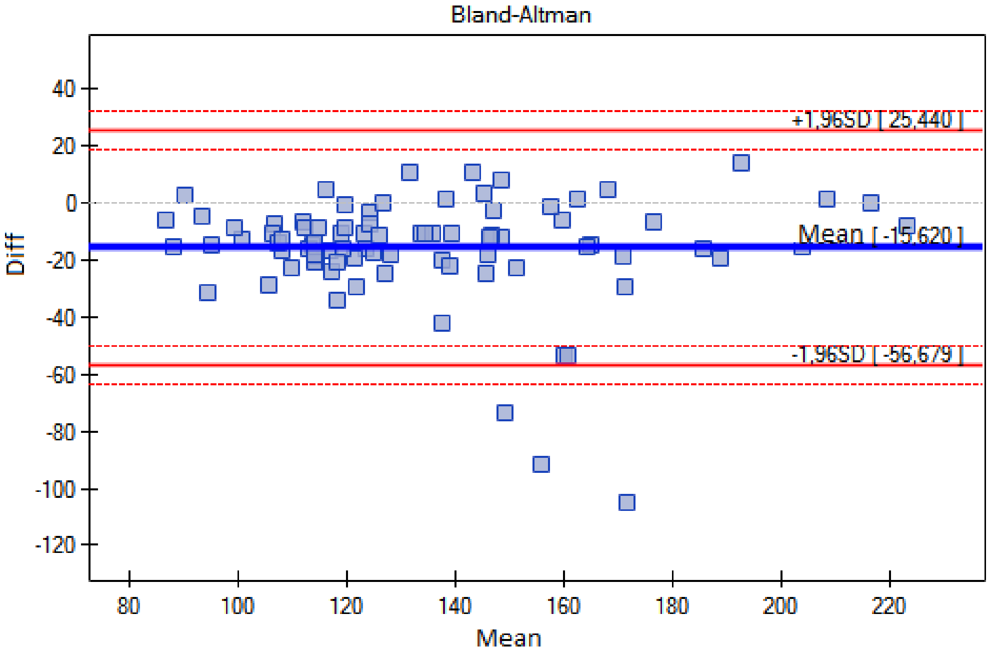
Example (6.2) continued *(preassure guage.pqs file)*

In the comparison of the agreement between the measurements taken by the compared blood pressure monitors, the repeatability of both methods was also taken into account. Therefore, the study was repeated two more times and finally 3 measurements were obtained for each patient using a semi-automatic blood pressure monitor and 3 measurements were obtained using a classic blood pressure monitor. A portion of the data is presented 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 agreement intervals are slightly wider than when using a single measurement for each method - the span of the agreement interval is as high as 82.11mmHG. This is because we take into account the degree of repeatability of the measurements. Unfortunately, taking several repetitions into account increases the width of the interval, but the presented results better represent reality. Without taking into account the repeatability, we assume that the repeatability is 100 percent, which is almost impossible under real conditions.



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

6.5 Correlation matrix

When we are interested in correlation between many variables, a convenient way to visualize it is to present correlation coefficients in the form of a chart. Depending on the scale on which the data was collected, in PQStat we have a choice of these coefficients:

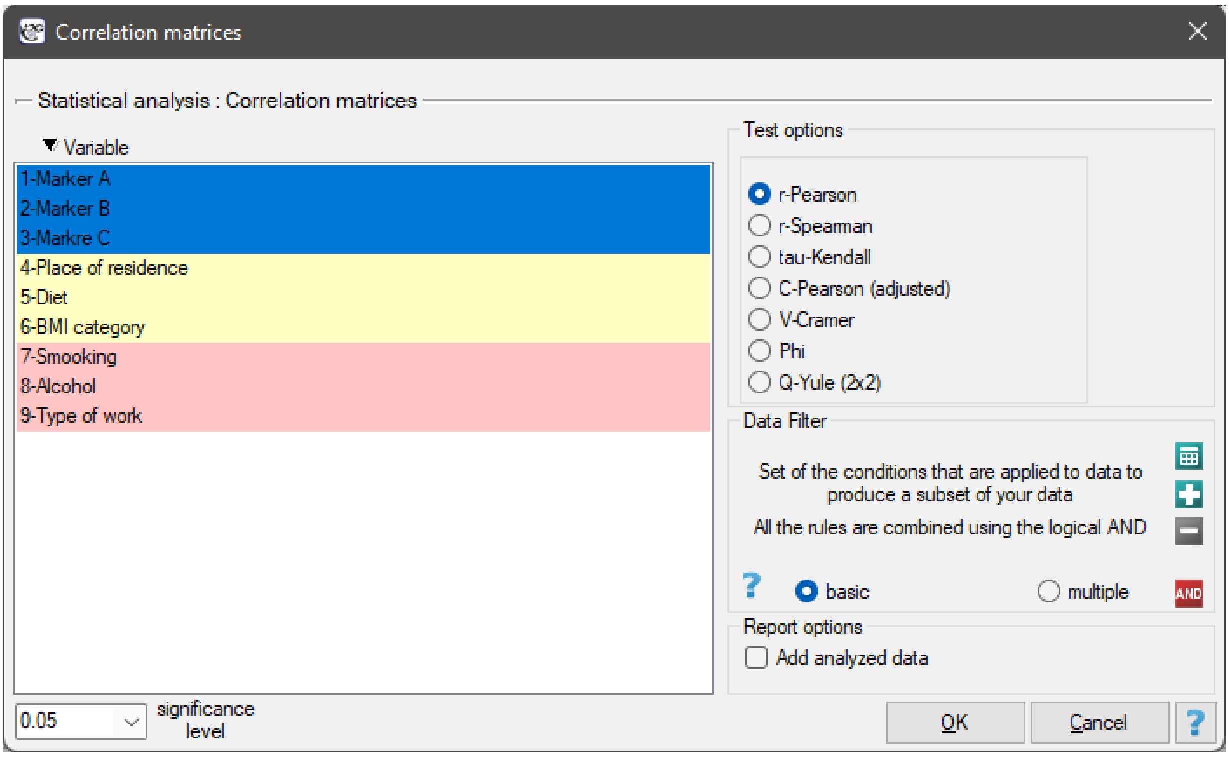
- r-Pearson (interval scale)
- r-Spearman (ordinal scale or stronger)
- tau-Kendall (ordinal scale or stronger)
- C-Pearson (nominal scale or stronger)

- V-Cramer (nominal scale or stronger)
- Phi (nominal scale or stronger)
- Q-Yule (nominal scale or stronger)

As a result of the analysis two matrices are created, i.e. a matrix of correlation coefficients and a matrix of p-values for the test determining the statistical significance of a given coefficient (for r-Pearson, r-Spearman and tau-Kendall coefficients these were the tests dedicated for them, for nominal scale in was the chi-square test). In the matrix of correlation coefficients , at the intersection of two variables, the coefficient of their correlation is given, and its p-value is in the corresponding place in the other matrix. The cell color in the coefficient matrix is graded from blue (negative coefficients) to red (positive coefficients).

The analysis determines the correlation for each pair of variables, so missing data are omitted in pairs. If we want to perform the analysis by omitting missing data in other variables (i.e., not those included in the pair), then we should do so by using advanced filter.

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


Correlation matrices

Statistical analysis : Correlation matrices

Variable
1-Marker A
2-Marker B
3-Markre C
4-Place of residence
5-Diet
6-BMI category
7-Smooking
8-Alcohol
9-Type of work

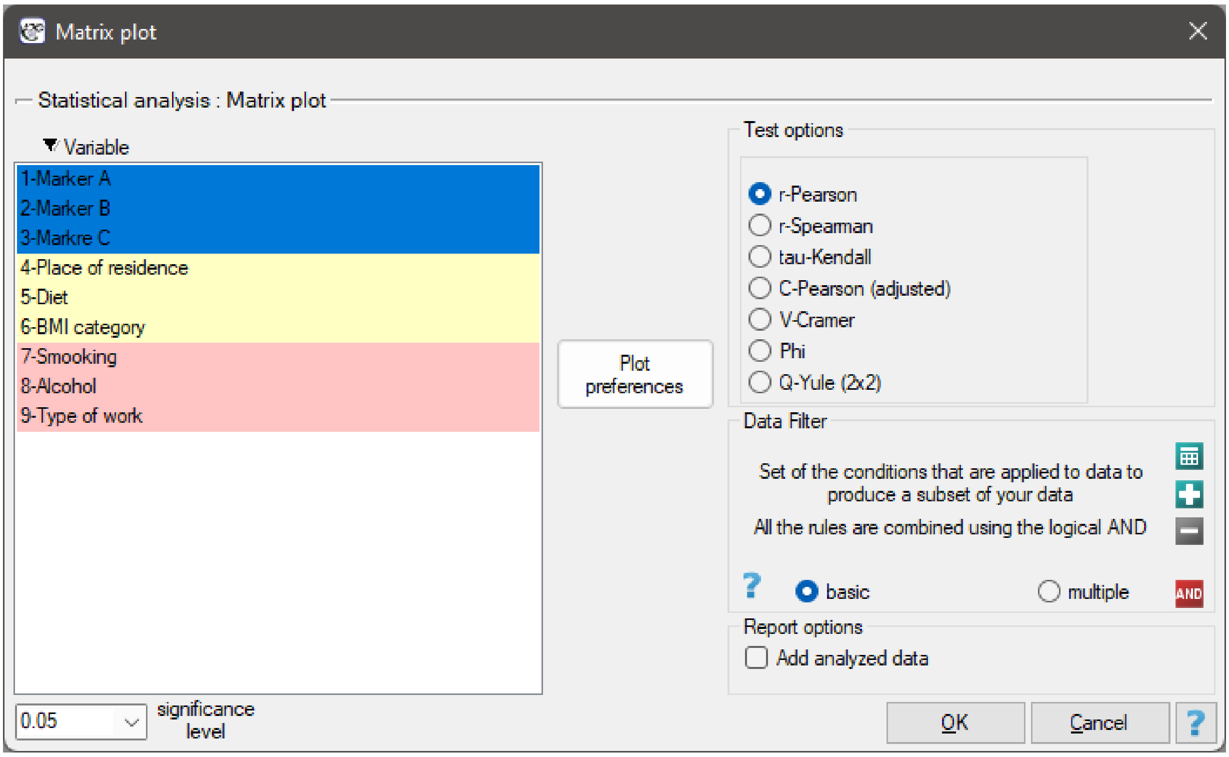
Test options
● r-Pearson
○ r-Spearman
○ tau-Kendall
○ C-Pearson (adjusted)
○ V-Cramer
○ Phi
○ Q-Yule (2x2)

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
□ Add analyzed data

0.05 significance level

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


Matrix plot

Statistical analysis : Matrix plot

Variable

- 1-Marker A
- 2-Marker B
- 3-Markre C
- 4-Place of residence
- 5-Diet
- 6-BMI category
- 7-Smooking
- 8-Alcohol
- 9-Type of work

Test options

● r-Pearson
○ r-Spearman
○ tau-Kendall
○ C-Pearson (adjusted)
○ V-Cramer
○ Phi
○ Q-Yule (2x2)

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data

0.05 significance level

EXAMPLE 6.3. (file markers others.pqs)

An excerpt from a larger study on cancer is given. The data taken represents a group of 100 people. The study measured, among other things, values of tumor markers (interval scale), determined BMI categories for patients and asked for opinions on the possible influence of their place of living and diet on health (ordinal scale), as well as recorded patients' answers on questions about smoking, alcohol consumption and type of work (nominal scale).

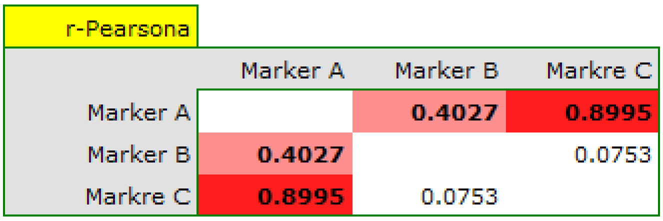
Conducting multivariate analyses often implies the need to first check for intercorrelations of variables. For the purposes of further analysis:

(1) We will examine the correlation within each of these groups.

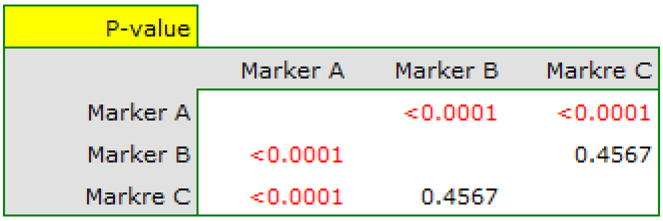
(2) We will test the correlation between all variables.

(1)

For the interval scale, assuming normality of distribution, correlation can be tested by Pearson's linear correlation coefficient. The strongest correlation is for marker A and marker C (r=0.8995, p<0.0001) the weakest and not statistically significant is for marker B and marker C (r=0.0753, p=0.4567).

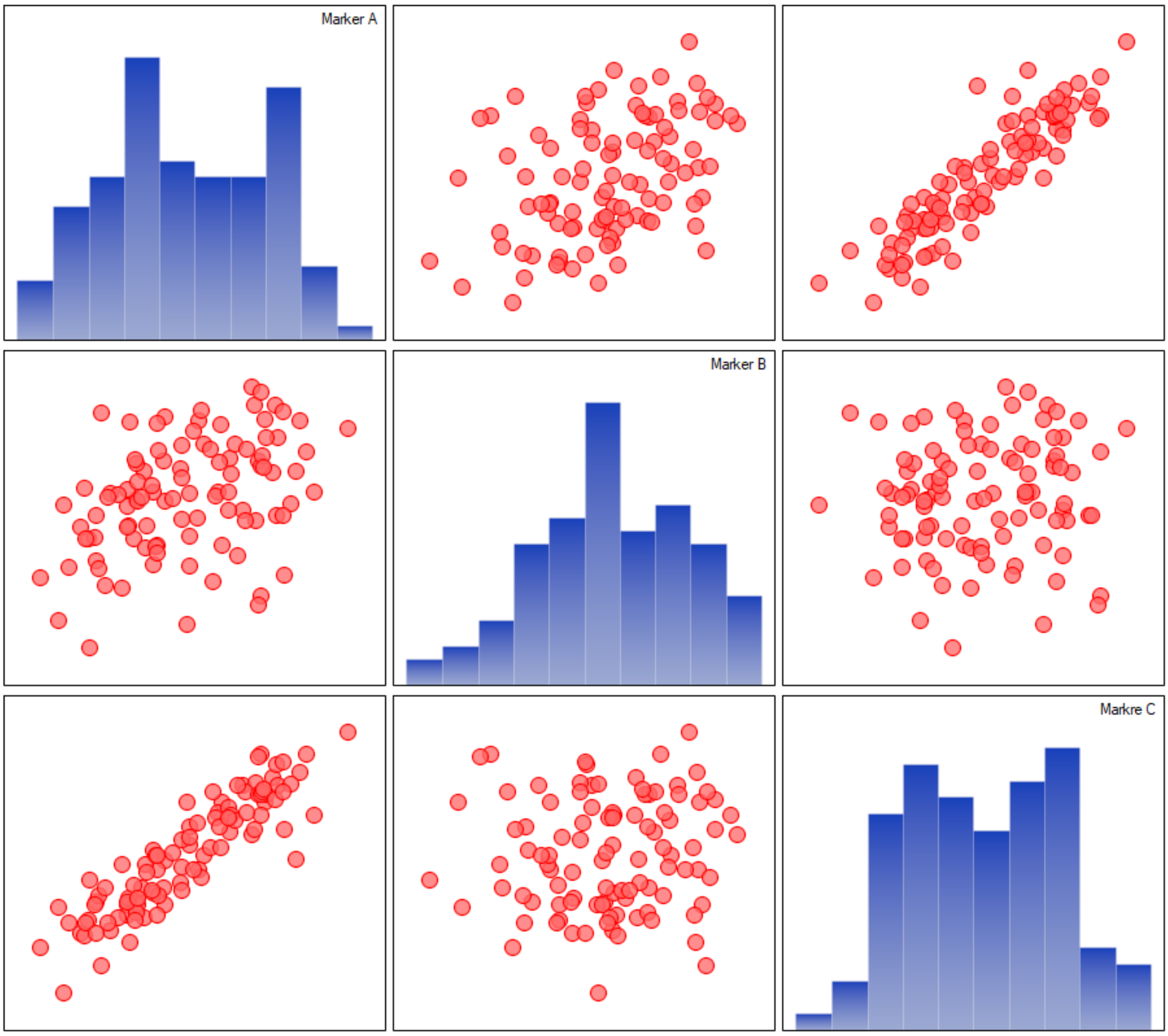


r-Pearsona	Marker A	Marker B	Markre C
Marker A		0.4027	0.8995
Marker B	0.4027		0.0753
Markre C	0.8995	0.0753	

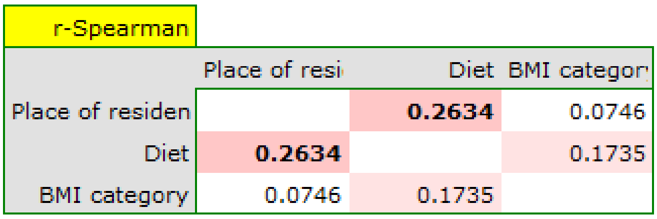


P-value	Marker A	Marker B	Markre C
Marker A		<0.0001	<0.0001
Marker B	<0.0001		0.4567
Markre C	<0.0001	0.4567	

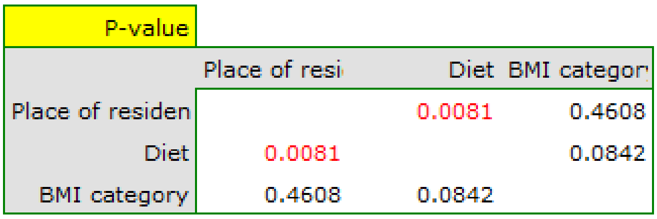
The described correlations can be observed in scatter plots (the X-axis of these plots is the variable described in columns, the Y-axis in rows), and the distributions of individual variables in column plots.



For the ordinal scale, we will check the correlation using the Spearman correlation coefficient. The only significant correlation is between diet and place of living (r=0.2634, p=0.0081).

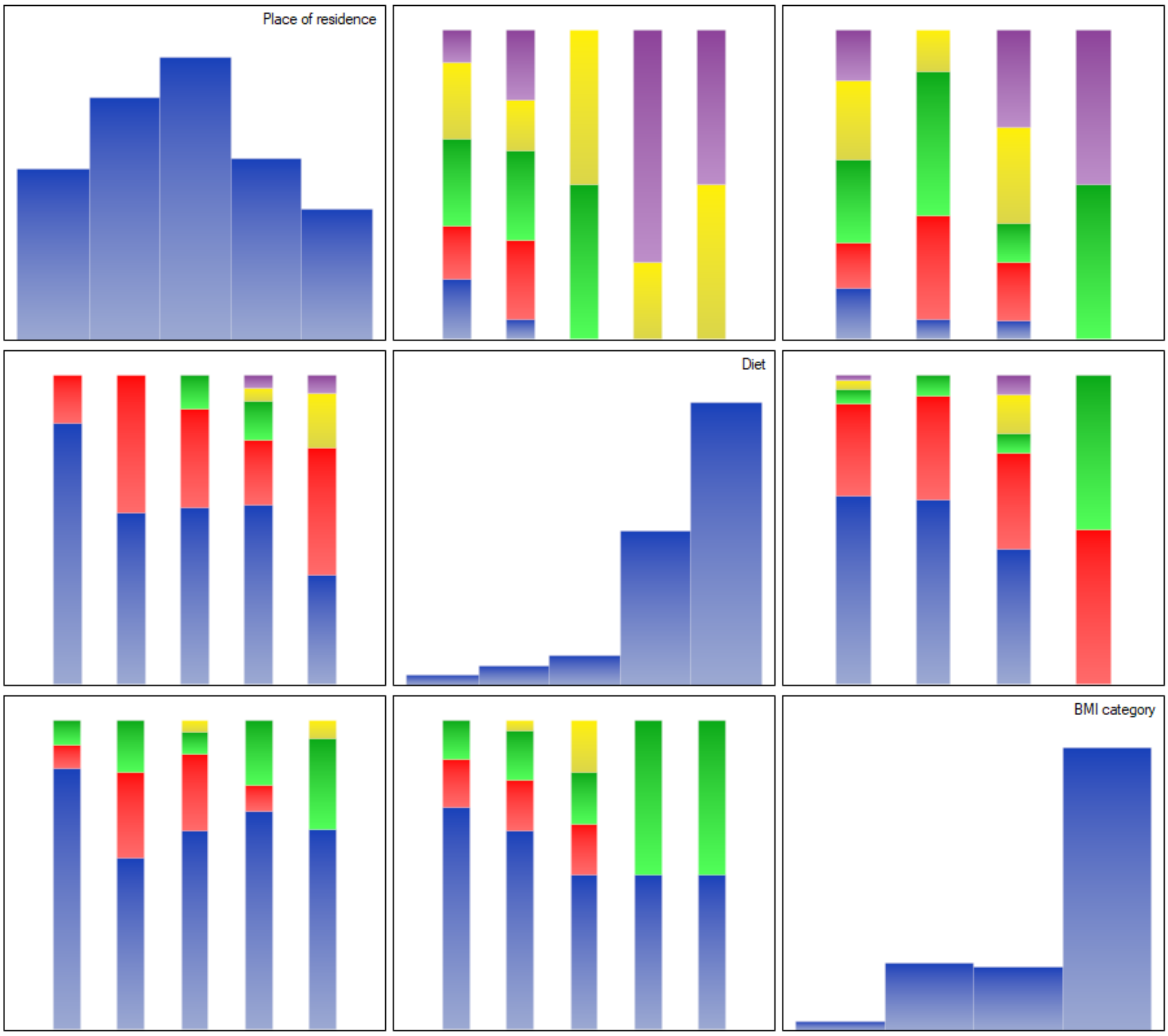


r-Spearman	Place of resi	Diet	BMI categor
Place of residen		**0.2634**	0.0746
Diet	**0.2634**		0.1735
BMI category	0.0746	0.1735	

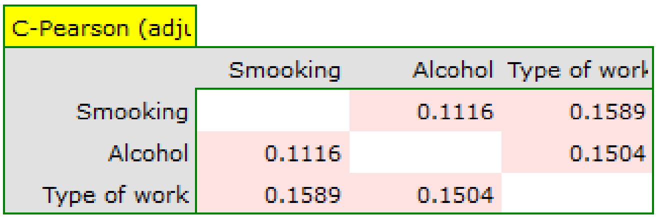


P-value	Place of resi	Diet	BMI categor
Place of residen		0.0081	0.4608
Diet	0.0081		0.0842
BMI category	0.4608	0.0842	

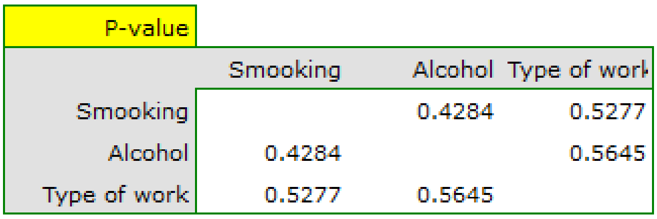
The correlations described can be observed in cumulative column plots (the X-axis of these plots is the variable described in the columns, the legend is the variable described in the rows), and the distributions of the individual variables in the column plots located on the main diagonal.



For the nominal scale, we check the correlation using the C-Pearson coefficient adjusted for chart size. We did not obtain statistically significant correlations.

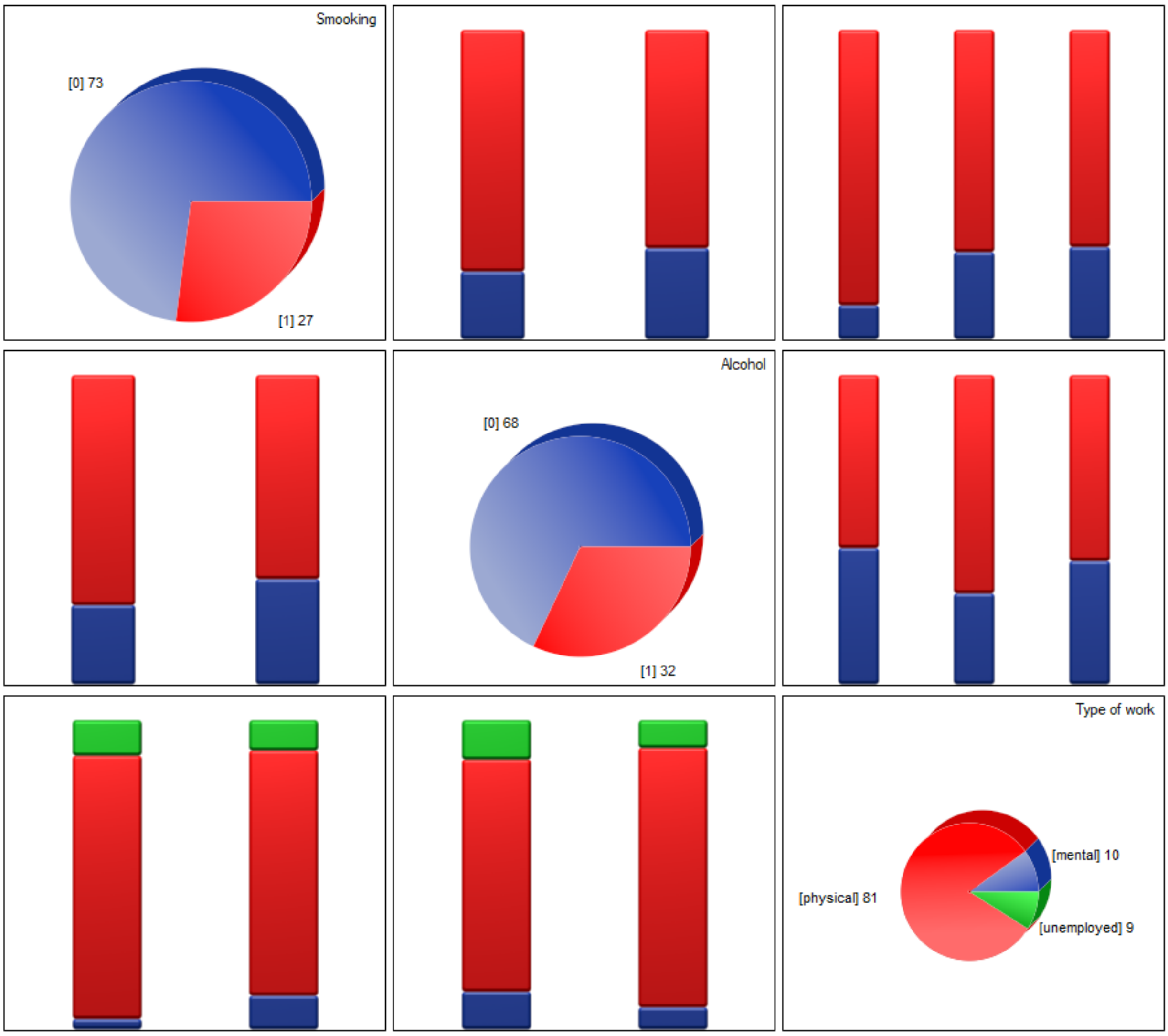


C-Pearson (adjι			
	Smooking	Alcohol	Type of worl
Smooking		0.1116	0.1589
Alcohol	0.1116		0.1504
Type of work	0.1589	0.1504	



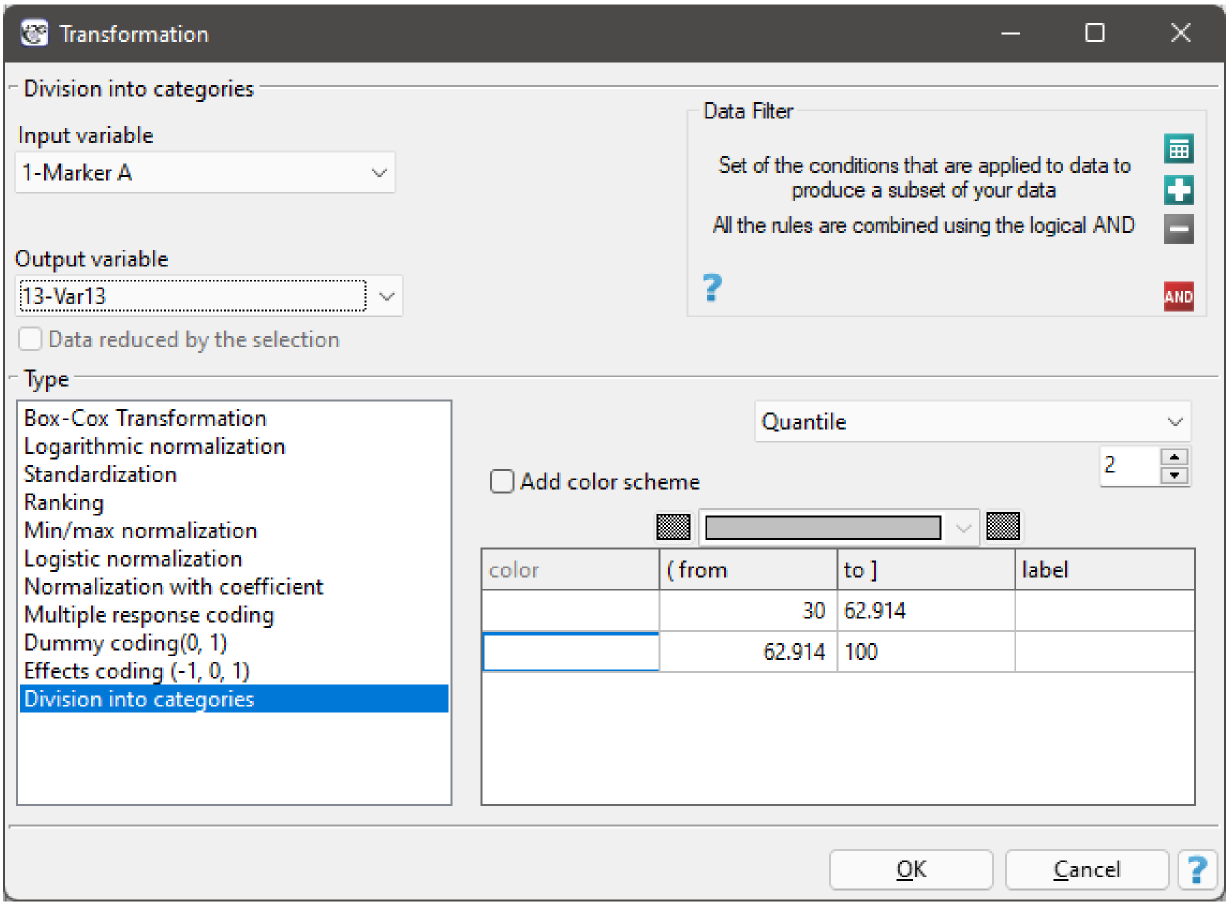
P-value			
	Smooking	Alcohol	Type of worl
Smooking		0.4284	0.5277
Alcohol	0.4284		0.5645
Type of work	0.5277	0.5645	

Correlations, if any, can be observed in cumulative column plots (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 individual variables can be observed in column plots located on the main diagonal.



(2) The easiest way to determine correlations between variables measured on different scales is to bring

them to the same scale. To do this, we will record 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 menu Data→Transformation.



Transformation

Division into categories

Input variable: 1-Marker A

Output variable: 13-Var13

□ Data reduced by the selection

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

Type

- Box-Cox Transformation
- Logarithmic normalization
- Standardization
- Ranking
- Min/max normalization
- Logistic normalization
- Normalization with coefficient
- Multiple response coding
- Dummy coding(0, 1)
- Effects coding (-1, 0, 1)
- Division into categories

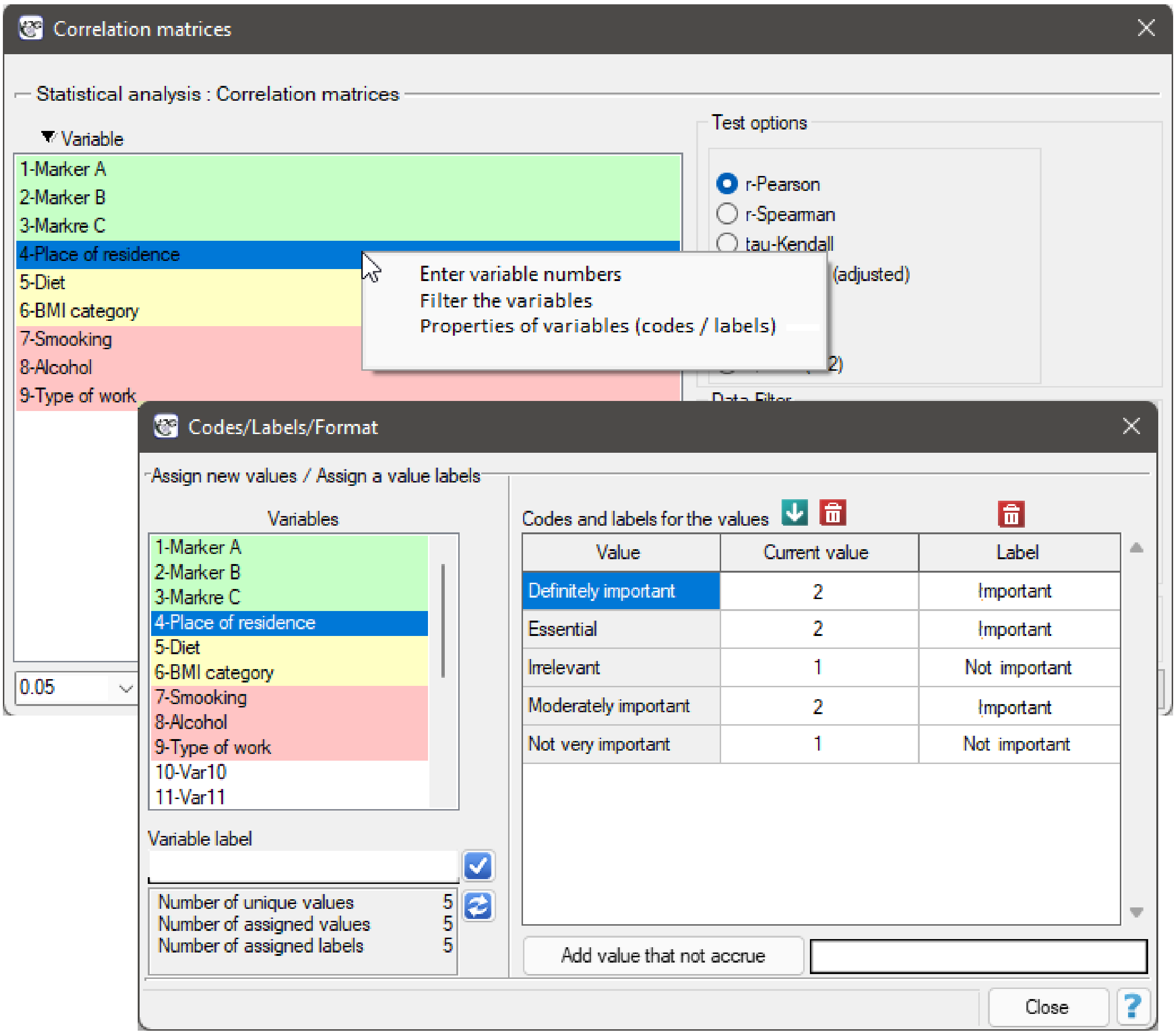
Quantile

2

□ Add color scheme

color	(from	to]	label
	30	62.914	
	62.914	100	

The ordinal data will also be divided into two categories, but the 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 the two labels.


Correlation matrices
Statistical analysis : Correlation matrices
Variable
Test options
Enter variable numbers
Filter the variables
Properties of variables (codes / labels)
Data Filter
Codes/Labels/Format
Assign new values / Assign a value labels
Variables
Codes and labels for the values
Variable label
Add value that not accrue
Close

Variable: 1-Marker A, 2-Marker B, 3-Markre C, 4-Place of residence, 5-Diet, 6-BMI category, 7-Smooking, 8-Alcohol, 9-Type of work, 10-Var10, 11-Var11

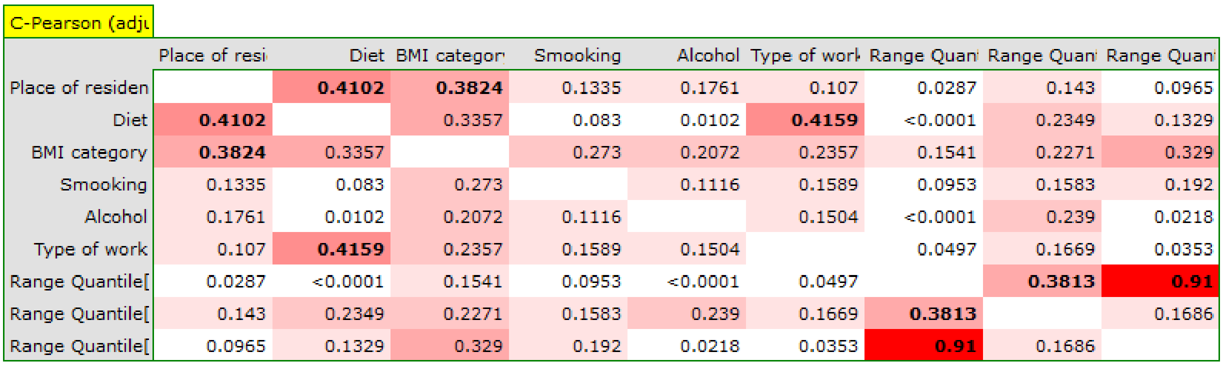
Test options: ● r-Pearson ○ r-Spearman ○ tau-Kendall

0.05

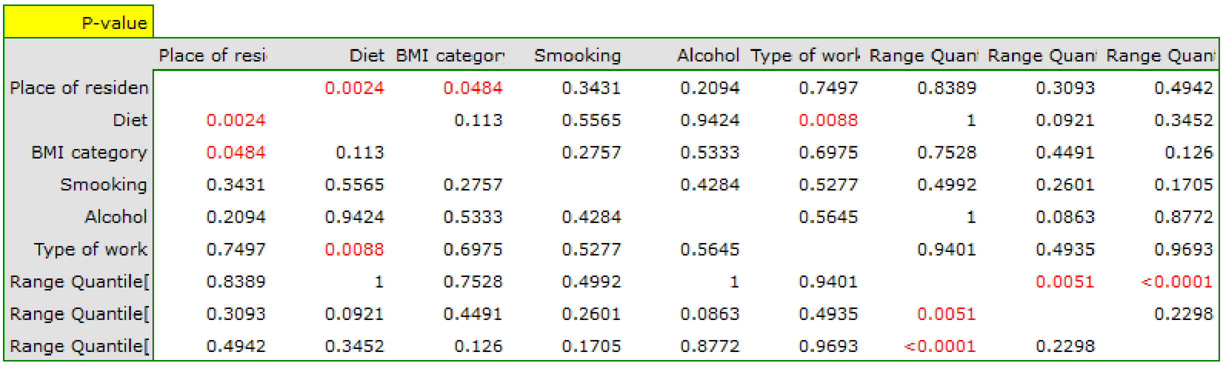
Value	Current value	Label
Definitely important	2	Important
Essential	2	Important
Irrelevant	1	Not important
Moderately important	2	Important
Not very important	1	Not important

Number of unique values	5
Number of assigned values	5
Number of assigned labels	5

As a result, we will only show the correlation matrix (without a graph), since a graph for so many variables will not be clear enough.



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	

7 TEST POWER AND SAMPLE SIZE

There are several ways we can approach determining the sample size. One possibility is to estimate how large the sample should be to reflect the population. Another possibility is to estimate the sample size for the situation of applying specific statistical tests. Then, in addition to the necessary sample size, we may be interested in the power of those tests. The first, and seemingly easier approach, is presented in subsection Sample size determination, the second in subsection Power and size for a test.

7.1 Sample size determination

For the margin of error of the proportion and the mean
Since it is usually neither practical nor possible to study the entire population, a subset of it - the sample - is chosen. The sample is of course correspondingly smaller than the population, but it should reflect it well. One of the key aspects in planning a study, besides the randomness of the sample, is the assumption of its size. The size should be chosen so that the inference about the population is true.

If we are interested in ensuring that the proportions of certain characteristics, or their mean values, calculated for a sample reflect the proportions or mean values in the population with as little bias as possible, we can estimate the necessary sample size accordingly.

Sample size for estimating population proportions

Assuming the possibility of an error in estimating the size of E, we can determine the necessary 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 specified by the user, whereas - if this quantity is not known, the estimated necessary size will be increased to be sufficient for each possible proportion, so the value $p = 0.5$ will be used.

When we know the population size (and in particular - when the size is relatively small with respect to n_0, i.e. when $n_0/N > 5\%$) we should use the so-called finite population correction (FPC) (Lenth (2001)[98], Armitage and Colton (2009)[12]) given by the formula:

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

Sample size for estimating the population mean

Assuming the possibility of an error in estimating the size of E, we can determine the necessary 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 – population standard deviation - known from previous studies.

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

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

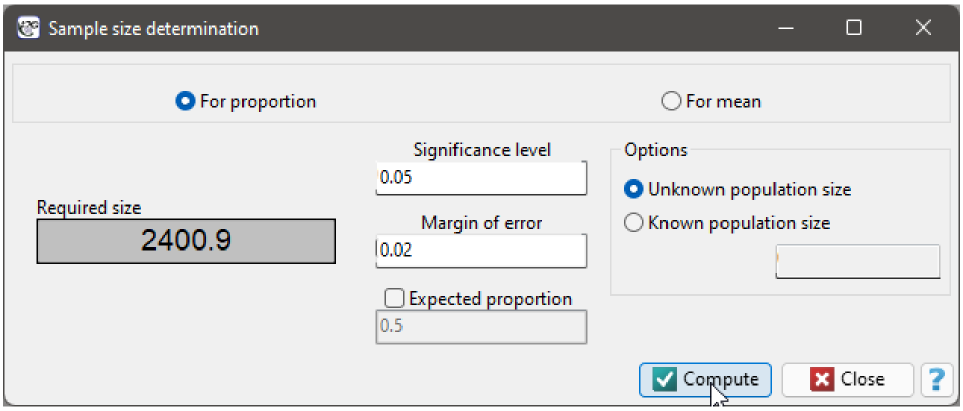
The window with the Sample size determination settings is opened via menu Advanced statistics→test power and sample size→Sample size determination

Example 7.1. Estimation of proportions

Population: Eligible to vote for President of Poland.

We are interested in endorsements of individual candidates.

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



Sample size determination

● For proportion ○ For mean

Significance level: 0.05

Margin of error: 0.02

□ Expected proportion: 0.5

Required size: 2400.9

Options: ● Unknown population size ○ Known population size

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

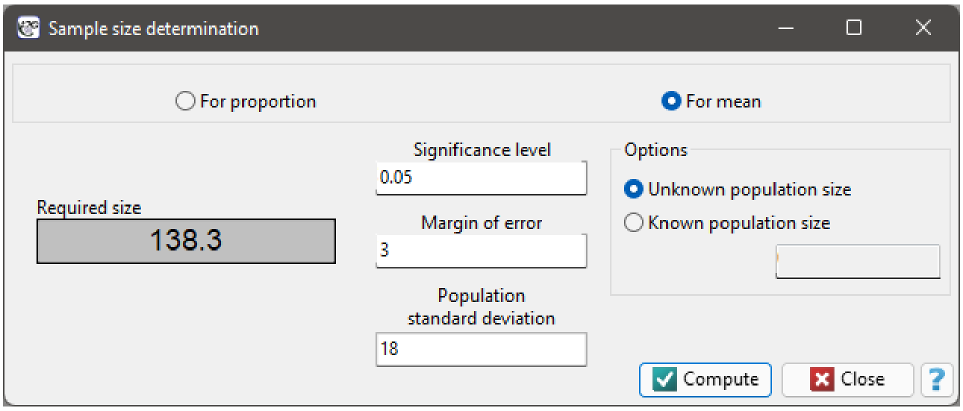
When choosing the size of the acceptable bias, one should pay attention to the fact whether there is not a situation in which candidates with small support (on the limit of the assumed estimation bias) take part in the election. If this is the case, it is worth reducing the value of the estimated bias - the consequence of reducing the bias will then be an increase in the necessary sample size.

Example 7.2. Estimating the mean value

Population: Individuals with hypertension in Poland in 2005-2010, aged 20-40 years.

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

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



Sample size determination

○ For proportion ● For mean

Required size: 138.3

Significance level: 0.05

Margin of error: 3

Population standard deviation: 18

Options

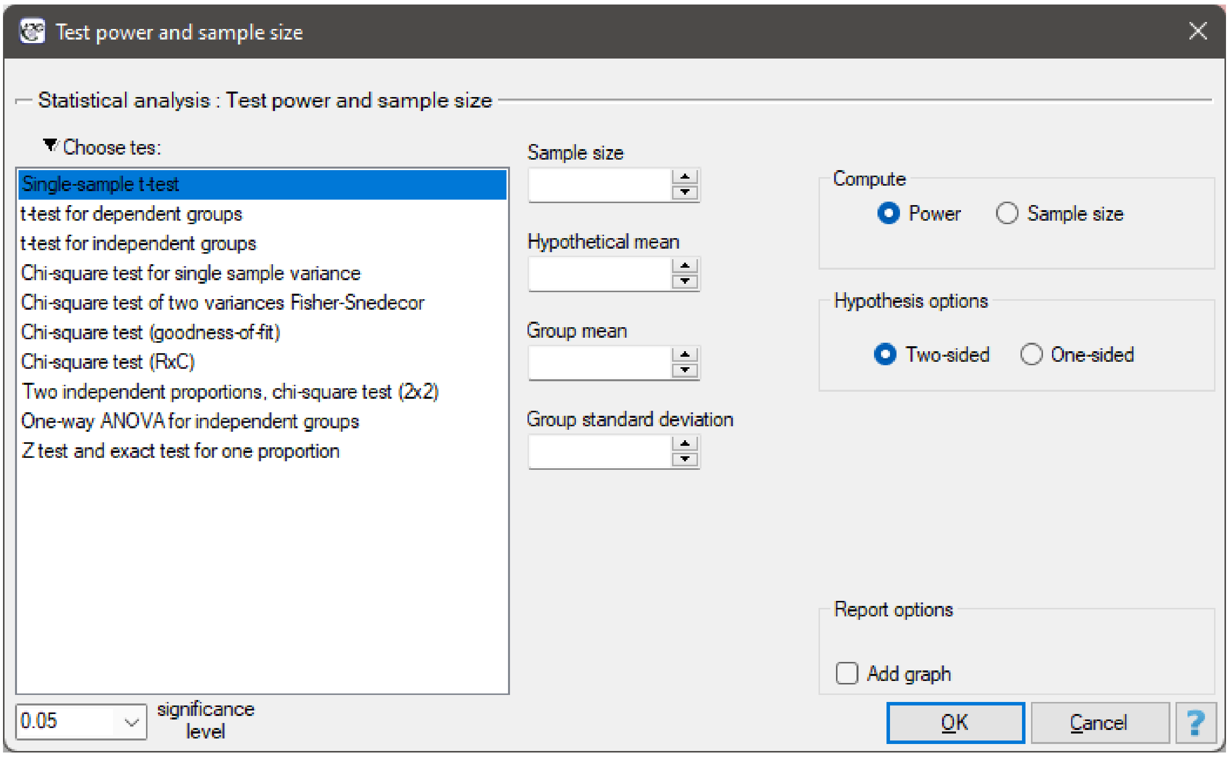
● Unknown population size

○ Known population size

To be 95% sure that there is a population mean within the bias (Δkg) built around the mean of our sample we need to select at least 139 individuals.

7.2 Power and sample size for test

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


Test power and sample size
Statistical analysis : Test power and sample size
Choose tes:
Single-sample t-test
t-test for dependent groups
t-test for independent groups
Chi-square test for single sample variance
Chi-square test of two variances Fisher-Snedecor
Chi-square test (goodness-of-fit)
Chi-square test (RxC)
Two independent proportions, chi-square test (2x2)
One-way ANOVA for independent groups
Z test and exact test for one proportion
Sample size
Hypothetical mean
Group mean
Group standard deviation
Compute
● Power ○ Sample size
Hypothesis options
● Two-sided ○ One-sided
Report options
□ Add graph
0.05
significance level
OK
Cancel
?

Power analysis is directly related to testing hypotheses, and therefore to specific statistical tests. Tests vary in their power. Some tests are stronger, others weaker. Because of this fact, if there are several tests available to solve a given statistical problem, it is better to choose the test which is more powerful. Such a test is stronger, so it will more easily reject the null hypothesis, and therefore it will be easier for us to prove the alternative hypothesis - which is, after all, the goal.

Power of statistical test is the ability of a test to detect differences, relations, correlations and any kind of dependencies which are described in alternative hypothesis. In technical language, the power of a test is called the probability of accepting an alternative hypothesis when it is in fact true.

The power of the test can be checked a priori, i.e. before collecting the data for the actual test, but often it is the reviewers of the papers or ourselves already during the analyses, i.e. a'posteriori, i.e. after collecting the actual sample, who are interested in the power of the analyses we perform. If the power of the test is low, then the results obtained may be ambiguous, if it is high - we may expect that in the future it will be difficult for other researchers to obtain different results, i.e. to undermine our results. For example, when we show using a test with a power of 80% that two groups of students are statistically significantly different from each other in terms of the number of correctly solved tasks, this means that when other researchers repeat this experiment under the same conditions as we do, they will also prove the alternative hypothesis in 80% of the random samples of the same size as ours, using the same test and assuming the same significance level.
Power of the test is determined by the formula:

$$Power = 1 - \beta,$$

where:
β - Type II error, which is the probability of accepting the null hypothesis when it is false.

The power of a test is directly related to the sample size n - the larger the sample, the greater the power, i.e., the more students we collect to run the test, the easier it will be to argue that the detected differences between groups are not due to chance, but actually occur between populations. Hence, using the same approach, you may be interested in determining the **required sample size** for a given statistical test while keeping its power at a given level.

In PQStat, we can calculate the power of a test by specifying the sample size, or we can calculate the sample size by specifying the test power we want to achieve. Unfortunately, both test power and sample size, in addition to being related to each other, also depend on other additional information about the collected sample that needs to be determined, these are:

- The effect size ES that we consider important. The larger the effect, the greater the power and the smaller the sample needed to obtain this power. For example, obtaining a difference of five correctly solved tasks will be easier to argue that the groups do in fact differ (we will have a more powerful argument) than if we tried to argue for a real advantage of one of the student groups by stating that they differ by only one correctly solved task.
  To determine power or the required size, usually the effect under study must be standardized, so in many situations it is necessary to provide additional information, e.g., standard deviation, correlation coefficient, and other coefficients to standardize such an effect.

- The level of statistical significance α (error of type I) - the greater the significance level, the greater the power. Unfortunately, the significance level is the part of the analysis over which we have only apparent influence, that is, there are few situations when it can be changed, and if a change is allowed, it involves decreasing α (see Multicomparison). Normally we assume that $\alpha = 0.05$.

- Direction of hypothesis i.e. two-sided hypothesis (equality in null hypothesis) or one-sided hypothesis (< or > signs in null hypothesis).The one-sided hypothesis gives more power but is much less often chosen because in real life situations when applying a statistical test we rarely assume that, for example, students from the first population have no chance of beating students from the second population, but often we give equal chances to both groups at the start.

Before you can determine the power of a test, you should know how to use it, understand its hypotheses, and be able to determine the effect size, and if you have data from a study such as a pilot study, you should also perform that test.

7.2.1 Single-sample t-test

Before determining the power or the required sample size of the Single-sample t-test, it is worthwhile to familiarize yourself with the rules of its application.

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

Power of the test	Group size
Group size	Power of the test
Hypothetical mean	
Group mean	
Group standard deviation	
α significance level	

The set effect size, in this case, is the size of the difference between the set mean of the study population and the hypothetical mean value.

The power of the test and the required sample size are calculated based on the noncentral t-test distribution.

Example 7.3. We want to test (at the 0.05 significance level) whether the wait time for a certain delivery company to deliver a package is on average days (i.e., 72 hours).

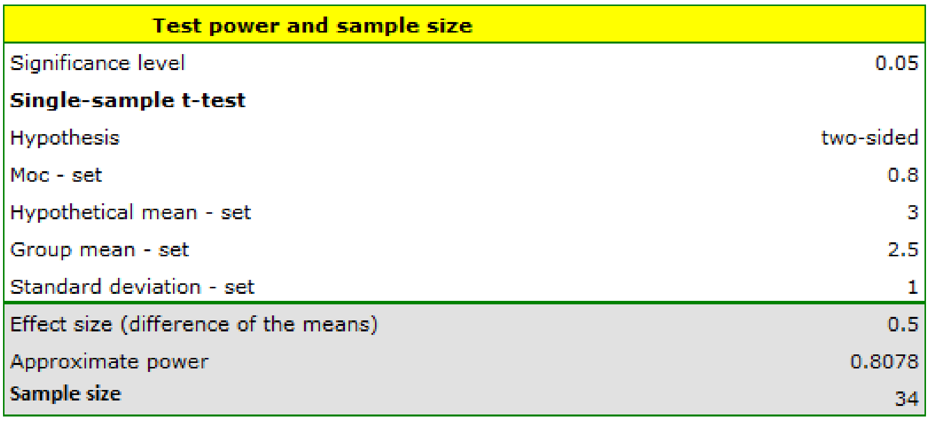
- **Testing a priori** - We plan to conduct a study, but have not collected data yet.

  We set the test power we want to obtain at 80%.
  The standard deviation that we provide should reflect the difference in delivery time that we expect to obtain in the planned study - we assume (based on the experience of the employees of this company) that it will be 1 day (24 hours).

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

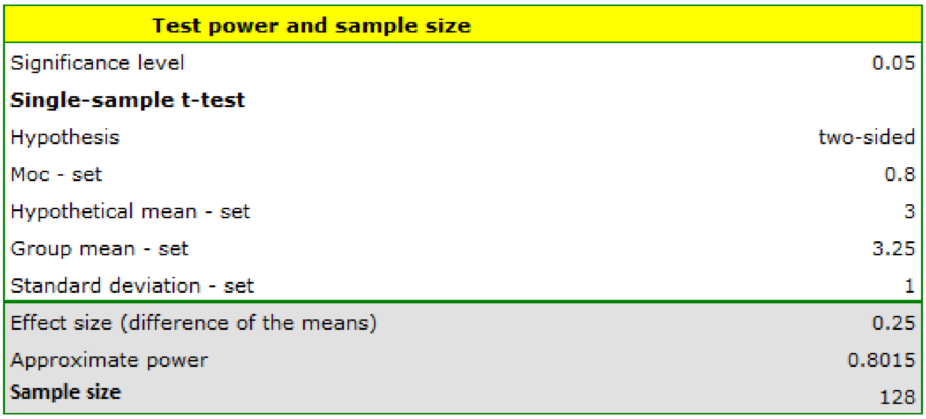
  Answer a) We know that the range from 2.5 to 3.5 days is within acceptable limits. Therefore, we use 3 days as the hypothetical mean and 2.5 days (or 3.5 days) as the test group mean.



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

  The resulting required sample size to prove that an effect exceeding 12 hours is statistically significant is 34 deliveries.

  Answer b) We know that the range from 2.75 to 3.25 days is within acceptable error. Therefore, we use 3 days as the hypothetical mean and 2.75 days (or 3.25 days) as the test group mean.



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

  The resulting required count to prove that an effect exceeding 6 hours is statistically significant is 128 deliveries.

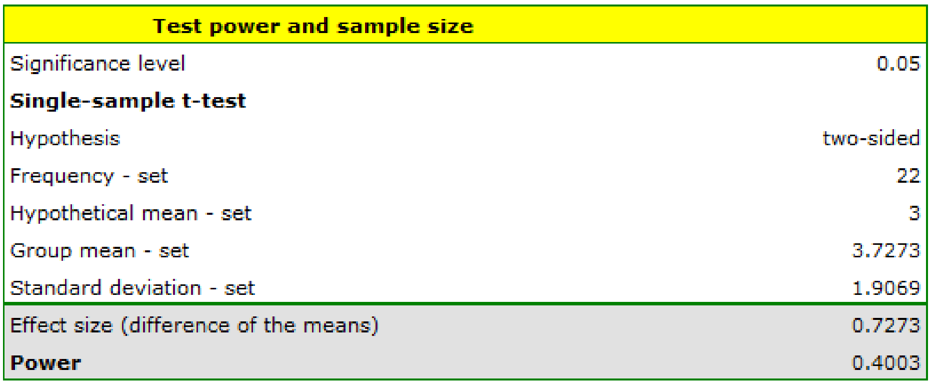
- **Testing a posteriori** - We collected data for the study - our sample includes 22 deliveries (data in file kurier.pqs).

  Based on the collected data, we determine the mean number of days to wait for delivery and the standard deviation of the group. In this case mean=3.727273, deviation=1.906925.

a) What is the power of the analysis performed?

b) What would the power look like if we increased the sample size to 100 elements while leaving the other assumptions unchanged?

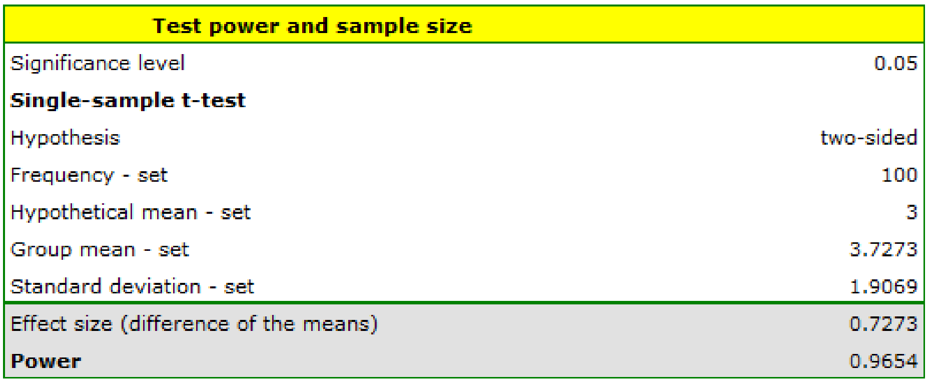
Answer a) The power of the analysis performed is only 0.400302.



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

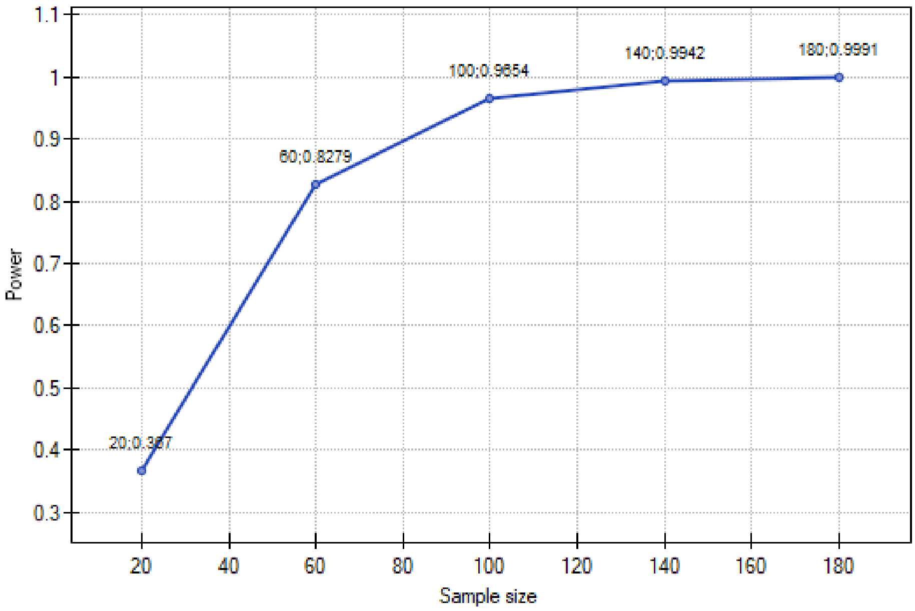
From this we know that many random samples with a sample size of 22 (about 60% of such samples) will not lead to confirmation of the alternative hypothesis.

Answer 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.



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

We can see how the power of the analysis will change with the sample size changing and the other assumptions unchanged in the chart.



7.2.2 T-test for dependent groups

Before determining the power or the required sample size of the t-test for dependent groups, it is worthwhile to familiarize yourself with the rules of its application.

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

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

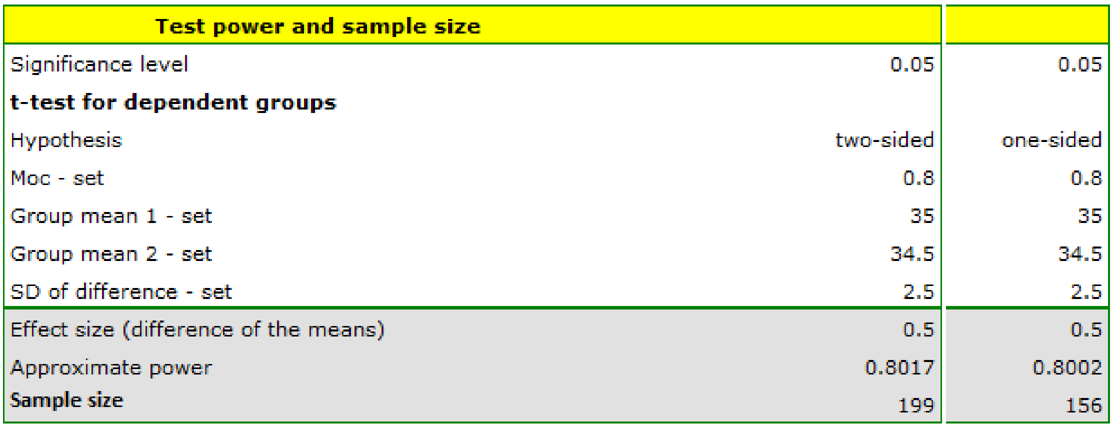
The set effect size, in this case, is the the difference between means that we expect to obtain in the population.

The power of the test and the required sample size are calculated based on the noncentral t-test distribution.

Example 7.4. We want to test (at the 0.05 significance level) 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?

Because we do not have data from the pilot study, we will provide the basic data for the calculations based on the experience and estimates of the clinic staff.
We assume that the average BMI of the treated person is 35 - such a value is entered in the box for the first mean. Since a change in BMI of less than half a unit is clinically insignificant, only a decrease below 34.5 (or an increase above 35.5) will be considered significant. Thus, we report a value of 34.5 (or 35.5) as the second mean. We presume that the standard deviation of the difference (BMI before and BMI

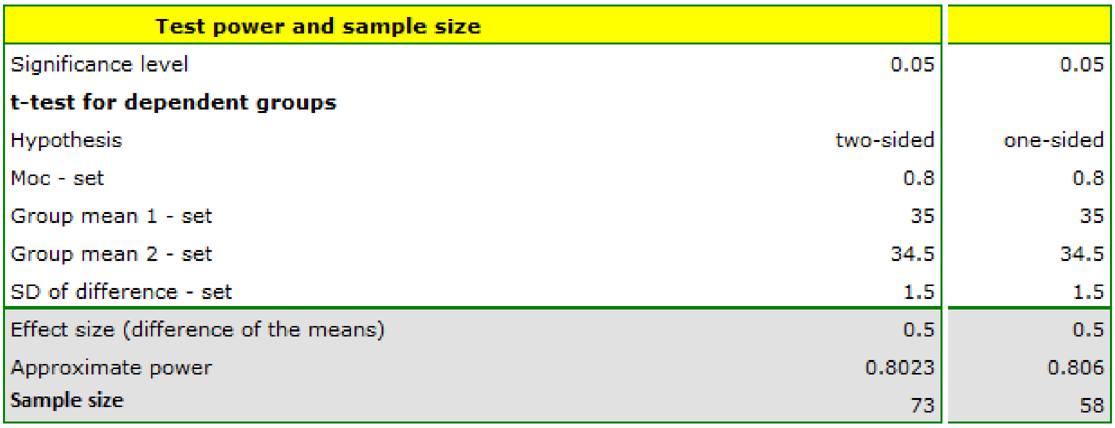
after) may be quite large, because usually the group will include people who are disciplined to follow a diet and those who still enjoy extra snacks between meals. Therefore, we set the deviation to 2.5. The power of the analysis we want to obtain is 80.



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

The resulting required sample size is 199 individuals when the hypothesis is two-sided (i.e., we assume that BMI may decrease or increase as a result of diet) or 156 individuals when the hypothesis is one-sided (i.e., we assume only a decrease in BMI).

If 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 i.e. 73 individuals for the two-sided hypothesis test and 58 individuals for the one-sided hypothesis.



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

7.2.3 T-test for independent groups

Before determining the power or the required sample size of the t-test for independent groups, it is worthwhile to familiarize yourself with the rules of its application.

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

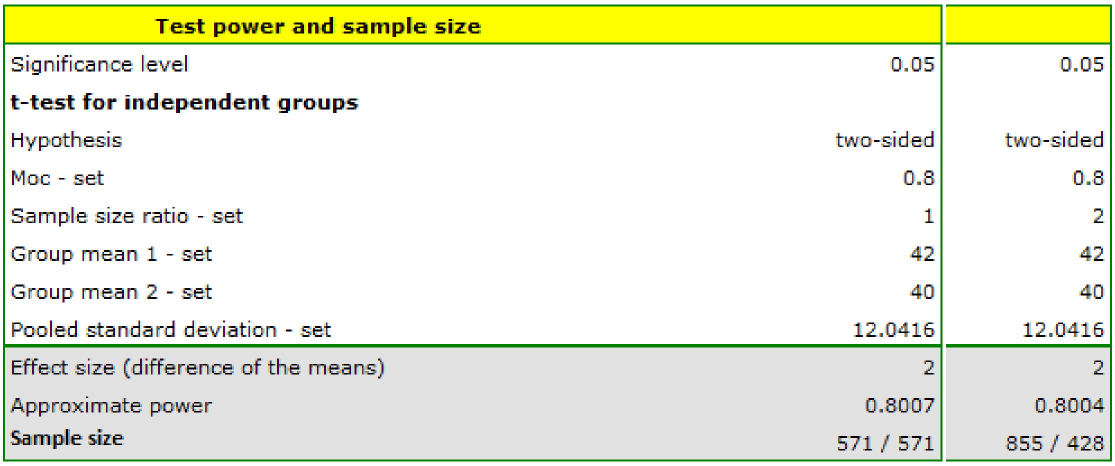
Power of the test	Group size
Group size 1 Group size 2	Power of the test
Group mean 1	
Group mean 2	
Common standard deviation (or deviations of both groups)	
Significance level α	

The set effect size, in this case, is the difference between means that we expect to obtain between populations.

The power of the test and the required sample size are calculated based on the noncentral t-test distribution.

Example 7.5. We are examining men with disease X and healthy men. We want to test (at the 0.05 significance level) whether the patients differ from the healthy ones in their HDL cholesterol levels. 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 using t-test for independent groups?

We report 40 mg/dL as the mean HDL for ill subjects and 42 mg/dL for healthy subjects. The ratio of the sample size of both groups is 1 because we assumed equal groups. We have data from the pilot study, hence we report the standard deviation for sick (13 mg/dl) and healthy (11 mg/dl) indicating 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

The resulting required sample size is 571 individuals when the groups are equolous (i.e., we assume due to n1/n2=1) or n1 = 855 and n2 = 428 when they are not equolous (i.e., we assume a sample size ratio of n1/n2=2)

7.2.4 Chi-square test for single sample variance

Before determining the power or the required sample size of the chi-square test for single sample variance, it is worthwhile to familiarize yourself with the rules of its application.

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

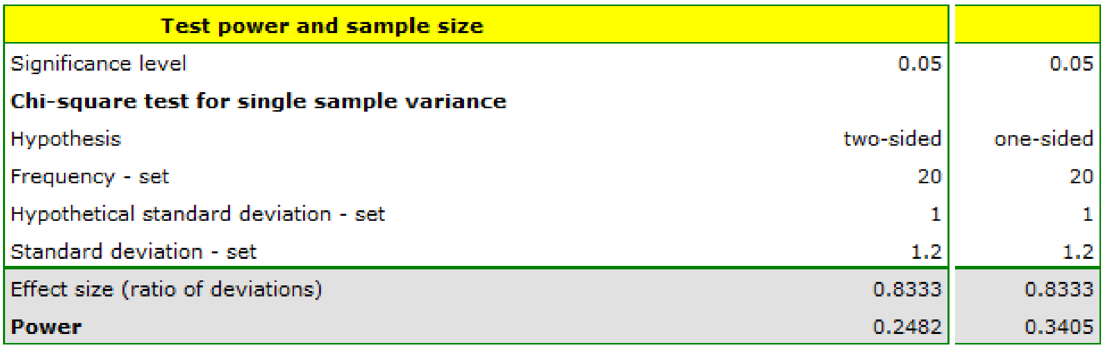
Power of the test	Group size
Group size	Power of the test
Hypothetical standard deviation	
Group standard deviation	
Significance level α	

The set effect size, in this case, is the quotient of the standard deviation of the test population and the hypothesized deviation.

The power of the test and the required sample size is calculated based on chi-square distribution.

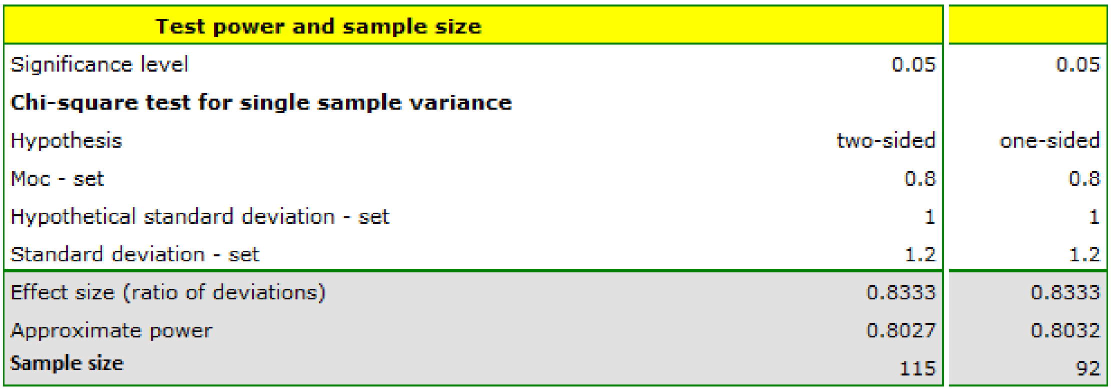
Example 7.6. Before producing another batch of certain cough syrup, control measurements should be taken of the amount of syrup poured into the bottles. The bottles should contain 200 ml of syrup. The technical documentation of the dosing device shows that the permissible variation in syrup volume measured by the standard deviation is 1 ml. Verify (at the 0.05 significance level) that the device under test is functioning properly. Will a sample of 20 bottles be sufficient to demonstrate excessive device error, if any? A standard deviation greater than 1.2 ml is considered excessive device error.

Since we expect the standard deviation for the dispenser to be as documented we enter a value of 1 ml as a hypothetical value. We will get too big error if the deviation exceeds 1.2 ml. We enter this value in the standard deviation box for the group of bottles we are going to test.
If the sample is 20 bottles, the resulting power using the two-sided hypothesis is only 0.25, and assuming the one-sided hypothesis is only 0.34. These are low values because less than 35% random samples of will detect a device error of 0.2 ml.



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

It must be recognized that 20 bottles, is too small of a group to prove a too high device standard deviation, if indeed there is one. 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, which in this case will be 115 for the two-sided hypothesis and 92 for the one-sided hypothesis.



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

7.2.5 Chi-square test of two variances Fisher-Snedecor

Before determining the power or the required sample size of the Fisher-Snedecor test, it is worthwhile to familiarize yourself with the rules of its application.

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

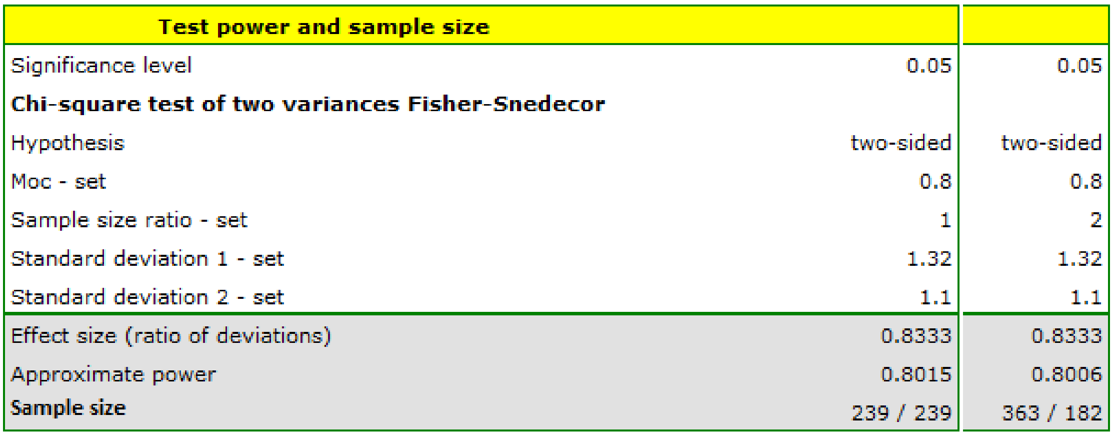
Power of the test	Group size
Group size 1 Group size 2	Power of the test
Group 1 standard deviation	
Group 2 standard deviation	
Significance level α	

The set effect size, in this case, is the quotient of the standard deviation of populations one and two.

The power of the test and the required sample size is calculated based on the F-Snedecor distribution.

Example 7.7. Before producing another batch of certain cough syrup, control measurements should be taken of the amount of syrup poured into the bottles. There should be 300 ml of syrup in the bottles. Two dispensing devices are used in the bottling plant. We want to test (at the 0.05 significance level) whether the distribution of syrup volume as measured by the standard deviation for the two devices is the same. A small pilot study was conducted and the standard deviation for the first device was found to be 1.32 and for the second device 1.1. If the difference is small, i.e., the quotient 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 show that a ratio of 1.2 is statistically significant?

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



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

The resulting sample size for each device is n1=n2=239, assuming equal groups (i.e., ratio n1/n2=1) and n1=363 and n2=182 assuming twice the sample size for the first device (i.e., ratio n1/n2=2).

7.2.6 Chi-square test (goodness of fit)

Before determining the power or the required sample size of the chi-square test (goodness of fit), it is worthwhile to familiarize yourself with the rules of its application.

To determine the power of the test and the required sample size, we need[45]:

Power of the test	Group size
Group size Number of categories	Power of the test
Effect size ϕ	
Significance level α	

The set effect size, or ϕ, in this case is the root of the quotient of the chi-square test statistic and the group size:

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

The power of the test and the required sample size are calculated based on a noncentral chi-square distribution.

7.2.7 Chi-square test (RxC)

Before determining the power or the required sample size of the Chi-square test RxC it is worthwhile to familiarize yourself with the rules of its application.

To determine the power of the test and the required sample size, we need[45][127]:

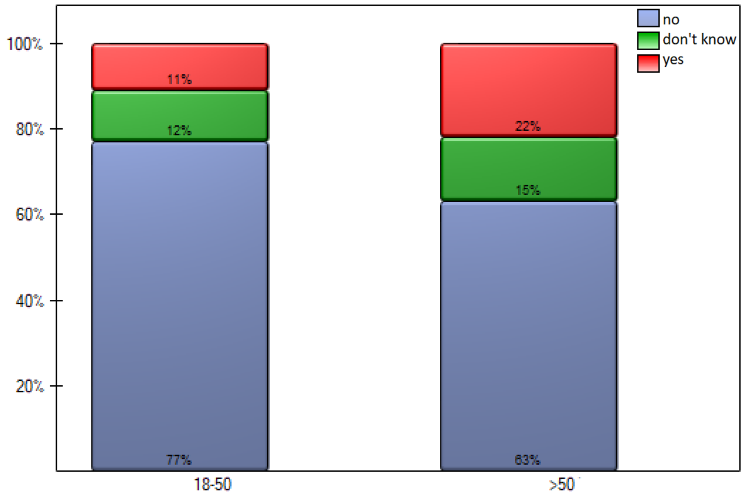
Power of the test	Group size
Group size Number of categories (rows) Number of categories (columns)	Power of the test
Effect size ϕ	
Significance level α	

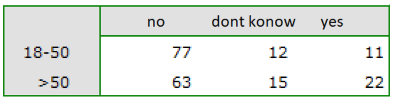
The set effect size, or ϕ, in this case is the root of the quotient of the chi-square test statistic and the group size:

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

The power of the test and the required sample size are calculated based on a noncentral chi-square distribution.

Example 7.8. There are plans to conduct a large survey showing the knowledge of the Polish population about ways of fighting common viruses. The project should determine whether educational activities informing about the ineffectiveness of antibiotic therapy in viral infections were as effective in the older age group (i.e. over 50 years) as in younger adults (18-50 years). A pilot study was conducted and a random sample of 200 people was asked the question, ”Do antibiotics fight viruses?” Respondents were asked the choose one of three answers: ”yes”, ”no” or ”don’t know” . The results of the pilot study were prepared for publication. The following is an excerpt from the description included in the paper:





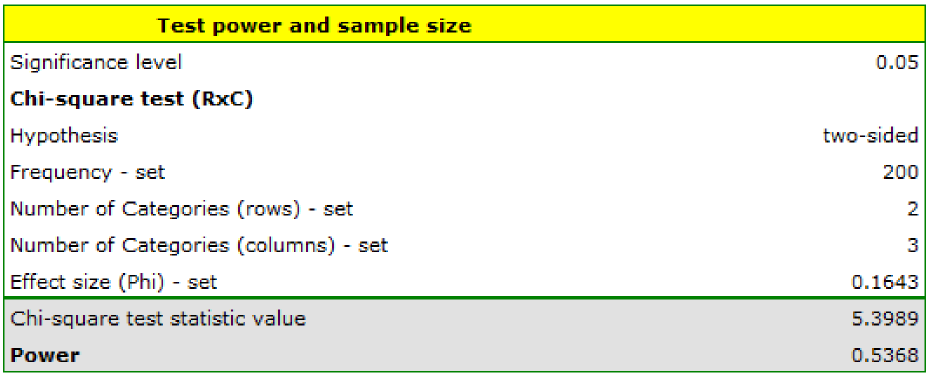
	no	dont konow	yes
18-50	77	12	11
>50	63	15	22

The obtained p-value in the chi-square test was statistically insignificant p=0.0672.

The paper reviewer was rightfully surprised to learn that twice as many people over the age of 50 incorrectly indicated that antibiotics fight viruses (22% vs. 11%), but this difference was not statistically significant.

As suggested by the reviewer, one should check whether the lack of statistical significance for this difference is due to the power of the test being too low, and state how large the sample should be to obtain a chi-square test power of 80% for the same percentages?

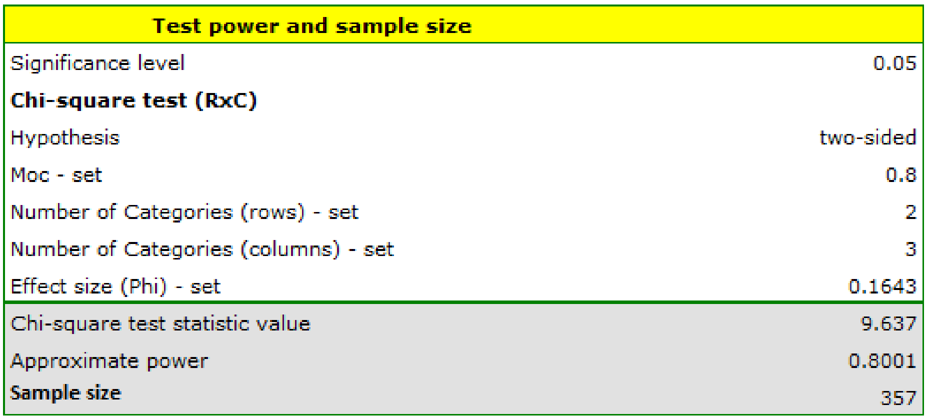
Preparing a response for the reviewer
We will determine the ϕ coefficient for the test (menu: Chi-square, Fisher, OR/RR→Correlation coefficients...→Phi). We obtain ϕ=0.1643.
Using a 200-element sample, with data placed in a table with two rows and three columns, and a given coefficient value of ϕ, we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

The power obtained in this analysis is low at 0.5368, which seems to confirm concerns about undersampling.

If we get the same distribution of data for a sample with a different sample size, that means we also get the same coefficient ϕ. To determine the sample size that would give us an 80% power of chi-square test, we again give 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

We obtain information that a sample size of 357 respondents will be needed. Since this is only a pilot study, we plan to increase the sample size to 357 respondents in the actual study.
However, we can already see that when omitting the undecided (i.e., those who chose the answer "don't know") and redoing the analysis, significant differences can be found (chi-square, p=0.0251). In the decided group, the percentages choosing the wrong answer are more than doubled to the detriment of those aged >50 years (12.5% vs 25.9%).

8 Two independent proportions, chi-square (2x2)

Before determining the power or the required sample size of the chi-square (2x2) and Z test for two independent proportions It is worthwhile to familiarize yourself with the rules of their application.

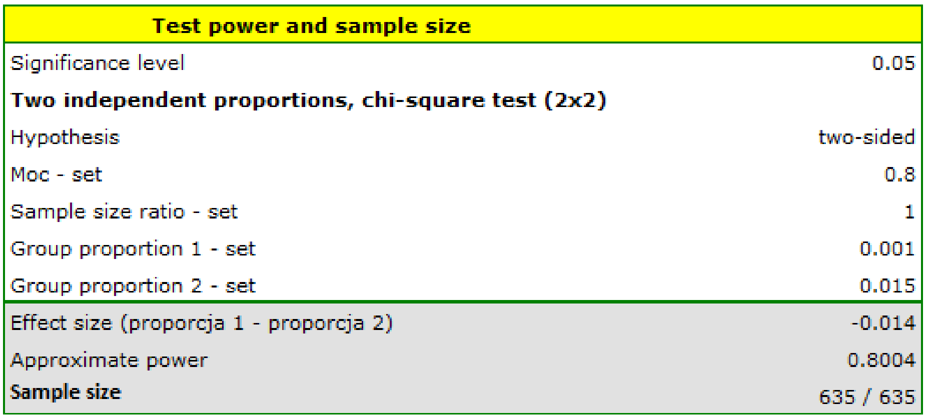
To determine the power of the test and the required sample size, we need[35]:

Power of the test	Group size
Group size	Power of the test
The proportion in the first group	
The proportion in the second group	
Significance level α	

The set effect size, is the difference between the highlighted proportions.

The test value and the required sample size are calculated based on the normal distribution normal distribution.

Example 8.1. Consider a study evaluating the effectiveness of aspirin in reducing mortality from myocardial infarction. Previous studies indicate that the rate of death from myocardial infarction is 0.015 for nonusers and 0.001 for aspirin users. The researchers want to determine the minimum sample size required to detect an absolute difference | 0.001-0.015 | = 0.014 at 80% power using a two-sided test with 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
```

Assuming the groups are equal, 635 people need to be picked for each group.

8.0.1 One-way ANOVA for independent groups

Before determining the power or the required sample size of the One-way ANOVA for independent groups It is worthwhile to familiarize yourself with the rules of its application.

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

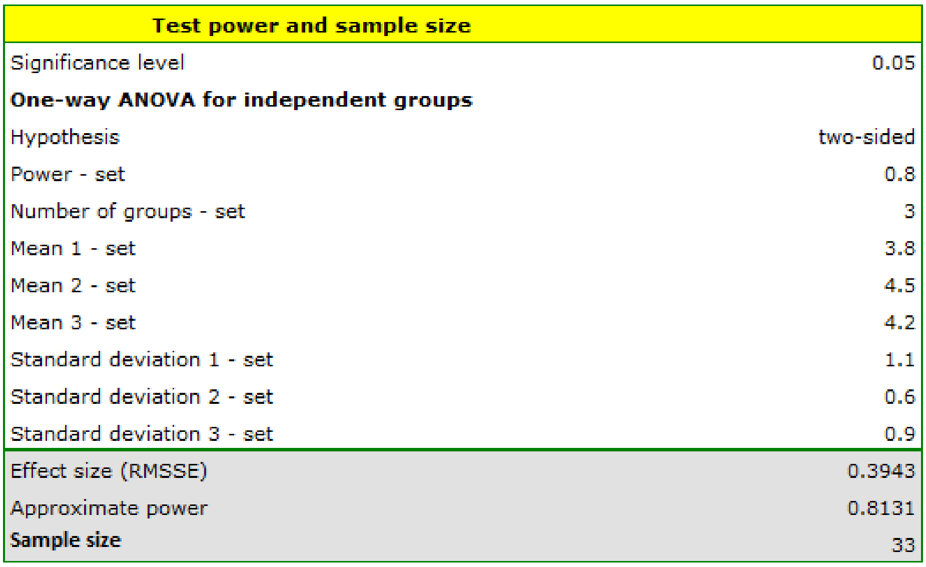
Power of the test	Group size
Group size	Power of the test
Group 1 mean	
Group 2 mean	
Group 3 mean	
Group 4 mean	
Group 1 standard deviation	
Group 2 standard deviation	
Group 3 standard deviation	
Group 4 standard deviation	
Significance level α	

The set effect size, in this case, is the RMSSE, which is the standardized measure used in ANOVA to describe the overall level of effect in the population.

The power of the test and the requiredy sample size are calculated based on the non-central F-Snedecor distribution.

Example 8.2. 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 test this, a pilot study was first conducted. Based on the results of this study, the predicted effect sizes were determined, 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 gathered if the quantities remain the same to prove that there are statistically significant differences?

We enter values for means and standard deviations. The resulting sample size for each of the study groups is 33, assuming an 80% power of the test.



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

8.0.2 Test for one proportion

Before determining the power or the required sample size of the test for one proportion it is worthwhile to familiarize yourself with the rules of its application.

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

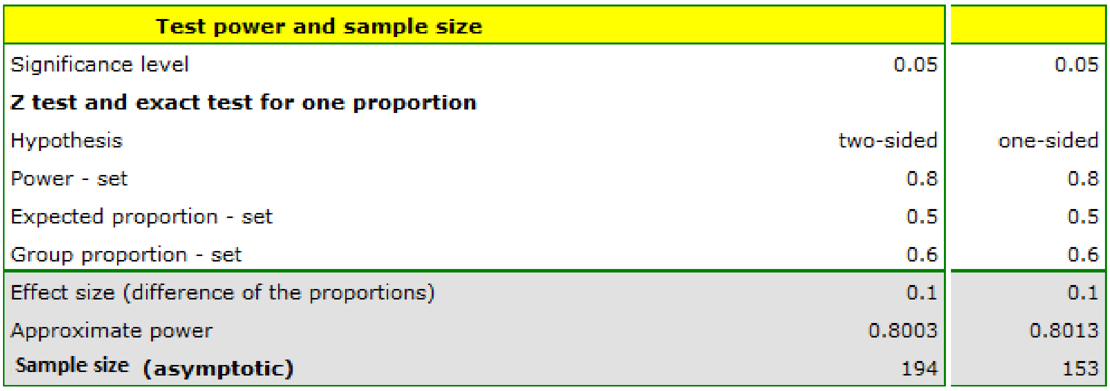
Power of the test	Group size
Group size	Power of the test
Expected proportion	
Group proportion	
Significance level α	

The set effect size, in this case, is the amount of difference between the set proportion in the study population and the hypothetical expected proportion.

The power of the test 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.

Example 8.3. 10 randomly selected families with children under the age of 10 and residing in Poznań were asked about the plans for their children's education. Of these, 6 families planned to educate their children at university. At a significance level of 0.05 and a test power of 0.8, how large a sample would we need to collect to conclude that more than 50% of the families in Poznań with children under the age of 10 are already planning for their children to attend university in the future?

We enter 0.5 as the expected proportion and 0.6 as the group proportion. The resulting sample size is 194 families-when the hypothesis being tested is two-sided or 153 families-when the hypothesis is one-sided.



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

9 DESCRIPTIVE ANALYSES

10 DESCRIPTIVE ANALYSES

The data collected by the researcher should first be described. Depending on how the measurements are made (on the measurement scale), different measures will be used to describe the variable.

10.1 MEASURING SCALES

The correct determination of the type of analysis to be performed depends on the scale used to represent the collected data. There are 3 main measurement scales:

1. **interval scale**

   A variable is represented on an interval scale when:

   - it can be organized,
   - it can be calculated by how much one element is larger than the other and the difference of these elements has a real world interpretation. Usually a unit of measurement is specified.

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

2. **Ordinal scale**

   A variable is represented on an ordinal scale when:

   - it can be organized, i.e. the order in which the elements appear matters,
   - there is no meaningful way to determine the difference or quotient between two values.

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

Note!
If a variable is expressed on an ordinal scale, then to be able to perform calculations on it properly it should be saved using numbers. These numbers are conventional identifiers that tell you the order of the elements.

3. **Nominal scale**

   A variable is expressed on a nominal scale when:

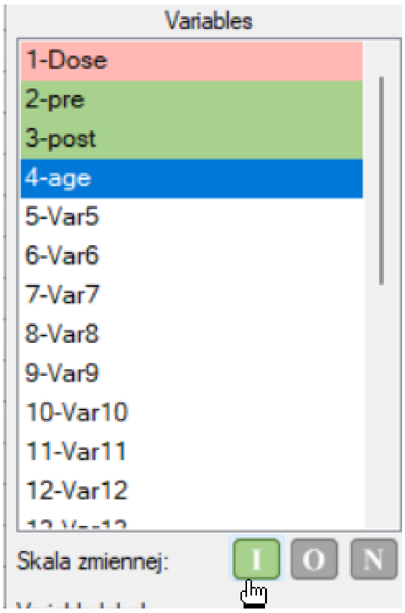
   - it cannot be ordered, i.e. there is no order resulting from the nature of a given event.,
   - The difference or quotient between two values cannot be determined in a meaningful way.

   Example: gender, country of residence, etc.

   Note!
   If a variable is expressed on a nominal scale, it can be stored using text labels. Even if the values of a nominal variable are expressed numerically, the numbers are only conventional identifiers, so you cannot perform arithmetic operations on them or compare them.

Before proceeding with analysis, it is recommended to assign measurement scales to individual variables. Such assignment will result in the variable headers gaining the corresponding color for the scale, i.e., green = **interval scale**, yellow = **ordinal scale**, red = **nominal scale**. The color of the variables (and therefore their scale) will be visible in the data sheet and in the list of variables in the analysis windows.

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



Quantitative data are 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.
Qualitative data are descriptive in nature, expressed in linguistic terms rather than in numerical values. Qualitative data include a nominal scale and sometimes also an ordinal scale.

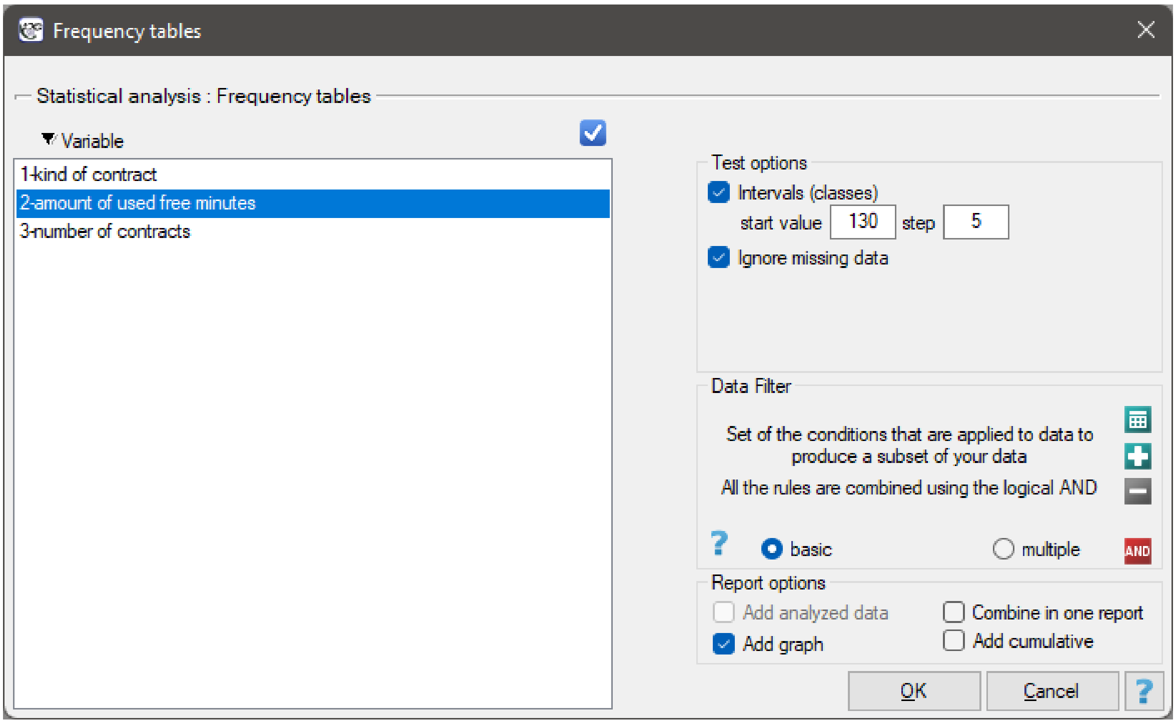
An ordinal scale that has a possible number of objectified categories is quantitative data, e.g., the SF-36 quality of life scale (from 0 to 100 points), but if there are so few categories that can be described by text, e.g., education (primary, secondary, tertiary), then it is qualitative data.

10.2 TABLES

10.2.1 FREQUENCY TABLES AND EMPIRICAL DISTRIBUTION OF THE DATA

The basis of statistical research is the determination of the **empirical distribution**, i.e., the distribution of a feature observed in a sample. The empirical distribution is determined by assigning a frequency of occurrence to successive values of the feature. Such distribution can be presented in the form of **frequency table** or as a graph (histogram). For small data sets, frequency tables can present all data - the so-called point distribution series, while for larger data sets the so-called interval distribution series are created.

To represent the data distribution in table form, bring up the Frequency tables window by selecting menu Statistics→Descriptive analysis→Frequency tables.



Frequency tables

Statistical analysis : Frequency tables

Variable ☑

1-kind of contract
2-amount of used free minutes
3-number of contracts

Test options
☑ Intervals (classes)
start value 130 step 5
☑ Ignore missing data

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
□ Add analyzed data □ Combine in one report
☑ Add graph □ Add cumulative

In this window we choose a variable to analyze and options for analysis. You can sort the output as a text or as a number by selecting the appropriate options. If there are empty cells in the analyzed column, they may be included or omitted in the analysis. The result of analysis will be placed in report attached to datasheet, for which analysis has been done.

In addition, if you want the data to be visualized with a column chart or histogram, then in the Frequency table window, check the Add graph option..

Example 10.1. (distribution.pqs file)

A mobile operator conducts a series of surveys on how customers use the number of "free minutes" they are given in their subscription. Customers can use up to 190 such minutes each month. The study was based on a random sample of 200 customers. Information analyzed included:
- type of subscription purchased,
- number of free minutes used,

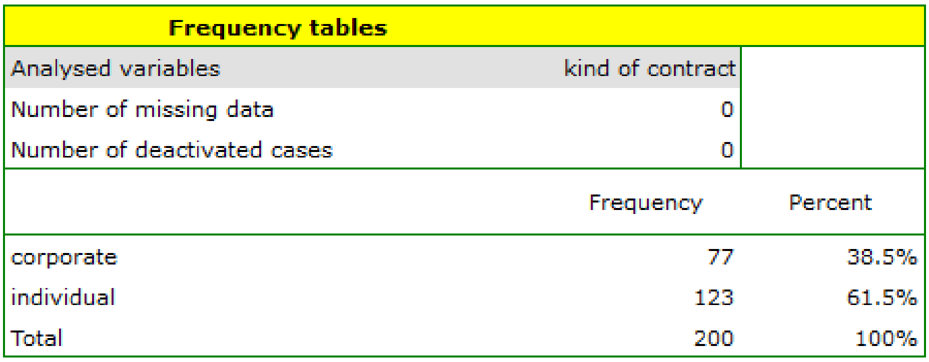
- number of subscriptions registered for a given customer (does not apply to companies).

We want to present the distribution of:

1. type of subscription,
2. number of free minutes used,
3. number of registered subscriptions for an individual person.

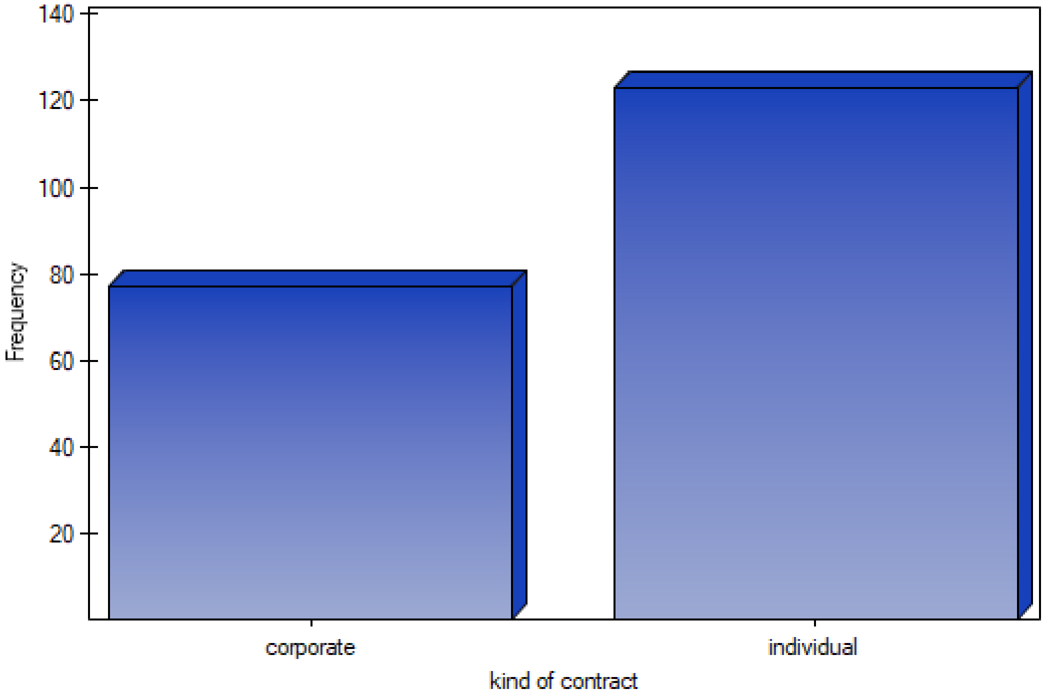
Open the Frequency table window..

1. Select the Variable to analyze: ”type of subscription” and Add graph. Then confirm the selected settings with OK button and the result is obtained as a report:

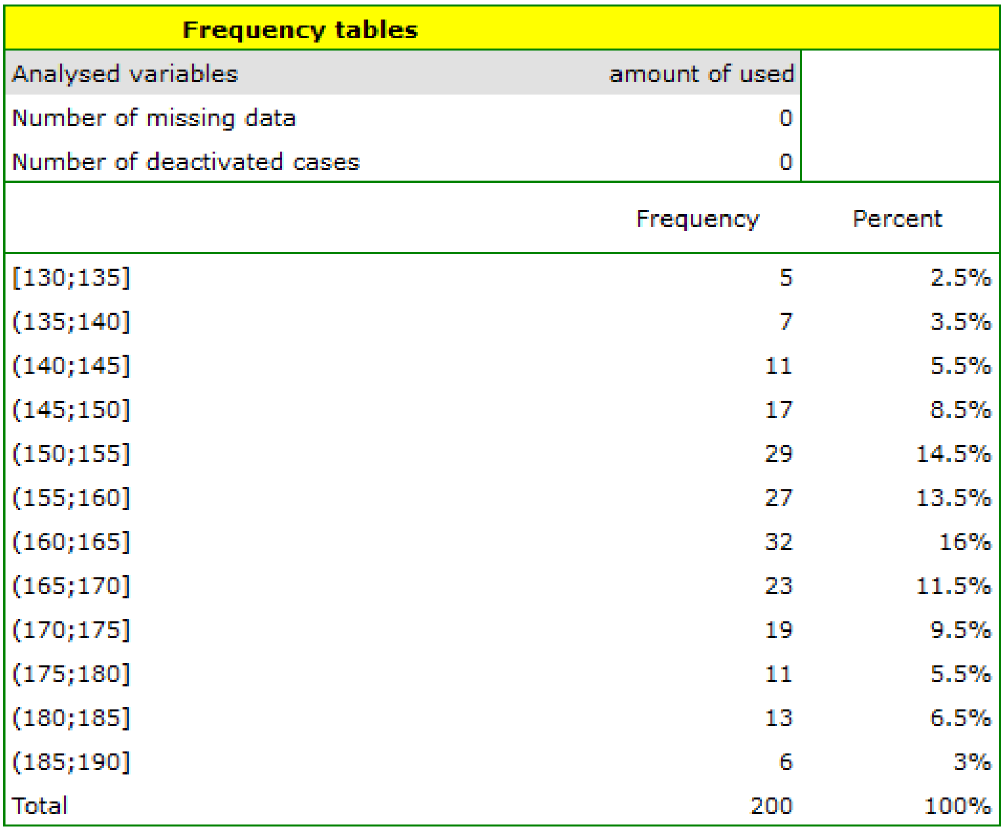


Frequency tables

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%



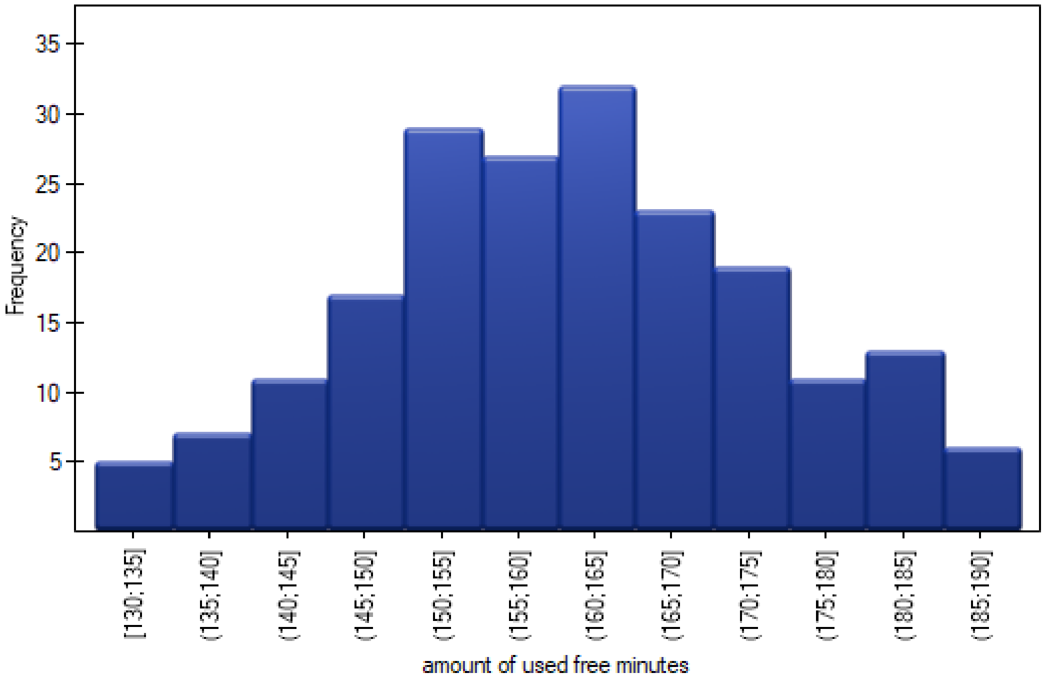
2. Resume Analysis by pressing Run the recent test. We select the variable to analyze: ”amount of used free minutes” and check the option Intervals (classes), set start value for example to 130 and step to 5. We can also check the option Add graph. Then confirm the selected options with OK and the result is obtained as a report:



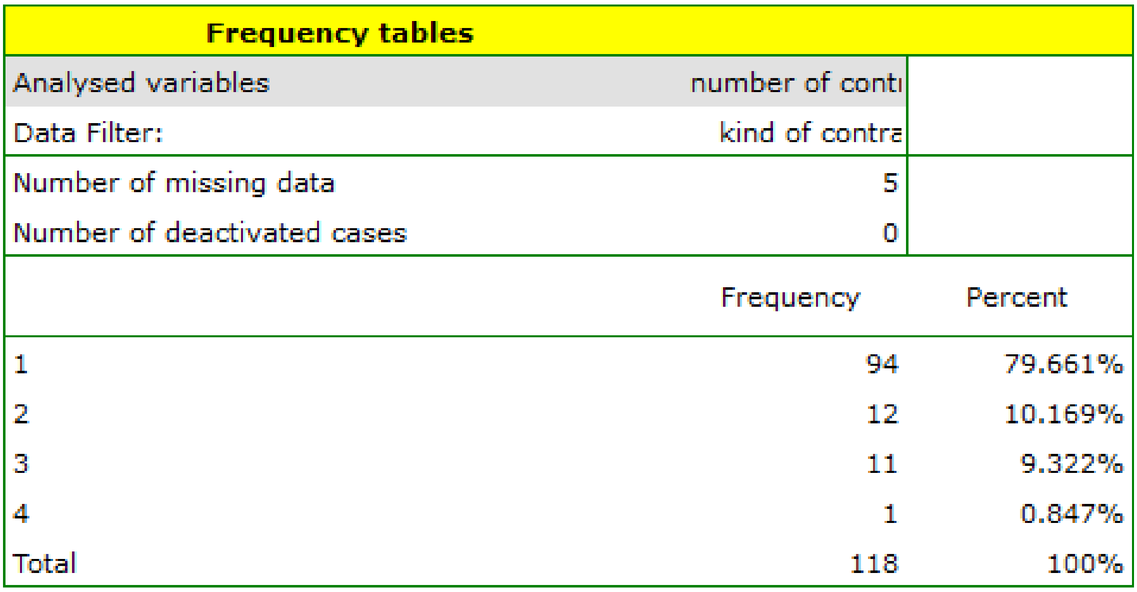
Frequency tables

Analysed variables	amount of used
Number of missing data	0
Number of deactivated cases	0

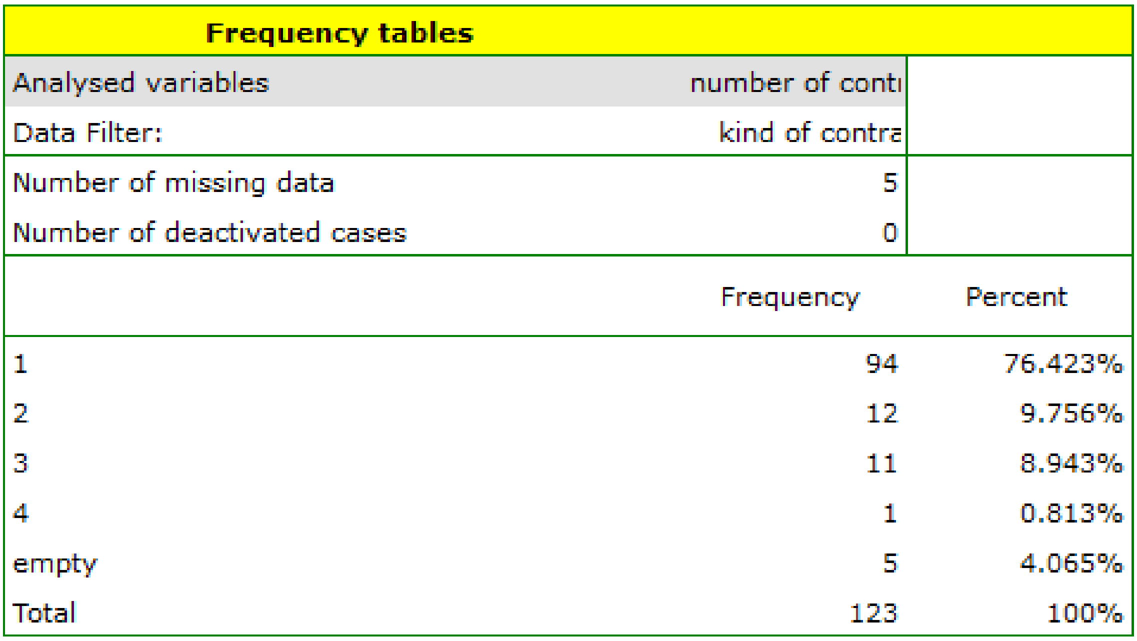
	Frequency	Percent
[130;135]	5	2.5%
(135;140]	7	3.5%
(140;145]	11	5.5%
(145;150]	17	8.5%
(150;155]	29	14.5%
(155;160]	27	13.5%
(160;165]	32	16%
(165;170]	23	11.5%
(170;175]	19	9.5%
(175;180]	11	5.5%
(180;185]	13	6.5%
(185;190]	6	3%
Total	200	100%



3. Resume Analysis by pressing Run the recent test. We set filter so that the analysis is performed only for individuals. We select the variable to analyze: "Number of subscriptions". Since this variable also contains missing data, the result obtained may or may not include these missing cases in the analysis, depending on the option selected:



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%

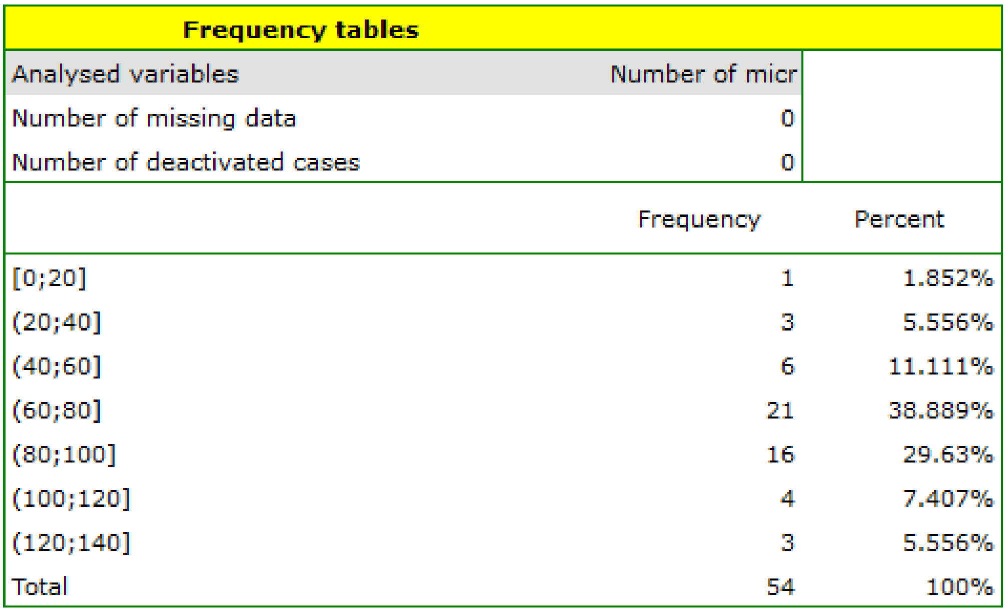


Frequency tables		
Analysed variables	number of conti	
Data Filter:	kind of contra	
Number of missing data	5	
Number of deactivated cases	0	
	Frequency	Percent
1	94	76.423%
2	12	9.756%
3	11	8.943%
4	1	0.813%
empty	5	4.065%
Total	123	100%

Example 10.2. (fertiliser.pqs file)

An experiment was conducted to study the microbiological condition of soil under perennial ryegrass cultivation supplied with biologically active fertilizers. Soils were fertilized with different types of microbial preparations and fertilizers and then the number of microorganisms present per gram of soil dry matter was calculated. We want to know the frequency of actinomycetes per 1 gram of dry nitrogen fertilized soil. We are interested in how often 0 to 20 actinomycetes were present in the sample, more than 20 to 40 actinomycetes, more than 40 to 60 actinomycetes, etc. We select only the first 54 rows in the datasheet that match the assumptions of the analysis (these are nitrogen-fertilized actinomycetes) and open the Frequency Tables.

In the options window, we select the variable to be analyzed: Number of microorganisms, and then set the class intervals so that the start value is 0 and the step is 20. You should see a message at the top of the window: Data limited by selection . Confirm the selection with the OK button and the result should appear as a report:



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%

10.2.2 TABLE REPORT

Using a table report, you can prepare a simultaneous summary of a large amount of data in the form of bivariate tables (tables of two features). For example, we can present the distribution of age groups by place of residence, education, etc. in the form of a table. Each table is presented in the form of frequency in particular categories, and additionally, it can be summarized by calculating percentages from a row, from a column, or from the total sum, and determining the frequency table expected. In addition, automatic summaries in the form of a column chart are possible for such tables. The window with the table report settings is opened via menu Statistics→Descriptive analysis→Table report

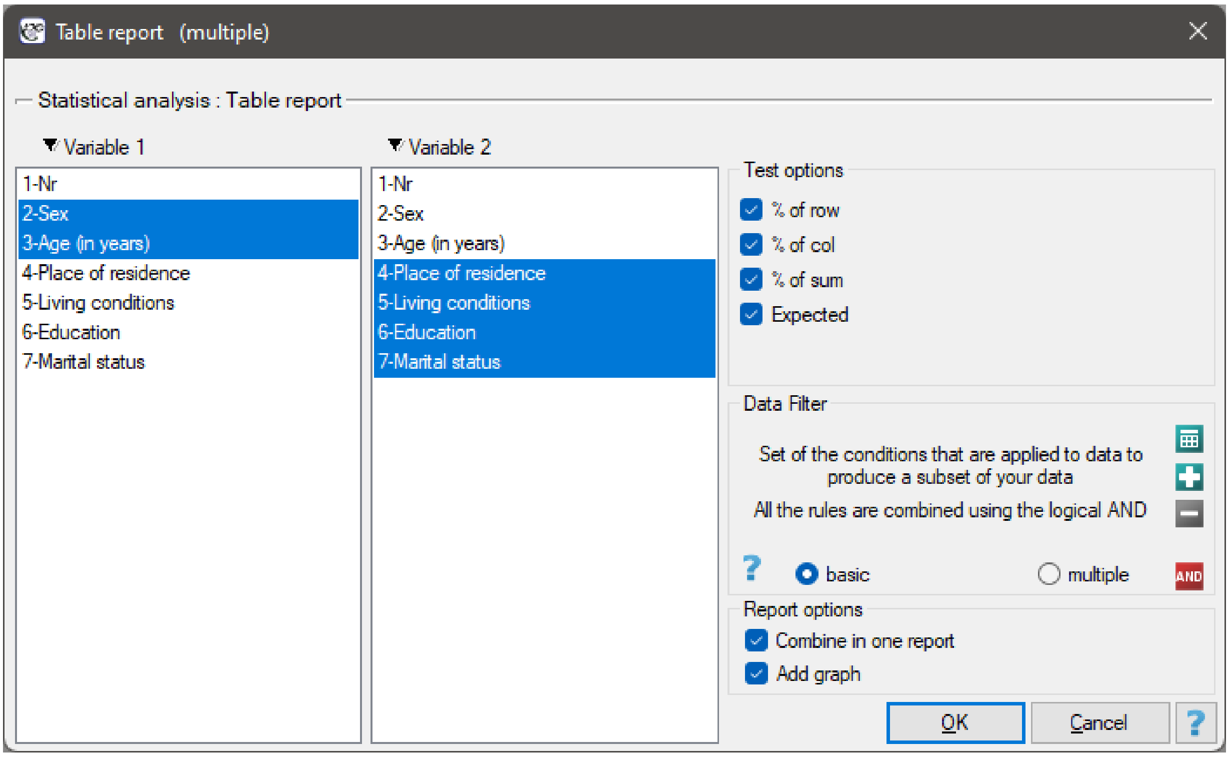


Table report (multiple)

Statistical analysis : Table report

Variable 1: 1-Nr, **2-Sex**, **3-Age (in years)**, 4-Place of residence, 5-Living conditions, 6-Education, 7-Marital status

Variable 2: 1-Nr, 2-Sex, 3-Age (in years), **4-Place of residence**, **5-Living conditions**, **6-Education**, **7-Marital status**

Test options
- ☑ % of row
- ☑ % of col
- ☑ % of sum
- ☑ Expected

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

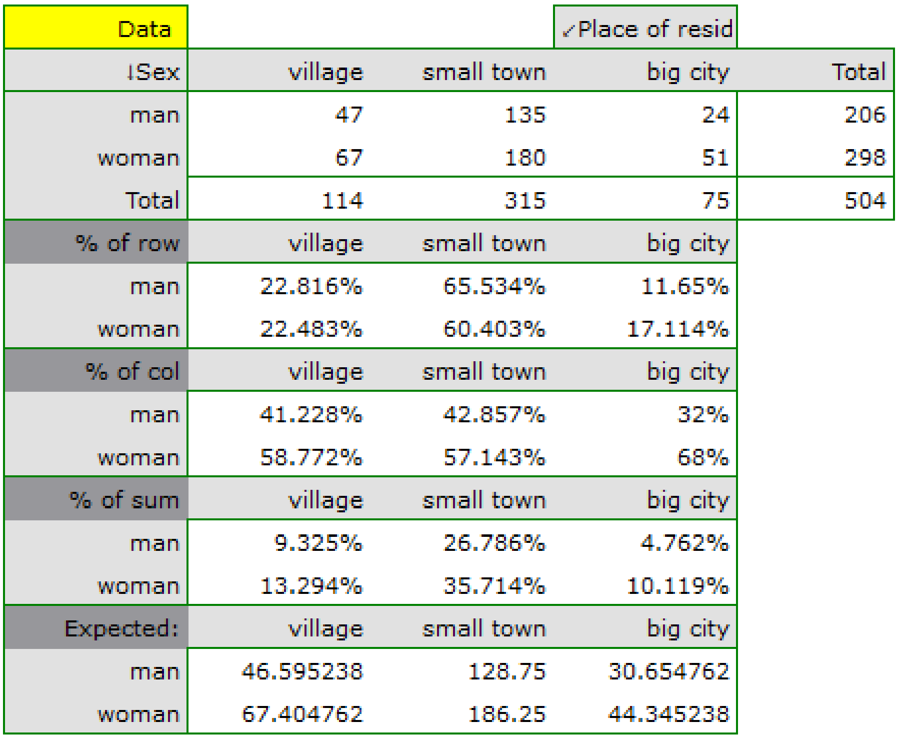
All the rules are combined using the logical AND

● basic ○ multiple

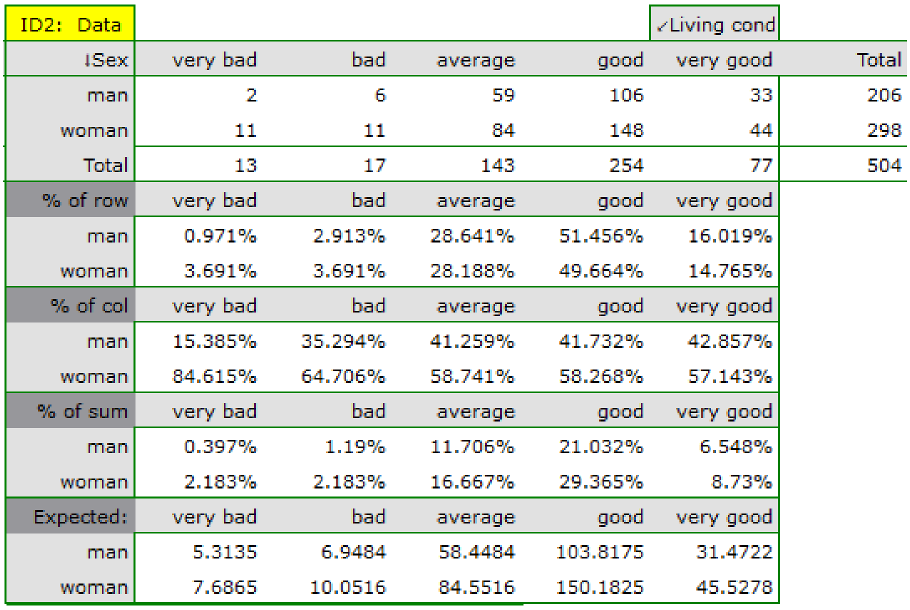
Report options
- ☑ Combine in one report
- ☑ Add graph

EXAMPLE 10.3. (Tables.pqs file)

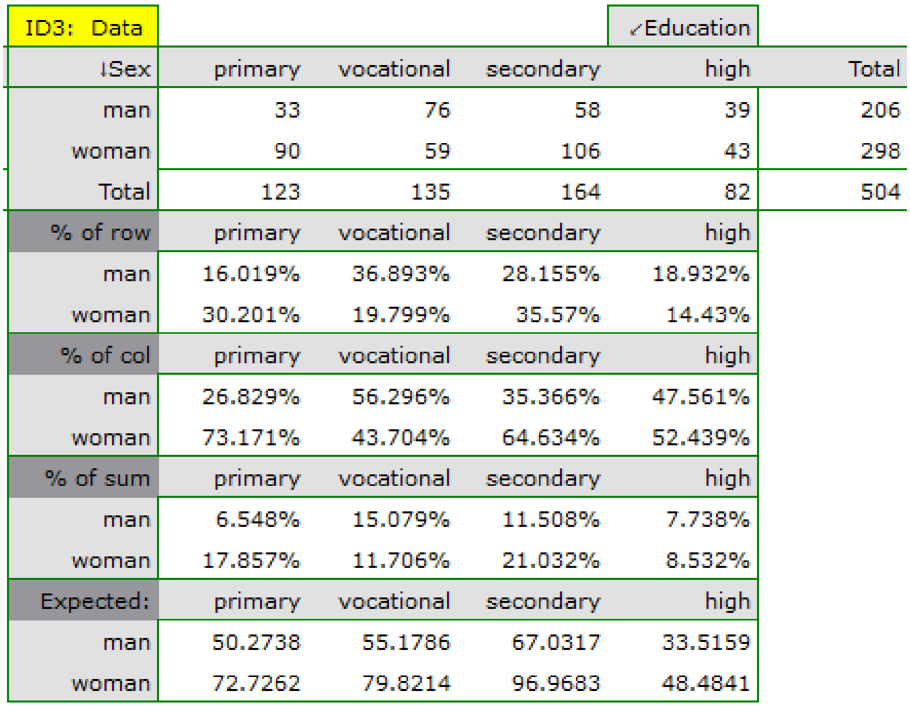
In the form of tables, we need to summarize the distribution of gender by place of residence, social and living conditions, education, marital status, and the distribution of age groups with respect to the same characteristics. This will result in 4 tables for each pair of traits, or 8 tables for all pairs and corresponding graphs. Only the distribution with respect to gender is presented below:



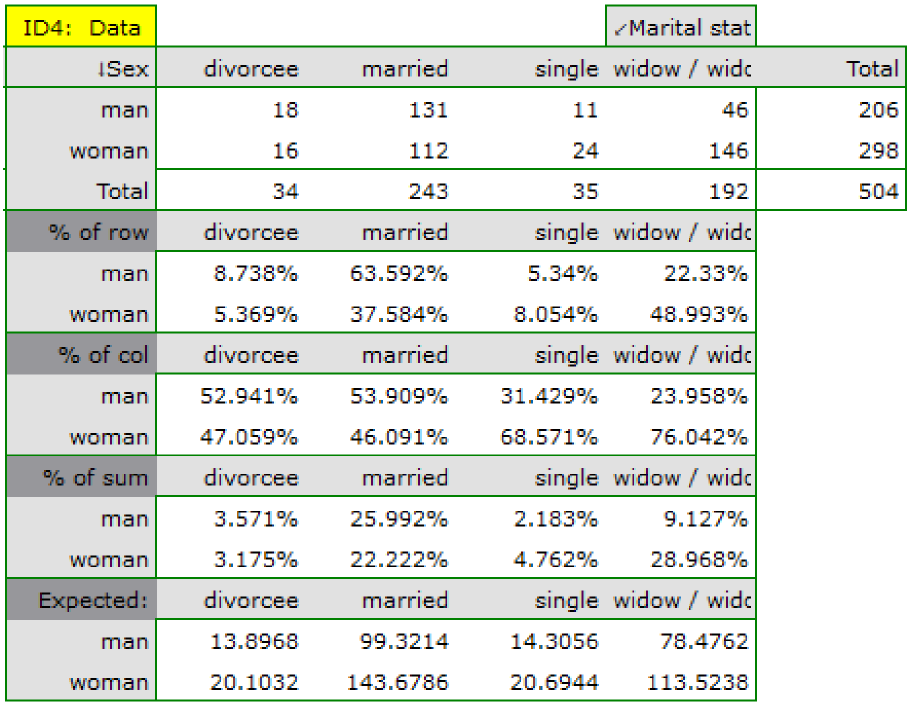
Data			✓Place of resid	
↓Sex	village	small town	big city	Total
man	47	135	24	206
woman	67	180	51	298
Total	114	315	75	504
% of row	village	small town	big city	
man	22.816%	65.534%	11.65%	
woman	22.483%	60.403%	17.114%	
% of col	village	small town	big city	
man	41.228%	42.857%	32%	
woman	58.772%	57.143%	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	



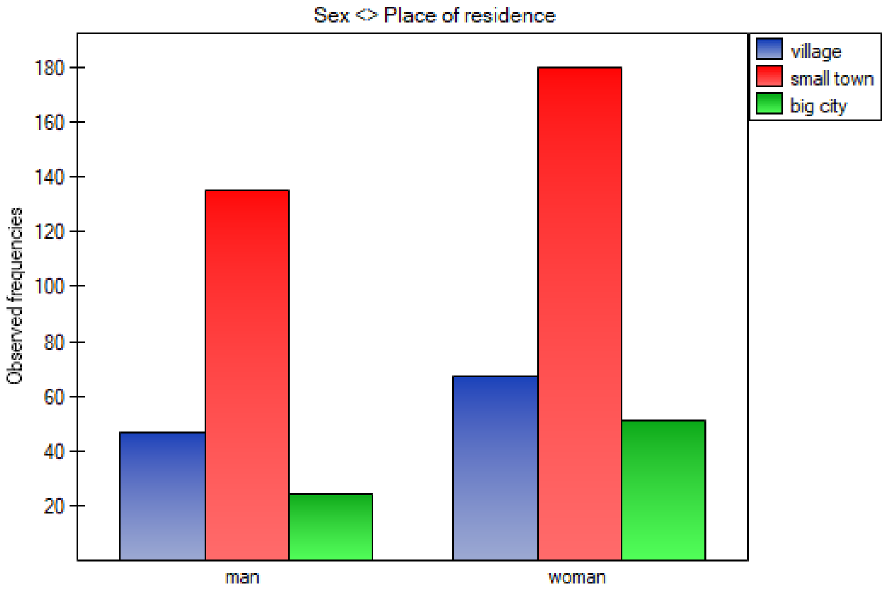
ID2: Data					✓Living cond	
↓Sex	very bad	bad	average	good	very good	Total
man	2	6	59	106	33	206
woman	11	11	84	148	44	298
Total	13	17	143	254	77	504
% of row	very bad	bad	average	good	very good	
man	0.971%	2.913%	28.641%	51.456%	16.019%	
woman	3.691%	3.691%	28.188%	49.664%	14.765%	
% of col	very bad	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	

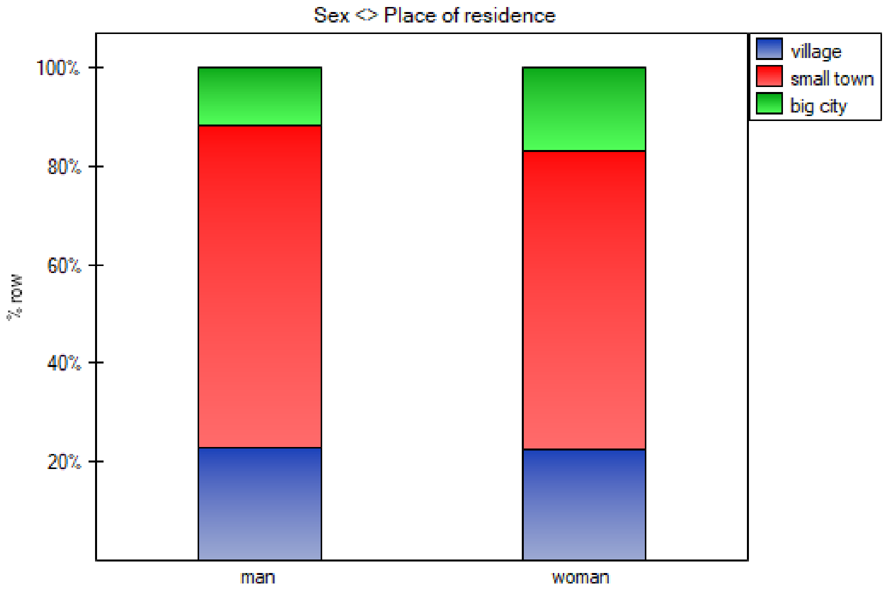


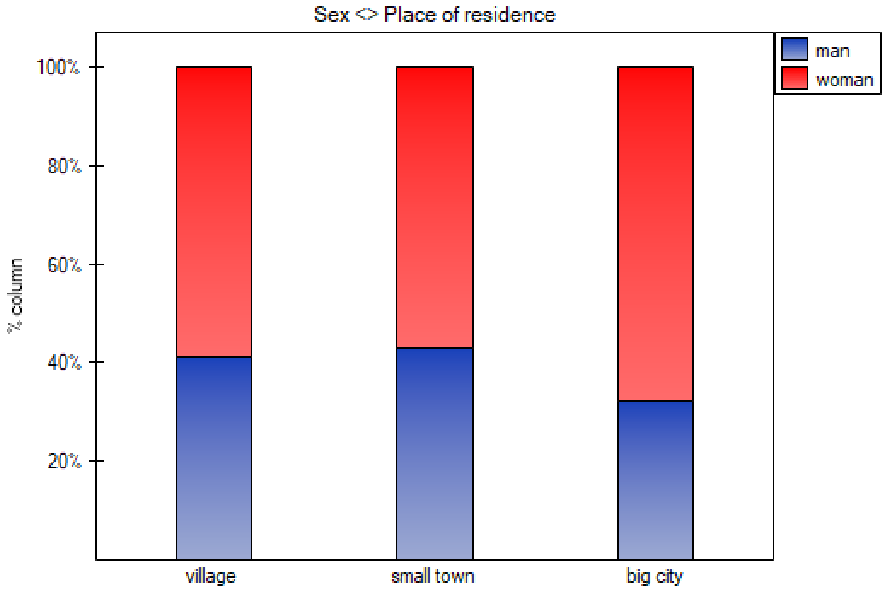
ID3: Data				✓Education	
↓Sex	primary	vocational	secondary	high	Total
man	33	76	58	39	206
woman	90	59	106	43	298
Total	123	135	164	82	504
% of row	primary	vocational	secondary	high	
man	16.019%	36.893%	28.155%	18.932%	
woman	30.201%	19.799%	35.57%	14.43%	
% of col	primary	vocational	secondary	high	
man	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	

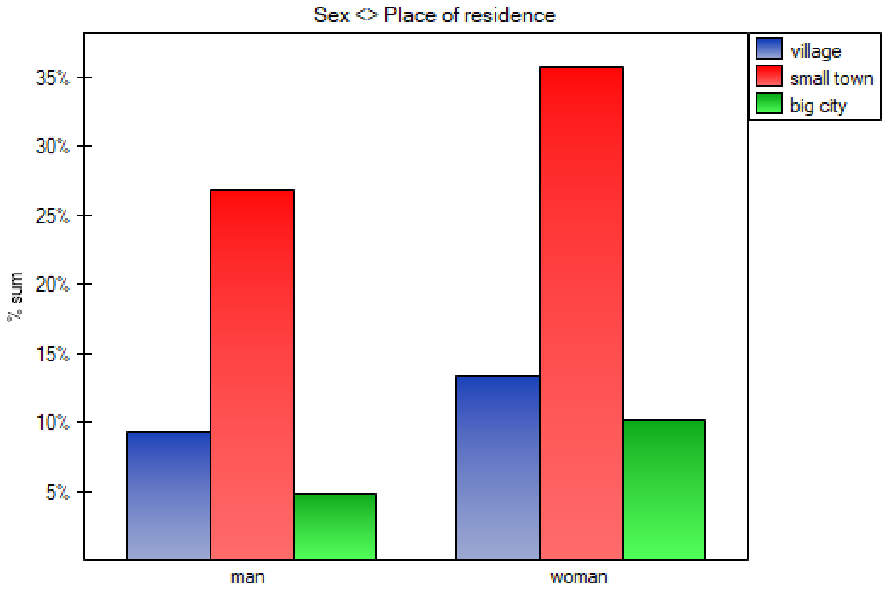


ID4: Data				↙Marital stat	
↓Sex	divorcee	married	single	widow / widc	Total
man	18	131	11	46	206
woman	16	112	24	146	298
Total	34	243	35	192	504
% of row	divorcee	married	single	widow / widc	
man	8.738%	63.592%	5.34%	22.33%	
woman	5.369%	37.584%	8.054%	48.993%	
% of col	divorcee	married	single	widow / 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	




Sex <> Place of residence
village
small town
big city
% row
100%
80%
60%
40%
20%
man
woman


Sex <> Place of residence
man
woman
% column
100%
80%
60%
40%
20%
village
small town
big city



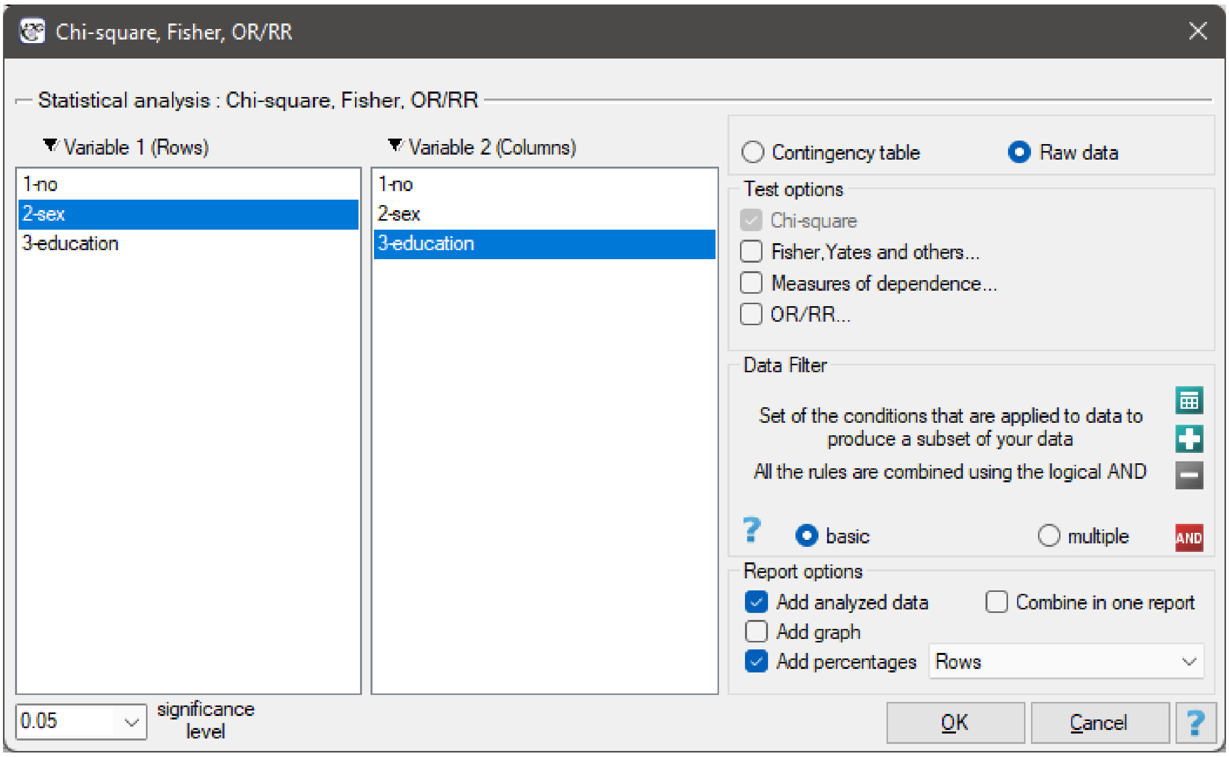
For the distribution with respect to age groups, age categories were first created through codes/labels/format.

10.2.3 ANALYSES FOR CONTINGENCY TABLES

Analyses for the contingency tables can be computed from data collected in the contingency tables or directly i.e., from raw data. Whereby it is possible to transform the data from the contingency table to the raw form or vice versa.

Example 10.4. (sex-education.pqs file)
Consider a sample consisting of 34 individuals ($n = 34$). We examine 2 traits of these individuals (X=sex, Y=education). Gender appears in 2 categories (X_1=female, X_2=male) education in 3 categories, (Y_1=primary + vocational Y_2=medium, Y_3=higher).

In the case of raw data, when you open the test options window, e.g., the χ^2 for the $C \times R$ tables, the raw data option will automatically be selected..



Chi-square, Fisher, OR/RR

Statistical analysis : Chi-square, Fisher, OR/RR

Variable 1 (Rows): 1-no, **2-sex** (selected), 3-education

Variable 2 (Columns): 1-no, 2-sex, **3-education** (selected)

○ Contingency table ● Raw data

Test options

☑ Chi-square
□ Fisher,Yates and others...
□ Measures of dependence...
□ OR/RR...

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

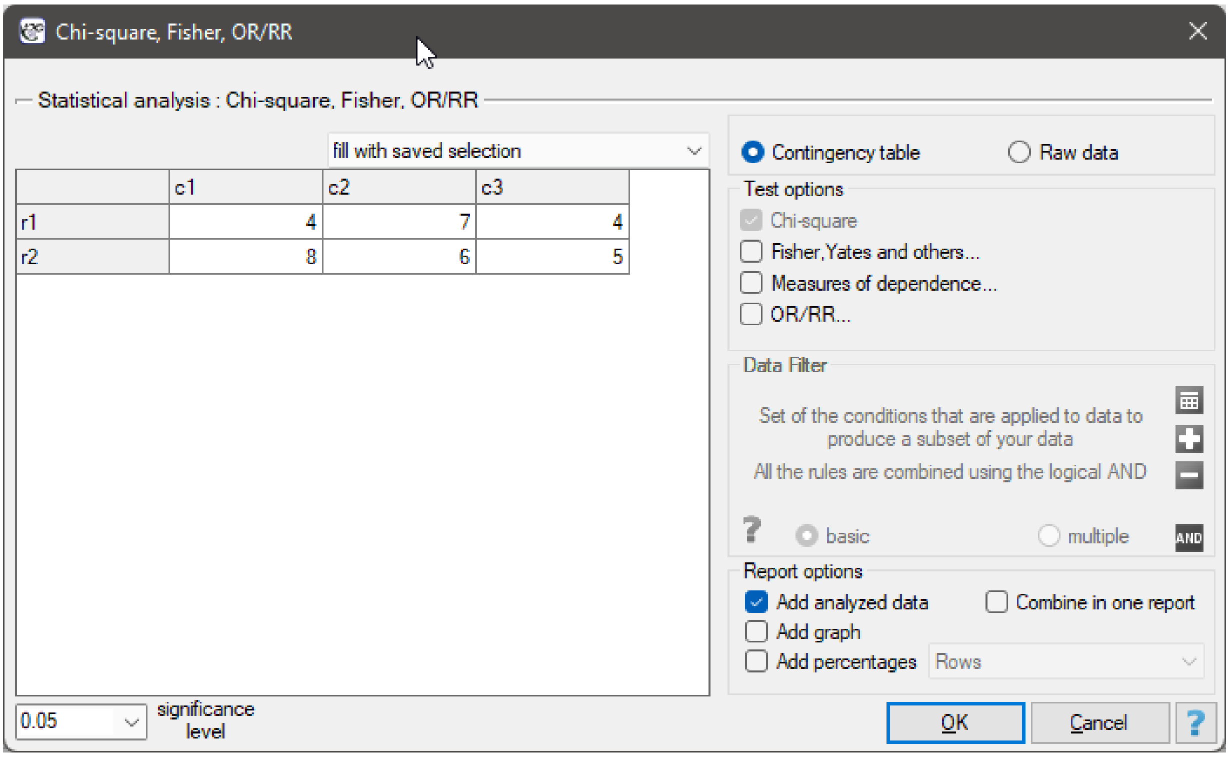
● basic ○ multiple

Report options

☑ Add analyzed data □ Combine in one report
□ Add graph
☑ Add percentages Rows

0.05 significance level

For data collected in a contingency table, it is a good idea to select this data (numerical values without headers) before opening the test window. Then, when you open the test window, the contingency table option will automatically be selected and the data from the selection will be displayed.



Chi-square, Fisher, OR/RR

Statistical analysis : Chi-square, Fisher, OR/RR

fill with saved selection

	c1	c2	c3
r1	4	7	4
r2	8	6	5

● Contingency table ○ Raw data

Test options

☑ Chi-square
□ Fisher,Yates and others...
□ Measures of dependence...
□ OR/RR...

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

○ basic ○ multiple

Report options

☑ Add analyzed data □ Combine in one report
□ Add graph
□ Add percentages Rows

0.05 significance level

In the test window, we can always change the automatically detected setting regarding the form of data organization, as well as enter data into the contingency table from the window.

Cochran's condition
This is a basic condition for using many statistical tests based on contingency tables, e.g., the chi-square test. This condition implies a large expectred frequencies. According to Cochran's 1952 interpretation[40], none of the expected frequencies can be < 1 and no more than 20% can be < 5. Information about whether this condition is met (or not) by the data collected in the table can be returned to the report.

Basic tests for contingency tables:

- Chi-square goodnes-of-fit test
- RxC (2x2) chi-square test and its corrections
- Chi-square test for trend for Rx2 tables
- McNemar test, Bowker internal symmetry test
- Chi-square test for multidimensional contingency tables
- Q-Cochran's ANOVA
- Mantel-Haenszel method for 2×2 tables

Coefficients for contingency tables:

- Relative Risk and Odds Ratio
- Contingency coefficients: Q-Yule, Phi, V -Cramer, C-Pearson
- Cohen's Kappa Coefficient of agreement
- Kappa Fleiss coefficient
- Sensitivity and specificity, PPV, NPV, LR(+), LR(-), prevalence, accuracy

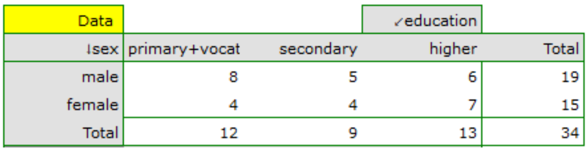
You can also include a basic summary of the tables in the results report:

- **Contingency table of observed frequencies** – that is, data in the form of a contingency table. Such a table shows the distribution of observations for several traits (several variables). Table for 2 traits (X, Y), of which the first has possible r and the second c categories are shown below (table(10.1)).

  Tabela 10.1. Contingency table $r \times c$ of observed frequencies

Frequencies observed O_{ij}		Trait Y				
		Y_1	Y_2	...	Y_c	Total
Trait 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}	$\sum_{j=1}^{c} O_{rj}$
	Suma	$\sum_{i=1}^{r} O_{i1}$	$\sum_{i=1}^{r} O_{i2}$	...	$\sum_{i=1}^{r} O_{ic}$	$n = \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij}$

  Frequencies observed O_{ij} ($i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c$) represent the frequency of each category for both traits.
  In order for such a table to be returned by the program, the option include analyzed data should be selected in the test window. For the data from the example (10.4), 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

- **A contingency table of expected frequencies** $-$ for each contingency table of observed frequencies, a corresponding table of **expected frequencies:** E_{ij} can be created (tabela(10.2)).

  Tabela 10.2. Contingency table $r \times c$ expected frequencies

frequencies expected E_{ij}		Trait Y			
		Y_1	Y_2	...	Y_c
Trait 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}

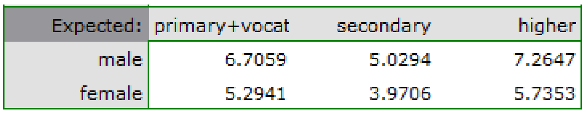
where:

$$E_{11}=\frac{\sum_{i=1}^r O_{i1}\times\sum_{j=1}^c O_{1j}}{n},\ E_{12}=\frac{\sum_{i=1}^r O_{i2}\times\sum_{j=1}^c O_{1j}}{n},\ E_{1c}=\frac{\sum_{i=1}^r O_{ic}\times\sum_{j=1}^c O_{1j}}{n}$$

$$E_{21}=\frac{\sum_{i=1}^r O_{i1}\times\sum_{j=1}^c O_{2j}}{n},\ E_{22}=\frac{\sum_{i=1}^r O_{i2}\times\sum_{j=1}^c O_{2j}}{n},\ E_{2c}=\frac{\sum_{i=1}^r O_{ic}\times\sum_{j=1}^c O_{2j}}{n}$$

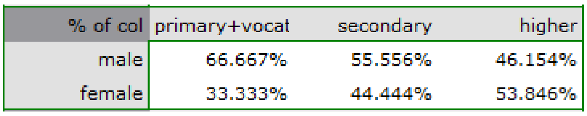
$$E_{r1}=\frac{\sum_{i=1}^r O_{i1}\times\sum_{j=1}^c O_{rj}}{n},\ E_{r2}=\frac{\sum_{i=1}^r O_{i2}\times\sum_{j=1}^c O_{rj}}{n},\ E_{rc}=\frac{\sum_{i=1}^r O_{ic}\times\sum_{j=1}^c O_{rj}}{n}.$$

For the data in the example (10.4) The contingency table of expected frequencies is as follows:



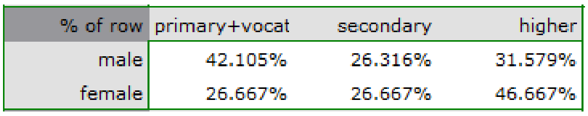
Expected:	primary+vocat	secondary	higher
male	6.7059	5.0294	7.2647
female	5.2941	3.9706	5.7353

- **Contingency table of percentages calculated from the sum of columns**. For the data in the example (10.4) this table is as follows:



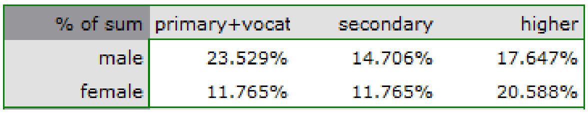
% of col	primary+vocat	secondary	higher
male	66.667%	55.556%	46.154%
female	33.333%	44.444%	53.846%

- **Tcontingency table of percentages calculated from the sum of the rows**. For the data in the example (10.4) 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%

- **A contingency table of percentages calculated from the sum of the total rows and columns**. For the data in the example (10.4) this table is as follows:

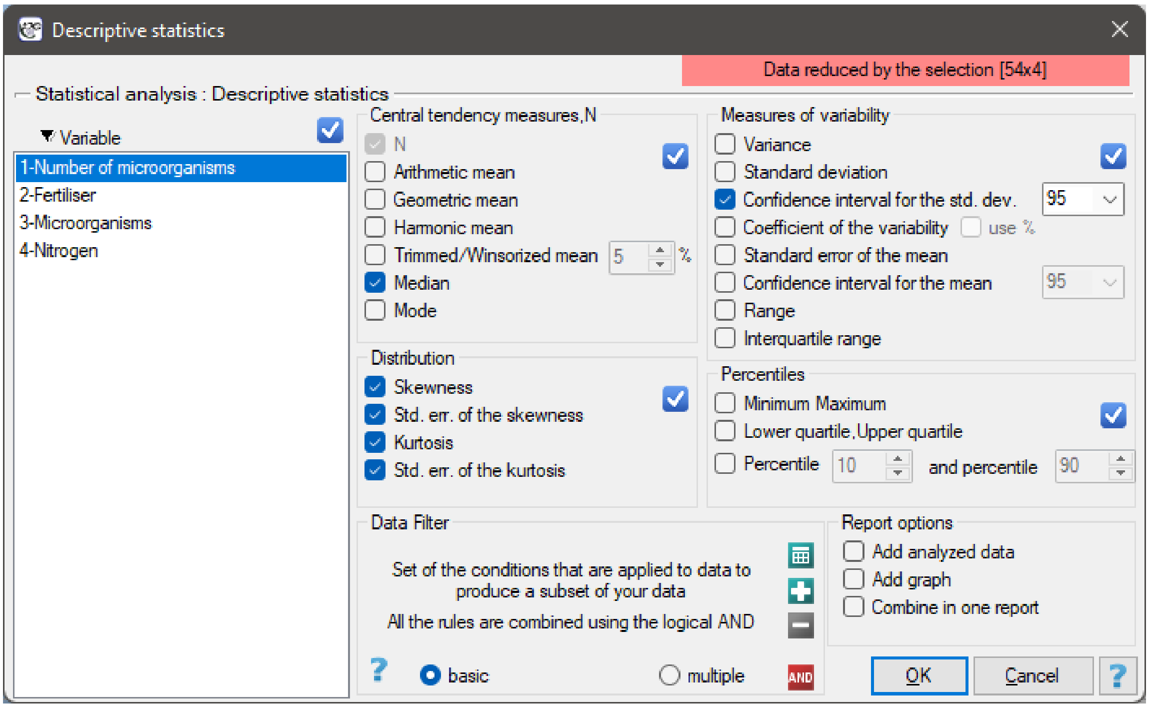


% of sum	primary+vocat	secondary	higher
male	23.529%	14.706%	17.647%
female	11.765%	11.765%	20.588%

10.3 DESCRIPTIVE STATISTICS

The purpose of using descriptive statistical methods is to summarize a set of data by certain characteristics, e.g., by the value of the mean, median, or standard deviation, and to draw some basic conclusions and generalizations about the dataset.

To calculate descriptive statistics for the data collected in the datasheet, open the Descriptive statistics window via menu Statistics→Descriptive analysis→Descriptive statistics.



Data reduced by the selection [54x4]

Statistical analysis : Descriptive statistics

Variable ☑

1-Number of microorganisms
2-Fertiliser
3-Microorganisms
4-Nitrogen

Central tendency measures,N ☑

☑ N
□ Arithmetic mean
□ Geometric mean
□ Harmonic mean
□ Trimmed/Winsorized mean 5 %
☑ Median
□ Mode

Measures of variability ☑

□ Variance
□ Standard deviation
☑ Confidence interval for the std. dev. 95
□ Coefficient of the variability □ use %
□ Standard error of the mean
□ Confidence interval for the mean 95
□ Range
□ Interquartile range

Distribution ☑

☑ Skewness
☑ Std. err. of the skewness
☑ Kurtosis
☑ Std. err. of the kurtosis

Percentiles ☑

□ Minimum Maximum
□ Lower quartile,Upper quartile
□ Percentile 10 and percentile 90

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data
□ Add graph
□ Combine in one report

In this window, we select the variable to be analyzed and the analysis settings and select the descriptive statistics measures we are interested in. You may select individual statistics or groups of statistics by clicking on the ☑. Confirm the selection by pressing OK. The result of the analysis will be in a report attached to the datasheet for which the analysis was performed.

In addition, if you want the data to be visualized with a box-and-whisker chart, then in the Descriptive statistics window select the Add graph.

10.3.1 LOCATION MEASURES

10.3.2 MEASURES OF CENTRAL TENDENCY

Measures of central tendency are so-called average measures that characterize the average or typical level of a trait's values.

Arithmetic mean is expressed by the formula:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

where x_i is the consecutive values of the variable and n is the sample size.

The arithmetic mean is used for interval scale. For a sample it is taken to be denoted by $\overline{x}$ and for a population by μ.

Trimmed mean - is determined as the arithmetic mean calculated after removing from the sample a given percentage of the smallest and largest measurements, e.g. if we cut off 5 per cent of the measurements, it means that we cut off 2.5 per cent of the largest and 2.5 per cent of the smallest values. At the same time, if the number of measurements to be removed obtained from the conversion is not an integer, it is rounded down to the nearest whole number.

Winsor mean - is determined as the arithmetic mean calculated after replacing the appropriate percentage of extreme measurements with the smallest and largest value that remains of the reduced set of values. If we choose to calculate the Winsor average by pruning, say, 5% of the measurements, then those discarded 5% will be replaced by the smallest and largest value determined from the remaining 95% of the measurements. As with the pruned average, when converting the percentage of values to be replaced to the number of measurements to be replaced does not result in an integer, then we round down to the nearest integer. **Geometric mean** is expressed by the formula:

$$\overline{x}_G=\sqrt[n]{x_1x_2...x_n}=\sqrt[n]{\prod_{i=1}^{n}x_i}.$$

This mean is used for the interval scale, when the variable has a log-normal distribution (the logarithm of the variable has a normal distribution).

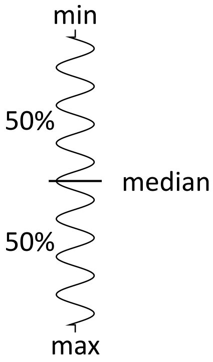
Harmonic mean is expressed by the formula:

$$\overline{x}_H=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}.$$

This mean is used for the interval scale.

Median

In an ordered data set, the median is the value that divides the data set into two equal parts. Half of all observations are below and half are above the median.



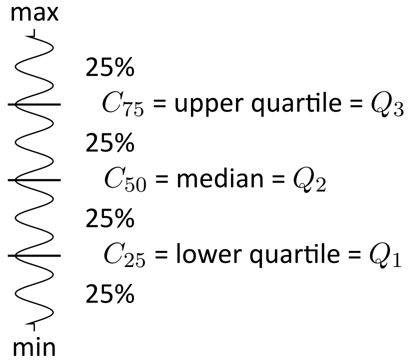
The median can be used in interval and ordinal scale.

Mode

Mode – is the value that occurs most frequently among the measurements obtained. Fashion can be used at any measurement scale.

10.3.3 OTHER MEASURES OF LOCATION

Quartiles, deciles, centiles


max
25%
C_{75} = upper quartile = Q_3
25%
C_{50} = median = Q_2
25%
C_{25} = lower quartile = Q_1
25%
min

Quartiles (Q_1, Q_2, Q_3) divide the 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, fifth decile, and fiftieth centile are equal to the median. These measures can be used in the interval and ordinal scale.

10.3.4 MEASURES OF VARIABILITY (DISPERSION)

Central tendency measures knowledge is not enough to fully describe a statistical data collection structure. The researched groups may have various variation levels of a feature you want to analyse. You need some formulas then, which enable you to calculate values of variability of the features.

Measures of variability are calculated only for an interval scale, because they are based on the distance between the points.

Range is formulated:

$$I = \max x_i - \min x_i,$$

where x_i are values of the analyzed variable

$$IQR = \text{Interquartile range} = Q_3 - Q_1,$$

where Q_1, Q_3 are the lower and the upper quartile.

Ranges for a percentile scale (decile, centile)

Ranges between percentiles are one of the dispersion measures. They define a percentage of all observations, which are located between the chosen percentiles.

Variance $-$ measures a degree of spread of the measurements around arithmetic mean

sample variance:

$$sd^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1},$$

where x_i are following values of variable and $\overline{x}$ is an arithmetic mean of these values,
n - sample size;

population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N},$$

where x_i are following values of variables and μ is an arithmetic mean of these values, N - population size;

Variance is always positive, but it is not expressed in the same units as measuring results.

Standard deviation $-$ measures a degree of spread of the measurements around arithmetic mean.

sample standard deviation:

$$sd=\sqrt{sd^2},$$

population standard deviation:

$$\sigma=\sqrt{\sigma^2}.$$

The higher standard deviation or a variance value is, the more diversed is the group in relation to an analyzed feature.

Note
The sample standard deviation is a kind of approximation (estimator) of the population standard deviation. The population standard deviation value is included in a range which contains the sample standard deviation. This range is called a **confidence interval** for standard deviation.

Coefficient of variation

Coefficient of variation, just like standard deviation, enables you to estimate the homogeneity level of an analyzed data collection. It is formulated as:

$$V=\frac{sd}{\overline{x}}100\%,$$

where sd means standard deviation, $\overline{x}$ means arithmetic mean.

This is a unitless value. It enables you to compare a diversity of several different datasets of a one feature. And also, you are able to compare a diversity of several features (expressed in different units). It is assumed, if V coefficient does not exceed 10%, features indicate a statistically insignificant diversity.

Standard errors $-$ they are not measures of a measurement dispersion. They measure an accuracy level, you can define the population parameters value, having just the sample estimators.
Standard error of the mean is defined by:

$$SEM=\text{standard error of the mean}=\frac{sd}{\sqrt{n}}.$$

Note
On the basis of a sample estimator you can calculate a confidence interval for a population parameter.

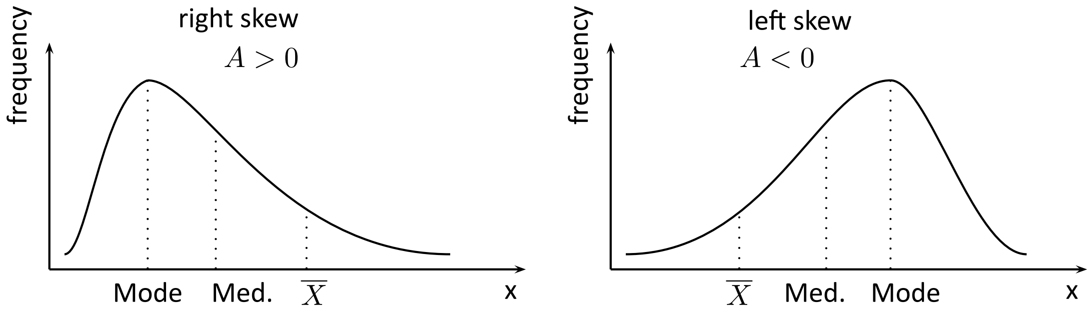
10.3.5 ANOTHER DISTRIBUTION CHARACTERISTICS

Skewness or **asymmetry coefficient** in other words

This measure tells us how data distribution differs from symmetrical distribution. The closer the value of skewness is to zero, the more symmetrically around the mean the data are spread. Usually the value of this coefficient is included in a range [-1, 1], but in the case of a very big asymmetry, it may occur outside the above-mentioned range. A positive skew value indicates that the right skew occurs (the tail on the right side is longer), whereas the negative skew indicates that the left skew occurs (the tail on the left side is longer). Skewness is defined by:

$$A=\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{sd}\right)^3,$$

where:
x_i- the following values of a variable,
$\overline{x}$, $sd-$ adequately - arithmetic mean and standard deviation x_i,
$n-$ sample size.

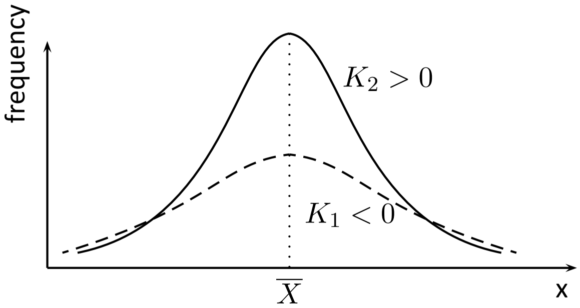


Kurtosis or **coefficient of concentration**

This measure tells us how much the spread of data around the mean is similar to the spread of data in normal distribution. The greater than zero the value of kurtosis is, the more narrow the tested distribution than normal one is. And inversely, the lower than zero the value of kurtosis is, the flatter the tested distribution than the normal one is. Kurtosis is defined by:

$$K=\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{sd}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)},$$

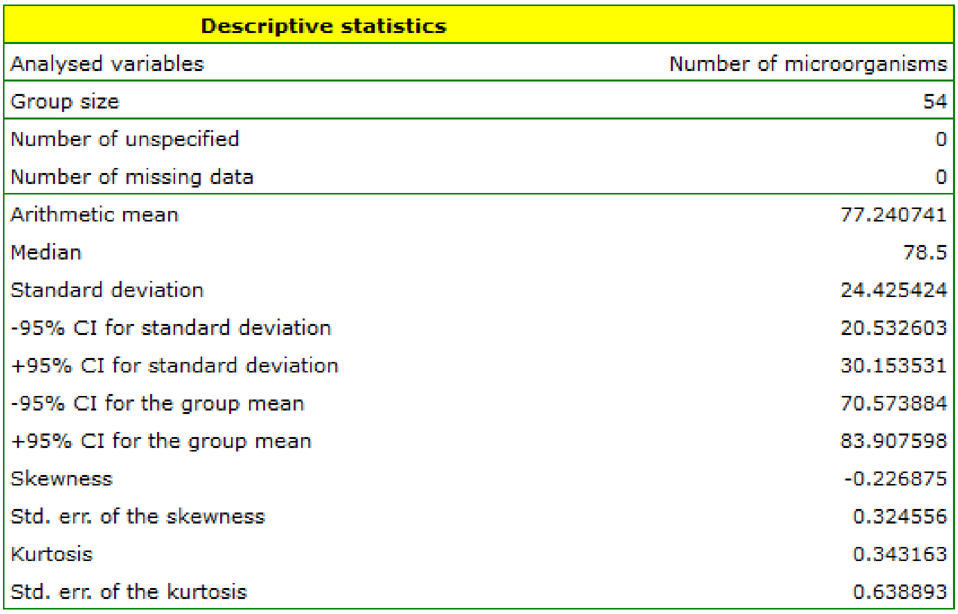
where:
x_i- the following values of a variable,
$\overline{x}$, $sd-$ adequately - arithmetic mean and standard deviation of x_i,
$n-$ sample size.



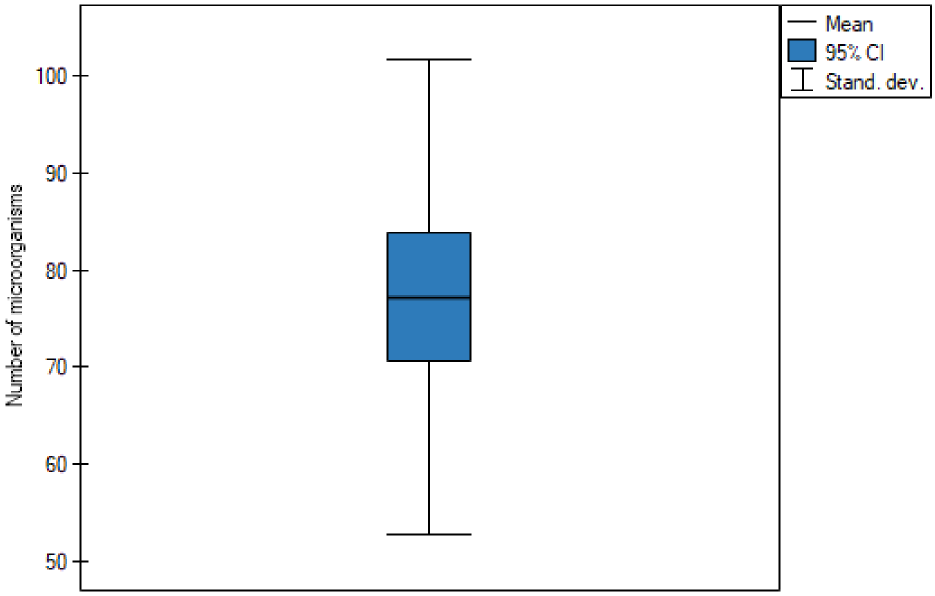
Example 10.5. (fertilisers.pqs file)

In an experiment related to a soil fertilising the with various sorts of microbiological specimens and fertilisers it was calculated how many microorganisms occur in a 1 gramme of dry mass of soil. Now we would like to calculate descriptive statistics of the amount of actinomycetes for the sample fertilised with nitrogen. Additionally, we want the data to be illustrated in the Box-Whiskers plot. In a datasheet, we select only the 54 first rows, which are relevant to the assumptions of the analysis (there

are actinomycetes fertilised with nitrogen). Then we open Descriptive statistics window in Statistics menu→Descriptive analysis→Descriptive statistics.
In the window of descriptive statistics options, select a variable to analyse: the number of microorganisms, and then all the procedures you want to follow (for example arithmetic mean altogether with the confidence interval, median, standard deviation altogether with the confidence interval, and an information about the skewness and kurtosis of distribution altogether with errors). At the top of the window you should see the following message: Data limited by the selected area. To add a graph to the report, we select Add graph option and chose the Box-Whiskers plot type . Confirm your choice by clicking OK and you get the result in a report:

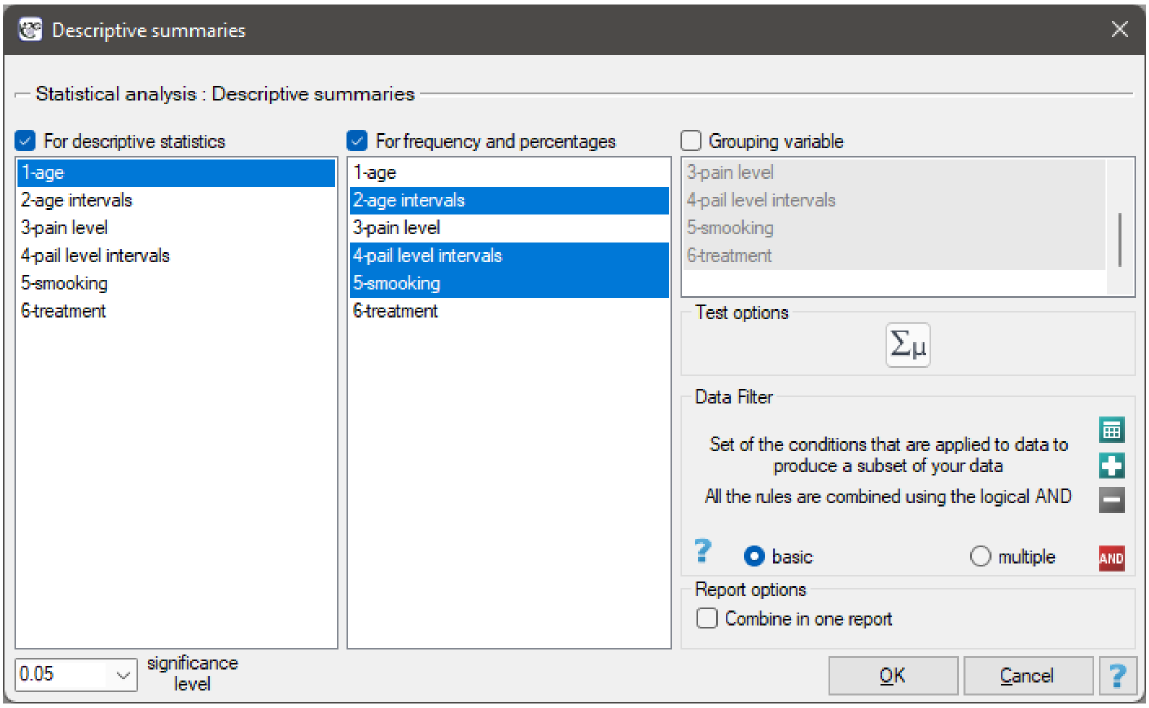


Descriptive statistics	
Analysed variables	Number of microorganisms
Group size	54
Number of unspecified	0
Number of missing data	0
Arithmetic mean	77.240741
Median	78.5
Standard deviation	24.425424
-95% CI for standard deviation	20.532603
+95% CI for standard deviation	30.153531
-95% CI for the group 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



10.4 DESCRIPTIVE SUMMARIES

Descriptive summaries are a quick way to prepare a report showing a description of your data that is ready to be inserted directly into a research paper. It is a tool that makes it easy to compile all forms of basic data description in one place.
The Descriptive summaries option settings window is invoked via the menu textsfStatistics menu→Descriptive analysis→Descriptive summaries.



Descriptive summaries

Statistical analysis : Descriptive summaries

☑ For descriptive statistics
- 1-age (selected)
- 2-age intervals
- 3-pain level
- 4-pail level intervals
- 5-smooking
- 6-treatment

☑ For frequency and percentages
- 1-age
- 2-age intervals (selected)
- 3-pain level
- 4-pail level intervals (selected)
- 5-smooking (selected)
- 6-treatment

☐ Grouping variable
- 3-pain level
- 4-pail level intervals
- 5-smooking
- 6-treatment

Test options

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

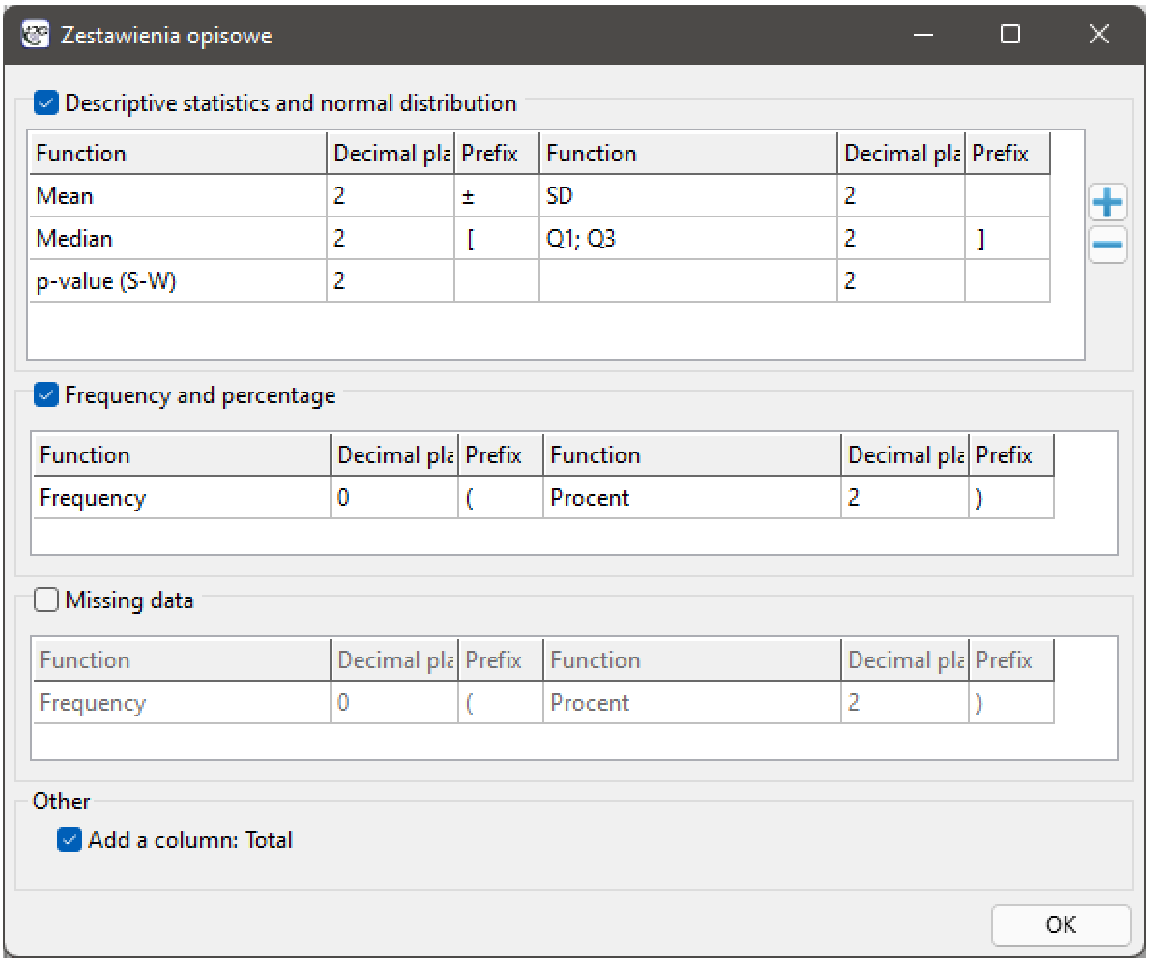
All the rules are combined using the logical AND

● basic ○ multiple

Report options

☐ Combine in one report

0.05 significance level



Zestawienia opisowe

☑ Descriptive statistics and normal distribution

Function	Decimal plą	Prefix	Function	Decimal plą	Prefix
Mean	2	±	SD	2	
Median	2	[	Q1; Q3	2	]
p-value (S-W)	2			2	

☑ Frequency and percentage

Function	Decimal plą	Prefix	Function	Decimal plą	Prefix
Frequency	0	(	Procent	2	)

☐ Missing data

Function	Decimal plą	Prefix	Function	Decimal plą	Prefix
Frequency	0	(	Procent	2	)

Other

☑ Add a column: Total

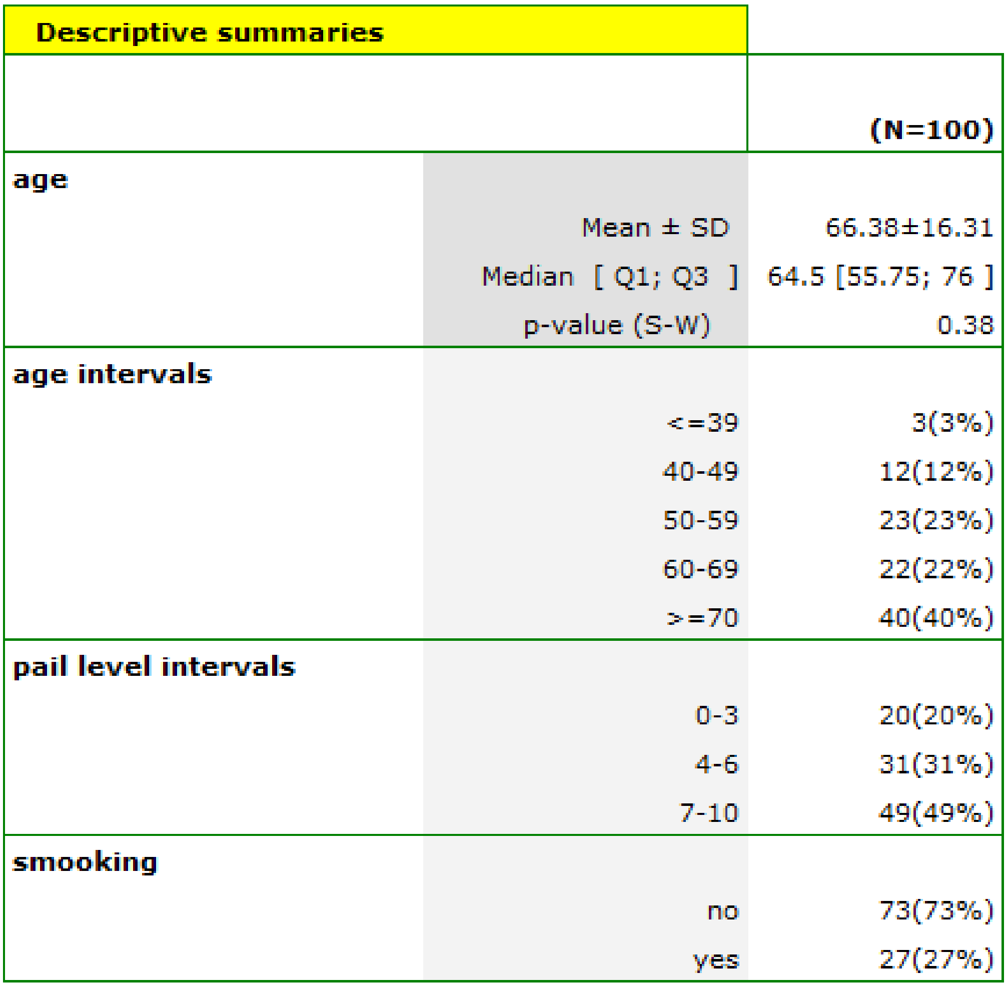
Depending on how we measure, we usually use one of three types of measures to describe a variable:

- mean ± standard deviation, abbreviated $\overline{x} \pm sd$
- median [lower quartile; upper quartile], abbreviated $Me[Q_1; Q_3]$ or median [min; max], abbreviated $Me[min; max]$
- the number (percentage) of each category, abbreviated $n(\%)$

Depending on the need and fulfillment of additional assumptions

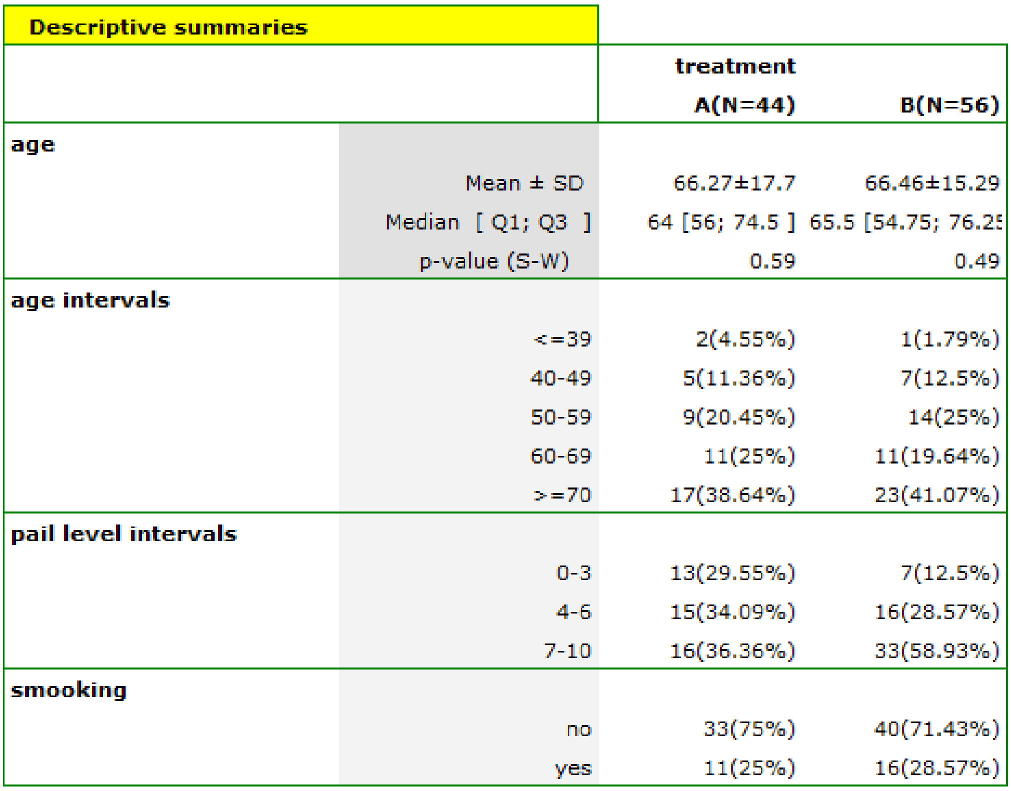
- in the interval scale, data can be described using any measure, in addition, often the researcher wants to check the normality of the distribution of such data,
- in the ordinal scale we have medians (with quartiles or the smallest and the largest value) or counts and percentages of particular categories,
- at the nominal scale, only the counts and percentages of each category.

Example 10.6. (Descriptive summaries.pqs file)
An example of how the data are described is shown in the obtained report summarizing age, pain level, and smoking status. These data are summarized in table one describing all 100 subjects and in table two 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%)

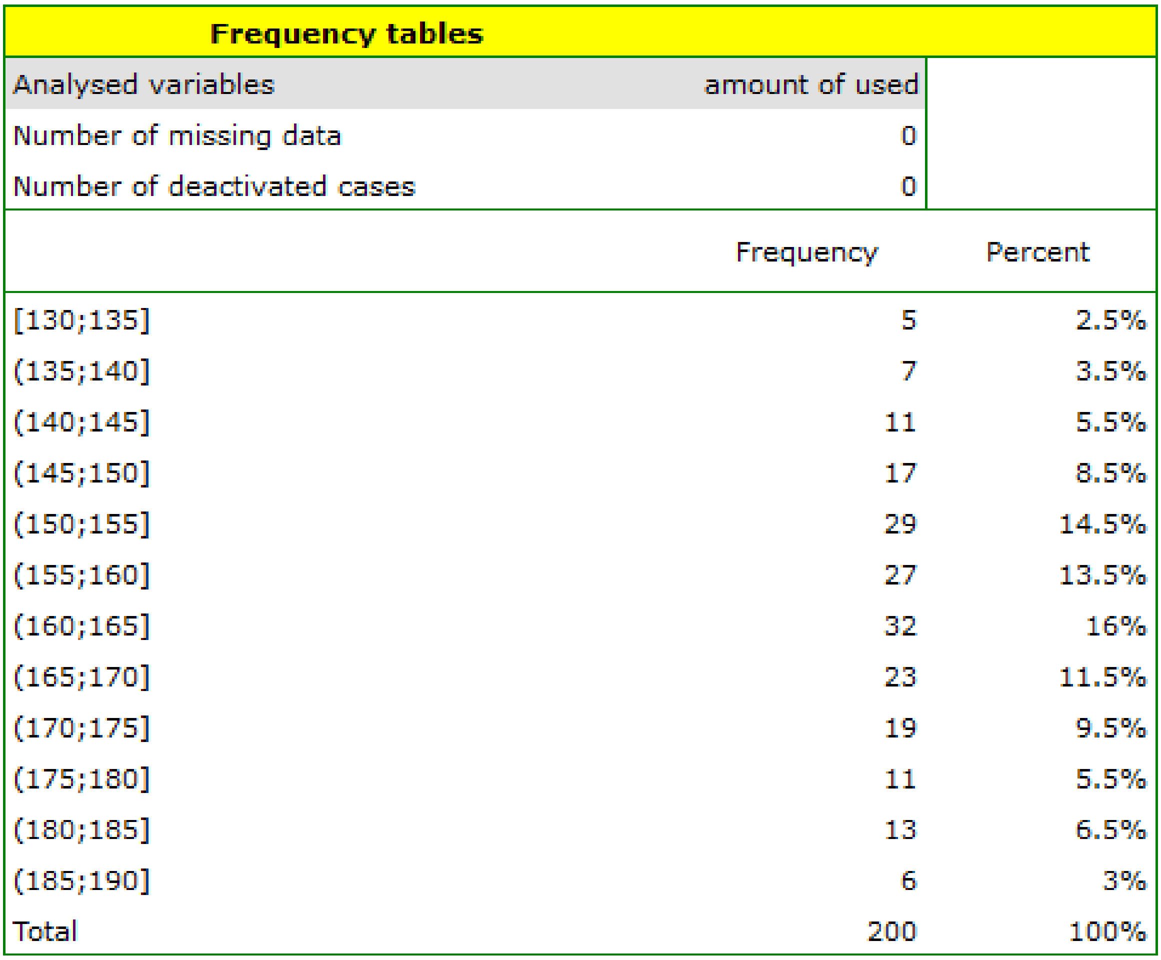


Descriptive summaries

		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%)

11 PROBABILITY DISTRIBUTIONS

A real data distribution from a sample - **empirical data distribution** may be carried out in a mean of a frequency tables (by selecting Statistic menu→Descriptive analysis)→Frequency tables). For example, a distribution of the amount of used free minutes by subscribers of some mobile network operator (*example (10.1), distribution.pqs file*) performs the following table:

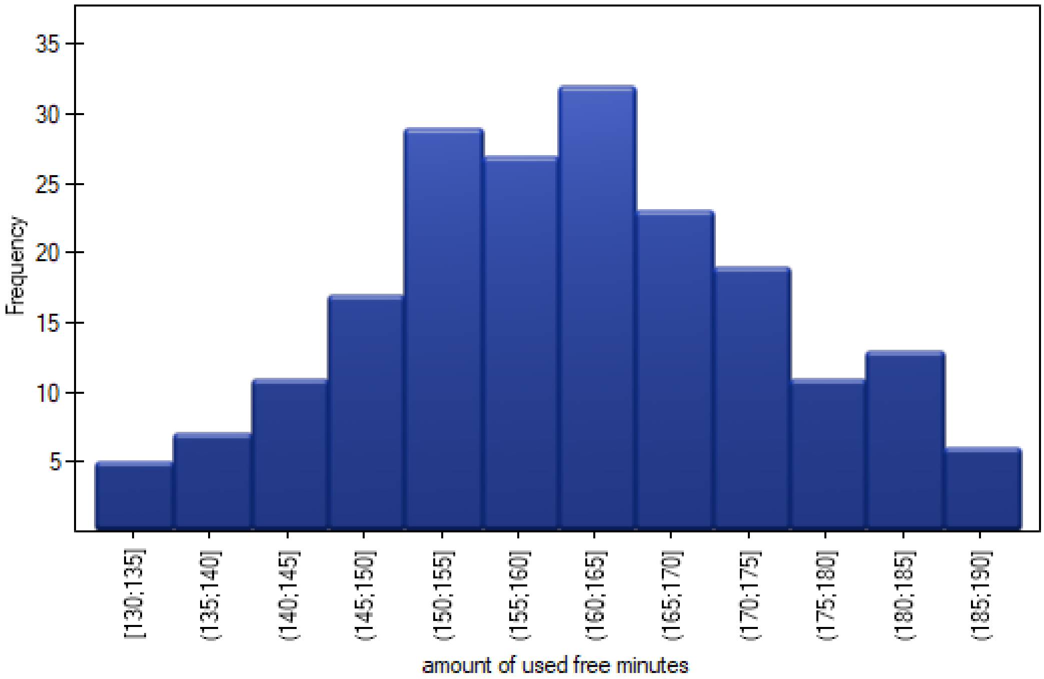


Frequency tables

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 results included in a table is usually done using a histogram or a bar plot.



Such graph can be created by selecting Add graph option in the Frequency tables window.

Theoretical data distribution which is also called a **probability distribution** is usually presented graphically by means of a line graph. Such line is described by a function (mathematical model) and it is called a **density function**. You can replace the empirical distribution with the adequate theoretical distribution.

Note
To replace an empirical distribution with the adequate theoretical distribution it is not enough to draw conclusions upon similarity of their shapes intuitively. To check it, you should use specially created compatibility tests.

The kind of probability distribution which is used the most often is a normal distribution (Gaussian distribution). Such distribution with a mean of 161.15 and a standard deviation 13.03 is presented by the data relating to the amount of used free minutes (*example (10.1), distribution.pqs file*).

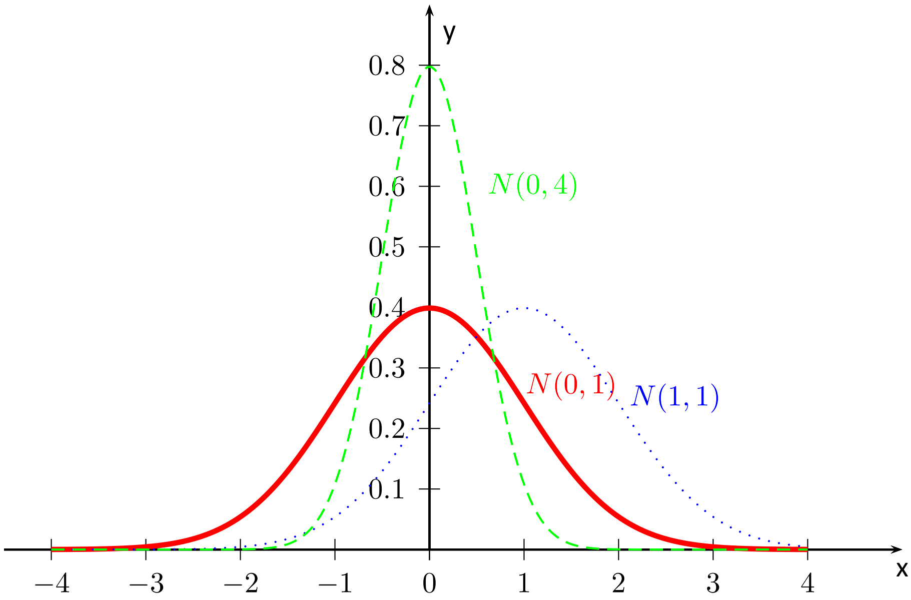
11.1 CONTINUOUS PROBABILITY DISTRIBUTIONS

- **Normal distribution** which is also called the Gaussian distribution or a bell curve, is one of the most important distribution in statistics. It has very interesting mathematical features and occurs very often in nature. It is usually designated with $N(\mu, \sigma)$.

  A density function is defined by:

  $$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \text{exp}\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

  where:
  $-\infty < x < +\infty$,
  μ – an expected value of population (its measure is mean),
  σ – standard deviation.

  Normal distribution is a symmetrical distribution for a perpendicular line to axis of abscissae going through the points designating the mean, mode and median.

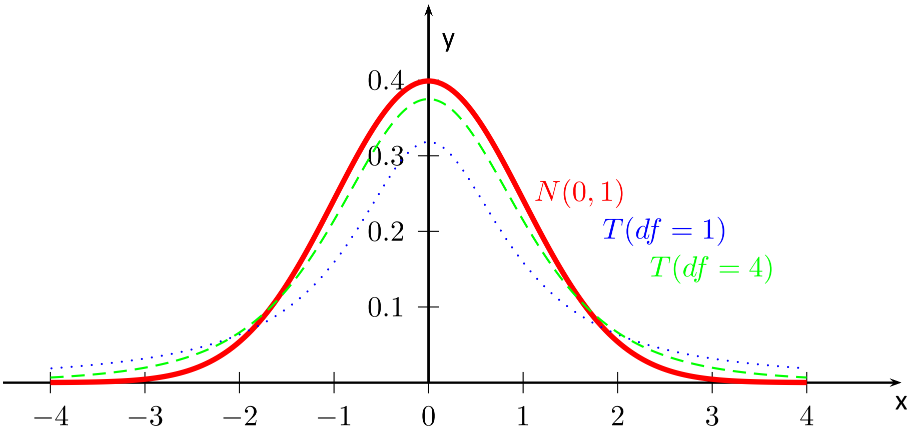
  Normal distribution with a mean of $\mu = 0$ and $\sigma = 1$ ($N(0, 1)$), is so called a **standardised normal distribution**.

- **t-Student distribution** – the shape of t-Student distribution is similar to standardised normal distribution, but its tails are longer. The higher the number of degrees of freedom (df), the more similar the shape of t-Student distribution to normal distribution.

  A density function is defined by:

  $$f(x, df) = \frac{\Gamma(\frac{df+1}{2})}{\Gamma(\frac{df}{2})\sqrt{df\pi}} \left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}},$$

  where:
  $-\infty < x < +\infty$,
  df – degrees of freedom (sample size is decreased by the number of limitations in given calculations),
  Γ is a Gamma function.

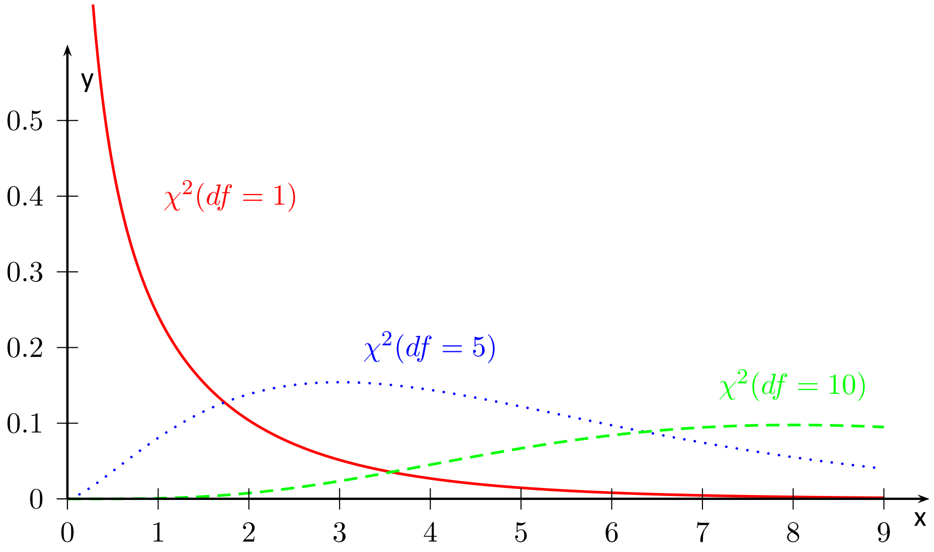


- **Chi-square (χ^2) distribution**, this is a right-skewed distribution with a shape depending on the number of degrees of freedom df. The higher the number of degrees of freedom, the more similar the shape of χ^2 distribution to the normal distribution.

  Density function is defined by:

$$f(x,df)=\frac{1}{2^{\frac{df}{2}}\Gamma\frac{df}{2}}x^{\frac{df}{2}-1}e^{-\frac{x}{2}},$$

  where:
  $x>0$,
  df – degrees of freedom (sample size is decreased by the number of limitations in given calculations),
  Γ is a Gamma function.

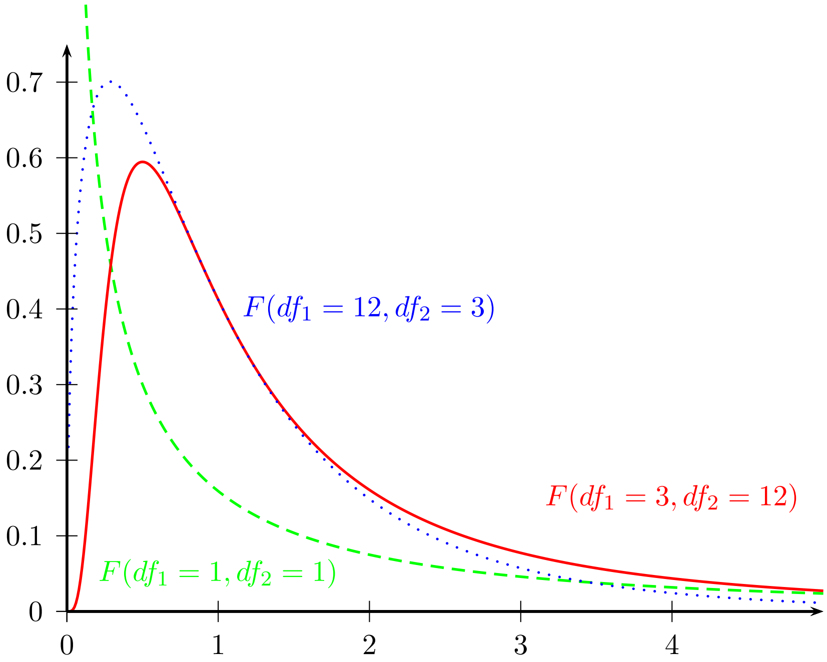


- **Fisher-Snedecor distribution**, this is a distribution which has a right tail that is longer and a shape that depends on the number of degrees of freedom df_1 and df_2.

A density function is defined by:

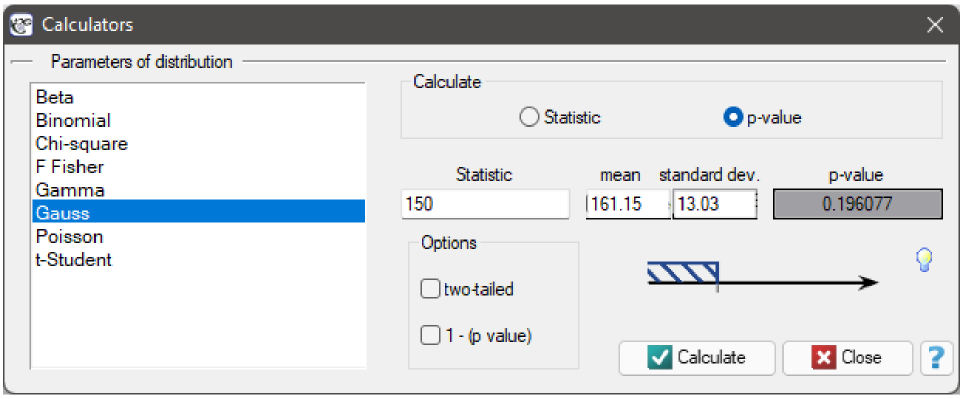
$$F(x,df_1,df_2)=\frac{\sqrt{\frac{(df_1x)^{df_1}d_2^{df_2}}{(df_1x+df_2)^{df_1+df_2}}}}{xB\left(\frac{df_1}{2},\frac{df_2}{2}\right)},$$

where:
$x>0$,
df_1, df_1 – degrees of freedom (it is assumed that if X i Y are independent with a χ^2 distribution with adequately df_1 and df_2 degrees of freedom, than $F=\frac{X/df_1}{Y/df_2}$ has a F Snedecor distribution $F(df_1,df_2)$),
B is a Beta function.



11.2 PROBABILITY DISTRIBUTION CALCULATOR

The area under a curve (density function) is p **probability** of occurrence of all possible values of an analysed random variable. The whole area under a curve comes to $p = 1$. If you want to analyse just a part of this area, you must put the border value, which is called the **critical value** or Statistic. To do this, you need to open the Probability distribution calculator window. In this window you can calculate not only a value of the area under the curve (p value) of the given distribution on the basis of Statistic, but also Statistic value on the basis of p value. To open the window of Probability distribution calculator, you need to select Probability distribution calculator from the Statistics→Calculators menu.


Calculators

Parameters of distribution: Beta, Binomial, Chi-square, F Fisher, Gamma, Gauss (selected), Poisson, t-Student

Calculate: ○ Statistic ● p-value

Statistic: 150
mean: 161.15
standard dev.: 13.03
p-value: 0.196077

Options: □ two-tailed □ 1 - (p value)

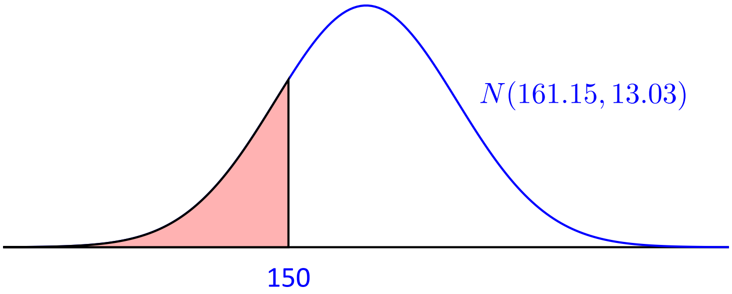
Example 11.1. Probability distribution calculator

Some mobile network operator did the research, which was supposed to show the usage of ”free minutes” given to his clients on a pay-monthly contract. On the basis of the sample, which consists of 200 of the above-mentioned network clients (where the distribution of used free minutes is of the shape of normal distribution) is calculated the mean value $\overline{x} = 161.15min.$ and standard deviation $sd = 13.03min.$ We want to calculate the probability, that the chosen client used:

1. 150 minutes or less,
2. more than 150 minutes,
3. the amount of minutes coming from the range $[\overline{x} - sd, \overline{x} + sd] = [148.12min., 174.18min.]$,
4. the amount of minutes out of the range $\overline{x} \pm sd$.

Open the Probability distribution calculator window, select Gaussian distribution and write the mean $\overline{x} = 161.15min.$ and standard deviation $sd = 13.03min.$ and select the option which indicates, that you are going to calculate the p value.

1. To calculate (using normal distribution (Gauss)) the probability that the client you have chosen used 150 free minutes or less, put the value of 150 in the Statistic field. Confirm all selected settings by clicking Calculate.

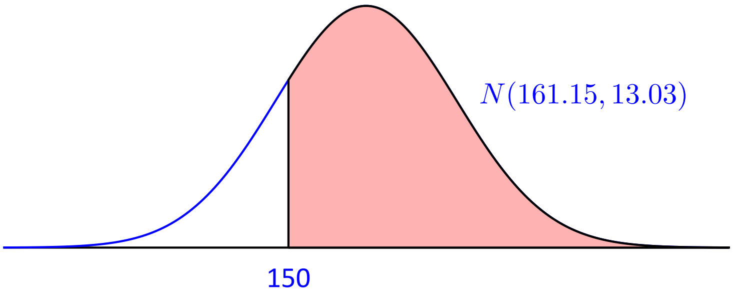


The obtained p value is 0.193961.

Note

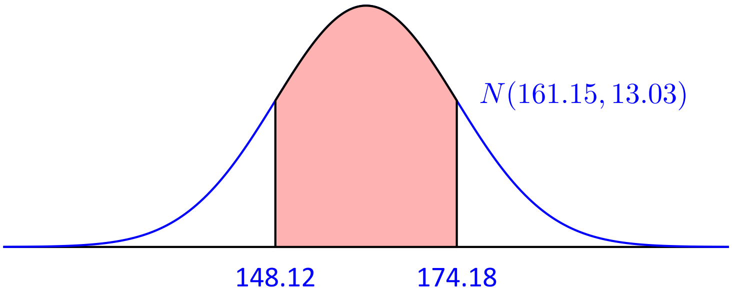
Similar calculations you can carry out on the basis of empirical distribution. The only thing you should do is to calculate a percentage of clients who use 150 minutes or less *(example (10.1)* by using the Frequency tables window. In the analysed sample (which consists of 200 clients) there are 40 clients who use 150 minutes or less. It is 20% of the whole sample, so the probability you are looking for is $p = 0.2$.

2. To calculate the probability (using the normal distribution (Gauss)), that the client who you have chosen used more than 150 free minutes, you need to put the value of 150 in the Statistic field and than select the option 1 - $(p$ value). Confirm all the chosen settings by clicking Calculate.



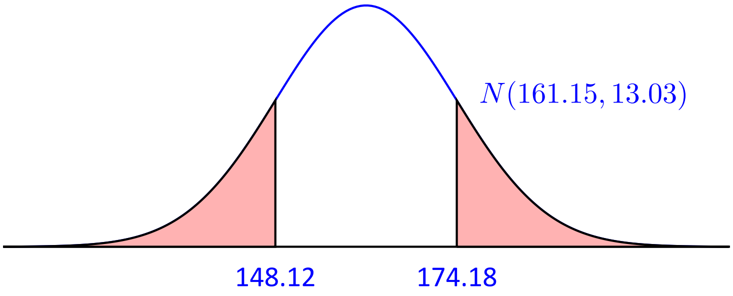
The obtained p value is 0.806039.

3. To calculate (using the normal distribution (Gauss)) a probability that the client you have chosen used free minutes which come from the range $[\overline{x} - sd, \overline{x} + sd] = [148.12min., 174.18min.]$ in the Statistic field, put one of the final range values and than select the option two-sided. Confirm all the chosen settings by clicking Calculate.



The obtained p value is 0.682689.

4. To calculate (using the normal distribution (Gauss)) a probability, that the client you have chosen used free minutes out of the range $[\overline{x} - sd, \overline{x} + sd] = [148.12min., 174.18min.]$ in the Statistic field put one of the final range values and than select the option: two-sided and 1 - $(p$ value). Confirm all the chosen settings by clicking Calculate.



The obtained p value is 0.317311.

12 HYPOTHESIS TESTING

The process of generalization of the results obtained from the sample for the whole population is divided into 2 basic parts:

- **estimation** $-$ estimating values of the parameters of the population on the basis of the statistical sample,
- **verification of statistical hypotheses** $-$ testing some specific assumptions formulated for the parameters of the general population on the basis of sample results.

12.0.1 POINT AND INTERVAL ESTIMATION

In practice, we usually do not know the **parameters** (characteristics) of the whole population. There is only a sample chosen from the population. **Point estimators** are the characteristics obtained from a random sample. The exactness of the estimator is defined by its **standard error**. The real parameters of population are in the area of the indicated point estimator. For example, the population parameter arithmetic mean μ is in the area of the estimator from the sample which is $\overline{x}$.

If you know the estimators of the sample and their theoretical distributions, you can estimate values of the population parameters with the **confidence level** $(1-\alpha)$ defined in advance. This process is called **interval estimation**, the interval: **confidence interval**, and α is called a **significance level**.

The most popular significance level comes to 0.05, 0.01 or 0.001.

12.0.2 VERIFICATION OF STATISTICAL HYPOTHESES

To verify a statistical hypotheses, follow several steps:

The 1st step: Make a hypotheses, which can be verified by means of statistical tests.

Each statistical test gives you a general form of the null hypothesis $\mathcal{H}_0$ and the alternative one $\mathcal{H}_1$:

$\mathcal{H}_0$: there is **no** statistically significant **difference** among **populations** (means, medians, proportions distributions etc.),

$\mathcal{H}_1$: there **is** a statistically significant **difference** among **populations** (means, medians, proportions, distributions etc.).

Researcher must formulate the hypotheses in the way, that it is compatible with the reality and statistical test requirements, for example:

$\mathcal{H}_0$: the percentage of women and men running their own businesses in an analysed population is exactly the same.

If you do not know, which percentage (men or women) in an analysed population might be greater, the alternative hypothesis should be two-sided. It means you should not assume the direction:

$\mathcal{H}_1$: the percentage of women and men running their own businesses in an analysed population is different.

It may happen (but very rarely) that you are sure you know the direction in an alternative hypothesis. In this case you can use one-sided alternative hypothesis.

The 2nd step: Verify which of the hypotheses $\mathcal{H}_0$ or $\mathcal{H}_1$ is more probable. Depending on the kind of an analysis and a type of variables you should choose an appropriate statistical test.

Note 1
Note, that choosing a statistical test means mainly choosing an appropriate measurement scale (interval, ordinal, nominal scale) which is represented by the data you want to analyse. It is also connected with choosing the analysis model (dependent or independent)

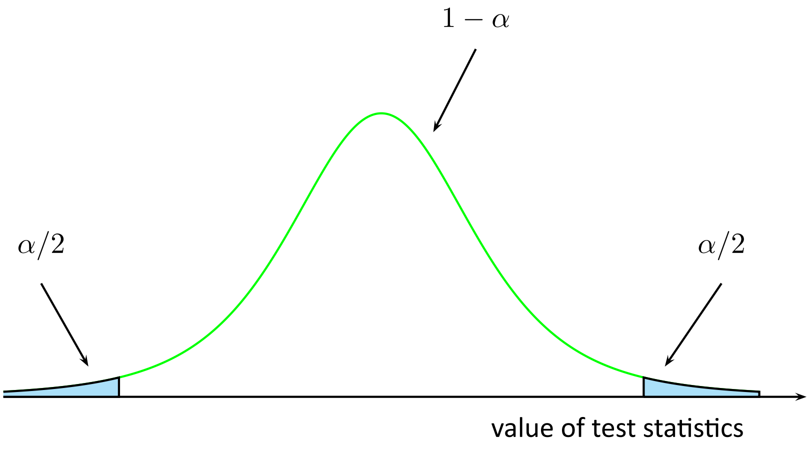
Measurements of the given feature are called **dependent (paired)**, when they are made a couple of times for the same objects. When measurements of the given feature are performed on the objects which belong to different groups, these groups are called **independent (unpaired)** measurements.

Some examples of researches in dependent groups:
Examining a body mass of patients before and after a slimming diet, examining reaction on the stimulus within the same group of objects but in two different conditions (for example - at night and during the day), examining the compatibility of evaluating of credit capacity calculated by two different banks but for the same group of clients etc.

Some examples of researches in independent groups:
Examining a body mass in a group of healthy patients and ill ones, testing effectiveness of fertilising several different kinds of fertilisers, testing gross domestic product (GDP) sizes for the several countries etc.

Note 2
A graph which is included in the Wizard window makes the choice of an appropriate statistical test easier.

Test statistic of the selected test calculated according to its formula is connected with the adequate theoretical distribution.



The application calculates a value of test statistics and also a p value for this statistics (a part of the area under a curve which is adequate to the value of the test statistics). The p value enables you to choose a more probable hypothesis (null or alternative). But you always need to assume if a null hypothesis is the right one, and all the proofs gathered as a data are supposed to supply you with the enough number of counterarguments to the hypothesis:

$$\begin{array}{lcl} \text{if } p \leq \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

There is usually chosen significance level $\alpha = 0.05$, accepting that for 5 % of the situations we will reject a null hypothesis if there is the right one. In specific cases you can choose other significance level for example 0.01 or 0.001.

Note
Note, that a statistical test may not be compatible with the reality in two cases:

		reality	
		$\mathcal{H}_0$: true	$\mathcal{H}_0$: false
test result	$\mathcal{H}_0$: true	OK	β
	$\mathcal{H}_0$: false	α	OK

We may make two kinds of mistakes:

α = 1st type of error (probability of rejecting hypothesis $\mathcal{H}_0$, when it is the right one),

β = 2nd type of error (probability of accepting hypothesis $\mathcal{H}_0$, when it is the wrong one).

Power of the test is $1 - \beta$.

Values α and β are connected with each other. The approved practice is to assume the significance level in advance α and minimalization β by decreasing a sample size.

The 3rd step: Description of results of hypotheses verification.

13 NORMALITY DISTRIBUTION TESTS

13.0.1 One-dimensional normality tests

A variety of tests may be applicable in testing the normality of a distribution, each of which pays attention to slightly different aspects of the Gaussian distribution. It is impossible to identify a test that is good for every possible data set.

The basic condition for using tests of normality of distribution:

- measurement on the interval scale.

Test hypotheses for normality of distribution:

$$\begin{array}{cl}\mathcal{H}_0: & \text{the distribution of the characteristic under study in the population is normal,}\\ \mathcal{H}_1: & \text{the distribution of the examined characteristic in the population is different from the normal .}\end{array}$$

The value p, determined on the basis of test statistics, we compare with the significance level α :

$$\begin{array}{ccl}\text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ by adopting } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no basis to reject } \mathcal{H}_0.\end{array}$$

Note!
Testing for normality of distribution can be done for variables or for differences determined from two variables.

Kolmogorov-Smirnov test for normality

The test proposed by Kolmogorov (1933)[92] is a relatively conservative test (it is more difficult to prove the non-normality of the distribution using it). It is based on the determination of the distance between the empirical and theoretical normal distribution. It is recommended to use it for large samples, but it should be used when the mean (μ) and standard deviation (σ) for the population from which the sample is drawn are known. We can then check that the distribution conforms to the distribution defined by the given mean and standard deviation.

Based on the sample data collected in the 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:
$F_n(x)$ – empirical cumulative distribution of the normal curve computed at individual points of the distribution, for n-element sample ,
$F(x)$ – theoretical cumulative distribution of the normal curve.

Statystyka testu podlega rozkładowi Kołmogorova-Smirnova.

Lilliefors test for normality

A test proposed by Lilliefors (1967, 1969, 1973)[101][102][103]. It is a correction of the Kolmogorov-Smirnov test when the mean (μ) and standard deviation (σ) are unknown for the population from which the sample is drawn. It is considered slightly less conservative than the Kolmogorov-Smirnov test.

The test statistic D is determined by the same formula used by the Kolmogorov-Smirnov test, but follows a Lilliefors distribution.

Shapiro-Wilk test for normality

Proposed by Shapiro and Wilk (1965)[147] for sparse groups, and adapted for more numerous groups (up to 5000 objects) by Royston (1992)[140][141]. This test has a relatively high power, which makes it easier to prove the non-normality of the distribution.
The idea of how the test works is shown in the Q-Q plot.

The Shapiro-Wilk test statistic has the form:

$$W = \frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

where:
a_i – coefficients determined based on expected values for ordered statistics, assigned weights, and covariance matrix,
$\overline{x}$ – average value of sample data.

This statistic is transformed to a statistic with a normal distribution:

$$Z = \frac{g(W) - \mu}{\sigma},$$

where:
$g(W)$, μ i σ – depend on the sample size:
– for small sample sizes $n \in < 4; 12)$:

$g(W) = -\ln(\gamma - \ln(1 - W))$,
$\gamma = 0.459n - 2.273$,
$\mu = -0.0006714n^3 + 0.025054n^2 - 0.39978n + 0.5440$,
$\sigma = \exp(-0.0020322n^3 + 0.062767n^2 - 0.77857n + 1.3822)$;

– for large sample sizes $n \in < 12; 5000 >$:

$g(W) = \ln(1 - W)$,
$\mu = 0.0038915u^3 - 0.083751u^2 - 0.31082u - 1.5851$,
$\sigma = \exp(0.0030302u^2 - 0.082676u - 0.4803)$,
$u = \ln(n)$.

D'Agostino-Pearson test for normality

Different types of statistical analyses that assume normality are sensitive to different degrees to different types of departure from this assumption. Tests that refer to means in their hypotheses are assumed to be more sensitive to skewness, and tests that compare variances are assumed to depend more on kurtosis.
A normal distribution should be characterized by zero skewness and zero kurtosis g2 (or b2 close to the value three). If the distribution is not normal, as found by the D'Agostino (1973)[3] test, one can check whether this is the result of high skewness or kurtosis by the skewness test and the kurtosis test.
Like the Shapiro-Wilk test, the D'Agostino test has higher power than the Kolmogorov-Smirnov test and the Lilliefors test (D'Agostino 1990[4]).

The test statistic is of the form:

$$K^2 = Z_A^2 + Z_K^2,$$

where:
Z_A^2 – test statistic for the skewness,
Z_K^2 – test statistic for kurtosis.

This statistic has an asymptotically χ^2 distribution with two degrees of freedom.

- **D'Agostino skewness test**
  Hypotheses:

  $\mathcal{H}_0$: distribution is not skewed (skewness in the population is zero),
  $\mathcal{H}_1$: the distribution is skewed (skewness in the population deviates from zero).

  The test statistic has the form:

$$Z_A = \delta \ln \left(\frac{Y}{\alpha} + \sqrt{\frac{Y^2}{\alpha^2} + 1} \right),$$

  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 has asymptotically (for large suple size) a normal distribution.

- **D'Agostino kurtosis test**
  Hypotheses:

  $\mathcal{H}_0$: kurtosis in the population corresponds to the kurtosis of a normal distribution,
  $\mathcal{H}_1$: the kurtosis in the population differs from the kurtosis of a normal distribution.

  The test statistic has the form:

$$Z_K = \frac{\left(1 - \frac{2}{9H}\right) - \left(\frac{1 - \frac{2}{A}}{1 + x\sqrt{\frac{2}{H-4}}} \right)^{1/3}}{\sqrt{\frac{2}{9H}}},$$

  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 has asymptotically (for large suple size) a normal distribution.

Quantile-Quantile plot *(Q-Q plot)*

A quantile-quantile type plot is used to show the correspondence of two distributions. When testing the fit of a normal distribution, it checks the fit of the data distribution (empirical distribution) to a Gaussian theoretical distribution. From it, you can visually see how well the normal distribution curve fits the data. If the quantiles of the theoretical distribution and the empirical distribution match, then the points are distributed along the line $y = x$. The horizontal axis represents the quantiles of the normal distribution, the vertical axis the quantiles of the data distribution

Various deviations from the normal distribution are possible – the interpretation of some of the most common ones is described in the diagram:

- data spread out on the line, but a few points deviate strongly from the line

  – there are outliers in the data
- points on the left side of the graph are above the line and on the right side are below the line

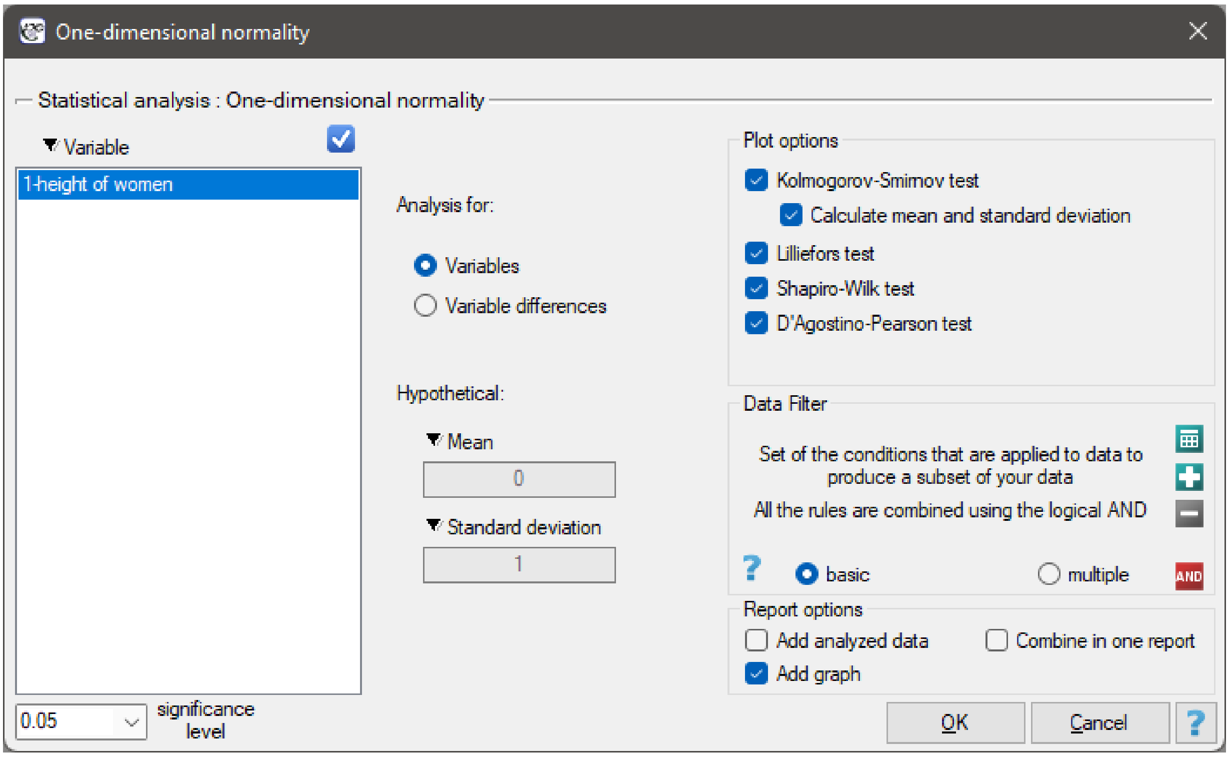
  – the distribution is characterized by a greater presence of outliers from the mean than is the case in a normal distribution (negative kurtosis)
- points on the left side of the graph are below the line and the points on the right side are above the line

  – the distribution is characterized by a smaller presence of values away from the mean than is the case in a normal distribution (positive kurtosis)
- points on the left and right sides of the graph are above the line

  – right-skewed distribution (positive skewness);
- points on the left and right sides of the graph are below the line

  – left-skewed distribution (negative skewness).

The window with the settings for the normality tests options is invoked via the Statistics→Normality tests→One-dimensional normality menu or via Wizard.



One-dimensional normality

Statistical analysis : One-dimensional normality

Variable ☑

1-height of women

Analysis for:

● Variables
○ Variable differences

Hypothetical:

Mean: 0

Standard deviation: 1

Plot options
☑ Kolmogorov-Smirnov test
☑ Calculate mean and standard deviation
☑ Lilliefors test
☑ Shapiro-Wilk test
☑ D'Agostino-Pearson test

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options
☐ Add analyzed data ☐ Combine in one report
☑ Add graph

0.05 significance level

Example 13.1. (Gauss.pqs file)

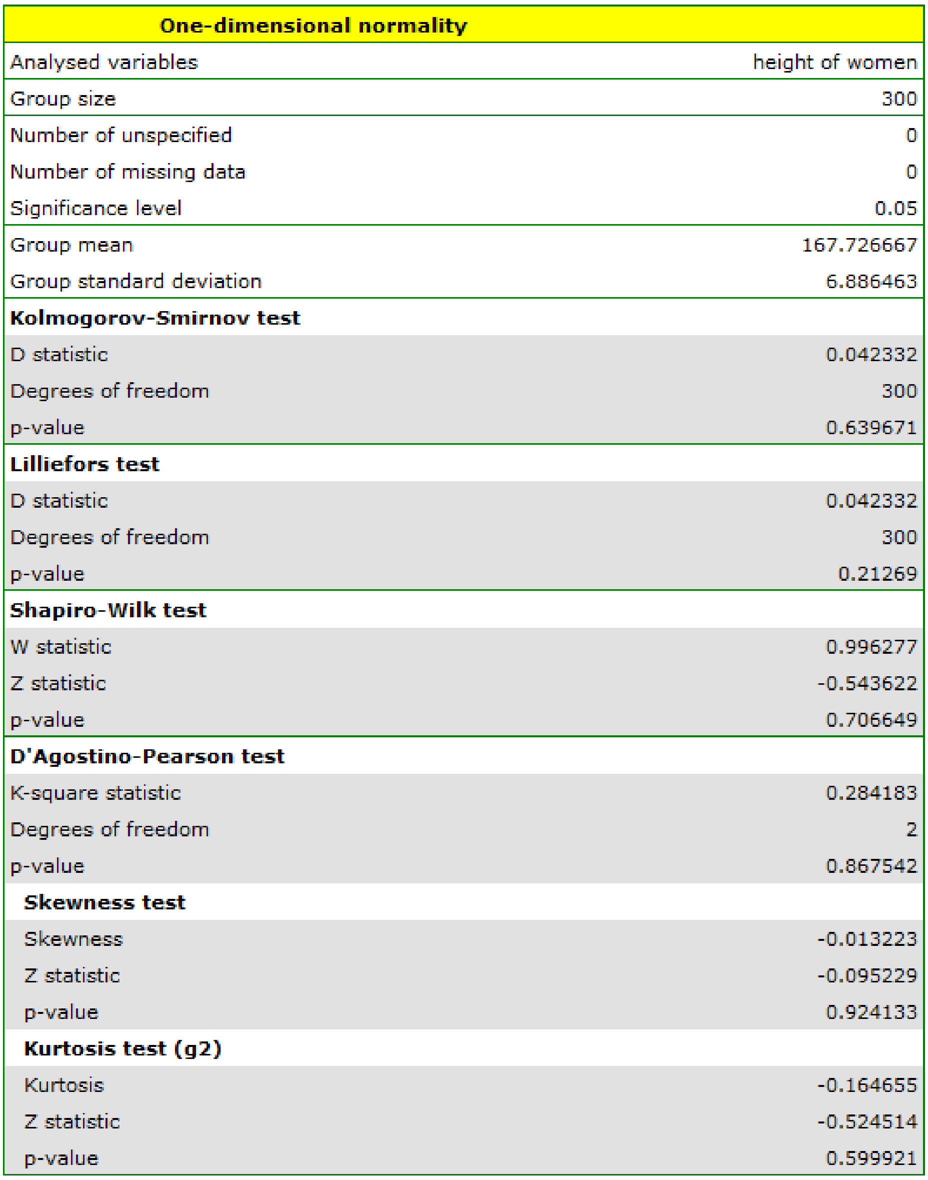
- **Women's growth**

  Let us assume that the height of women is such a characteristic, for which the average value is 168cm. 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 completely short and also very tall, but relatively rarely. Since very low and very high values occur rarely, and average values often, we can expect that the distribution of height is normal. To find out, 300 randomly selected women were measured.

  Hypotheses:

  $\mathcal{H}_0$: the height distribution of women in the study population is normal ,
  $\mathcal{H}_1$: the height distribution of women in the study population is not normal.

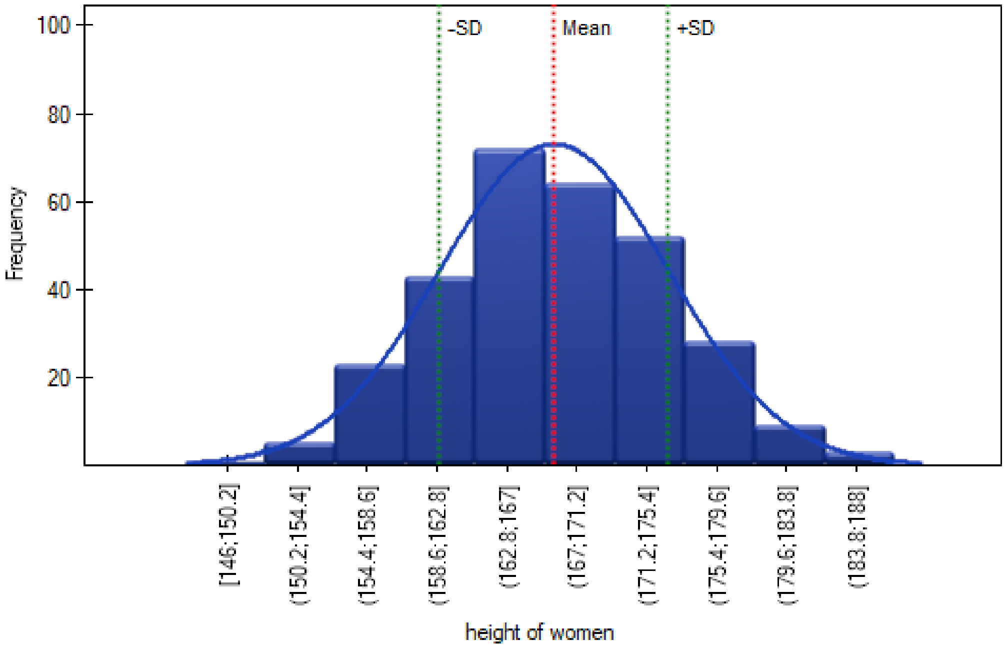
  Since we do not know the mean or standard deviation for female height, but only have an assumption about these quantities, they will be determined from the sample.



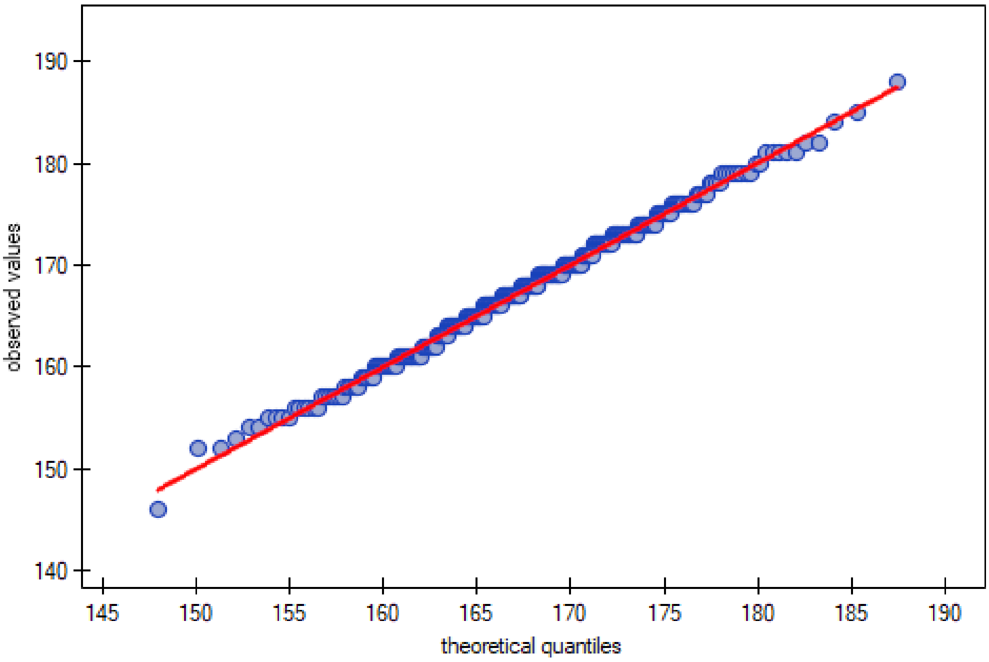
One-dimensional normality	
Analysed variables	height of women
Group size	300
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Group mean	167.726667
Group standard deviation	6.886463
Kolmogorov-Smirnov test	
D statistic	0.042332
Degrees of freedom	300
p-value	0.639671
Lilliefors 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

All of the designated tests indicate that there is no deviation from the normal distribution, as their p values are above the standard significance level of $\alpha = 0.05$. Also, the test that examines skewness and kurtosis shows no deviation.

In the column chart, we presented the height distribution as 10 columns. Women between 167 cm and 171 cm are the most numerous group, while women shorter than 150 cm or taller than 184 cm are the least numerous. The bell curve of the normal distribution seems to describe this distribution well.



In the quantile-quantile plot, the points lie almost perfectly on the line, which also indicates a very good fit of the normal distribution.



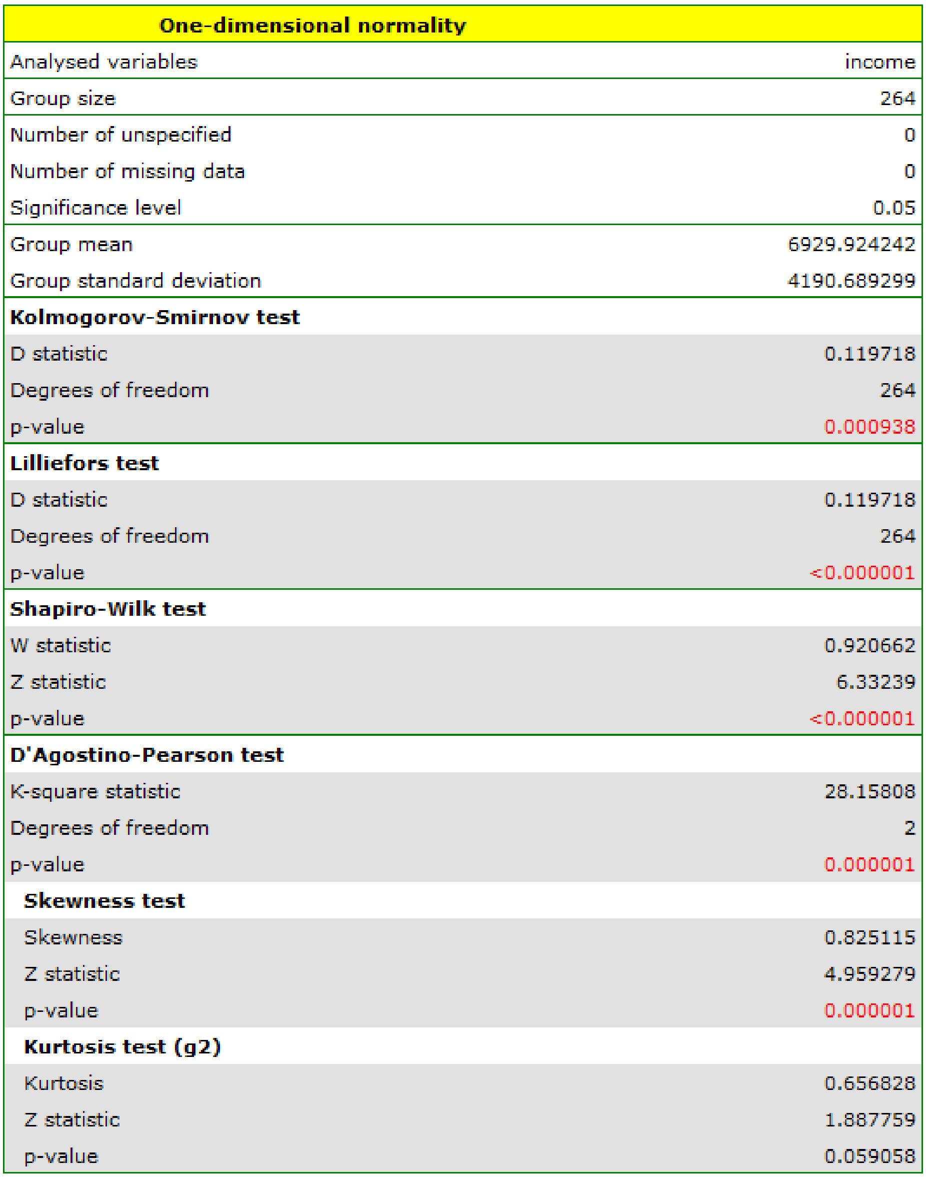
The normal distribution can therefore be regarded as the distribution that characterizes the growth of women in the population studied.

- **Income**
  Suppose we study the income of people in a certain country. We expect that the income of most people will be average, however, there will be no people earning very little (below the minimum salary imposed by the authorities), but there will be people earning very much (company presidents), who are relatively few in number. In order to check whether the income of people in

the examined country has a normal distribution, information about the income of 264 randomly selected people was collected.

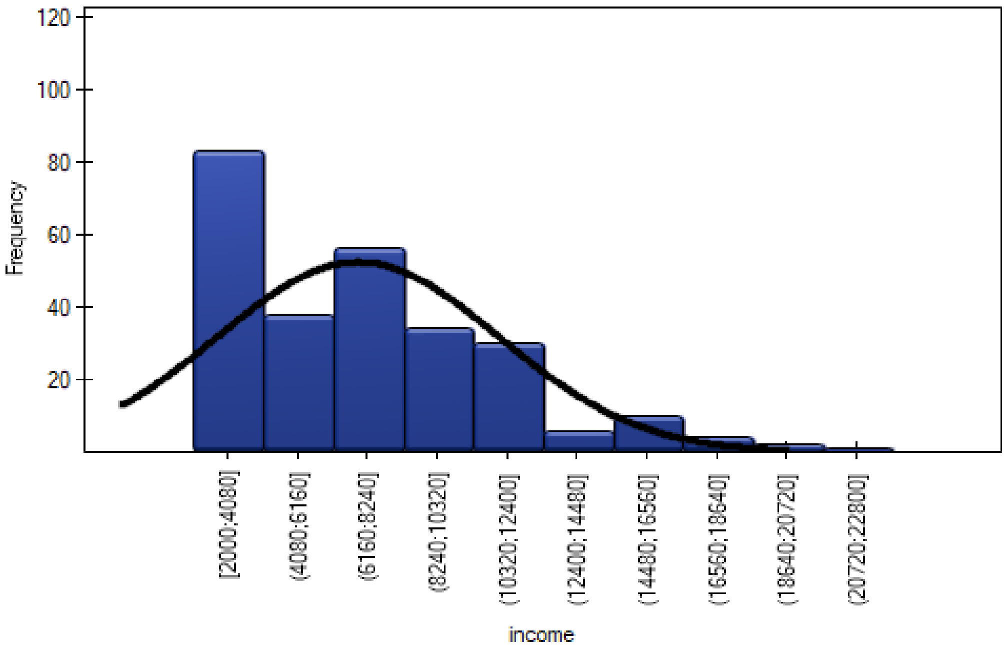
Hipotezy:

$\mathcal{H}_0$: the income distribution of individuals in the study population is a normal distribution,
$\mathcal{H}_1$: the income distribution of individuals in the study population is different than a normal distribution.

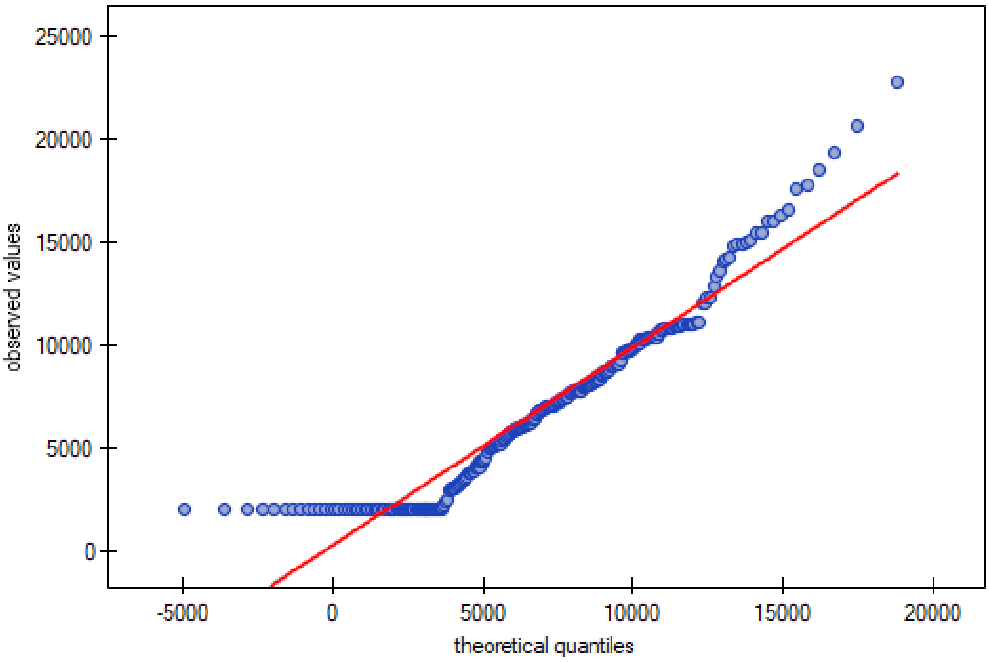


One-dimensional normality	
Analysed variables	income
Group size	264
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Group mean	6929.924242
Group standard deviation	4190.689299
Kolmogorov-Smirnov test	
D statistic	0.119718
Degrees of freedom	264
p-value	0.000938
Lilliefors test	
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

The distribution is not a normal distribution, as evidenced by all test results testing the normality of the distribution ($p < \alpha$). A positive and statistically significant ($p < \alpha$) skewness value indicates that the right tail of the function is too long. The function distribution is also more slender than the normal distribution, but this is not a statistically significant difference (kurtosis test).



In a quartile-quartile plot, the deviation from the normal distribution is illustrated by right-hand skewness, i.e., the location of the initial and final points of the plot significantly above the line.



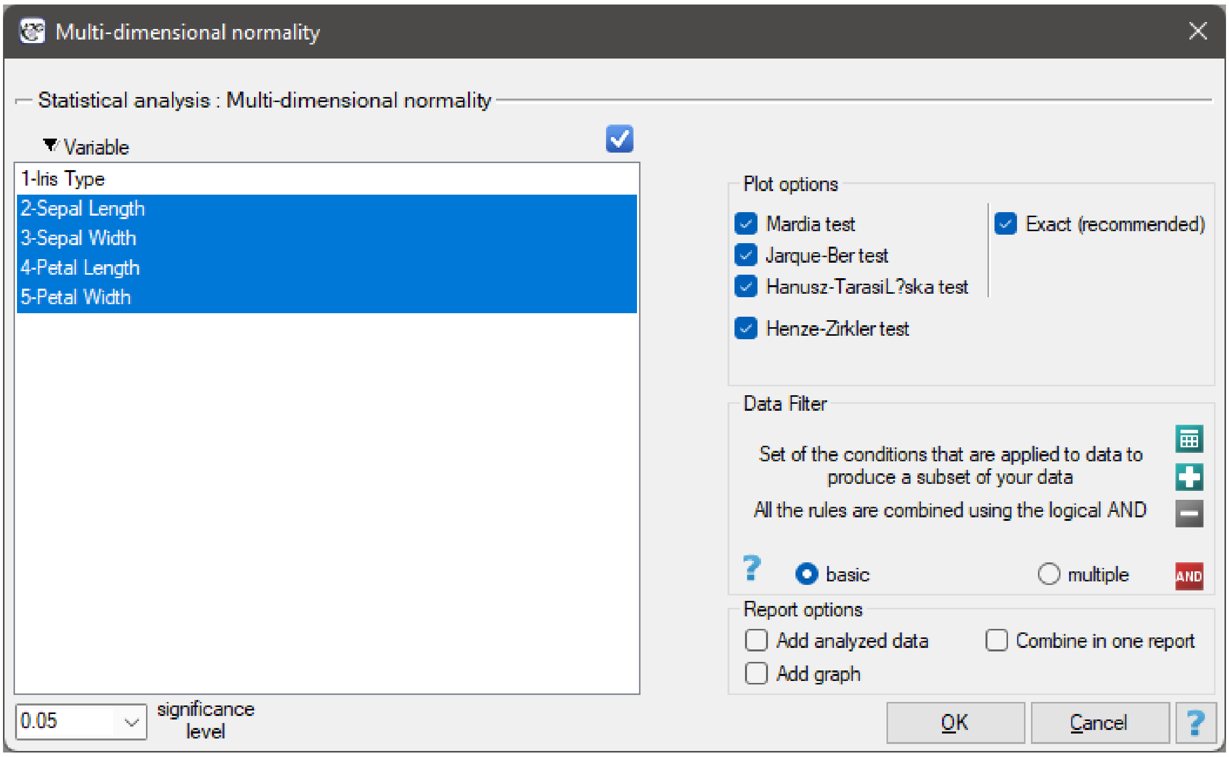
As a result, the data collected do not show that the income distribution is consistent with a normal distribution.

13.0.2 Multivariate normality tests

Many methods of multivariate analysis, including MANOVA, Hotelling tests, or regression models are based on the assumption of multivariate normality. If a set of variables is characterized by a multivariate normal distribution, then each variable can be assumed to have a normal distribution. However,

when all individual variables are characterized by a normal distribution, their set does not have to have a multivariate normal distribution. Therefore, testing the unidimensional normality of each variable may be helpful, but cannot be assumed to be sufficient.
Different types of statistical analyses that assume normality are sensitive to different degrees to different types of departure from this assumption. Tests that refer to means in their hypotheses are generally taken as more sensitive to skewness, while tests comparing covariances depend more heavily on kurtosis.

The window with the test of multivariate normality of distribution settings is opened via Statistics→Normality tests→Multi-dimensional normality.



Multi-dimensional normality

Statistical analysis : Multi-dimensional normality

Variable ☑

1-Iris Type
2-Sepal Length
3-Sepal Width
4-Petal Length
5-Petal Width

Plot options
☑ Mardia test
☑ Jarque-Ber test
☑ Hanusz-TarasiL?ska test
☑ Henze-Zirkler test
☑ Exact (recommended)

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
□ Add analyzed data □ Combine in one report
□ Add graph

0.05 significance level

Mardia's test for multivariate normality

The test proposed by Mardia in 1970 [113] and modified in 1974 [114] tests the normality of a distribution by analyzing separately the magnitude of multivariate skewness and multivariate kurtosis. Jarque and Bera [85] proposed combining these two Mardia measures into a single test. A similar way of combining skewness and kurtosis information into a single test is provided by the method of Hanusz and Tarasinska [75].
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

$$m_{ij}=\left(X_i-\bar{X}\right)^T S^{-1}\left(X_j-\bar{X}\right),$$
$$S=\frac{1}{n}\sum_{j=1}^{n}\left(X_i-\bar{X}\right)\left(X_i-\bar{X}\right)^T,$$
$\bar{X}$ -mean, S - covariance matrix.

For data derived 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 $(n/(n-1))^2$.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & \text{population data distribution} = \text{multivariate normal distribution},\\ \mathcal{H}_1: & \text{population data distribution} \neq \text{multivariate normal distribution},\end{array}$$

- **Mardia test of skewness:** When the sample is drawn from a population with a multivariate normal distribution (null hypothesis), the test statistic is in the form of (Mardia, 1970):

  $$\chi^2(M)=\frac{n}{6}skew$$

  or with correction of exact moments for groups with smaller numbers (<20) (Mardia, 1974):

  $$\chi^2_c(M)=\frac{n}{6}\frac{(n+1)(n+3)(k+1)}{n((n+1)(k+1)-6)}skew$$

  This statistic has asymptotically (for large numbers) distribution χ^2 with $df = f = \frac{k(k+1)(k+2)}{6}$ degrees of freedom.

- **Mardia test of kurtosis:** When the sample is drawn from a population with a multivariate normal distribution (null hypothesis), the test statistic is in the form of (Mardia, 1974):

  $$Z(M)=\frac{kurt-k(k+2)}{\sqrt{\frac{8k(k+2)}{n}}}$$

  or with 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 has asymptotically (for large numbers) normal distribution.

The value p, determined on the basis of test statistics, for both tests i.e. of the skewness test and the kurtosis test are compared with the significance level α :

$$\begin{array}{rcl}\text{jeżeli } p\le\alpha \text{ for at least one test} & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ adopting } \mathcal{H}_1,\\ \text{jeżeli } p>\alpha \text{ for both tests used} & \Longrightarrow & \text{there is no basis to reject } \mathcal{H}_0.\end{array}$$

Jarque-Bera test for multivariate normality
Jarque and Bera's (1987) [85] test is based on the skewness and kurtosis statistics of the Mardia test. The test statistic is in the form of:

$$\chi^2(JB)=\chi^2(M)+(Z(M))^2$$

or with correction (Mardia, 1974)

$$\chi^2_c(JB)=\chi^2_c(M)+(Z_c(M))^2$$

This statistic has asymptotically (for large numbers) χ^2 distribution with $df = f+1$ degrees of freedom.

The value p, determined on the basis of test statistics, we compare with the significance level α :

$$\begin{array}{rcl}\text{if } p\le\alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ by adopting } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no basis to reject } \mathcal{H}_0.\end{array}$$

Hanusz-Tarasinska test for multivariate normality

Zofia Hanusz and Joanna Tarasinska's (2014) [75] test is based on the skewness and kurtosis statistics of the Mardia test. The test statistic is in the form of:

$$t_c(HT) = \frac{Z_c(M)}{\sqrt{\frac{\chi_c^2(M)}{f}}}$$

The test statistic has t-Student distribution with $df = f$ degrees of freedom.

The value p, determined on the basis of test statistics, we compare with the significance level α :

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ by adopting } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there are no grounds to reject } \mathcal{H}_0. \end{array}$$

Henze-Zirkler test for multivariate normality

Henze and Zirkler (1990) [77] proposed a test to examine multivariate normality of the distribution extending the work of Baringhaus and Henze on the empirical characteristic function [54]. In the literature, this test is considered one of the strongest tests dedicated to multivariate normal distribution (Thode 2002) [161]. The test statistic has the form:

$$Z(HZ)_\beta = n\left(4I_E + D_{n,\beta}I_{E^c}\right)$$

I_E and I_{E^c} are indicator functions that depend on the singularities of the covariance matrix,

$D_{n,\beta} = \frac{1}{n^2}\sum exp\left(\frac{-\beta^2||Y_j-Y_k||^2}{2}\right) + (1+2\beta^2)^{-p/2} - 2(1+\beta^2)^{-p/2}\sum exp\left(\frac{-\beta^2||Y_j||^2}{2(1+\beta^2)}\right)$

$Y_i = S^{1/2}(X_i - \bar{X})$

$\beta* = 2^{-1/2}\left(\frac{n(2k+1)}{4}\right)^{1/(k+4)}$ - optimum parameter value β

The statistic $Z(HZ)_\beta$ has an asymptotically (for large sizes) normal distribution based on the mean and variance described by Henze and Zirkler and read one-sided.

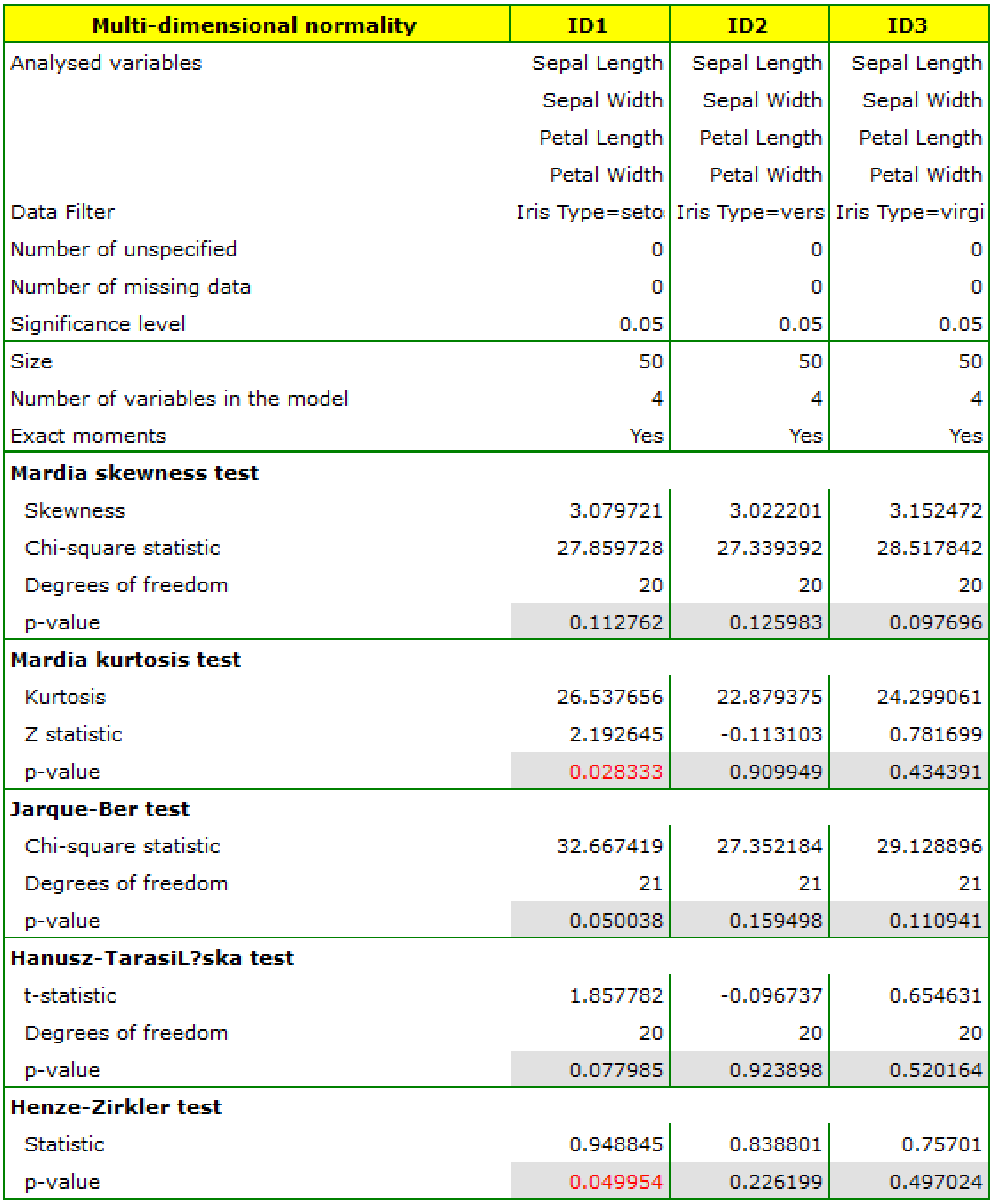
The value p, determined on the basis of test statistics, we compare with the significance level α :

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ by adopting } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there are no grounds to reject } \mathcal{H}_0. \end{array}$$

Example 13.2. (Iris.pqs file)

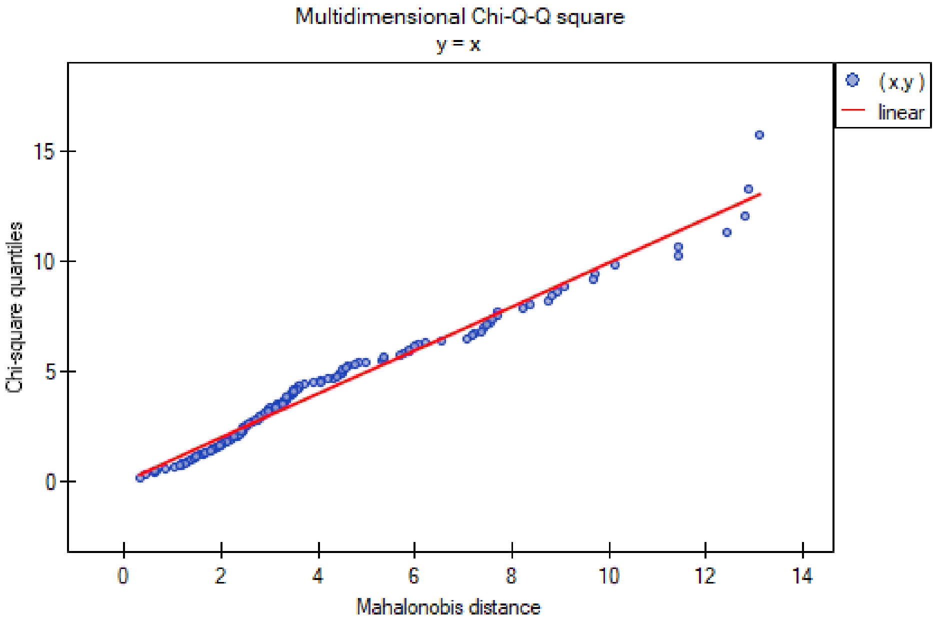
We examine the normality of the distribution for the classical data set of R.A. Fisher 1936 [58]. The file can be found in the manual included with the program and contains measurements of the length and width of the petals and calyx sepals for 3 varieties of iris flower. The analysis will be performed separately for each variety.

In the analysis window, select all tests and the graph, and set a multiple filter to repeat the analysis for each variety of iris. All the results will be returned to the same datasheet, so select Combine into one report.

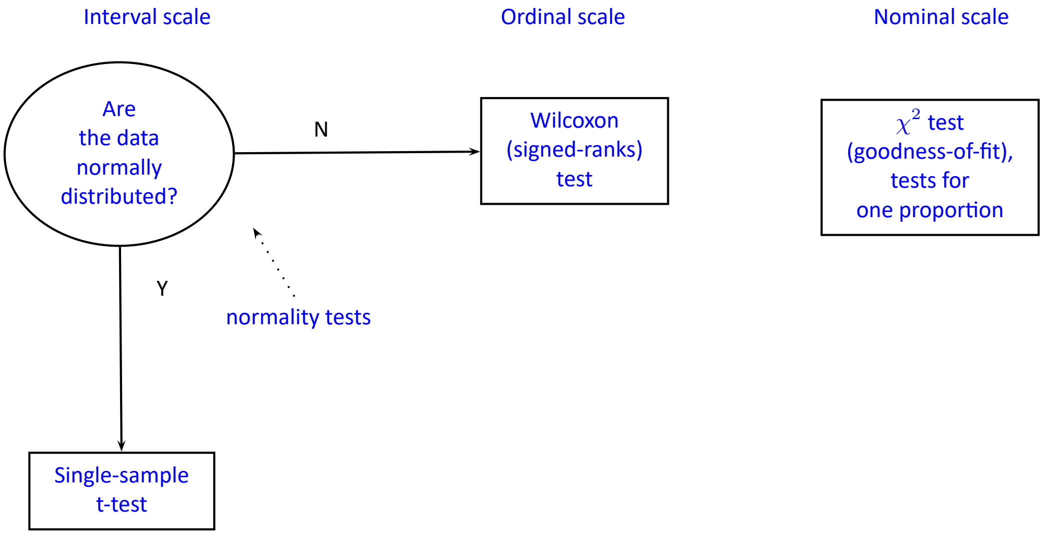


Multi-dimensional normality	ID1	ID2	ID3
Analysed variables	Sepal Length	Sepal Length	Sepal Length
	Sepal Width	Sepal Width	Sepal Width
	Petal Length	Petal Length	Petal Length
	Petal Width	Petal Width	Petal Width
Data Filter	Iris Type=seto	Iris Type=vers	Iris Type=virgi
Number of unspecified	0	0	0
Number 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 confirm the normality of the distribution for the versicolor and virginica varieties. For the setosa cultivar, the test results are on the verge of statistical significance, with the Mardia test for Kurtosis and the Henze-Zirkler test indicating deviations from the multivariate normal distribution. We can observe such deviations also in the first graph, where as the Mahalanobis distance increases, the points are further and further from the straight line.



14 COMPARISON - 1 GROUP


Interval scale
Ordinal scale
Nominal scale
Are the data normally distributed?
N
Wilcoxon (signed-ranks) test
χ^2 test (goodness-of-fit), tests for one proportion
Y
normality tests
Single-sample t-test

14.1 PARAMETRIC TESTS

14.1.1 The t-test for a single sample

The single-sample t test is used to verify the hypothesis, that an analysed sample with the mean ($\overline{x}$) comes from a population, where mean (μ) is a given value.

Basic assumptions:

- measurement on an interval scale,
- normality of distribution of an analysed feature.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \mu=\mu_0, \\ \mathcal{H}_1: & \mu\neq\mu_0, \end{array}$$

where:
μ – mean of an analysed feature of the population represented by the sample,
μ_0 – a given value.

The test statistic is defined by:

$$t=\frac{\overline{x}-\mu_0}{sd}\sqrt{n},$$

where:
sd – standard deviation from the sample,
n – sample size.

The test statistic has the t-Student distribution with $n-1$ degrees of freedom.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl} \text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note
Note, that: If the sample is large and you know a standard deviation of the population, then you can calculate a test statistic using the formula:

$$t=\frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}.$$

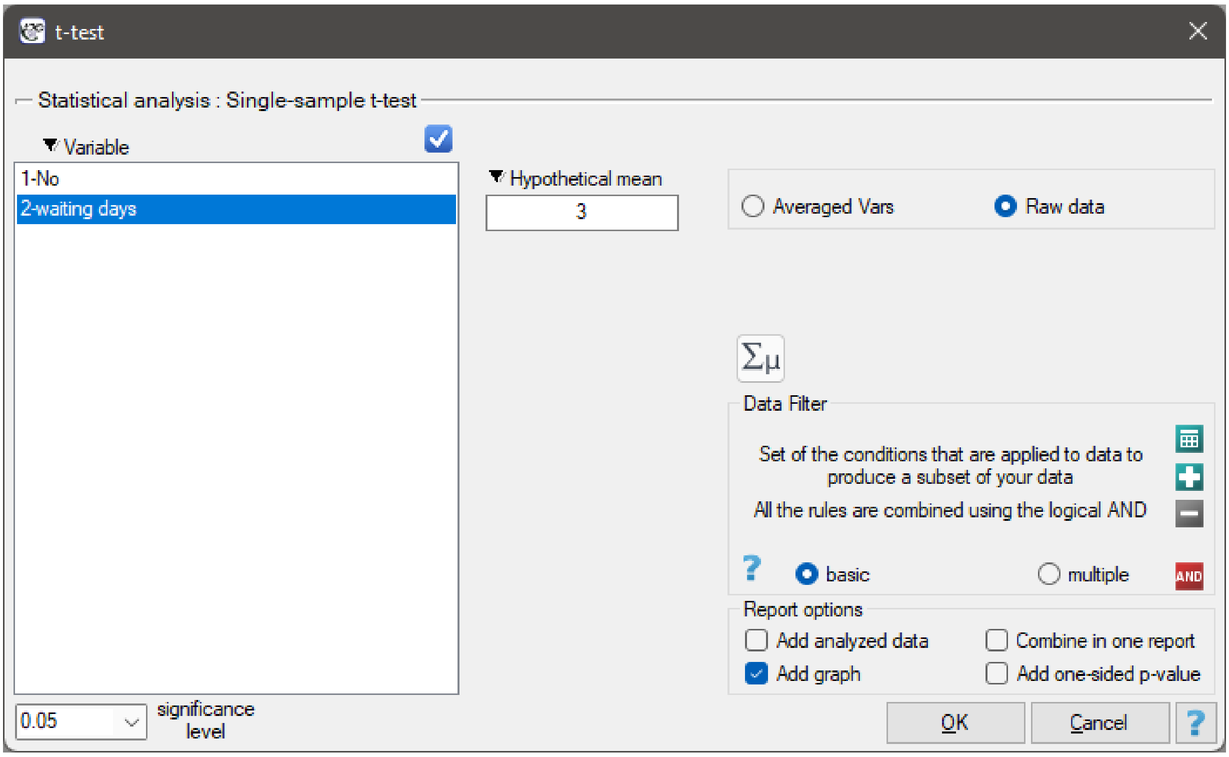
The statistic calculated this way has the normal distribution. If $n\rightarrow\infty$ t-Student distribution converges to the normal distribution $N(0,1)$. In practice, it is assumed, that with $n>30$ the t-Student distribution may be approximated with the normal distribution.

Standardized effect size
The **Cohen's d** determines how much of the variation occurring is the difference between the averages.

$$d=\left|\frac{\overline{x}-\mu_0}{sd}\right|$$

When interpreting an effect, researchers often use general guidelines defined by Cohen [45] defining small (0.2), medium (0.5) and large (0.8) effect sizes.

The settings window with the Single-sample t-test can be opened in Statistics menu→Parametric tests→t-test or in Wizard.



t-test

Statistical analysis : Single-sample t-test

Variable ☑

1-No
2-waiting days

Hypothetical mean: 3

○ Averaged Vars ● Raw data

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report
☑ Add graph □ Add one-sided p-value

0.05 significance level

OK Cancel

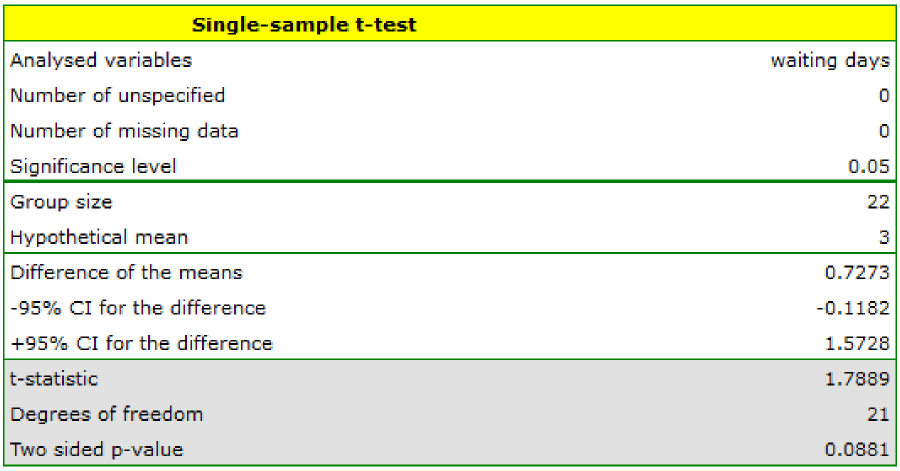
Note
Calculations can be based on raw data or data that are averaged like: arithmetic mean, standard deviation and sample size.

Example 14.1. (courier.pqs file)
You want to check if the time of awaiting for a delivery by some courier company is 3 days on the average $(\mu_0 = 3)$. In order to calculate it, there are 22 persons chosen by chance from all clients of the company as a sample. After that, there are written information about the number of days passed since the delivery was sent till it is delivered. There are following values: (1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7).

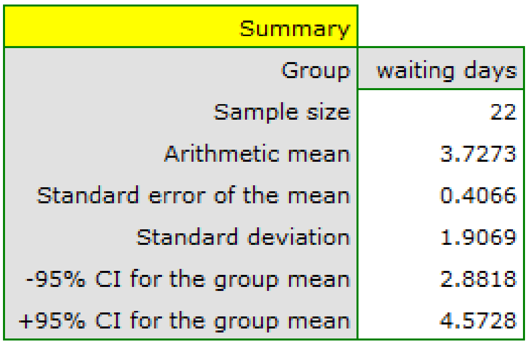
The number of awaiting days for the delivery in the analysed population fulfills the assumption of normality of distribution.

Hypotheses:

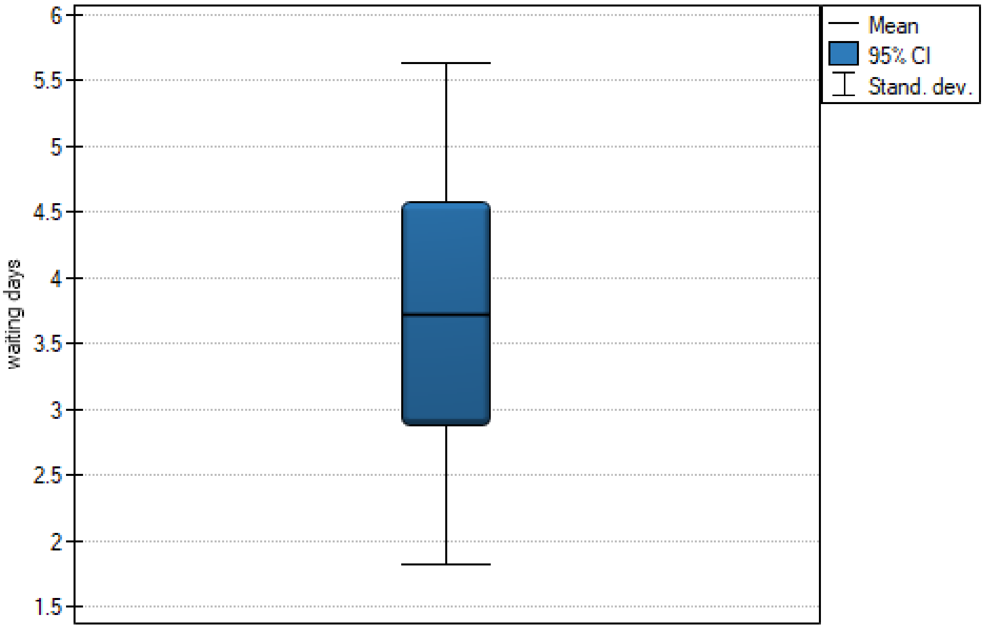
$\mathcal{H}_0$: mean of the number of awaiting days for the delivery, which is supposed to be delivered by the above-mentioned courier company is 3,
$\mathcal{H}_1$: mean of the number of awaiting days for the delivery, which is supposed to be delivered by the above-mentioned courier company is different from 3.



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



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



Comparing the p value = 0.0881 of the t-test with the significance level $\alpha = 0.05$ we draw the conclusion, that there is no reason to reject the null hypothesis which informs that the average time of awaiting for the delivery, which is supposed to be delivered by the analysed courier company is 3. For the tested sample, the mean is $\overline{x} = 3.73$ and the standard deviation is $sd = 1.91$.

14.1.2 The Single-Sample Chi-square Test for a Population Variance

The χ^2 test of the variance of a single sample is used to verify the hypothesis that the sample being tested comes from a population for which the variance (or standard deviation σ) is a given value. At the same time, hypotheses can refer to both the variance and equivalently the standard deviation.

Basic assumptions:

- measurement on an interval scale,
- normality of distribution of an analysed feature.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \sigma = \sigma_0, \\ \mathcal{H}_1: & \sigma \neq \sigma_0, \end{array}$$

where:
σ – standard deviation of a characteristic in the population represented by the sample,
σ_0 – setpoint.

The test statistic is defined by:

$$t = \frac{(n-1)sd^2}{\sigma_0^2},$$

where:
sd – standard deviation in the sample,
n – sample size.

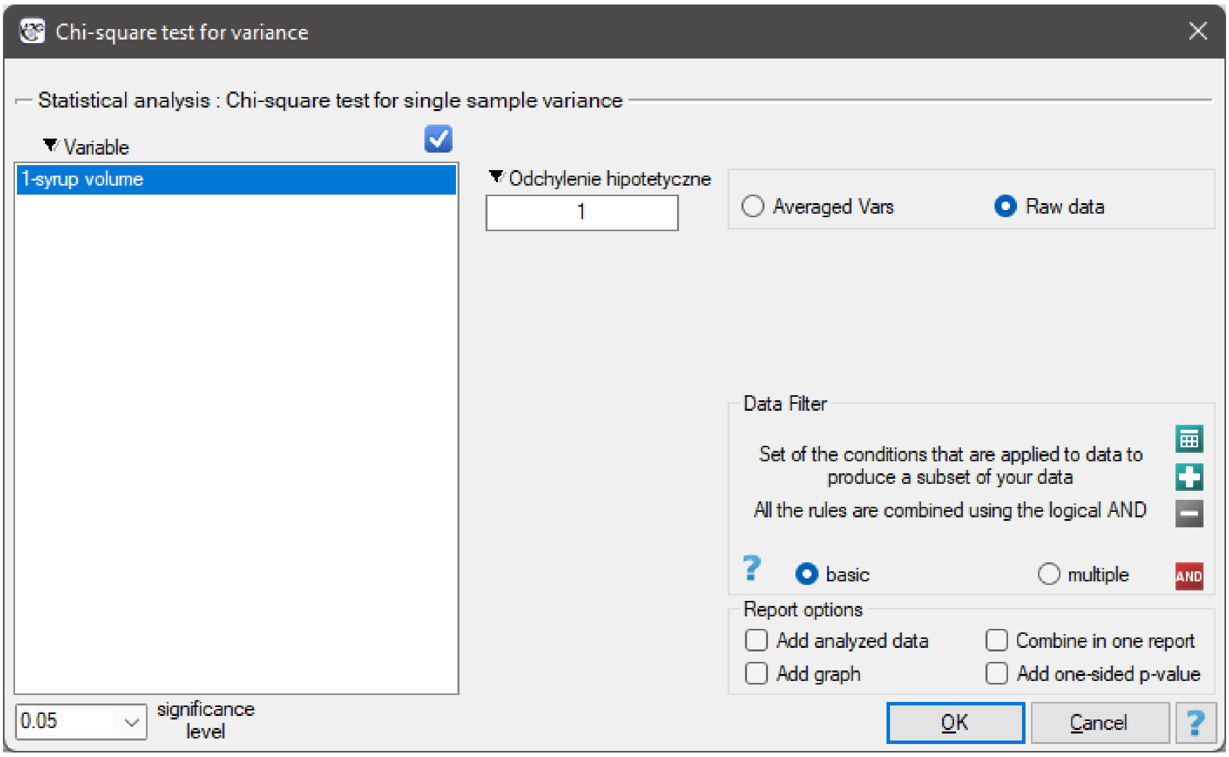
The test statistic has the χ^2 distribution with the degrees of freedom determined by the formula: $df = n-1$.

The p value, designated on the basis of the test statistic, is compared with the

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Whereby, if the standard deviation value is less than the setpoint, the p value is calculated as the doubled value of the area under the chi-square distribution curve to the left of the corresponding critical value, and if it is greater than the setpoint, it is the doubled value of the corresponding area to the right.

he settings window with the χ^2 test for variance w can be opened in Statistics menu→Parametric tests→Chi-square test for variance.



Chi-square test for variance

Statistical analysis : Chi-square test for single sample variance

Variable: 1-syrup volume

Odchylenie hipotetyczne: 1

○ Averaged Vars ● Raw data

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

□ Add graph □ Add one-sided p-value

0.05 significance level

Note!

Calculations can be based on raw data or data that are averaged like: standard deviation and sample size.

Example 14.2. (dispenser.pqs file)

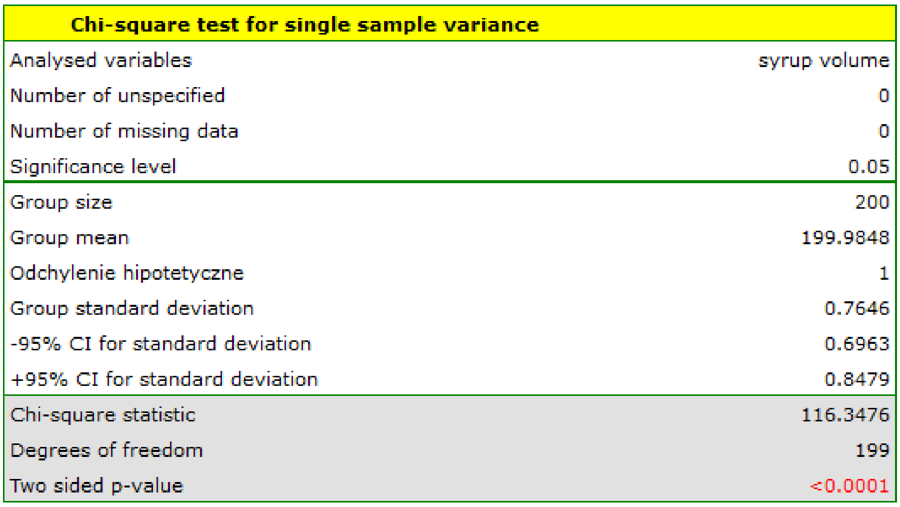
Before starting the production of another batch of a certain cough syrup, control measurements of the volume of syrup poured into the bottles were made. The technical documentation of the dosing device shows that the permissible variation in syrup volume measured by the standard deviation is 1ml. It should be verified that the tested device is working properly.

The distribution of the volume of syrup poured into the bottles was checked (with the Lilliefors test) obtaining a result consistent with this distribution. The analysis concerning the standard deviation can therefore be performed with the chi-square test for variance

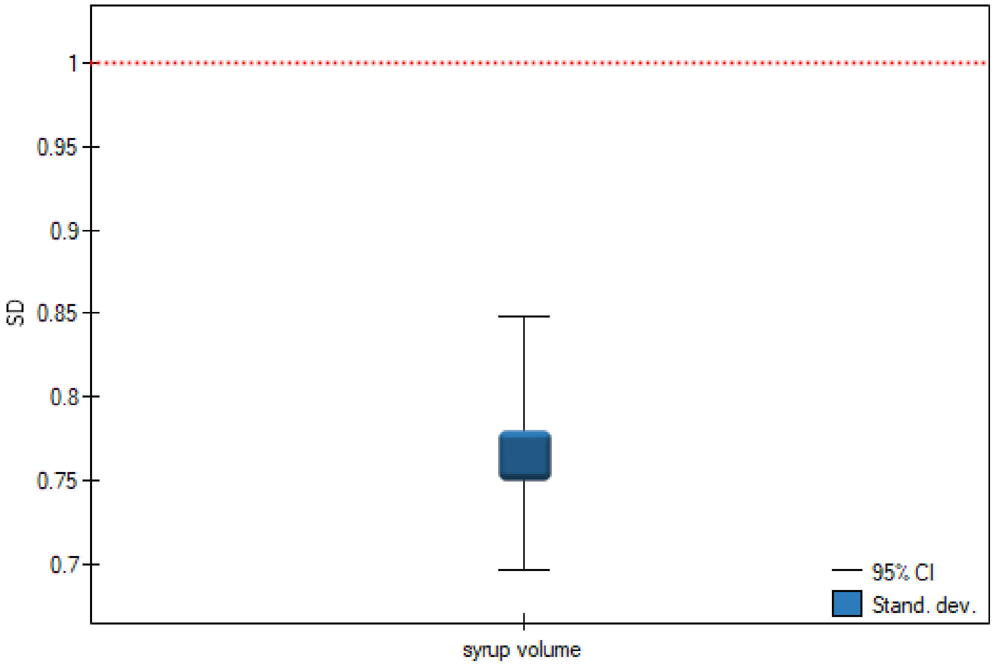
Hypotheses:

$\mathcal{H}_0$: the standard deviation of the volume of syrup poured by the dosing device is 1ml,

$\mathcal{H}_1$: the standard deviation of the volume of syrup poured by the dosing device is other than 1ml.



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



Comparing the $p < 0.0001$ value of the χ^2 test with the significance level $\alpha = 0.05$ we find that the scatter of the dispensing device is different from 1ml. However, we can consider the performance of the device as correct because the standard deviation of the sample is 0.76, which is significantly less than the acceptable value from the technical documentation.

14.2 NON-PARAMETRIC TESTS

14.2.1 The Wilcoxon test (signed-ranks)

The Wilcoxon signed-ranks test is also known as the Wilcoxon single sample test, Wilcoxon (1945, 1949)[169]. This test is used to verify the hypothesis, that the analysed sample comes from the population, where median (θ) is a given value.

Basic assumptions:

- measurement on an ordinal scale or on an interval scale.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \theta = \theta_0, \\ \mathcal{H}_1: & \theta \neq \theta_0. \end{array}$$

where:
θ – median of an analysed feature of the population represented by the sample,
θ_0 – a given value.

Now you should calculate the value of the test statistics Z (T – for the small sample size), and based on this p value.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note
Depending on the size of the sample, the test statistic takes a different form:

- for a small sample size

$$T = \min\left(\sum R_-, \sum R_+\right),$$

where:
$\sum R_+$ and $\sum R_-$ are adequately: a sum of positive and negative ranks.

This statistic has the Wilcoxon distribution

- 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 being included in the interlinked rank.

The test statistic formula Z includes the correction for ties. This correction should be used when ties occur (when there are no ties, the correction is not calculated, because $\left(\sum t^3 - \sum t\right)/48 = 0$.

Z statistic asymptotically (for a large sample size) has the normal distribution.

Continuity correction of the Wilcoxon test (Marascuilo and McSweeney (1977)[112])
A continuity correction is used to enable the test statistic to take in all values of real numbers, according to the assumption of the normal distribution. Test statistic with a continuity correction is defined by:

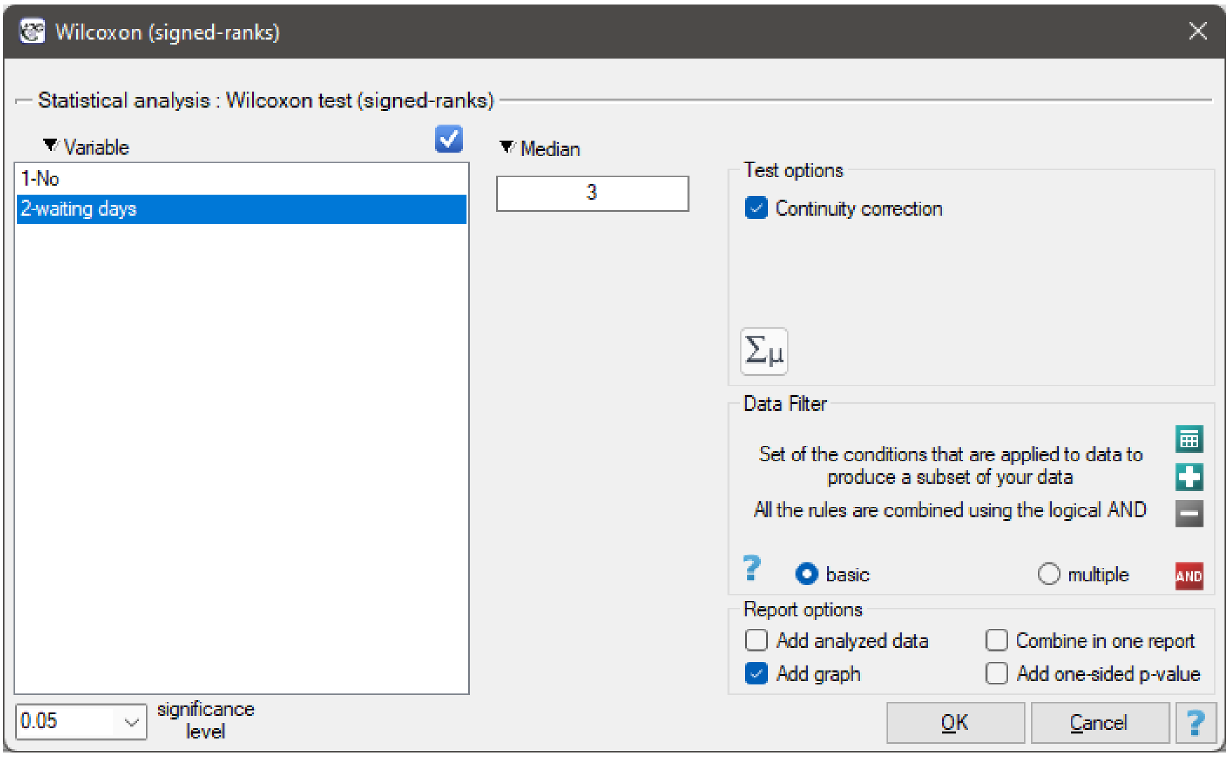
$$Z=\frac{\left|T-\frac{n(n+1)}{4}\right|-0.5}{\sqrt{\frac{n(n+1)(2n+1)}{24}-\frac{\sum t^3-\sum t}{48}}}.$$

Standardized effect size
The distribution of the Wilcoxon test statistic is approximated by the normal distribution, which can be converted to an effect size $r=|Z/n|$ [?] to then obtain the **Cohen's d** value according to the standard conversion used for meta-analyses:

$$d=\frac{2r}{\sqrt{1-r^2}}$$

When interpreting an effect, researchers often use general guidelines proposed by Cohen [45] defining small (0.2), medium (0.5) and large (0.8) effect sizes.

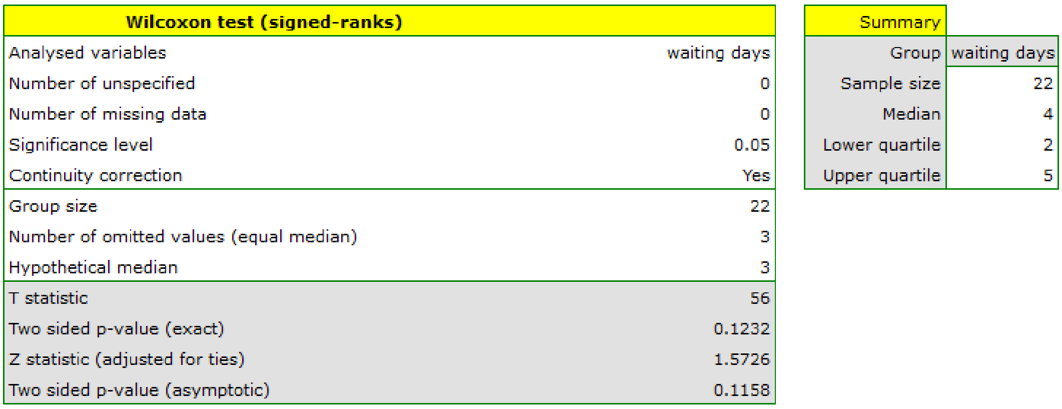
The settings window with the Wilcoxon test (signed-ranks) can be opened in Statistics menu→ Non-Parametric tests→Wilcoxon (signed-ranks) or in Wizard.



Example 14.1 cont. *(courier.pqs file)*

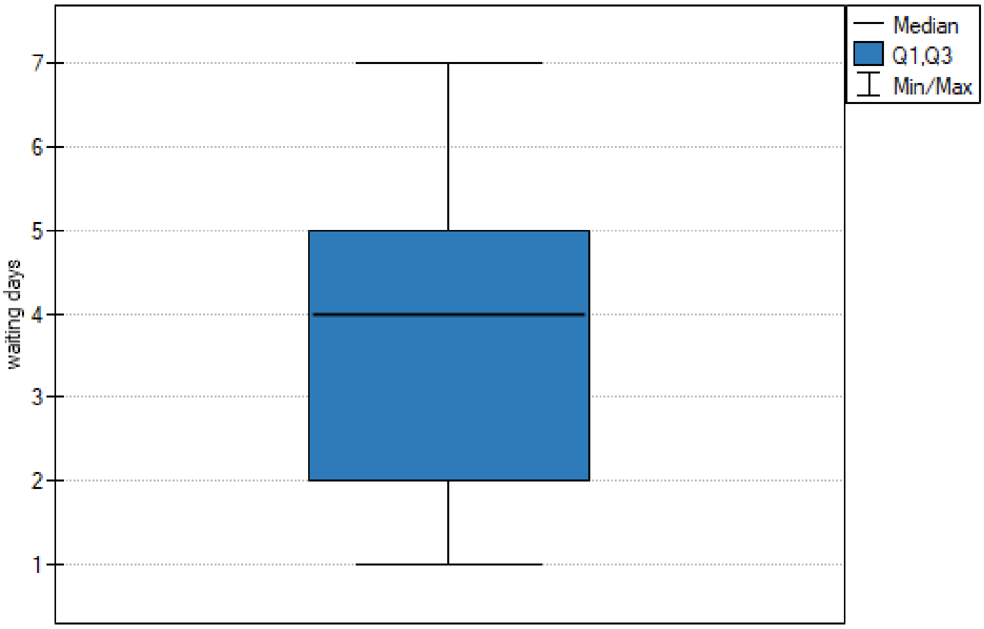
Hypotheses:

$\mathcal{H}_0$: median of the number of awaiting days for the delivery, which is supposed to be delivered by the analysed courier company is 3
$\mathcal{H}_1$: median of the number of awaiting days for the delivery, which is supposed to be delivered by the analysed courier company is different from 3



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



Comparing the p-value = 0.1232 of Wilcoxon test based on T statistic with the significance level $\alpha = 0.05$ we draw the conclusion, that there is no reason to reject the null hypothesis informing us, that usually the number of awaiting days for the delivery which is supposed to be delivered by the analysed courier company is 3. Exactly the same decision you would make basing on the p-value = 0.1112 or p-value = 0.1158 of Wilcoxon test based upon Z statistic or Z with correction for continuity.

14.2.2 The Chi-square goodness-of-fit test

The χ^2 test (goodnes-of-fit) is also called the one sample χ^2 test and is used to test the compatibility of values observed for r ($r >= 2$) categories $X_1, X_2, ..., X_r$ of one feature X with hypothetical expected values for this feature. The values of all n measurements should be gathered in a form of a table consisted of r rows (categories: $X_1, X_2, ..., X_r$). For each category X_i there is written the frequency of its occurence O_i, and its expected frequency E_i or the probability of its occurence p_i. The expected frequency is designated as a product of $E_i = np_i$. The built table can take one of the following forms:

X_i categories	O_i	E_i
X_1	O_1	E_i
X_2	O_2	E_2
...	...	...
X_r	O_r	E_r

X_i categories	O_i	p_i
X_1	O_1	p_1
X_2	O_2	p_2
...	...	...
X_r	O_r	p_r

Basic assumptions:

- measurement on a nominal scale - any order is not taken into account,
- large expected frequencies (according to the Cochran interpretation (1952)[40],
- observed frequencies total should be exactly the same as an expected frequencies total, and the total of all p_i probabilities should come to 1.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & O_i = E_i \text{ for all categories,}\\ \mathcal{H}_1: & O_i \neq E_i \text{ for at least one category.}\end{array}$$

Test statistic is defined by:

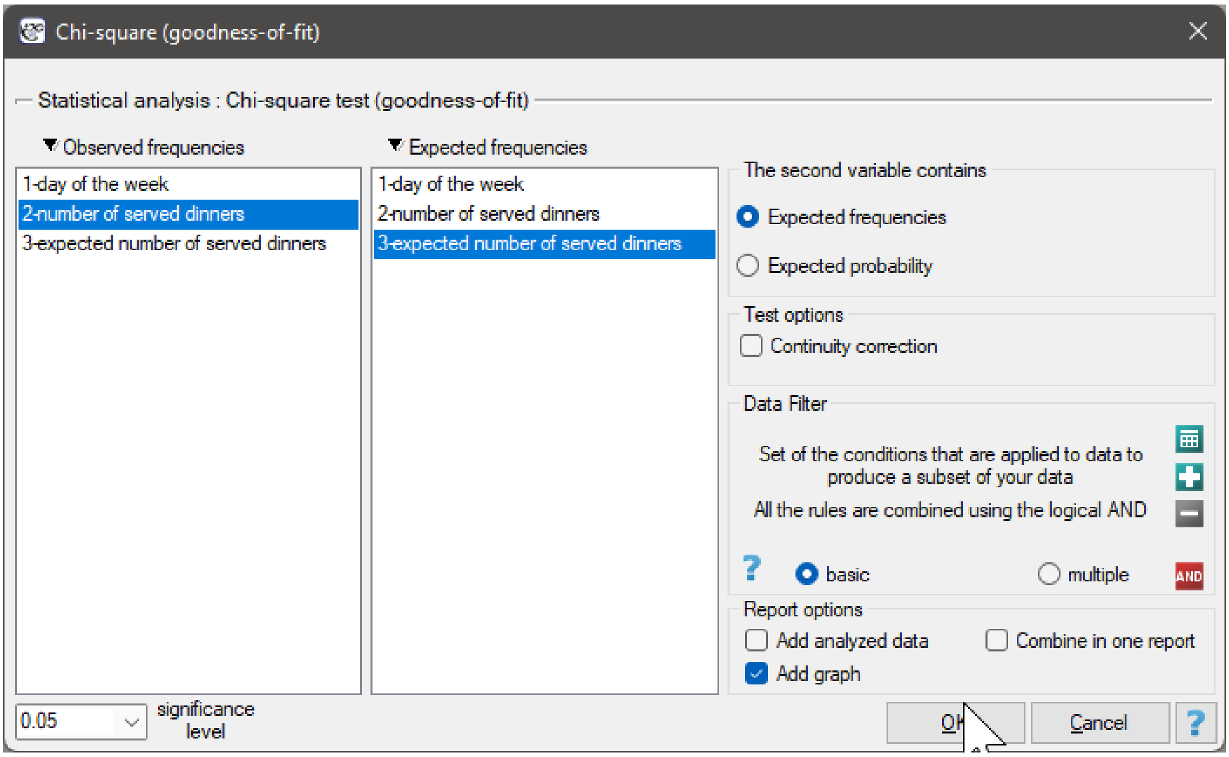
$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}.$$

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, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

The settings window with the Chi-square test (goodness-of-fit) can be opened in Statistics menu → NonParametric tests (unordered categories)→Chi-square (goodnes-of-fit) or in Wizard.


Chi-square (goodness-of-fit)

Statistical analysis : Chi-square test (goodness-of-fit)

Observed frequencies: 1-day of the week; 2-number of served dinners (selected); 3-expected number of served dinners

Expected frequencies: 1-day of the week; 2-number of served dinners; 3-expected number of served dinners (selected)

The second variable contains
● Expected frequencies
○ Expected probability

Test options
□ Continuity correction

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
□ Add analyzed data □ Combine in one report
☑ Add graph

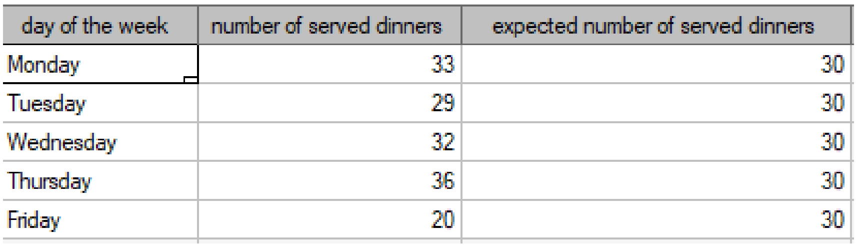
0.05 significance level

EXAMPLE 14.3. (dinners.pqs file)

We would like to get to know if the number of dinners served in some school canteen within a given frame of time (from Monday to Friday) is statistically the same. To do this, there was taken a one-week-sample and written the number of served dinners in the particular days: Monday - 33, Tuesday - 29, Wednesday - 32, Thursday -36, Friday - 20.

As a result there were 150 dinners served in this canteen within a week (5 days).

We assume that the probability of serving dinner each day is exactly the same, so it comes to $\frac{1}{5}$. The expected frequencies of served dinners for each day of the week (out of 5) comes to $E_i = 150 \cdot \frac{1}{5} = 30$.

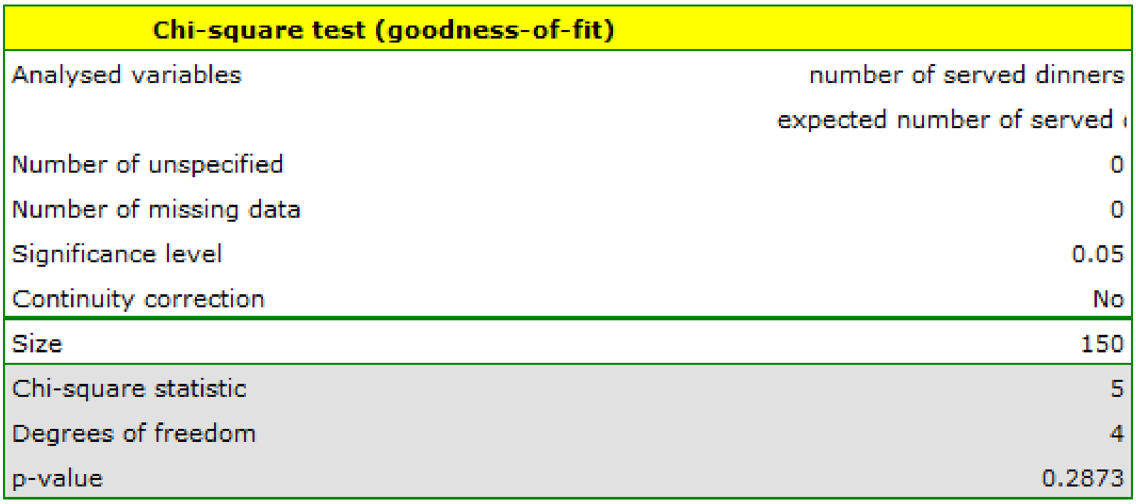


day of the week	number of served dinners	expected number of served dinners
Monday	33	30
Tuesday	29	30
Wednesday	32	30
Thursday	36	30
Friday	20	30

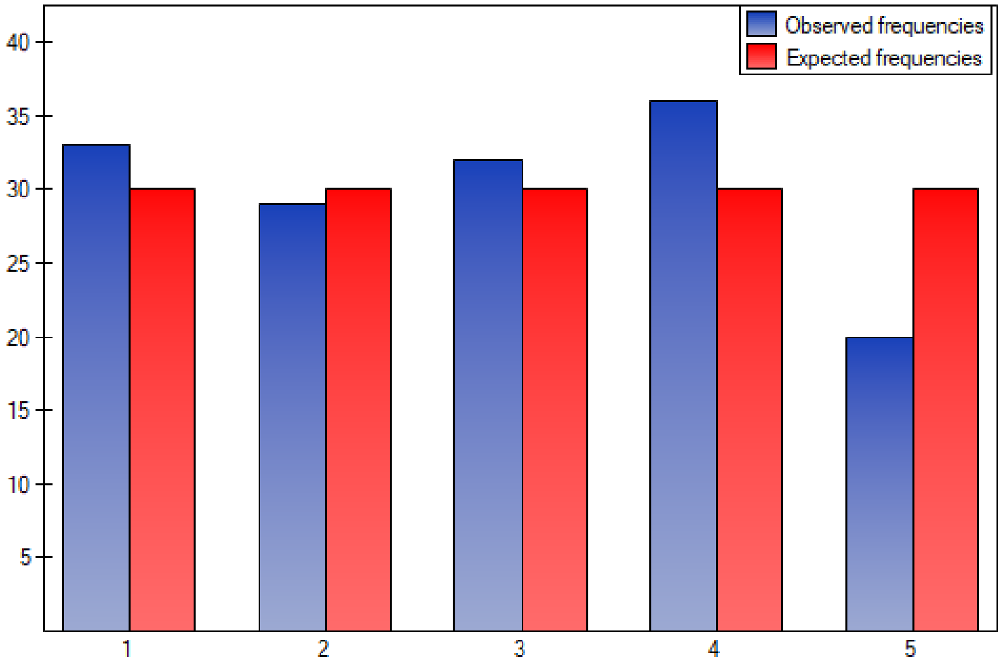
Hypotheses:

$\mathcal{H}_0$: the number of served dinners in the analysed school canteen within given days (of the week) is consistent with the expected number of given out dinners these days,

$\mathcal{H}_1$: the number of served out dinners in the analysed school canteen within a given week is not consistent with the expected number of dinners given out these days.



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



The p value from the χ^2 distribution with 4 degrees of freedom comes to 0.2873. So using the significance level $\alpha = 0.05$ you can estimate that there is no reason to reject the null hypothesis that informs about the compatibility of the number of served dinners with the expected number of dinners served within the particular days.

Note!
If you want to make more comparisons within the framework of a one research, it is possible to use the **Bonferroni correction**[2]. The correction is used to limit the size of I type error, if we compare the observed frequencies and the expected ones between particular days, for example:
Friday $\Longleftrightarrow$ Monday,
Friday $\Longleftrightarrow$ Tuesday,
Friday $\Longleftrightarrow$ Wednesday,
Friday $\Longleftrightarrow$ Thursday,

Provided that, the comparisons are made independently. The significance level $\alpha = 0.05$ for each comparison must be calculated according to this correction using the following formula: $\alpha = \frac{0.05}{r}$, where r is the number of executed comparisons. The significance level for each comparison according to the Bonferroni correction (in this example) is $\alpha = \frac{0.05}{4} = 0.0125.$

However, it is necessary to remember that if you reduce α for each comparison, the power of the test is increased.

14.2.3 Tests for one proportion

You should use tests for proportion if there are two possible results to obtain (one of them is an distinguished result with the size of m) and you know how often these results occur in the sample (we know a p proportion). Depending on a sample size n you can choose the Z **test for a one proportion** – for large samples and the exact **binominal test for a one proportion** – for small sample sizes . These tests are used to verify the hypothesis that the proportion in the population, from which the sample is taken, is a given value.

Basic assumptions:

– measurement on a nominal scale - any order is not taken into account.

The additional condition for the Z test for proportion

– large frequencies (according to Marascuilo and McSweeney interpretation (1977)[112] each of these values: $np > 5$ and $n(1-p) > 5$).

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & p = p_0, \\ \mathcal{H}_1: & p \neq p_0, \end{array}$$

where:
p – probability (distinguished proportion) in the population,
p_0 – expected probability (expected proportion).

The Z test for one proportion
The test statistic is defined by:

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where:

$p = \frac{m}{n}$ distinguished proportion for the sample taken from the population,
m – frequency of values distinguished in the sample,
n – sample size.

The test statistic with a continuity correction is defined by:

$$Z = \frac{|p - p_0| - \frac{1}{2n}}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

The Z statistic with and without a continuity correction asymptotically (for large sizes) has the normal distribution.

Binomial test for one proportion
The binomial test for one proportion uses directly the **binomial distribution** which is also called the Bernoulli distribution, which belongs to the group of discrete distributions (such distributions, where the analysed variable takes in the finite number of values). The analysed variable can take in $k = 2$ values. The first one is usually definited with the name of a success and the other one with the name of a failure. The probability of occurence of a success (distinguished probability) is p_0, and a failure $1 - p_0$.

The probability for the specific point in this distribution is calculated using the formula:

$$P(m) = \binom{n}{m} p_0^m (1 - p_0)^{n-m},$$

where:

$\binom{n}{m} = \frac{n!}{m!(n-m)!}$,
m – frequency of values distinguished in the sample,
n – sample size.

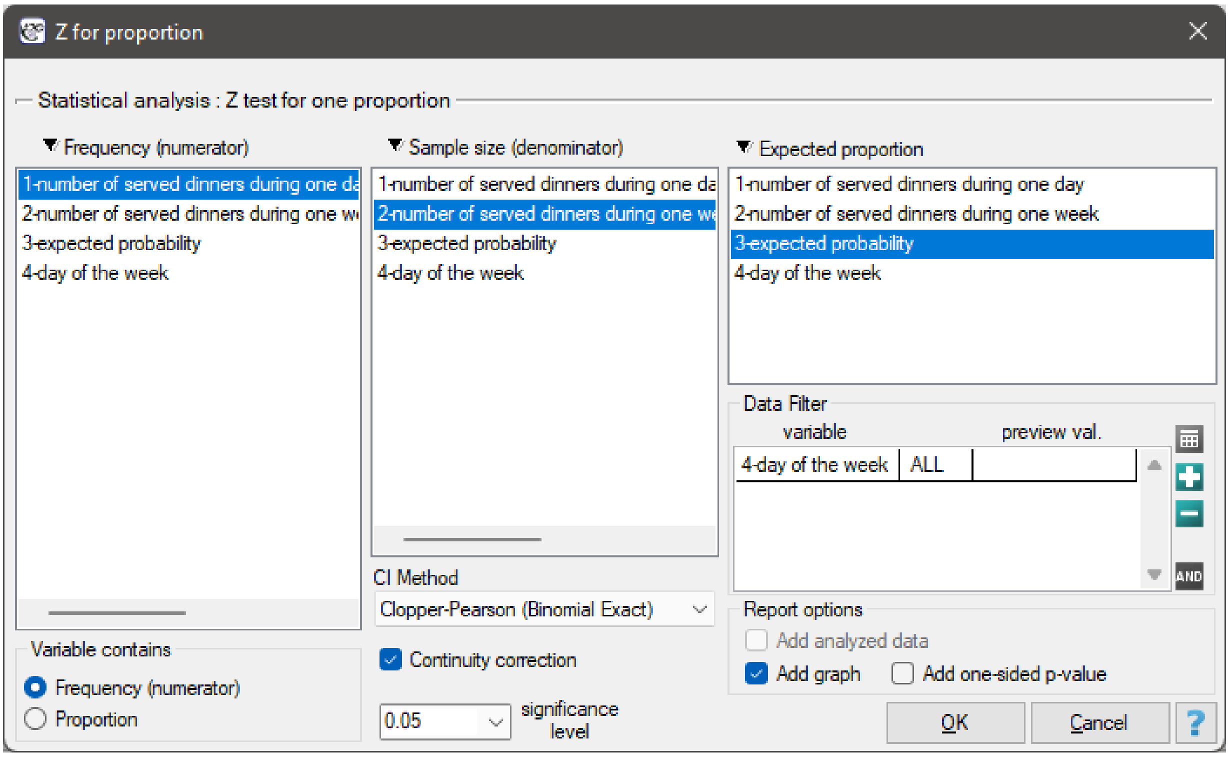
Based on the total of appropriate probabilities P a one-sided and a two-sided p value is calculated, and a two-sided p value is defined as a doubled value of the less of the one-sided probabilities.

The p value is compared with the significance level α:

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note
Note that, for the estimator from the sample, which in this case is the value of the p proportion, a confidence interval is calculated. The interval for a large sample size can be based on the normal distribution - so-called Wald intervals. The more universal are intervals proposed by Wilson (1927)[172] and by Agresti and Coull (1998)[5]. Clopper and Pearson (1934)[38] intervals are more adequate for small sample sizes.
Comparison of interval estimation methods of a binomial proportion was published by Brown L.D et al (2001)[29]

The settings window with the Z test for one proportion can be opened in Statistics menu→NonParametric tests (unordered categories)→Z for proportion.



Z for proportion

Statistical analysis : Z test for one proportion

Frequency (numerator): 1-number of served dinners during one da (selected), 2-number of served dinners during one w, 3-expected probability, 4-day of the week

Sample size (denominator): 1-number of served dinners during one da, 2-number of served dinners during one we (selected), 3-expected probability, 4-day of the week

Expected proportion: 1-number of served dinners during one day, 2-number of served dinners during one week, 3-expected probability (selected), 4-day of the week

Data Filter — variable: 4-day of the week; preview val.: ALL

CI Method: Clopper-Pearson (Binomial Exact)

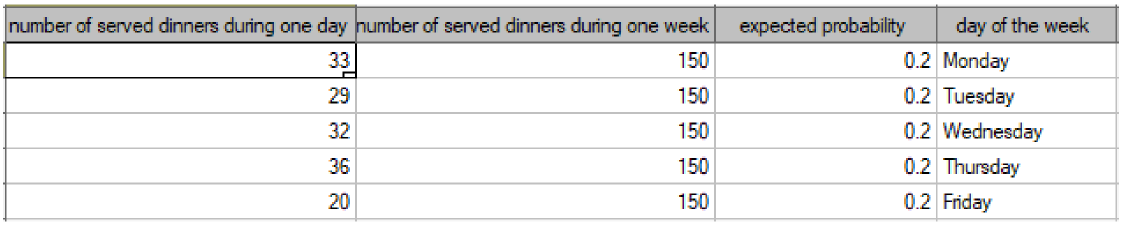
☑ Continuity correction

0.05 significance level

Variable contains: ● Frequency (numerator) ○ Proportion

Report options: □ Add analyzed data ☑ Add graph □ Add one-sided p-value

Example 14.3 cont. *(dinners.pqs file)*
Assume, that you would like to check if on Friday $\frac{1}{5}$ of all the dinners during the whole week are served. For the chosen sample $m = 20$, $n = 150$.



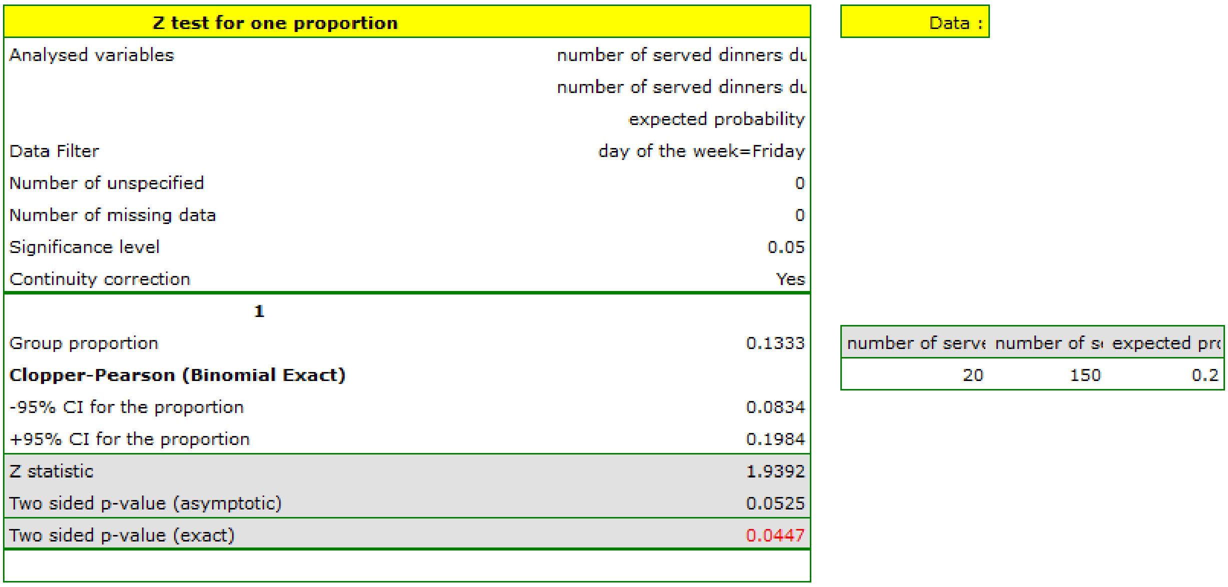
number of served dinners during one day	number of served dinners during one week	expected probability	day of the week
33	150	0.2	Monday
29	150	0.2	Tuesday
32	150	0.2	Wednesday
36	150	0.2	Thursday
20	150	0.2	Friday

Select the options of the analysis and activate a filter selecting the appropriate day of the week – Friday. If you do not activate the filter, no error will be generated, only statistics for given weekdays will be calculated.

Hypotheses:

$\mathcal{H}_0$: on Friday, in a school canteen there are served $\frac{1}{5}$ out of all dinners which are served within a week,

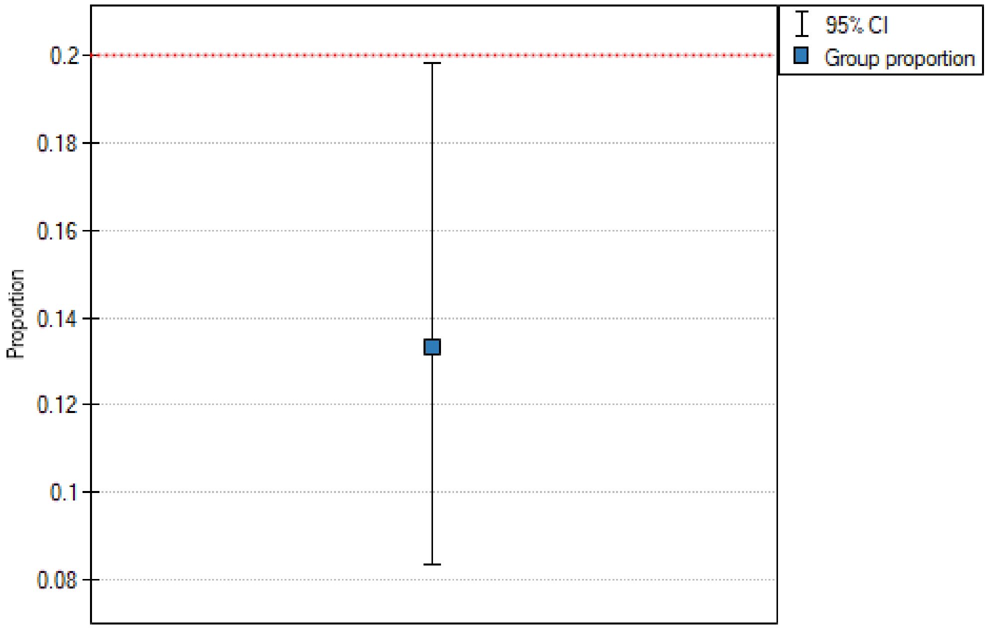
$\mathcal{H}_1$: on Friday, in a school canteen there are significantly more than $\frac{1}{5}$ or less than $\frac{1}{5}$ dinners out of all the dinners served within a week in this canteen.



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 :

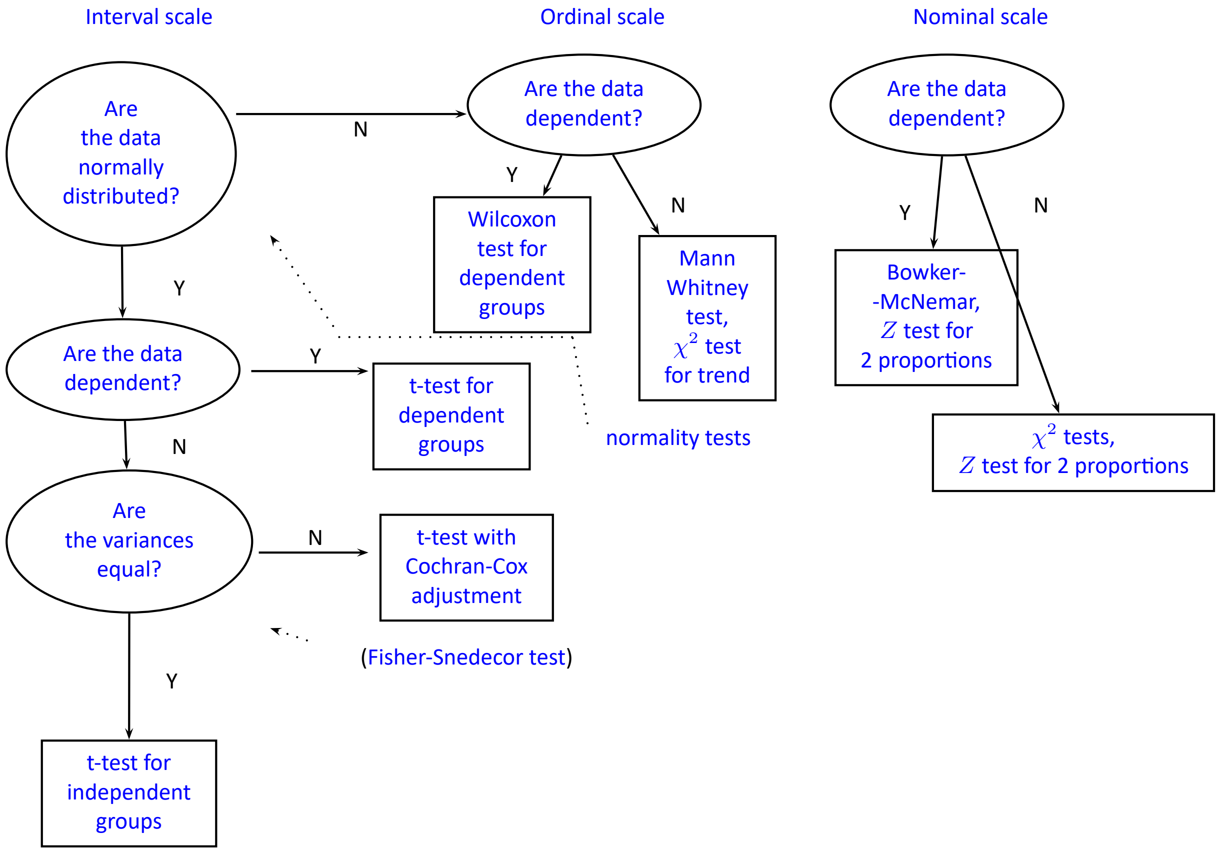
number of serv	number of s	expected pr
20	150	0.2



The proportion of the distinguished value in the sample is $p = \frac{m}{n} = 0.133$ and 95% Clopper-Pearson confidence interval for this fraction $(0.083, 0.198)$ does not include the hypothetical value of 0.2.

Based on the Z test without the continuity correction (p-value = 0.0412) and also on the basis of the exact value of the probability calculated from the binomial distribution (p-value = 0.0447) you can assume (on the significance level $\alpha = 0.05$), that on Friday there are statistically less than $\frac{1}{5}$ dinners served within a week. However, after using the continuity correction it is not possible to reject the null hypothesis p-value = 0.0525).

15 COMPARISON - TWO GROUPS


Interval scale
Ordinal scale
Nominal scale
Are the data normally distributed?
N
Y
Are the data dependent?
Y
N
t-test for dependent groups
Are the variances equal?
N
Y
t-test with Cochran-Cox adjustment
(Fisher-Snedecor test)
t-test for independent groups
Are the data dependent?
Y
N
Wilcoxon test for dependent groups
Mann Whitney test, χ^2 test for trend
normality tests
Are the data dependent?
Y
N
Bowker- -McNemar, Z test for 2 proportions
χ^2 tests, Z test for 2 proportions

15.1 PARAMETRIC TESTS

15.1.1 The Fisher-Snedecor test

The F-Snedecor test is based on a variable F which was formulated by Fisher (1924), and its distribution was described by Snedecor. This test is used to verify the hypothesis about equality of variances of an analysed variable for 2 populations.

Basic assumptions:

- measurement on an interval scale,
- normality of distribution of an analysed feature in both populations,
- an independent model.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & \sigma_1^2=\sigma_2^2,\\ \mathcal{H}_1: & \sigma_1^2\neq\sigma_2^2,\end{array}$$

where:
σ_1^2, σ_2^2 – variances of an analysed variable of the 1st and the 2nd population.

The test statistic is defined by:

$$F=\frac{sd_1^2}{sd_2^2},$$

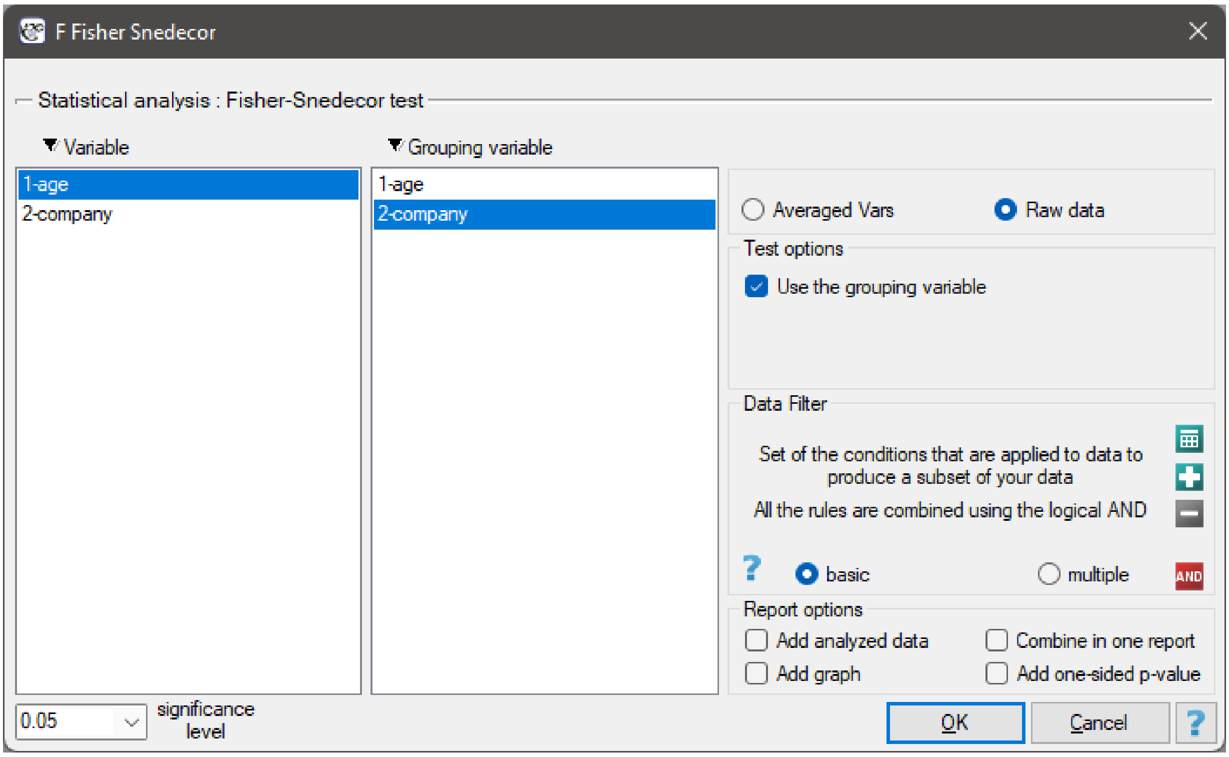
where:
sd_1^2, sd_2^2 – variances of an analysed variable of the samples chosen randomly from the 1st and the 2nd population.

The test statistic has the F Snedecor distribution with n_1-1 and n_2-1 degrees of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl}\text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

The settings window with the Fisher-Snedecor test can be opened in Statistics menu→Parametric tests→F Fisher Snedecor.



F Fisher Snedecor

Statistical analysis : Fisher-Snedecor test

Variable: 1-age (selected), 2-company

Grouping variable: 1-age, 2-company (selected)

○ Averaged Vars ● Raw data

Test options

☑ Use the grouping variable

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

□ Add graph □ Add one-sided p-value

0.05 significance level

Note!
Calculations can be based on raw data or data that are averaged like: arithmetic means, standard deviations and sample sizes.

15.1.2 The t-test for independent groups

The t-test for independent groups is used to verify the hypothesis about the equality of means of an analysed variable in 2 populations.

Basic assumptions:

- measurement on an interval scale,
- normality of distribution of an analysed feature in both populations,
- an independent model,
- equality of variances of an analysed variable in 2 populations.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \mu_1=\mu_2, \\ \mathcal{H}_1: & \mu_1\neq\mu_2. \end{array}$$

where:
μ_1, μ_2 – means of an analysed variable of the 1st and the 2nd population.

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 an analysed variable 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 an analysed variable 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.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note:

- **pooled standard deviation** is defined by:

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

- **standard error of difference of means** is defined by:

$$SE_{\overline{x}_1-\overline{x}_2} = \sqrt{\frac{(n_1-1)sd_1^2 + (n_2-1)sd_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

Standardized effect size
The **Cohen's d** determines how much of the variation occurring is the difference between the averages.

$$d = \left|\frac{\overline{x}_1-\overline{x}_2}{SD_p}\right|$$

When interpreting an effect, researchers often use general guidelines proposed by Cohen [45] defining small (0.2), medium (0.5) and large (0.8) effect sizes.

15.1.3 The t-test with the Cochran-Cox adjustment

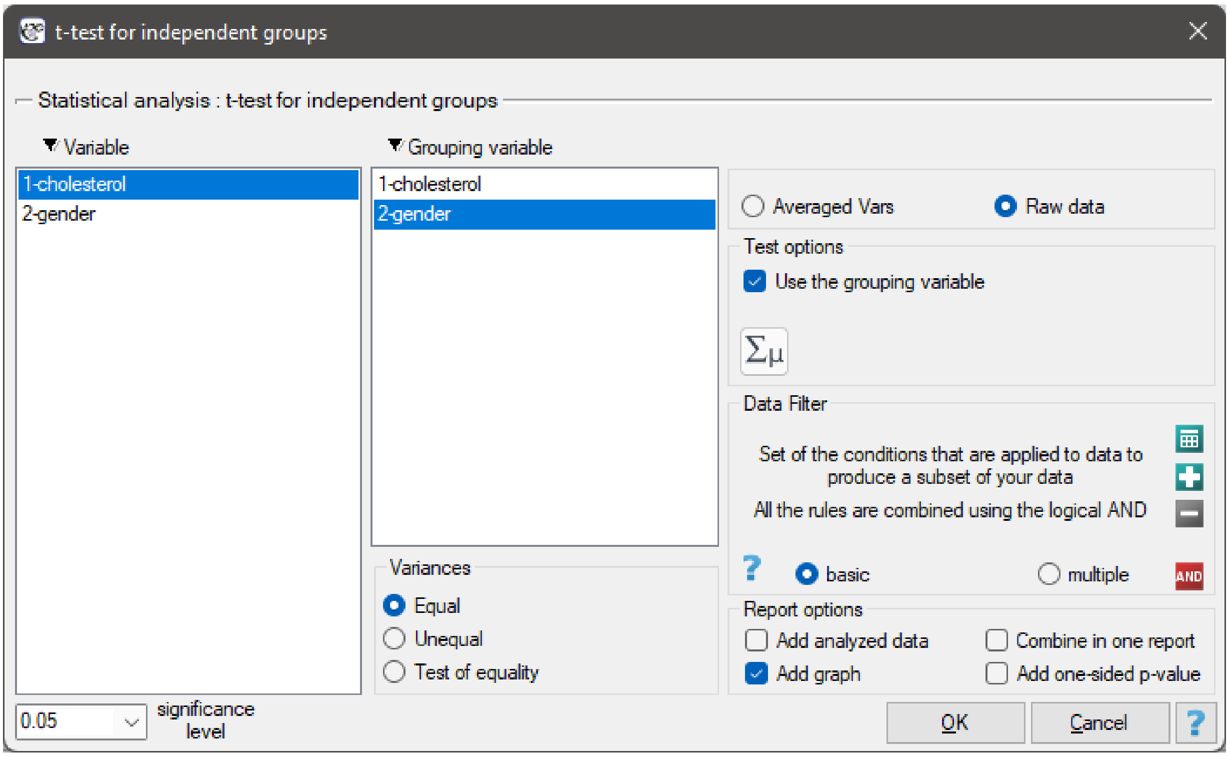
The Cochran-Cox adjustment relates to the t-test for independent groups (1957)[42] and is calculated when variances of analysed variables in both populations are different.
The test statistic is defined by:

$$t = \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{sd_1^2}{n_1}+\frac{sd_2^2}{n_2}}}.$$

The test statistic has the t-Student distribution with degrees of freedom proposed by Satterthwaite (1946)[143] and calculated using the formula:

$$df = \frac{\left(\frac{sd_1^2}{n_1}+\frac{sd_2^2}{n_2}\right)^2}{\left(\frac{sd_1^2}{n_1}\right)^2\cdot\frac{1}{(n_1-1)}+\left(\frac{sd_2^2}{n_2}\right)^2\cdot\frac{1}{(n_2-1)}}.$$

The settings window with the t- test for independent groups can be opened in Statistics menu→Parametric tests→t-test for independent groups or in Wizard.



t-test for independent groups

Statistical analysis : t-test for independent groups

Variable: 1-cholesterol (selected), 2-gender

Grouping variable: 1-cholesterol, 2-gender (selected)

○ Averaged Vars ● Raw data

Test options

☑ Use the grouping variable

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Variances

● Equal

○ Unequal

○ Test of equality

Report options

☐ Add analyzed data ☐ Combine in one report

☑ Add graph ☐ Add one-sided p-value

0.05 significance level

If, in the window which contains the options related to the variances, you have choosen:

- equal, the t-test for independent groups will be calculated ,
- different, the t-test with the Cochran-Cox adjustment will be calculated,
- check equality, to calculate the Fisher-Snedecor test, basing on its result and set the level of significance, the t-test for independent groups with or without the Cochran-Cox adjustment will be calculated.

Note

Calculations can be based on raw data or data that are averaged like: arithmetic means, standard deviations and sample sizes.

Example 15.1. (cholesterol.pqs file)

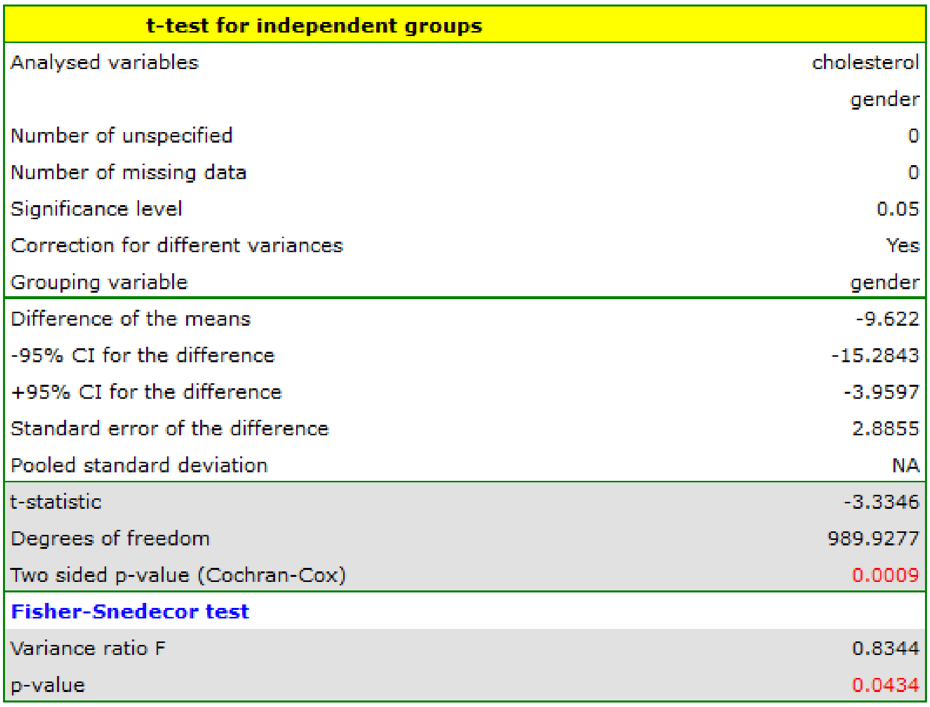
Five hundred subjects each were drawn from a population of women and a population of men over 40 years of age. The study concerned the assessment of cardiovascular disease risk. Among the parameters studied is the value of total cholesterol. The purpose of this study will be to compare men and women as to this value. We want to show that these populations differ on the level of total cholesterol and not only on the level of cholesterol broken down into its fractions.

The distribution of age in both groups is a normal distribution (this was checked with the Lilliefors test). The mean cholesterol value 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 analysis will use the Student's t-test with Cochran-Cox correction

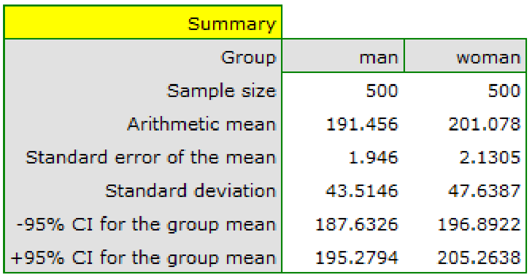
Hypotheses:

$\mathcal{H}_0$: The average total cholesterol of the female population is different from the average total cholesterol of the male population,

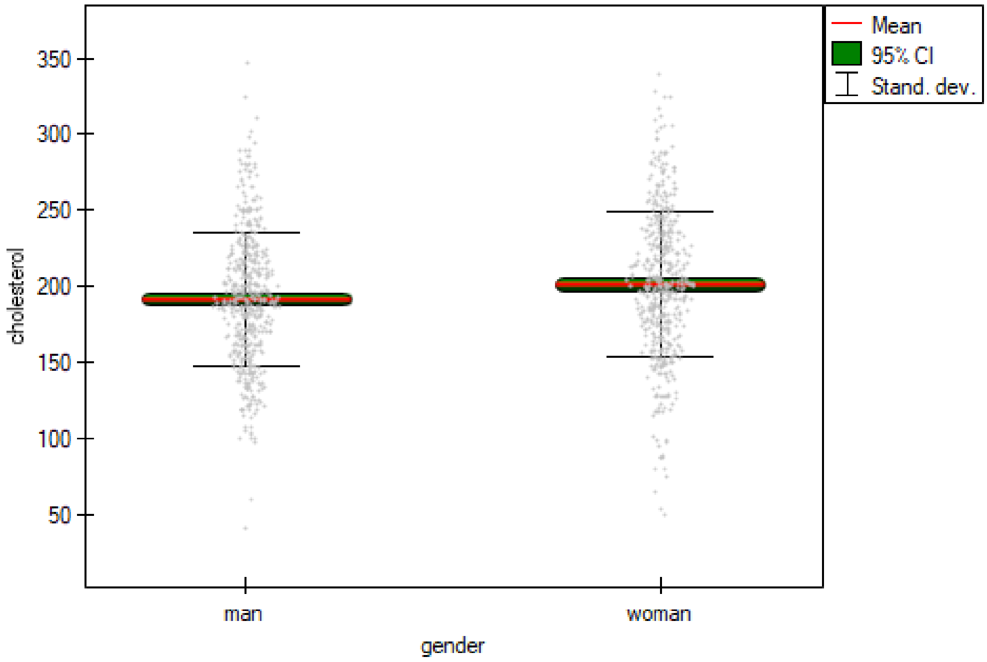
$\mathcal{H}_1$: The average total cholesterol of the female population equals the average total cholesterol of the male population.



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



Comparing $p = 0.0009$ with a significance level $\alpha = 0.05$ we find that women and men in Poland have statistically significant differences in total cholesterol values. The average Polish man over the age of 40 has higher total cholesterol than the average Polish woman by almost 10 units.

15.1.4 The t-test for dependent groups

The t-test for dependent groups is used when the measurement of an analysed variable you do twice, each time in different conditions (but you should assume, that variances of the variable in both measurements are pretty close to each other). We want to check how big is the difference between the pairs of measurements ($d_i = x_{1i} - x_{2i}$). This difference is used to verify the hypothesis informing us that the mean of the difference in the analysed population is 0.

Basic assumptions:

- measurement on an interval scale,
- normality of distribution of measurements d_i (or the normal distribution for an analysed variable in each measurement),
- a dependent model.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & \mu_0 = 0,\\ \mathcal{H}_1: & \mu_0 \neq 0,\end{array}$$

where:
μ_0, – mean of the differences d_i in a population.

The test statistic is defined by:

$$t = \frac{\overline{d}}{sd_d}\sqrt{n},$$

where:
$\overline{d}$ – mean of differences d_i in a sample,
sd_d – standard deviation of differences d_i in a sample,
n – number of differences d_i in a sample.

Test statistic has the t-Student distribution with $n-1$ degrees of freedom.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

Note

- **standard deviation of the difference** is defined by:

$$sd_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}},$$

- **standard error of the mean of differences** is defined by:

$$SEM_d = \frac{SD_d}{\sqrt{n}}.$$

Standardized effect size

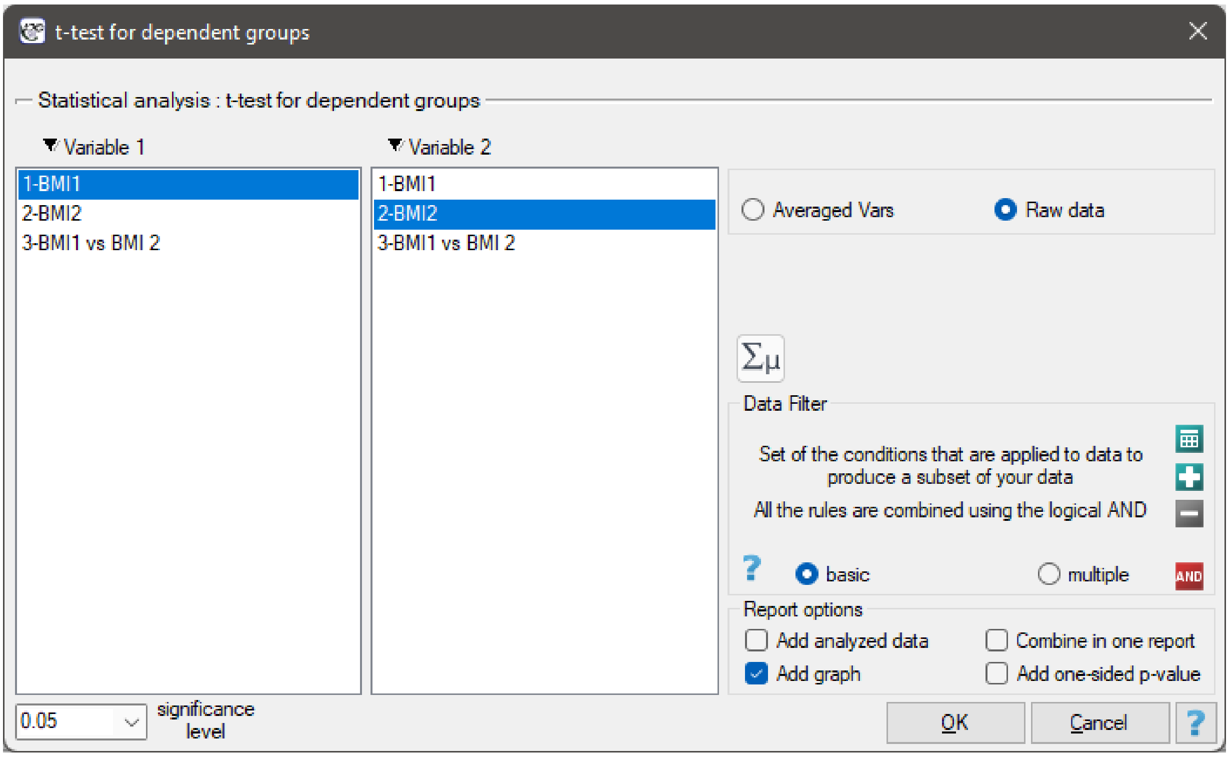
The **Cohen's d** determines how much of the variation occurring is the difference between the averages, while taking into account the correlation of the variables.

$$d = \frac{dz}{\sqrt{1-r_p}},$$

where:
$dz = \left|\frac{\overline{d}}{sd_d}\right|$,
r_p - the Pearson product-moment correlation coefficient.

When interpreting an effect, researchers often use general guidelines proposed by Cohen [45] defining small (0.2), medium (0.5) and large (0.8) effect sizes.

The settings window with the t-test for dependent groups can be opened in Statistics menu→Parametric tests→t-test for dependent groups or in Wizard.



t-test for dependent groups

Statistical analysis : t-test for dependent groups

Variable 1
- 1-BMI1
- 2-BMI2
- 3-BMI1 vs BMI 2

Variable 2
- 1-BMI1
- 2-BMI2
- 3-BMI1 vs BMI 2

○ Averaged Vars ● Raw data

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

☑ Add graph □ Add one-sided p-value

0.05 significance level

Note

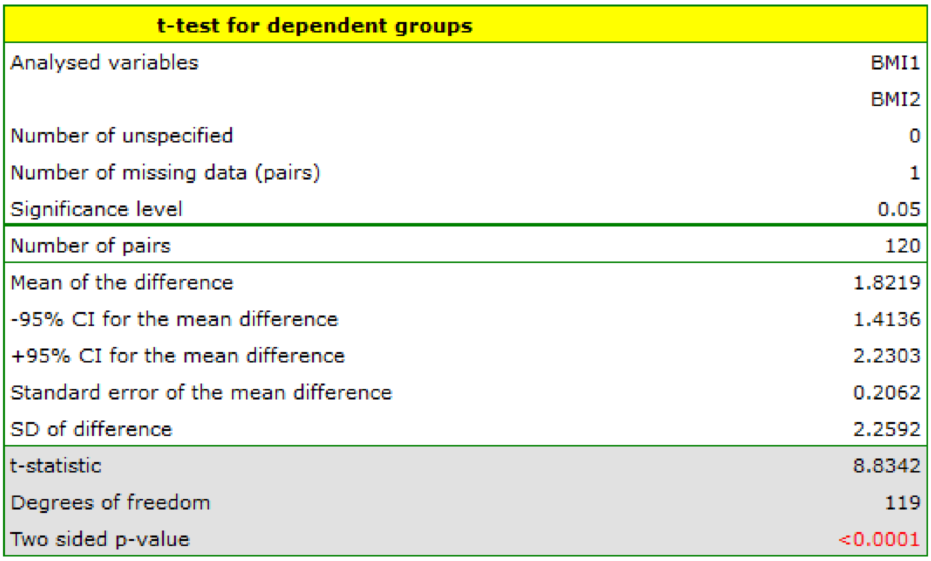
Calculations can be based on raw data or data that are averaged like: arithmetic mean of difference, standard deviation of difference and sample size.

Example 15.2. (BMI.pqs file)

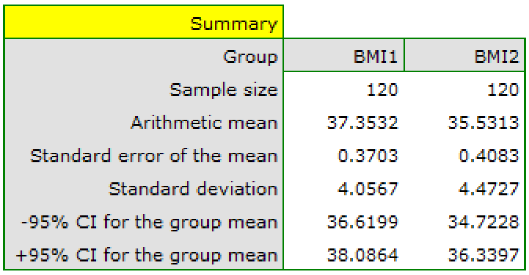
A clinic treating eating disorders studied the effect of a recommended ”diet A” on weight change. A sample of 120 obese patients were put on the diet. Their BMI levels were measured twice: before the diet and after 180 days of the diet. To test the effectiveness of the diet, the obtained BMI measurements were compared.

Hypotheses:

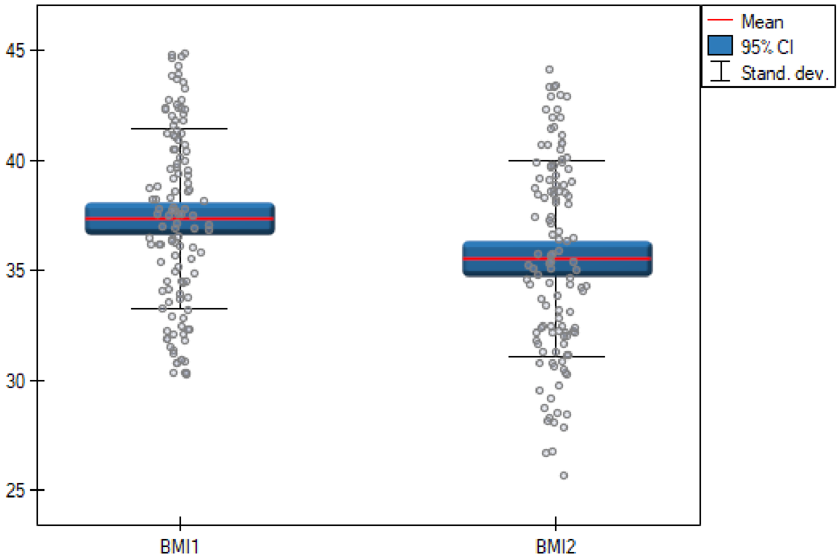
$$\begin{array}{ll} \mathcal{H}_0: & \text{Mean BMI values do not change with diet,} \\ \mathcal{H}_1: & \text{Mean BMI values change as a result of diet.} \end{array}$$



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



Comparing $p < 0.0001$ with a significance level $\alpha = 0.05$ we find that the mean BMI level changed significantly. Before the diet, it was higher by less than 2 units on average.

The study was able to use the Student's t-test for dependent groups because the distribution of the difference between pairs of measurements was a normal distribution (Lilliefors test, $p = 0.0837$).

15.2 NON-PARAMETRIC TESTS

15.2.1 The Mann-Whitney U test

The Mann-Whitney U test is also called as the Wilcoxon Mann-Whitney test (Mann and Whitney (1947)[108] and Wilcoxon (1949)[171]). This test is used to verify the hypothesis that there is no shift in the compared distributions, i.e., most often the insignificance of differences between medians of an analysed variable in 2 populations (but you should assume that the distributions of a variable are pretty similar to each other - comparison of rank variances can be performed with the Conover rank test).

Basic assumptions:

- measurement on an ordinal scaleor on an interval scale,
- an independent model.

Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & \phi_1=\phi_2,\\ \mathcal{H}_1: & \phi_1\neq\phi_2,\end{array}$$

where:
ϕ_1, ϕ_2 distributions of an analysed variable of the 1st and the 2nd population.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

Note
Depending on a sample size, the test statistic is calculated using by different formulas:

- For a small sample size:

$$U=n_1n_2+\frac{n_1(n_1+1)}{2}-R_1,$$

or

$$U'=n_1n_2+\frac{n_2(n_2+1)}{2}-R_2,$$

where n_1, n_2 are sample sizes, R_1, R_2 are rank sums for the samples.

This statistic has the Mann-Whitney distribution and it 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:

$$Z=\frac{U-\frac{n_1n_2}{2}}{\sqrt{\frac{n_1n_2(n1+n_2+1)}{12}-\frac{n_1n_2\sum(t^3-t)}{12(n_1+n_2)(n_1+n_2-1)}}},$$

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. This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $\frac{n_1n_2\sum(t^3-t)}{12(n_1+n_2)(n_1+n_2-1)}=0$)

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

The Mann-Whitney test with the continuity correction (Marascuilo and McSweeney (1977)[112])

The continuity correction should be used to guarantee the possibility of taking in all the values of real numbers by the test statistic, according to the assumption of the normal distribution. The formula for the test statistic with the continuity correction is defined as:

$$Z = \frac{\left|U - \frac{n_1 n_2}{2}\right| - 0.5}{\sqrt{\frac{n_1 n_2 (n1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum(t^3 - t)}{12(n_1+n_2)(n_1+n_2-1)}}}.$$

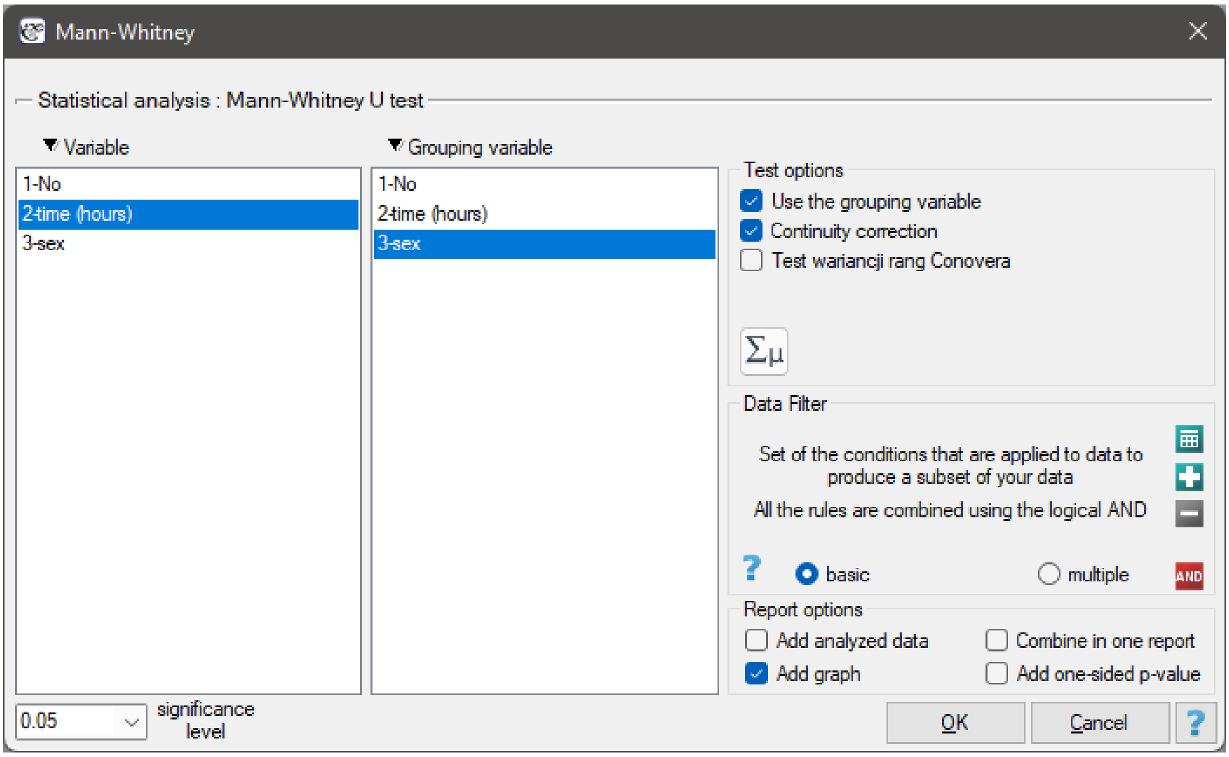
Standardized effect size

The distribution of the Mann-Whitney test statistic is approximated by the normal distribution, which can be converted to an effect size $r = |Z/(n_1 + n_2)|$ [?] to then obtain the **Cohen's d** value according to the standard conversion used for meta-analyses:

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

When interpreting an effect, researchers often use general guidelines proposed by Cohen [45] defining small (0.2), medium (0.5) and large (0.8) effect sizes.

The settings window with the Mann-Whitney U test can be opened in Statistics menu → NonParametric tests (ordered categories) → Mann-Whitney or in Wizard.



Example 15.3. (computer.pqs file)

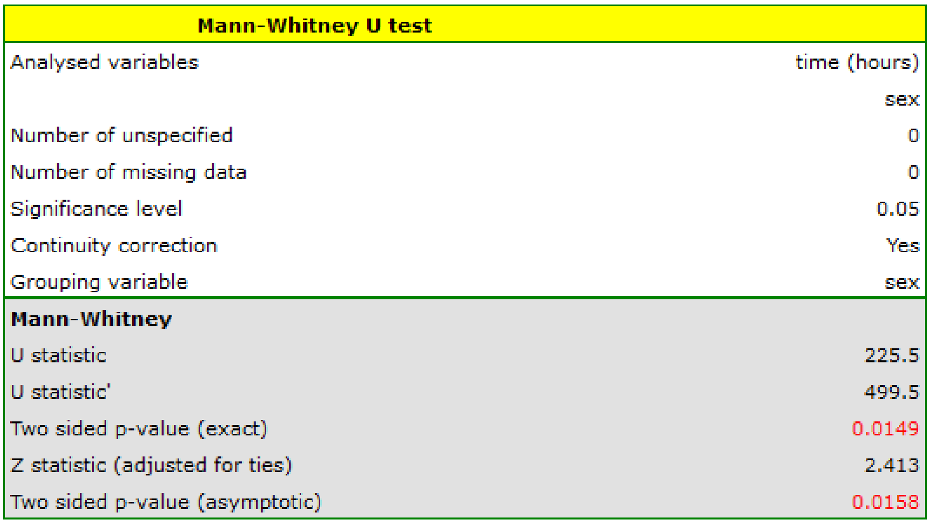
There was made a hypothesis that at some university male math students spend statistically more time in front of a computer screen than the female math students. To verify the hypothesis from the population of people who study math at this university, there was drawn a sample consisting of 54 people (25 women and 29 men). These persons were asked how many hours they spend in front of the computer screens daily. There were obtained the following results:

(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, 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) (6, m) (7, k) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (8, k) (8, m) (8, m).

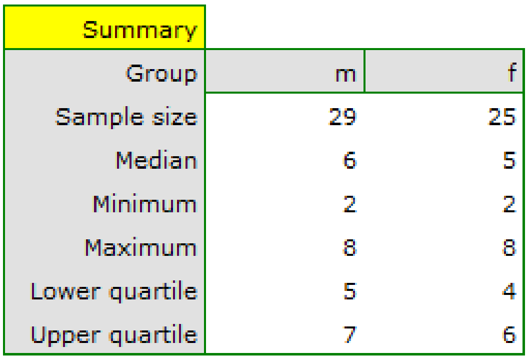
Hypotheses:

$\mathcal{H}_0$: the median of the time spent in front of a computer screen is exactly the same both in the male and the female population of students, at the analysed university,

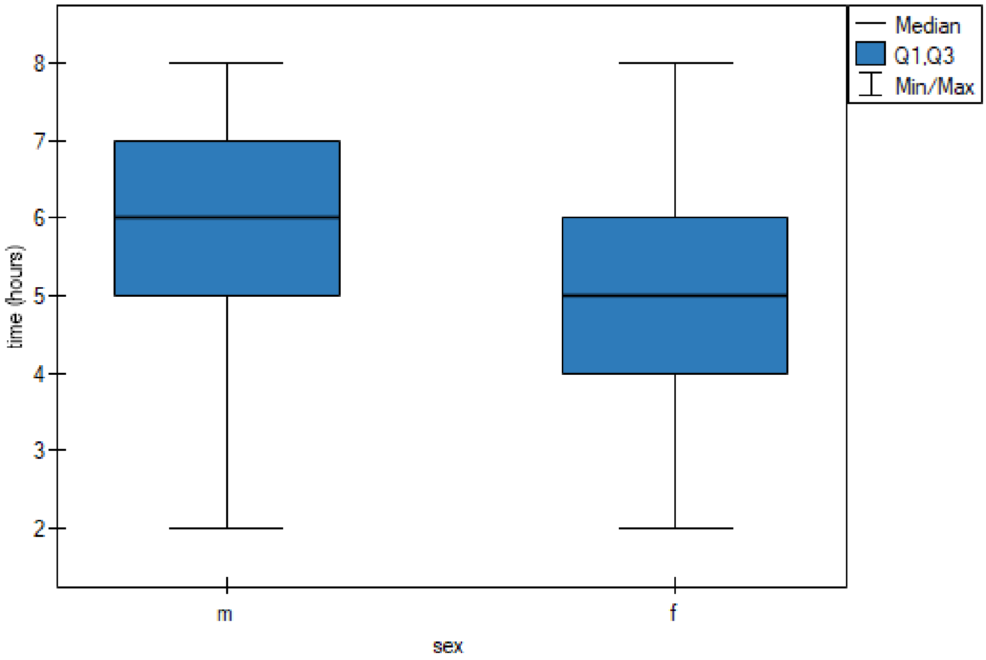
$\mathcal{H}_1$: the median of the time spent in front of a computer screen is different among the male population and the female population of students, at the analysed university.



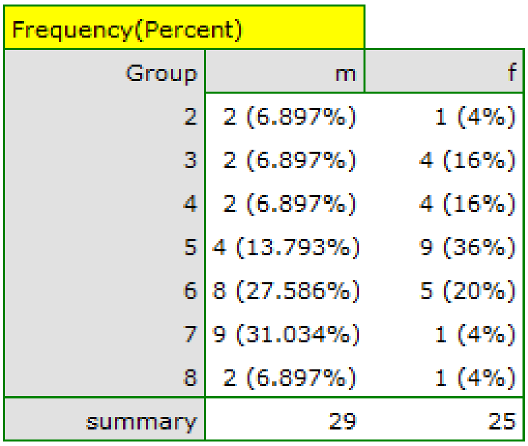
Mann-Whitney U test	
Analysed variables	time (hours) sex
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Grouping variable	sex
Mann-Whitney	
U statistic	225.5
U statistic'	499.5
Two sided p-value (exact)	0.0149
Z statistic (adjusted for ties)	2.413
Two 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

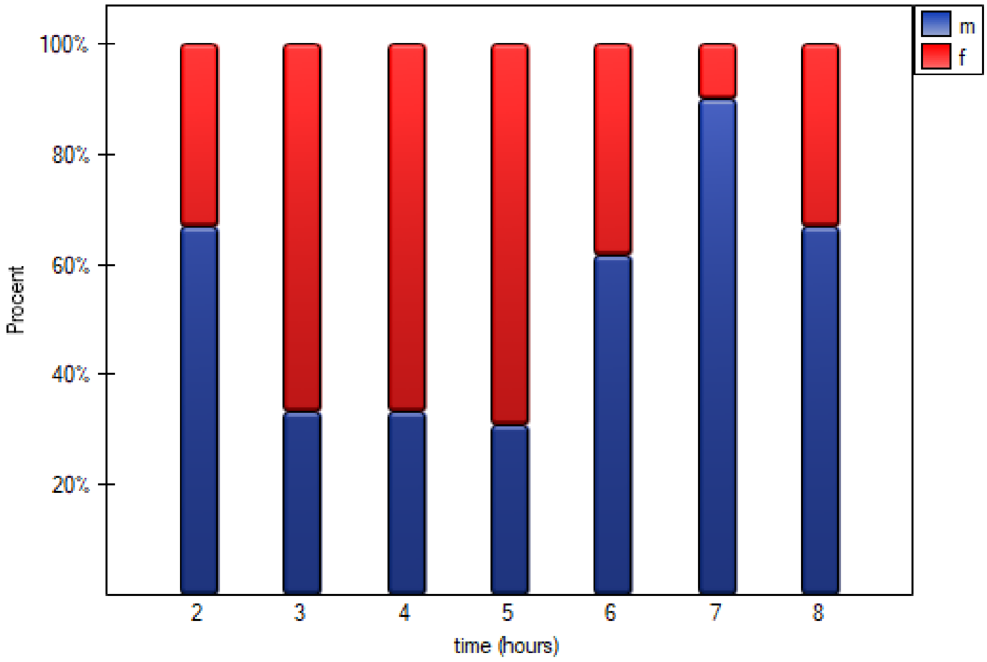


Based on the assumed $\alpha = 0.05$ and the Z statistic of the Mann-Whitney test without correction for continuity (p=0.0154) as well as with this correction $p = 0.0158$, as well as on the exact U statistic (p=0.0149) we can assume that there are statistically significant differences between female and male math students in the amount of time spent in front of the computer. These differences are that female students spend less time in front of the computer than male students. They can be described by the median, quartiles, and the largest and smallest value, which we also see in a box-and-whisker plot. Another way to describe the differences is to represent the time spent in front of the computer based on a table of counts and percentages (which we run in the analysis window by setting descriptive statistics includegraphics Σμ) or based on a column plot.



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



15.2.2 The Wilcoxon test (matched-pairs)

The Wilcoxon matched-pairs test, is also called as the Wilcoxon test for dependent groups (Wilcoxon 1945[**?**],1949[**?**]). It is used if the measurement of an analysed variable you do twice, each time in different conditions. It is the extension for the two dependent samples of the Wilcoxon test (signed-ranks) – designed for a one sample. We want to check how big is the difference between the pairs of measurements ($d_i = x_{1i} - x_{2i}$) for each of i analysed objects. This difference is used to verify the hypothesis determining that the median of the difference in the analysed population counts to 0.

Basic assumptions:

- measurement on an ordinal scale or on an interval scale,
- a dependent model.

Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & \theta_0 = 0,\\ \mathcal{H}_1: & \theta_0 \neq 0,\end{array}$$

where:
θ_0 – median of the differences d_i in a population.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p \leq \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

Note
Depending on the sample size, the test statistic is calculated by using different formulas:

- For small a sample size:

$$T = \min\left(\sum R_-, \sum R_+\right),$$

where:
$\sum R_+$ – sums of positive ranks,
$\sum R_-$ – sums of negative ranks.

This statistic has the Wilcoxon distribution and does not contain any correction for ties.

- For a large sample size

$$Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}},$$

where:
n – number of ranked signs (number of the ranks),
t – number of the cases included in a tie.

The formula for the Z statistic includes the correction for ties. This correction is used, when the ties occur (if there are no ties, the correction is not calculated, because of $\frac{\sum t^3 - \sum t}{48} = 0$).

The Z statistic (for large sample sizes) asymptotically has the normal distribution.

The Wilcoxon test with the continuity correction (Marascuilo and McSweeney (1977)[112])

The continuity correction is used to guarantee the possibility of taking in all the values of the real numbers by the test statistic, according to the assumption of the normal distribution. The test statistic with the continuity correction is defined by:

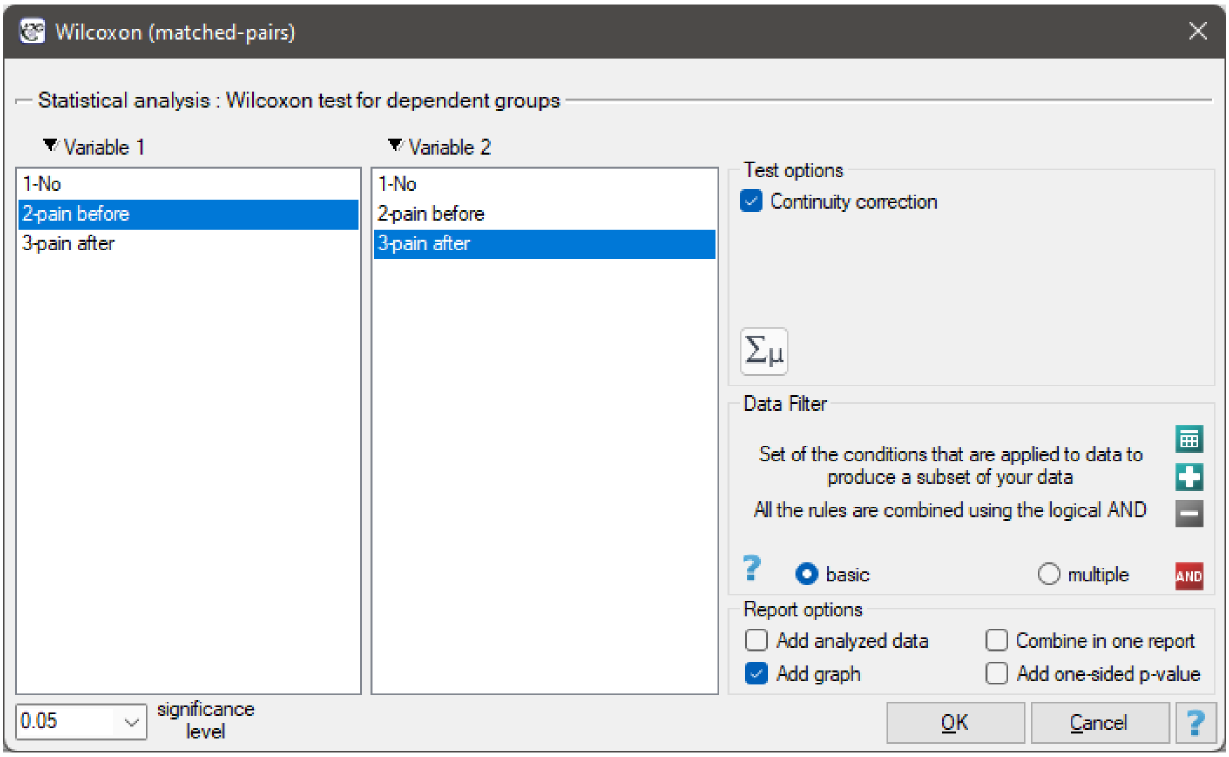
$$Z = \frac{\left|T - \frac{n(n+1)}{4}\right| - 0.5}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}}.$$

Note
The median calculated for the difference column includes all pairs of results except those with a difference of 0.
Standardized effect size
The distribution of the Wilcoxon test statistic is approximated by the normal distribution, which can be converted to an effect size $r = |Z/n|$ [**?**] to then obtain the **Cohen's d** value according to the standard conversion used for meta-analyses:

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

When interpreting an effect, researchers often use general guidelines proposed by Cohen [45] defining small (0.2), medium (0.5) and large (0.8) effect sizes.

The settings window with the Wilcoxon test for dependent groups can be opened in Statistics menu $\rightarrow$ NonParametric tests$\rightarrow$Wilcoxon (matched-pairs) or in Wizard.



Wilcoxon (matched-pairs)

Statistical analysis : Wilcoxon test for dependent groups

Variable 1: 1-No, **2-pain before**, 3-pain after

Variable 2: 1-No, 2-pain before, **3-pain after**

Test options

☑ Continuity correction

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

☐ Add analyzed data ☐ Combine in one report

☑ Add graph ☐ Add one-sided p-value

0.05 significance level

EXAMPLE 15.4. (pain.pqs file)

There was chosen a sample consisting of 22 patients suffering from a cancer. They were examined to check the level of felt pain (1 – 10 scale, where 1 means the lack of pain and 10 means unbearable pain). This examination was repeated after a month of the treatment with a new medicine which was supposed to lower the level of felt pain. There were obtained the following results:

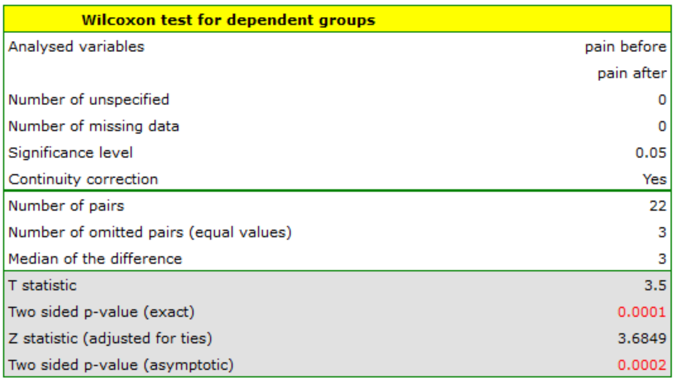
(pain before, pain after): (2, 2) (2, 3) (3, 1) (3,1) (3, 2) (3, 2) (3, 3) (4, 1) (4, 3) (4, 4) (5, 1) (5, 1) (5, 2) (5, 4) (5, 4) (6, 1) (6, 3) (7, 2) (7, 4) (7, 4) (8, 1) (8, 3).

Now, you want to check if this treatment has any influence on the level of felt pain in the population, from which the sample was chosen.

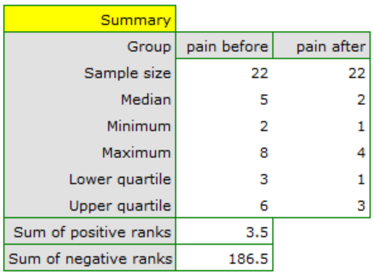
Hypotheses:

$\mathcal{H}_0$: the median of the differences between the level of pain before and after a month of treatment in the analysed population comes to 0,

$\mathcal{H}_1$: the median of the differences between the level of pain before and after a month of treatment in the analysed population is different from 0.

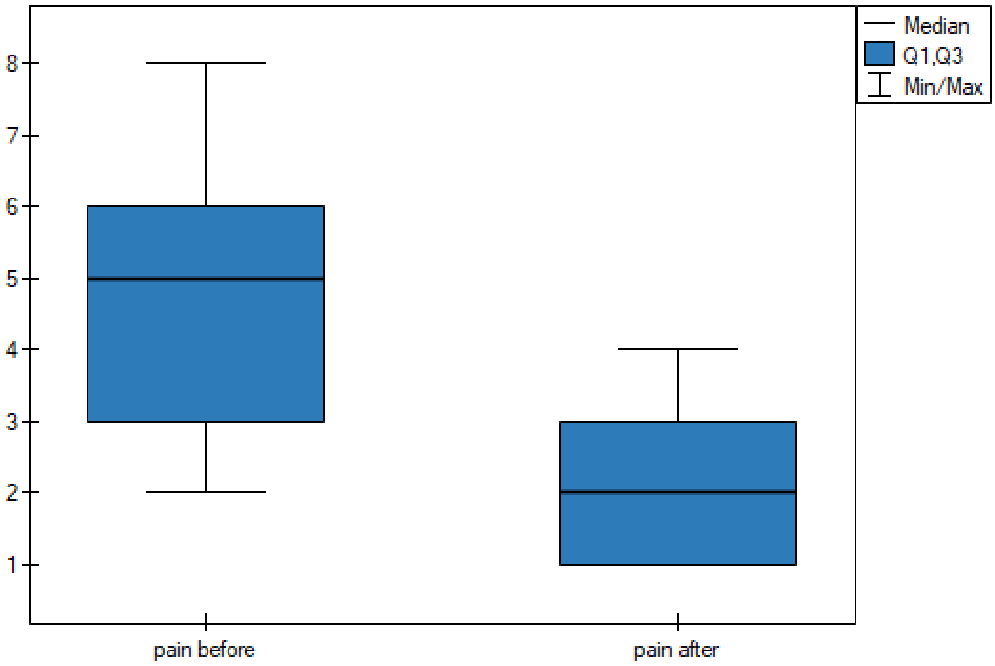


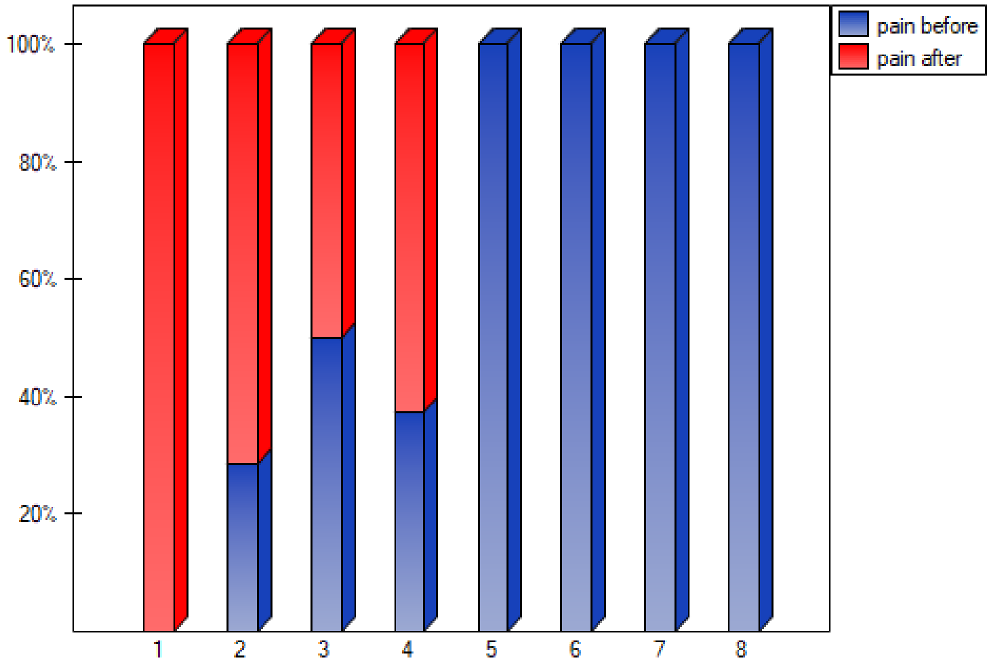
Wilcoxon test for dependent groups	
Analysed variables	pain before
	pain after
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Number of pairs	22
Number of omitted pairs (equal values)	3
Median of the difference	3
T statistic	3.5
Two sided p-value (exact)	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	

Comparing the p value = 0.0001 of the Wilcoxon test, based on the T statistic, with the significance level $\alpha = 0.05$ you assume, that there is a statistically significant difference if concerning the level of felt pain between these 2 examinations. The difference is, that the level of pain decreased (the sum of the negative ranks is significantly greater than the sum of the positive ranks). Exactly the same decision you would make on the basis of p value = 0.00021 or p value = 0.00023 of the Wilcoxon test which is based on the Z statistic or the Z statistic with the continuity correction. We can see the differences in a box-and-whisker plot or a column plot.





15.2.3 The Chi-square tests

These tests are based on data collected in the form of a contingency table of 2 traits, trait X and trait Y, the former having r and the latter c categories, so the resulting table has r rows and c columns. Therefore, we can speak of the 2x2 chi-square test (for tables with two rows and two columns) or the

RxC chi-square test (with multiple rows and columns). (See table (10.1)).

We can read the details of the chi-square test of the two features here:
chi-square test 2x2
chi-square test RxC.

Basic assumptions:

- measurement on a nominal scale - any order is not taken into account,
- an independent model.

The additional assumption for the χ^2 :

- large expected frequencies (according to Cochran interpretation (1952)[40].

- **General hypotheses:**

$$\begin{array}{cl} \mathcal{H}_0: & O_{ij}=E_{ij} \text{ for all categories,} \\ \mathcal{H}_1: & O_{ij}\neq E_{ij} \text{ for at least one category,} \end{array}$$

where:
O_{ij} – observed frequencies in a contingency table,
E_{ij} – expected frequencies in a contingency table.

- Hypotheses in the meaning of independence:

$\mathcal{H}_0$: there is no dependence between the analysed features of the population (both classifications are statistically independent according to X and Y feature),
$\mathcal{H}_1$: there is a dependence between the analysed features of the population.

Compare the p value, calculated on the basis of the test statistic, with the significance level α:

$$\begin{array}{ccl} \text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Additionally

- In addition to the chi-square test, another related test may need to be determined. In the event that Cochran's condition is not satisfied, one can determine:
Fisher's exact test for RxC tables
Fisher's exact test for 2x2 tables
chi-square test with Yates correction
mid-p test for 2x2 tables.
- If we obtain a table of Rx2, and the R categories can be ordered, it is possible to determine the trend:
chi-square test for trend for Rx2 tables
- When significant relationships or differences are found based on a test performed on a table larger than 2x2, then multiple comparisons can be performed with appropriate correction of the multiple comparisons to locate the location of these relationships/differences. This correction can be done automatically when the table has many columns. In such case, in test option window you should select Multiple column comparisons (RxC).

- In the case where we want to describe the strength of the relationship between feature X and feature Y, we can determine:
  measures of dependence
- In the case when we want to describe for 2x2 tables the effect size showing the impact of a risk factor, we can determine:
  Odds Ratio (OR) and Relative Risk (RR).

15.2.4 The Chi-square test for large tables

These tests are based on the data gathered in the form of a contingency table of 2 features (X, Y). One of them has possible r categories $X_1, X_2, ..., X_r$ and the other one c categories $Y_1, Y_2, ..., Y_c$ (look at the table (10.1)).

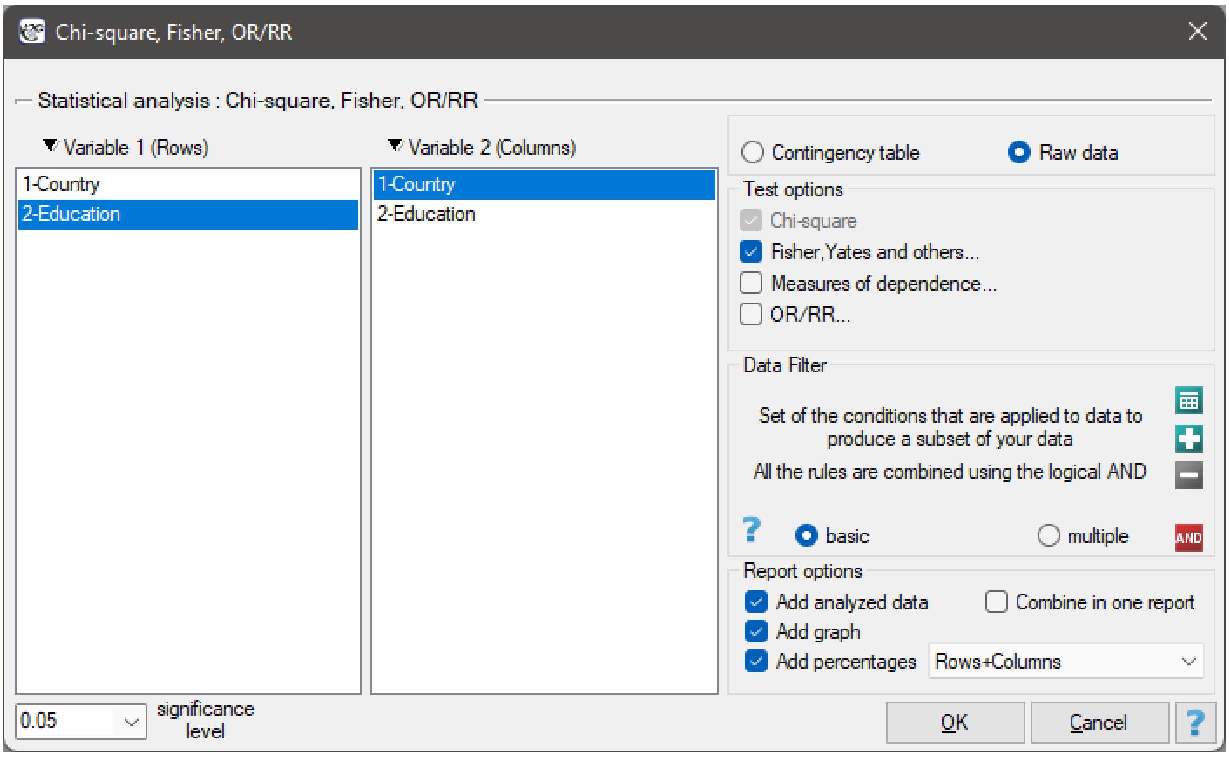
The χ^2 test for $r \times c$ tables is also known as the Pearson's Chi-square test (Karl Pearson 1900). This test is an extension on 2 features of the χ^2 test (goodness-of-fit).
The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}.$$

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with a number of degrees of freedom calculated using the formula: $df = (r-1)(c-1)$.

Compare the p value, calculateld on the basis of the test statistic, with the significance level α.

The settings window with the Chi-square test (RxC) can be opened in Statistics menu → NonParametric tests → Chi-square, Fisher, OR/RR or in Wizard



Chi-square, Fisher, OR/RR

Statistical analysis : Chi-square, Fisher, OR/RR

Variable 1 (Rows): 1-Country, 2-Education (selected: 2-Education)

Variable 2 (Columns): 1-Country (selected), 2-Education

○ Contingency table ● Raw data

Test options
☑ Chi-square
☑ Fisher,Yates and others...
☐ Measures of dependence...
☐ OR/RR...

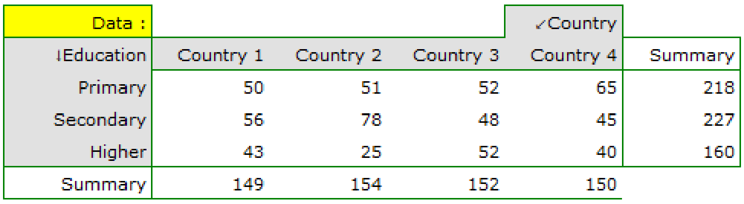
Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
☑ Add analyzed data ☐ Combine in one report
☑ Add graph
☑ Add percentages Rows+Columns

0.05 significance level

Example 15.5. (country-education.pqs file)

There is a sample of 605 persons ($n = 605$), who had 2 features analysed for (X=country of residence, Y=education). The first feature occurrs in 4 categories, and the second one 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 below, in the contingency table:



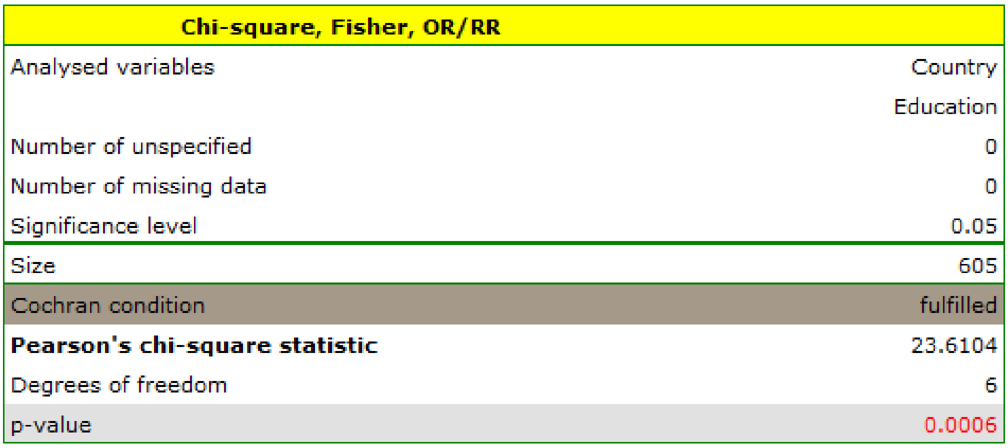
Data :				↙Country	
↓Education	Country 1	Country 2	Country 3	Country 4	Summary
Primary	50	51	52	65	218
Secondary	56	78	48	45	227
Higher	43	25	52	40	160
Summary	149	154	152	150	

Based on this sample, you would like to find out if there is any dependence between education and country of residence in the analysed population.

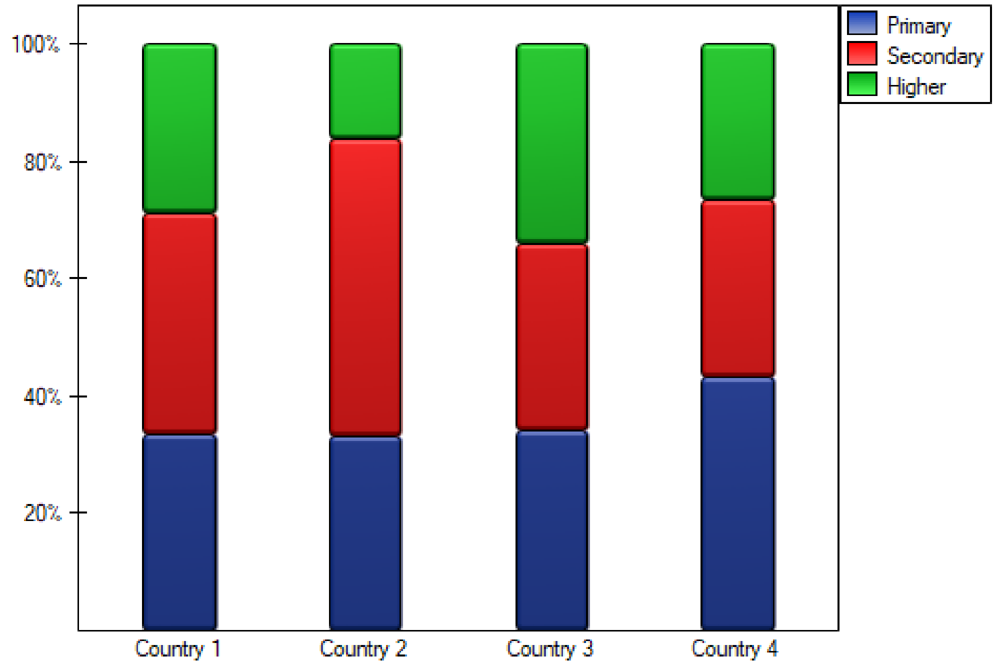
Hypotheses:

$\mathcal{H}_0$: there is no dependence between education and country of residence in the analysed population,

$\mathcal{H}_1$: there is a dependence between education and country of residence in the analysed population.



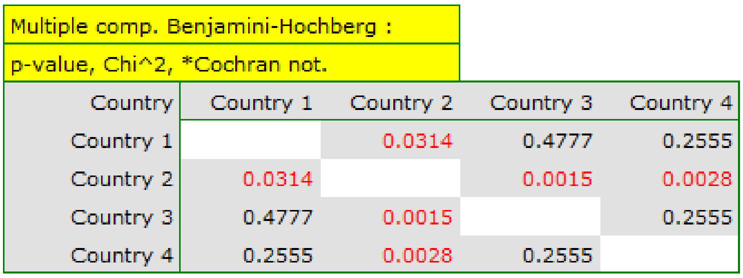
Chi-square, Fisher, OR/RR	
Analysed variables	Country
	Education
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	605
Cochran condition	fulfilled
Pearson's chi-square statistic	23.6104
Degrees of freedom	6
p-value	0.0006



Cochran's condition is satisfied.

The p value = 0.0006. So, on the basis of the significance level $\alpha = 0.05$ we can draw the conclusion that there is a dependence between education and country of residence in the analysed population.

If we are interested in more precise information about the detected dependencies, we will obtain it by determining multiple comparisons through the options Fisher, Yates and others... and then Multiple column comparisons (RxC) and one of the corrections e.g. Benjamini-Hochberg



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.

15.2.5 The Fisher's test for large tables

The Fisher test for $r \times c$ tables is also called the Fisher-Freeman-Halton test (Freeman G.H., Halton J.H. (1951)[62]). This test is an extension on $r \times c$ tables of the Fisher's exact test. It defines the exact probability of an occurrence specific distribution of numbers in the table (when we know n and we set the marginal totals).
If you define marginal sums of each row as:

$$W_i = \sum_{j=1}^{c} O_{ij},$$

where:
O_{ij} – observed frequencies in a table,

and the marginal sums of each column as:

$$K_i = \sum_{i=1}^{r} O_{ij}.$$

then, having defined the marginal sums for the different distributions of the observed frequencies represented by U_{ij}, you can calculate the P probabilities:

$$P = \frac{D^{-1}\prod_{j=1}^{c} K_j!}{U_{1j}!U_{2j}!\dots U_{rj}},$$

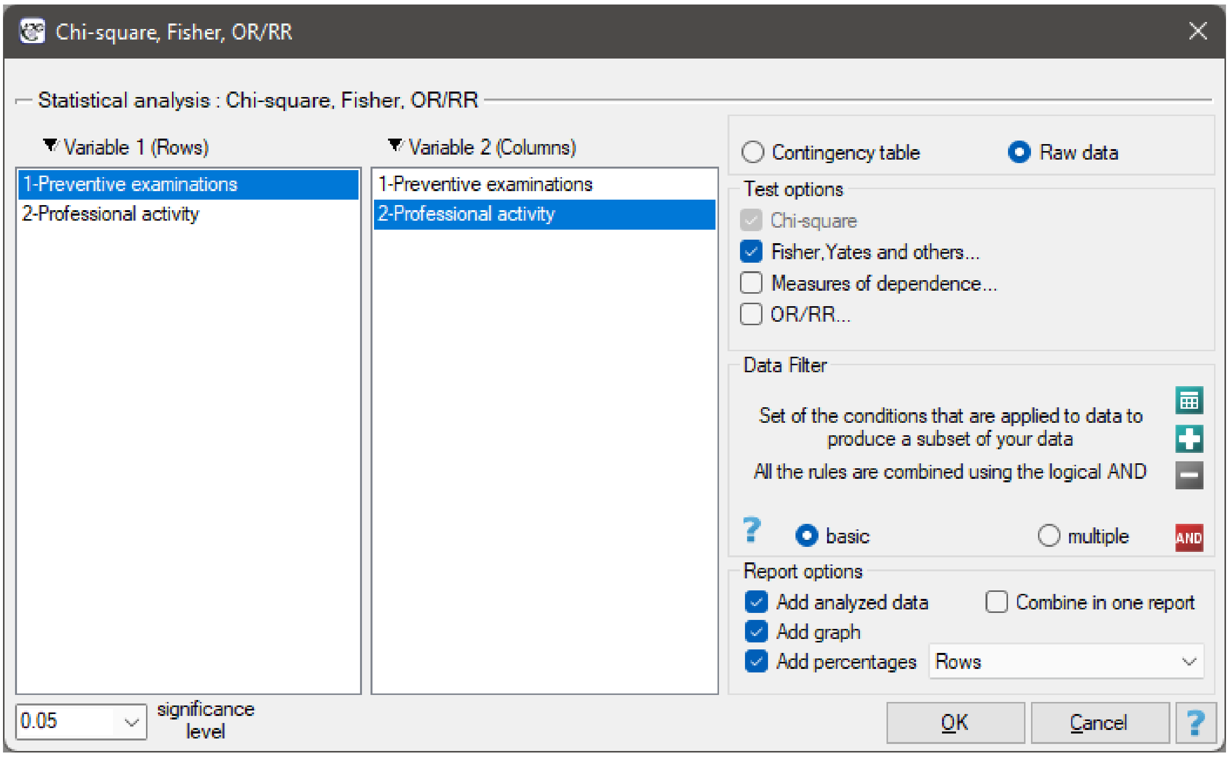
where

$$D = \frac{(W_1 + W_2 + \dots + W_r)!}{W_1!W_2!\dots W_r!}.$$

The exact significance level p: is the sum of P probabilities (calculated for new values U_{ij}), which are smaller or equal to P probability of the table with the initial numbers O_{ij}.

The exact p value is compared with the significance level α.

The settings window with the Fisher exact test (RxC) can be opened in Statistics menu → NonParametric tests → Chi-square, Fisher, OR/RR or in Wizard.



Chi-square, Fisher, OR/RR

Statistical analysis : Chi-square, Fisher, OR/RR

Variable 1 (Rows)
1-Preventive examinations
2-Professional activity

Variable 2 (Columns)
1-Preventive examinations
2-Professional activity

○ Contingency table ● Raw data

Test options
☑ Chi-square
☑ Fisher, Yates and others...
□ Measures of dependence...
□ OR/RR...

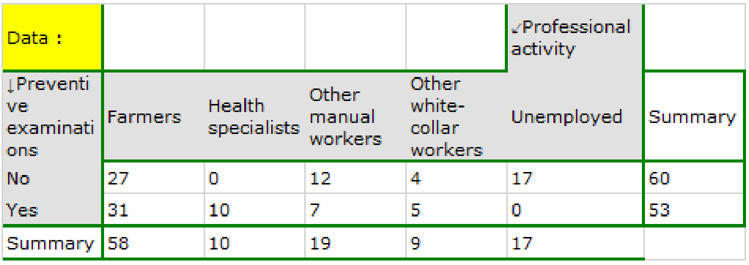
Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
☑ Add analyzed data □ Combine in one report
☑ Add graph
☑ Add percentages Rows

0.05 significance level

Info.
The process of calculation of p values for this test is based on the algorithm published by Mehta (1986)[117].

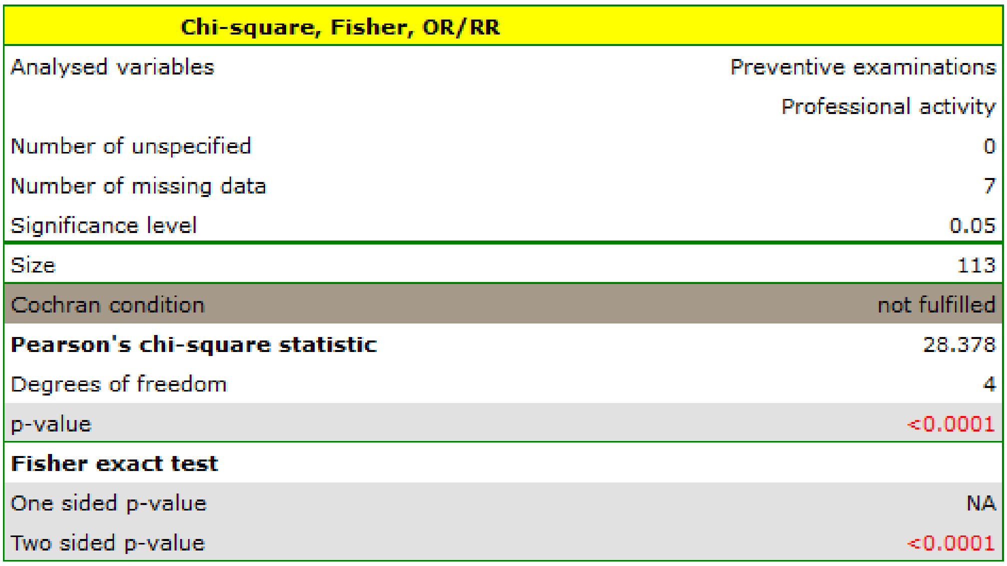
Example 15.6. (job prevention.pqs file)
In the population of people living in the rural areas of Komorniki municipality it was examined whether the performance of preventive health examinations depends on the type of occupational activity of the residents. A random sample of 120 people was collected and asked about their education and whether they perform preventive examinations. Complete answers were obtained from 113 persons.



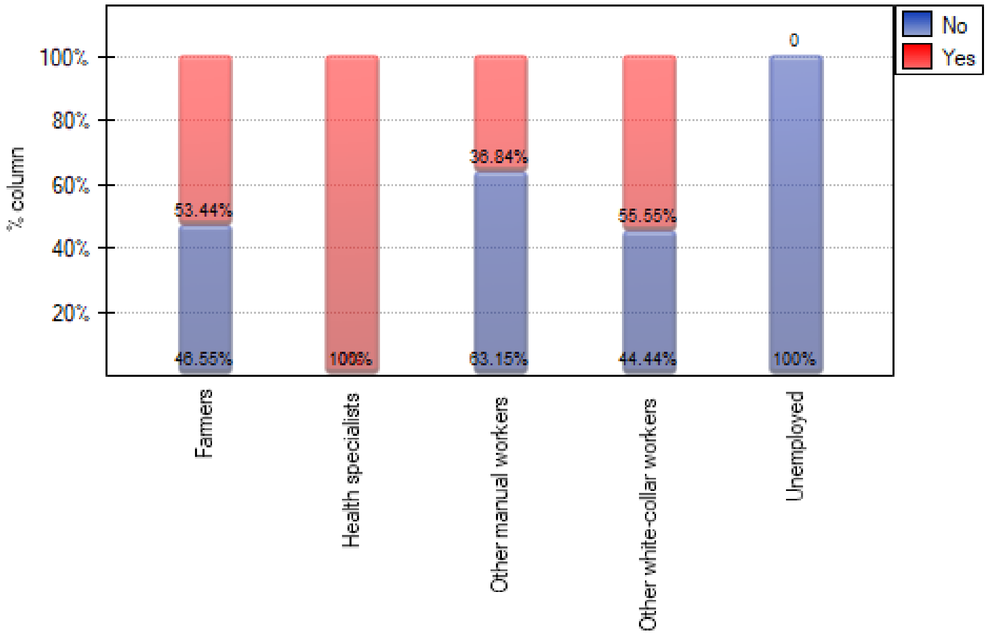
Data :					↙Professional activity	
↓Preventive examinations	Farmers	Health specialists	Other manual workers	Other white-collar workers	Unemployed	Summary
No	27	0	12	4	17	60
Yes	31	10	7	5	0	53
Summary	58	10	19	9	17	

Hypotheses:

$\mathcal{H}_0$: there is no correlation between performance of preventive examinations and the type of work performed by the residents of rural areas of the Komorniki commune,
$\mathcal{H}_1$: there is a correlation between performance of preventive examinations and the type of work performed by the residents of rural areas of the Komorniki commune.



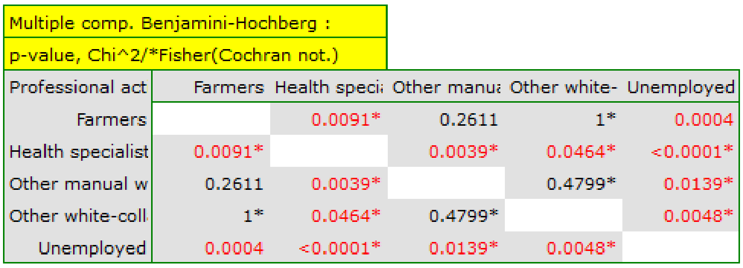
Chi-square, Fisher, OR/RR	
Analysed variables	Preventive examinations Professional activity
Number of unspecified	0
Number of missing data	7
Significance level	0.05
Size	113
Cochran condition	not fulfilled
Pearson's chi-square statistic	28.378
Degrees of freedom	4
p-value	<0.0001
Fisher exact test	
One sided p-value	NA
Two sided p-value	<0.0001



Cochran's condition is not satisfied, thus we should not use the chi-square test.

Value $p < 0.0001$. Therefore, at the significance level $\alpha = 0.05$ we can say that there is a relationship between the performance of preventive examinations and the type of work performed by residents of rural areas of Komorniki municipality.

If we are interested in more precise information about the correlations detected, we will obtain it by determining multiple comparisons through the options Fisher, Yates and others... and then Multiple column comparisons (RxC) and one of the corrections e.g. Benjamini-Hochberg.



Multiple comp. Benjamini-Hochberg :

p-value, Chi^2/*Fisher(Cochran not.)

Professional act	Farmers	Health speci	Other manua	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%, which means that these three groups are not statistically significantly different from each other. Part of the p-values obtained in the table is marked with an asterisk, it denotes those results which were obtained by using the Fisher's exact test with Benjamini-Hochberg correction, values not marked with an asterisk are the results of the chi-square test with Benjamini-Hochberg correction, in which Cochran's assumptions were fulfilled.

15.2.6 The Chi-square test for small tables

These tests are based on the data gathered in the form of a contingency table of 2 features (X, Y), each of them has 2 possible categories X_1, X_2 and Y_1, Y_2 (look at the table (10.1)).

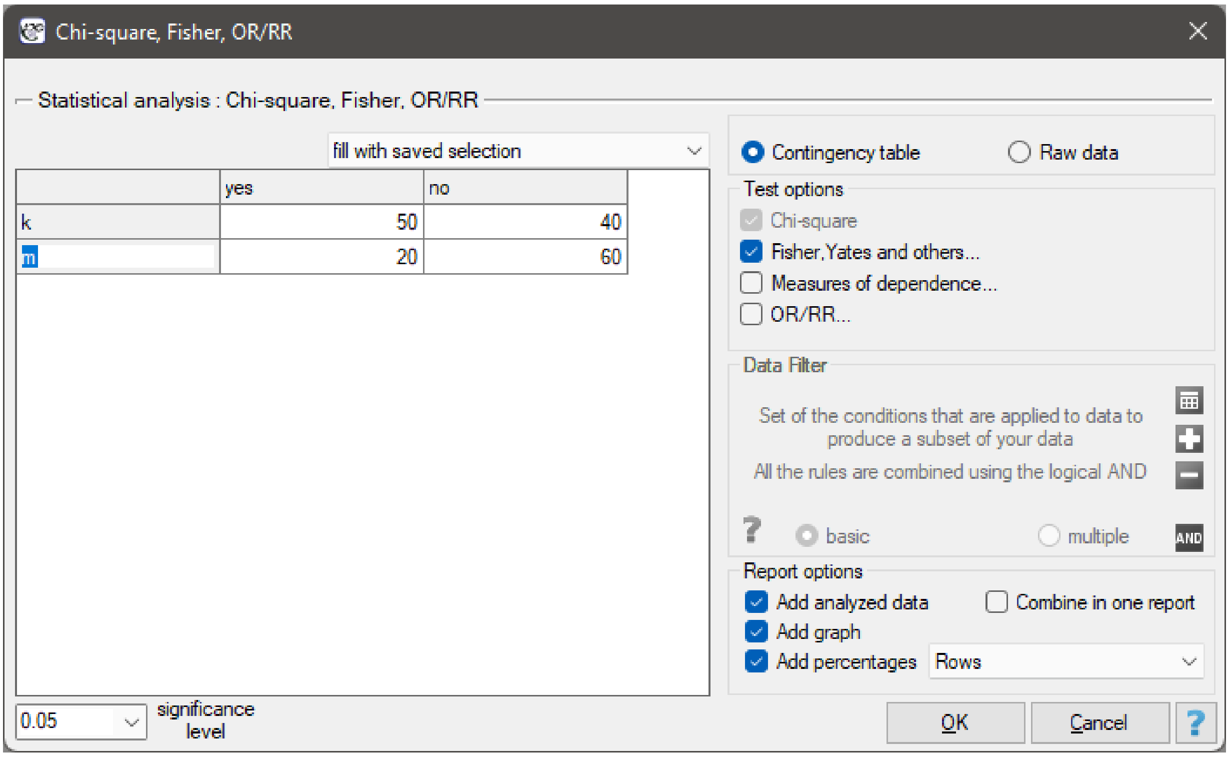
The χ^2 test for 2×2 tables – The Pearson's Chi-square test (Karl Pearson 1900) is constraint of the χ^2 test for $r \times c$ tables.

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 a 1 degree of freedom.

The settings window with the Chi-square test (2x2) can be opened in Statistics menu → NonParametric tests→Chi-square, Fisher, OR/RR or in Wizard.



Chi-square, Fisher, OR/RR

Statistical analysis : Chi-square, Fisher, OR/RR

fill with saved selection

	yes	no
k	50	40
m	20	60

● Contingency table ○ Raw data

Test options

☑ Chi-square

☑ Fisher,Yates and others...

□ Measures of dependence...

□ OR/RR...

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

☑ Add analyzed data □ Combine in one report

☑ Add graph

☑ Add percentages Rows

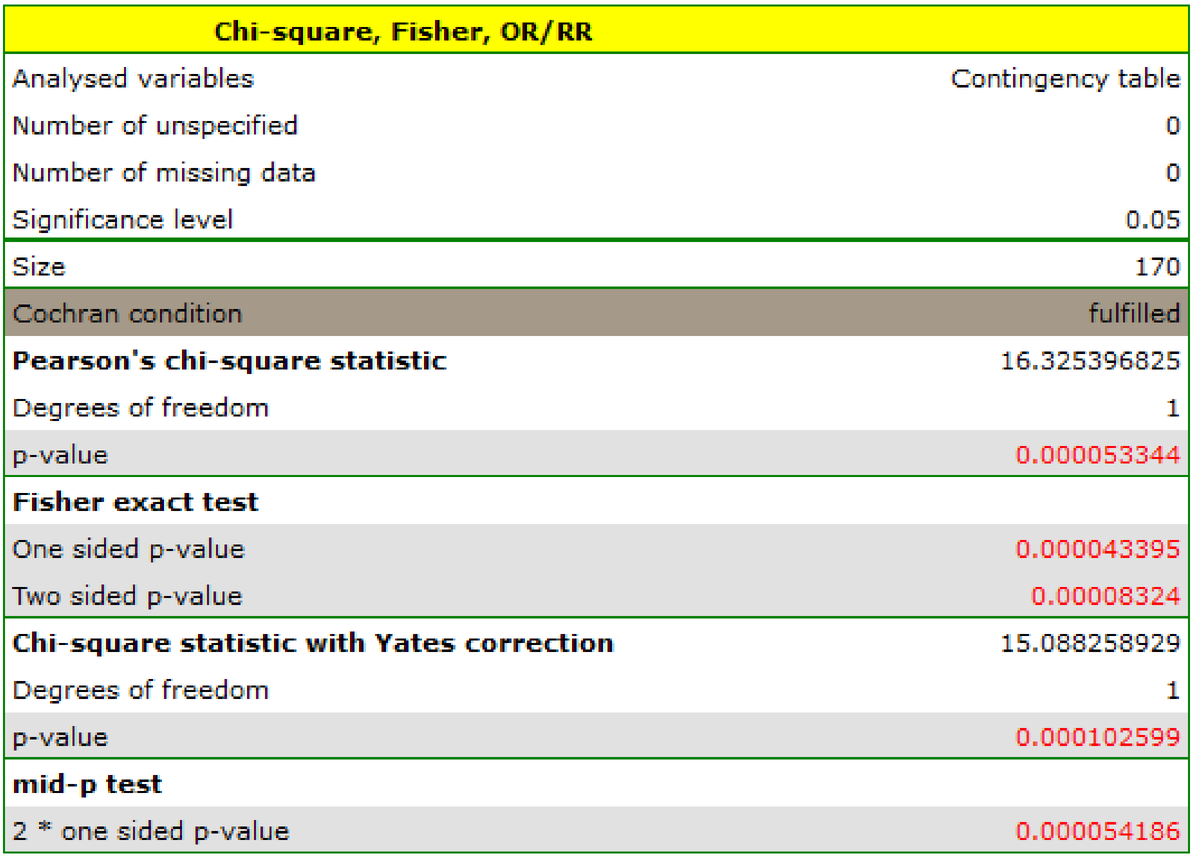
0.05 significance level

Example 15.7. (sex-exam.pqs file)
There is a sample consisting of 170 persons ($n = 170$). Using this sample, you want to analyse 2 features (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 the sample you want to get to know, if there is any dependence between sex and exam passing in the above population. The data distribution is presented in the contingency table below:

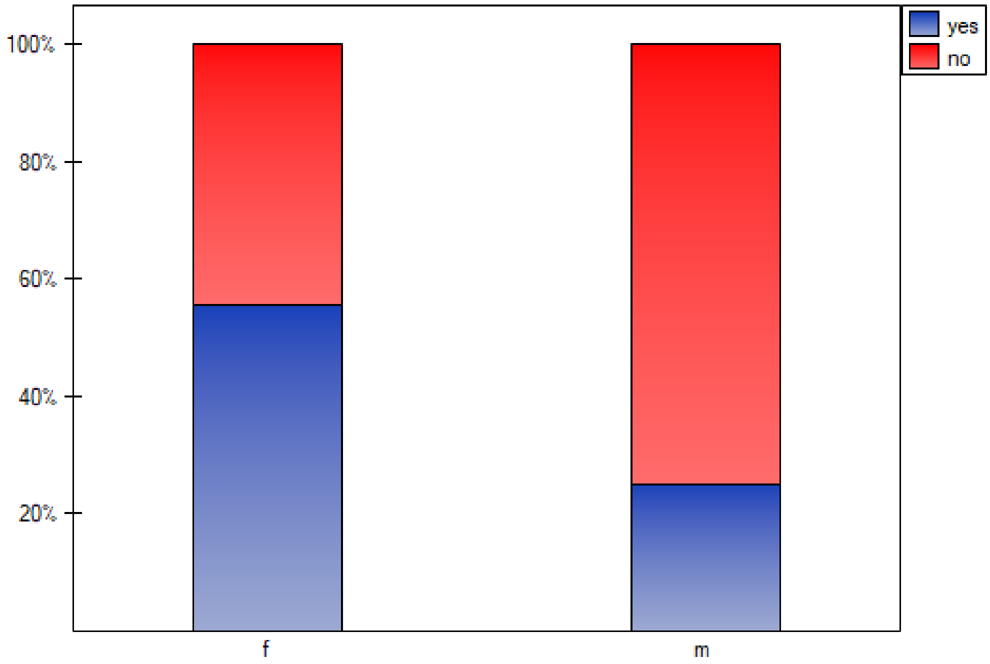
Observed frequencies O_{ij}		exam passing		
		yes	no	total
sex	f	50	40	90
	m	20	60	80
	total	70	100	170

Hypotheses:

$\mathcal{H}_0$: there is no dependence between sex and exam passing in the analysed population,
$\mathcal{H}_1$: there is a dependence between sex and exam passing in the analysed population.



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



The expectation count table contains no values less than 5. Cochran's condition is satisfied.

At the assumed significance level of $\alpha = 0.05$ all tests performed confirmed the truth of the alternative hypothesis:

- chi-square test, $p = 0.000053$,

- chi-square test with Yeates correction, $p = 0.000103$,
- Fisher’s exact test, $p = 0.000083$,
- mid-p test, $p = 0.000054$.

15.2.7 The Chi-square test corrections for small tables

These tests are based on data collected in the form of a contingency table of 2 features (X, Y), each of which has possible 2 categories X_1, X_2 and Y_1, Y_2 (look at the table(10.1)).

The Chi-square test with the Yate's correction for continuity

The χ^2 test with the Yate's correction (Frank Yates (1934)[174]) is a more conservative test than the χ^2 test (it rejects a null hypothesis more rarely than the χ^2 test). The correction for continuity guarantees the possibility of taking in all the values of real numbers by a test statistic, according to the χ^2 distribution assumption.
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 for 2×2 tables

The Fisher test for 2×2 tables is also called the Fisher exact test (R. A. Fisher (1934)[56], (1935)[57]). This test enables you to calculate the exact probability of the occurrence of the particular number distribution in a table (knowing n and defined marginal sums.

$$P=\frac{\binom{O_{11}+O_{21}}{O_{11}}\binom{O_{12}+O_{22}}{O_{12}}}{\binom{O_{11}+O_{12}+O_{21}+O_{22}}{O_{11}+O_{12}}}.$$

If you know each marginal sum, you can calculate the P probability for various configurations of observed frequencies. The exact p significance level is the sum of probabilities which are less or equal to the analysed probability.

The mid-p test

The mid-p is the Fisher exact test correction. This modified p value is recommended by many statisticians (Lancaster 1961[95], Anscombe 1981[9], Pratt and Gibbons 1981[133], Plackett 1984[132], Miettinen 1985[118] and Barnard 1989[15], Rothman 2008[138]) as a method used in decreasing the Fisher exact test conservatism. As a result, using the mid-p the null hypothesis is rejected much more qucikly than by using the Fisher exact test. For large samples a p value is calculated by using the χ^2 test with the Yate's correction and the Fisher test gives quite similar results. But a p value of the χ^2 test without any correction corresponds with the mid-p.

The p value of the mid-p is calculated by the transformation of the probability value for the Fisher exact test. The one-sided p value is calculated by using the following formula:

$$p_{I(mid-p)}=p_{I(Fisher)}-0.5\cdot P_{punktu(tabeli\quad zadanej)},$$

where:

$p_{I(mid-p)}$ – one-sided p value of mid-p,
$p_{I(Fisher)}$ – one-sided p value of Fisher exact test,

and the two-sided p value is defined as a doubled value of the smaller one-sided probability:

$$p_{II(mid-p)}=2p_{I(mid-p)},$$

where:

$p_{II(mid-p)}$ – two-sided p value of mid-p.

The settings window with the chi-square test and its corrections can be opened in Statistics menu → NonParametric tests→Chi-square, Fisher, OR/RR or in Wizard.

15.2.8 The Chi-square test for trend

The χ^2 test for trend (also called the Cochran-Armitage trend test[41][10]) is used to determine whether there is a trend in proportion for particular categories of an analysed variables (features). It is based on the data gathered in the contingency tables of 2 features. The first feature has the possible r ordered categories: $X_1, X_2, ..., X_r$ and the second one has 2 categories: G_1, G_2 (table (15.1)).

Tabela 15.1. The contingency table of $r \times 2$ observed frequencies

Observed frequencies O_{ij}		Feature 2 (group)		
		G_1	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	$C_1 = \sum_{i=1}^{r} O_{i1}$	$C_2 = \sum_{i=1}^{r} O_{i2}$	$n = C_1 + C_2$

Basic assumptions:

- measurement on an ordinal scale or on an interval scale,
- an independent model (the second feature – 2 independent groups).

Hypotheses:

$\mathcal{H}_0$: In the analysed population the trend in a proportion of $p_1, p_2, ..., p_r$ does not exist,
$\mathcal{H}_1$: There is the trend in a proportion of $p_1, p_2, ..., p_r$ in the analysed population.

where:
$p_1, p_2, ..., p_r$ are the proportions $p_1 = \frac{O_{11}}{W_1}$, $p_2 = \frac{O_{21}}{W_2}$,..., $p_r = \frac{O_{r1}}{W_r}$.

The test statistic is defined by:

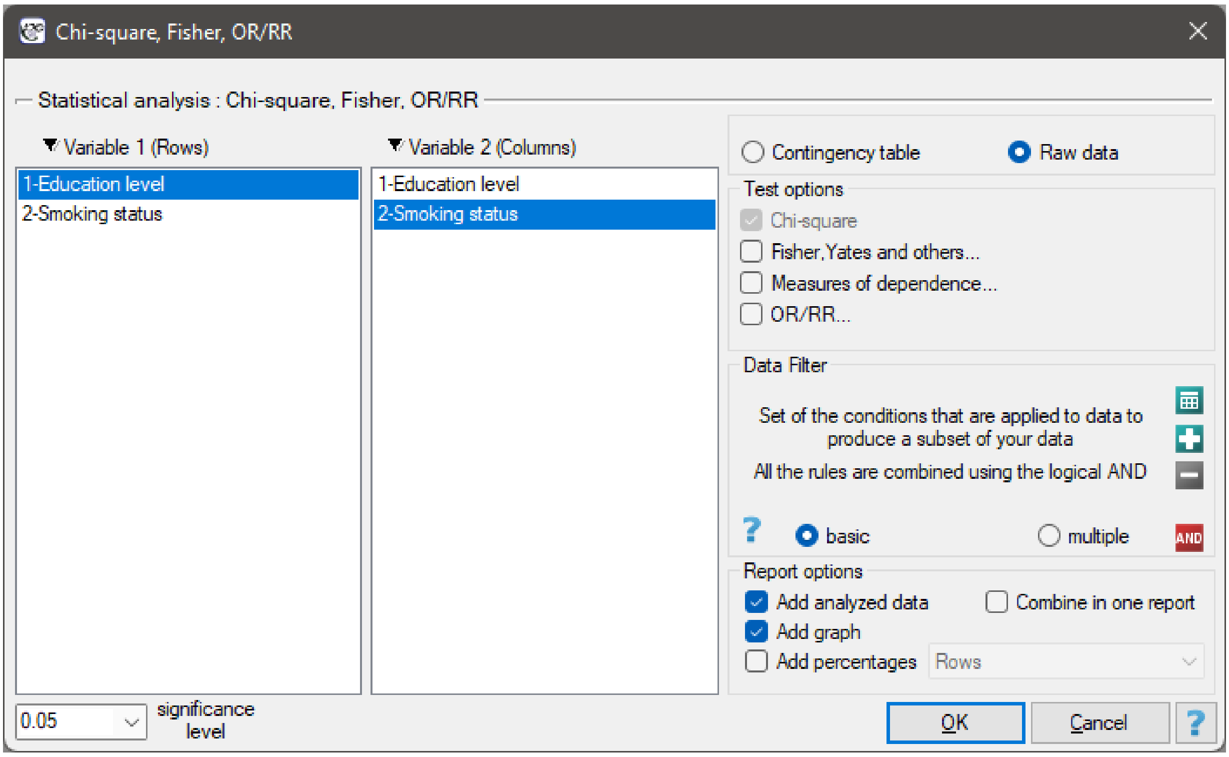
$$\chi^2 = \frac{\left[\left(\sum_{i=1}^{r} i \cdot O_{i1}\right) - C_1\left(\sum_{i=1}^{r} \frac{i \cdot W_i}{n}\right)\right]^2}{\frac{C_1}{n}\left(1 - \frac{C_1}{n}\right)\left[\left(\sum_{i=1}^{n} i^2 W_i\right) - n\left(\sum_{i=1}^{n} \frac{i \cdot W_i}{n}\right)^2\right]}.$$

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with 1 degree of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

if $p \le \alpha$ $\Longrightarrow$ reject $\mathcal{H}_0$ and accept $\mathcal{H}_1$,
if $p > \alpha$ $\Longrightarrow$ there is no reason to reject $\mathcal{H}_0$.

The settings window with the Chi-square test for trend can be opened in Statistics menu → NonParametric tests → Chi-square, Fisher, OR/RR → Chi-square for trend.



Chi-square, Fisher, OR/RR

Statistical analysis : Chi-square, Fisher, OR/RR

Variable 1 (Rows): 1-Education level (selected), 2-Smoking status

Variable 2 (Columns): 1-Education level, 2-Smoking status (selected)

○ Contingency table ● Raw data

Test options
☑ Chi-square
☐ Fisher,Yates and others...
☐ Measures of dependence...
☐ OR/RR...

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

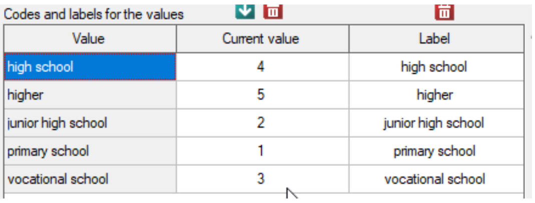
Report options
☑ Add analyzed data ☐ Combine in one report
☑ Add graph
☐ Add percentages Rows

0.05 significance level

EXAMPLE 15.8. (smoking-education.pqs file)
We examine whether cigarette smoking is related to the education of residents of a village. A sample of 122 people was drawn. The data were recorded in a file.

We assume that the relationship can be of two types i.e. the more educated people, the more often they smoke or the more educated people, the less often they smoke. Thus, we are looking for an increasing or decreasing trend.

Before proceeding with the analysis, we need to prepare the data, i.e., we need to indicate the order in which the education categories should appear. To do this, from the properties of the Education variable, we select Codes/Labels/Format... and assign the order by specifying consecutive natural numbers. We also assign labels.



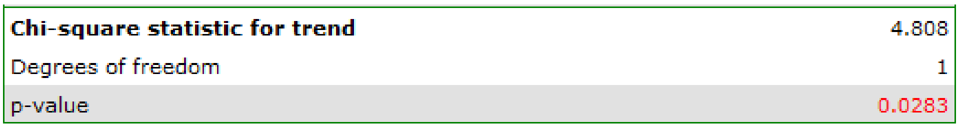
Codes and labels for the values

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

Hypotheses:

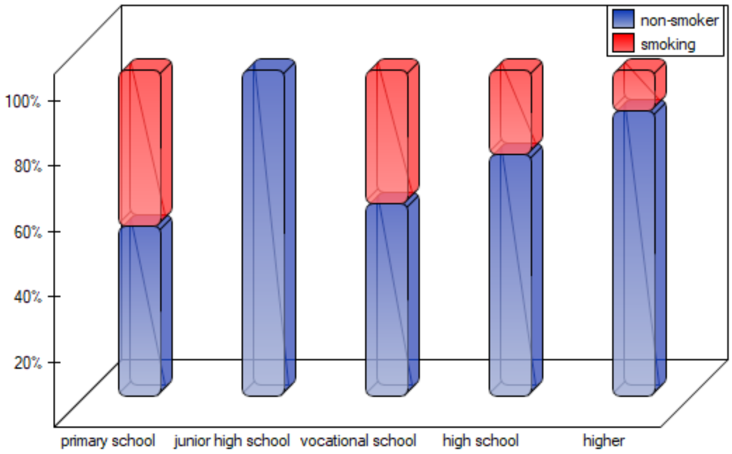
$\mathcal{H}_0:$ there is no trend in the rural population of increasing/decreasing wraz ze wzrostem wykształcenia,

$\mathcal{H}_1:$ there is a trend in the rural population of increasing/decreasing numbers of smokers with increasing education.

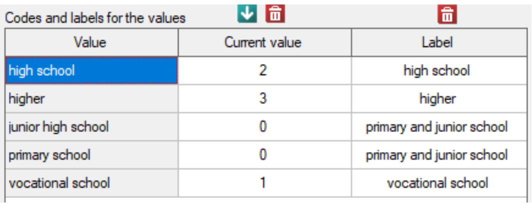


Chi-square statistic for trend	4.808
Degrees of freedom	1
p-value	0.0283

A value of $p = 0.0091$, which compared to a significance level of α=0.05 indicates that the alternative hypothesis that a trend exists is true.



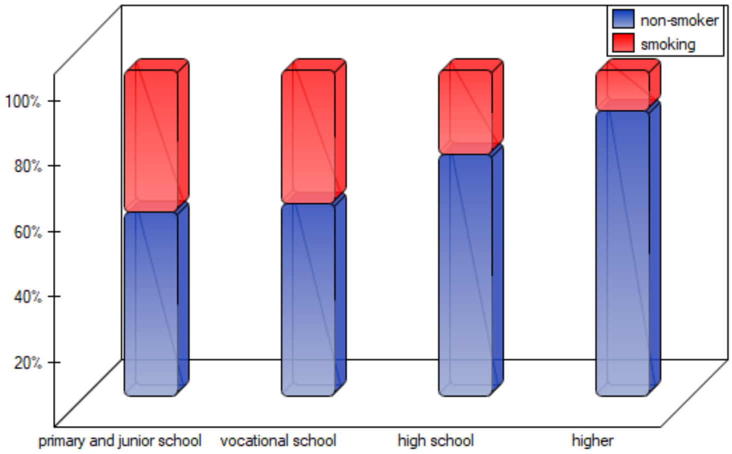
As the graph shows, the more educated people are, the less often they smoke. However, the result obtained by people with junior high school education deviates from this trend. Since there are only two people with lower secondary school education, it did not have much influence on the trend. Due to the very small size of this group, it was decided to repeat the analysis 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

A small value was again obtained $p = 0.0078$ and confirmation of a statistically significant trend.



Example 15.9. (viewers.pqs file)
Because of the decrease in people watching some particular soap opera there was carried out an opinion survey. 100 persons were asked, who has recently started watching this soap opera, and 300 persons were asked, who has watched it regularly from the beginning. They were asked about the level of preoccupation with the character's life. The results are written down 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

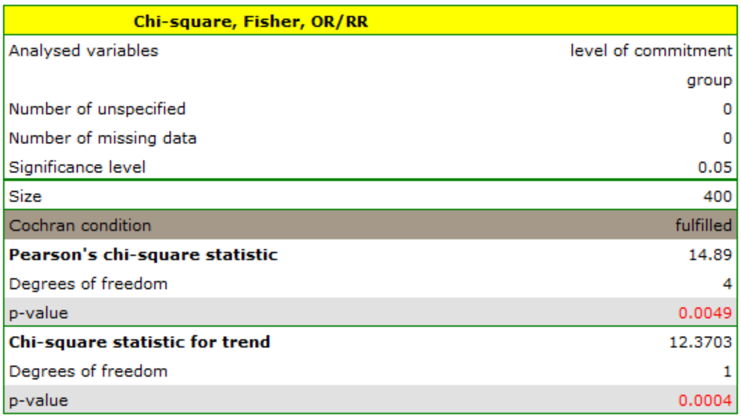
The new viewers consist of 25% of all the analysed viewers. This proportion is not the same for each level of commitment, but looks like this:

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%**

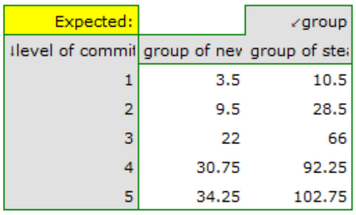
Hypotheses:

$\mathcal{H}_0$: in the population of the soap opera viewers, the trend in proportions of p_1, p_2, p_3, p_4, p_5 does not exist,

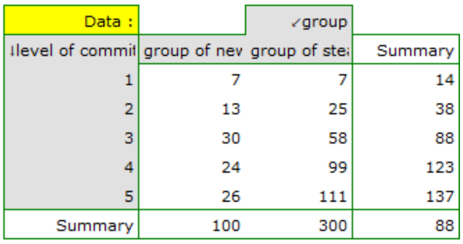
$\mathcal{H}_1$: in the population of the soap opera viewers, the trend in proportions of p_1, p_2, p_3, p_4, p_5 does exists.



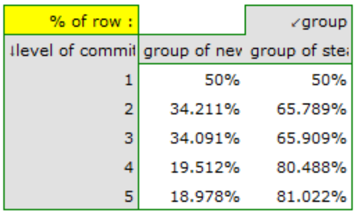
Chi-square, Fisher, OR/RR	
Analysed variables	level of commitment group
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	400
Cochran condition	fulfilled
Pearson's chi-square statistic	14.89
Degrees of freedom	4
p-value	0.0049
Chi-square statistic for trend	12.3703
Degrees of freedom	1
p-value	0.0004



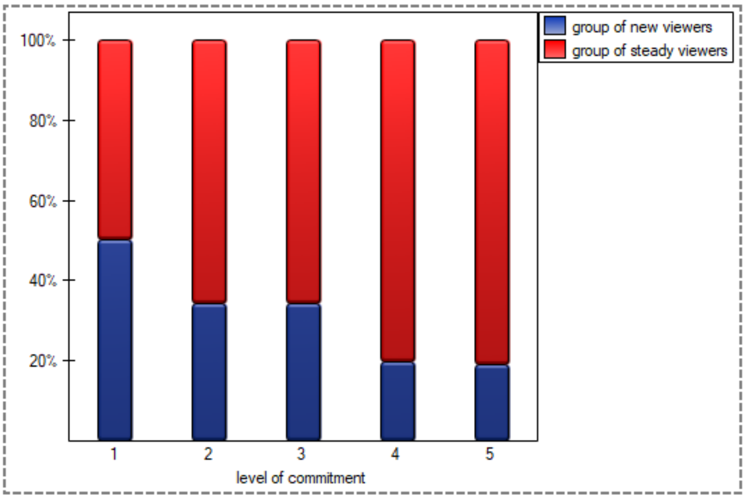
Expected:		✓group
level of commit	group of nev	group of ste
1	3.5	10.5
2	9.5	28.5
3	22	66
4	30.75	92.25
5	34.25	102.75



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



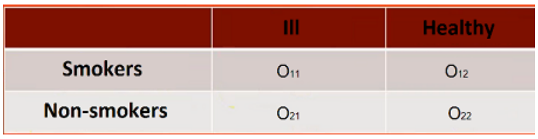
% of row :		✓group
level of commit	group of nev	group of ste
1	50%	50%
2	34.211%	65.789%
3	34.091%	65.909%
4	19.512%	80.488%
5	18.978%	81.022%



The $p - value = 0.0004$ which, compared to the significance level α=0.05, proves the truth of the alternative hypothesis that there is a trend in the proportions $p_1, p_2, ..., p_5$. As can be seen from the contingency table of the percentages calculated from the sum of the columns, this is a decreasing trend (the more interested the group of viewers is in the fate of the characters of the series, the smaller part of it is made up of new viewers).

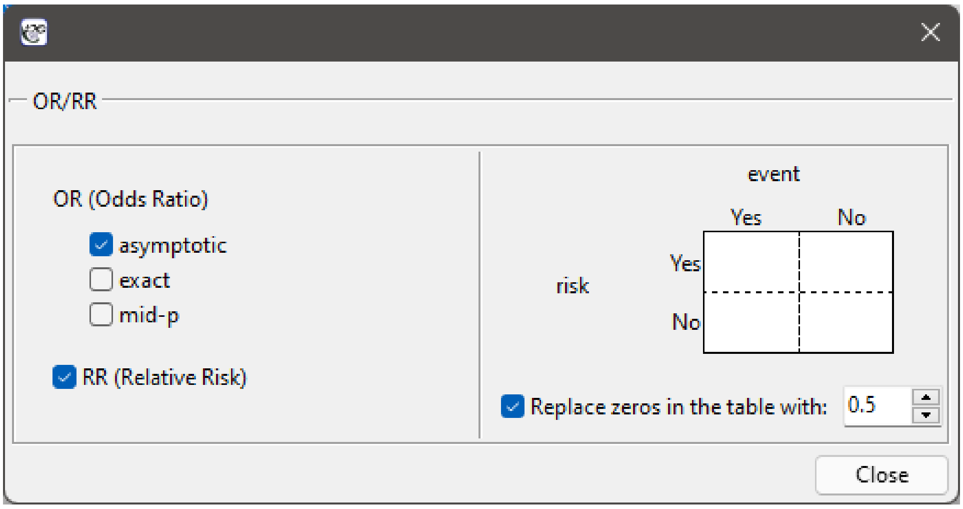
15.2.9 The Relative Risk and the Odds Ratio

The risk and odds designation of occurence an analysed phenomenon, on the basis of exposure to the factor that can cause it, is estimated according to data collected in the contingency table 2×2. For example, we can look at how cigarette smoking affects disease:



	Ill	Healthy
Smokers	O_{11}	O_{12}
Non-smokers	O_{21}	O_{22}

The window with the ability to determine these measures is called up via the menu can be opened in Statistics menu → NonParametric tests→chi-square, Fiser, OR/RR by selecting OR/RR or in Wizard.



OR/RR

OR (Odds Ratio)

☑ asymptotic
□ exact
□ mid-p

☑ RR (Relative Risk)

☑ Replace zeros in the table with: 0.5

If a study is a **case-control** study, the **odds ratio** of occurence the phenomenon is calculated for the table. Usually, they are retrospective studies – the researcher decides on his own about the sample size, with the phenomenon, and about the control sample (without the phenomenon).

If a study is a **cohort** study, the **relative risk** of occurence the phenomenon is calculated for the table. Usually, they are prospective studies – the researcher cares about experiment conditions, because of the structure of an analysed phenomenon in a sample and in a population should be similar.

The odds ratio (2×2 table)

For the designation of odds ratio, we calculate the probability of being a case in the exposed group and in the unexposed group, 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

This test is used to the hypothesis verification about the odds of occurence the analysed phenomenon is the same in the group of exposed and unexposed to the risk factor.
Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & OR = 1, \\ \mathcal{H}_1: & OR \neq 1. \end{array}$$

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 the $\ln(OR)$.

The test statistic asymptotically (for large sample size) has the normal distribution.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note

In the interpretation of odds ratio significance, we usually use the designated confidence interval. Then, we check if the interval contains the value of 1.

The relative risk (2×2 table)

In the cohort study, we can designate the **risk** of occurence the analysed phenomenon (because the structure of phenomenon, in the sample, should come closer to the population, from which the sample was taken) and calculate the relative risk (RR).
The estimated risk of occurence the analysed phenomenon is designated by the following formula $R = \frac{O_{11}+O_{21}}{n}$. However, the relative risk is designated by:

$$RR = \frac{O_{11}/(O_{11}+O_{12})}{O_{21}/(O_{21}+O_{22})}$$

The test of significance for the RR

This test is used to the hypothesis verification about the risk of occurence the analysed occurrence is the same in the group of exposed and unexposed to the risk factor.
Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & RR = 1, \\ \mathcal{H}_1: & RR \neq 1. \end{array}$$

The test statistic is defined by:

$$z = \frac{\ln(RR)}{SE},$$

where:
$SE = \sqrt{\frac{1}{O_{11}} - \frac{1}{0_{11}+0_{12}} + \frac{1}{O_{21}} - \frac{1}{0_{21}+0_{22}}}$ – standard error of the $\ln(RR)$.

The test statistic asymptotically (for large sample size) has the normal distribution.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note

In the interpretation of the relative risk significance, we usually use the designated confidence interval. Then, we check if the interval contains the value of 1.

Note

When zeros are present in the data table, then it may not be possible to calculate the odds ratio or relative risk. In such a situation, to ensure that the odds ratio can be determined, you can check the Replace zero in the table with: option in the analysis window. Selecting this option (adjustments for continuity) adds the set value to all cells in the table.

15.2.10 The Z test for 2 independent proportions

The Z test for 2 independent proportions is used in the similar situations as the chi^2 test (2×2). It means, when there are 2 independent samples with the total size of n_1 and n_2, with the 2 possible results to gain (one of the results is distinguished with the size of m_1 - in the first sample and m_2 - in the second one). For these samples it is also possible to calculate the distinguished proportions $p_1 = \frac{m_1}{n_1}$ and $p_2 = \frac{m_2}{n_2}$. This test is used to verify the hypothesis informing us that the distinguished proportions P_1 and P_2 in populations, from which the samples were drawn, are equal.

Basic assumptions:

- measurement on a nominal scale - any order is not taken into account,
- an independent model,
- large sample sizes.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & P_1 = P_2, \\ \mathcal{H}_1: & P_1 \neq P_2, \end{array}$$

where:
P_1, P_2 fraction for the first and the second population.

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 Z Statistic with and without the continuity correction asymptotically (for the large sample sizes) has the normal distribution.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Apart from the difference between proportions, the program calculates the value of the NNT.

NNT (*number needed to treat*) – indicator used in medicine to define the number of patients which have to be treated for a certain time in order to cure one person.

. NNT is calculated from the formula:

$$NNT = \frac{1}{|p_1 - p_2|}$$

and is quoted when the difference $p_1 - p_2$ is positive.

NNH (*number needed to harm*) – an indicator used in medicine, denotes the number of patients whose exposure to a risk over a specified period of time, results in harm to one person who would not otherwise be harmed. NNH is calculated in the same way as NNT, but is quoted when the difference $p_1 - p_2$ is negative.

Confidence interval – The narrower the confidence interval, the more precise the estimate. If the confidence interval includes 0 for the difference in proportions and ∞ for the NNT and/or NNH, then there is an indication to treat the result as statistically insignificant

Note

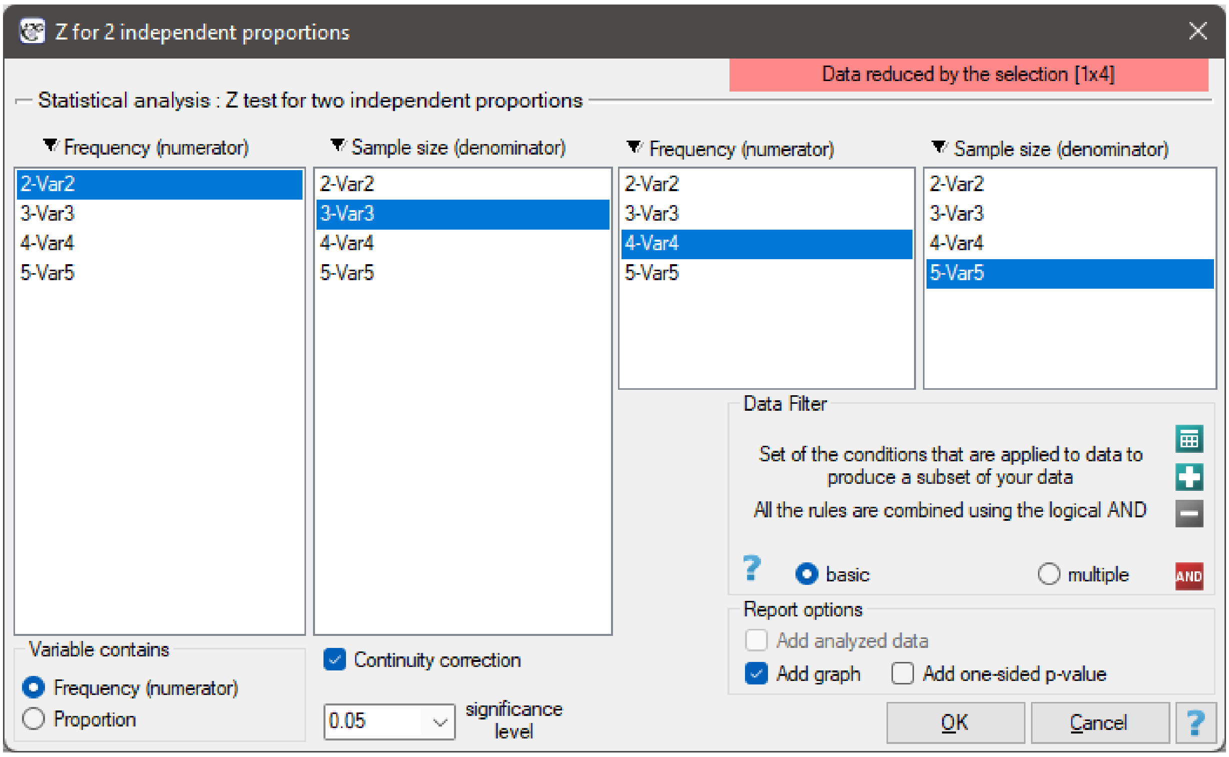
From PQStat version 1.3.0, the confidence intervals for the difference between two independent proportions are estimated on the basis of the Newcombe-Wilson method. In the previous versions it was estimated on the basis of the Wald method.

The justification of the change is as follows:
Confidence intervals based on the classical Wald method are suitable for large sample sizes and for the difference between proportions far from 0 or 1. For small samples and for the difference between proportions close to those extreme values, the Wald method can lead to unreliable results (Newcombe 1998[122], Miettinen 1985[119], Beal 1987[17], Wallenstein 1997[163]). A comparison and analysis of many methods which can be used instead of the simple Wald method can be found in Newcombe's study (1998)[122]. The suggested method, suitable also for extreme values of proportions, is the method first published by Wilson (1927)[172], extended to the intervals for the difference between two independent proportions.

Note

The confidence interval for NNT and/or NNH is calculated as the inverse of the interval for the proportion, according to the method proposed by Altman (Altman (1998)[6]).

The settings window with the Z test for 2 proportions can be opened in Statistics menu → NonParametric tests → Z for 2 independent proportions.



Z for 2 independent proportions

Data reduced by the selection [1x4]

Statistical analysis : Z test for two independent proportions

Frequency (numerator)	Sample size (denominator)	Frequency (numerator)	Sample size (denominator)
2-Var2	2-Var2	2-Var2	2-Var2
3-Var3	**3-Var3**	3-Var3	3-Var3
4-Var4	4-Var4	**4-Var4**	4-Var4
5-Var5	5-Var5	5-Var5	**5-Var5**

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data

☑ Add graph □ Add one-sided p-value

Variable contains

● Frequency (numerator)

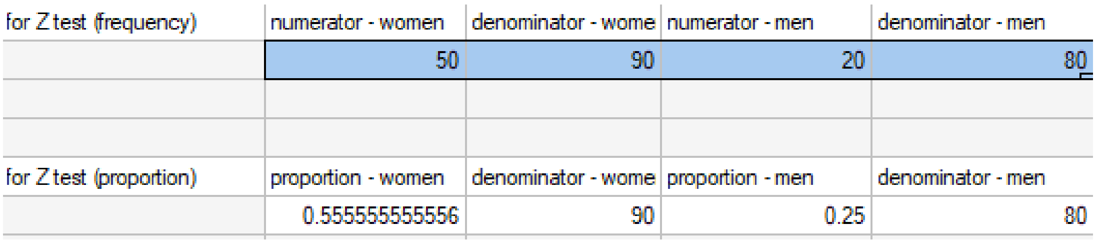
○ Proportion

☑ Continuity correction

0.05 significance level

Example (15.7) cont. *(sex-exam.pqs file)*

You know that $\frac{50}{90} = 55.56\%$ out of all the women in the sample who passed the exam and $\frac{20}{80} = 25.00\%$ out of all the men in the sample who passed the exam. This 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:

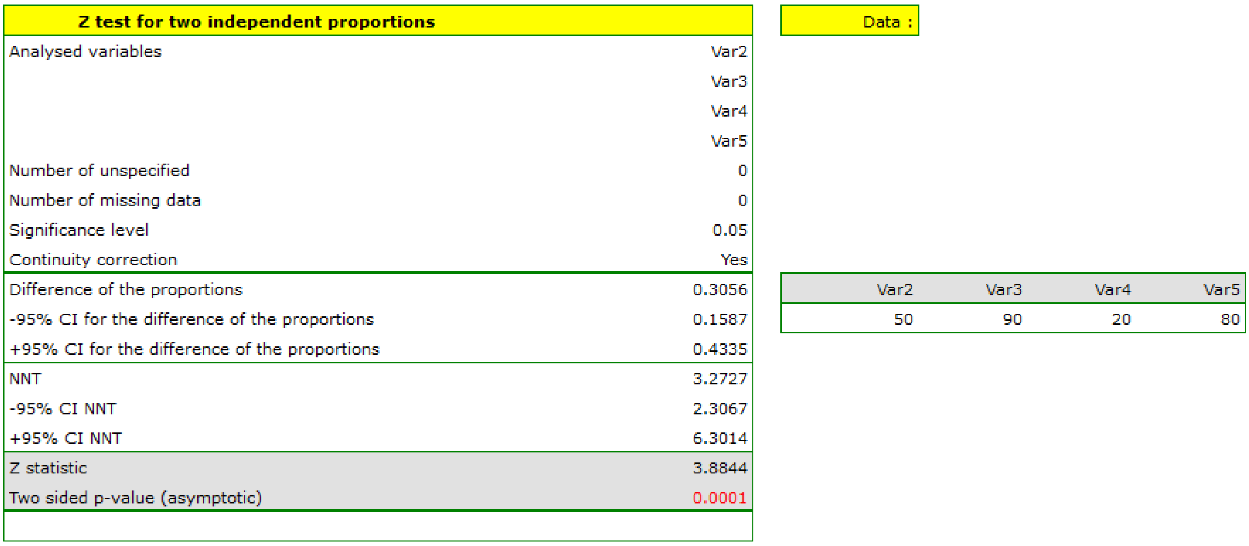


for Z test (frequency)	numerator - women	denominator - wome	numerator - men	denominator - men
	50	90	20	80
for Z test (proportion)	proportion - women	denominator - wome	proportion - men	denominator - men
	0.555555555556	90	0.25	80

Hypotheses:

$\mathcal{H}_0$: The proportion of the men who passed the exam is the same as the proportion of the women who passed the exam in the analysed population,

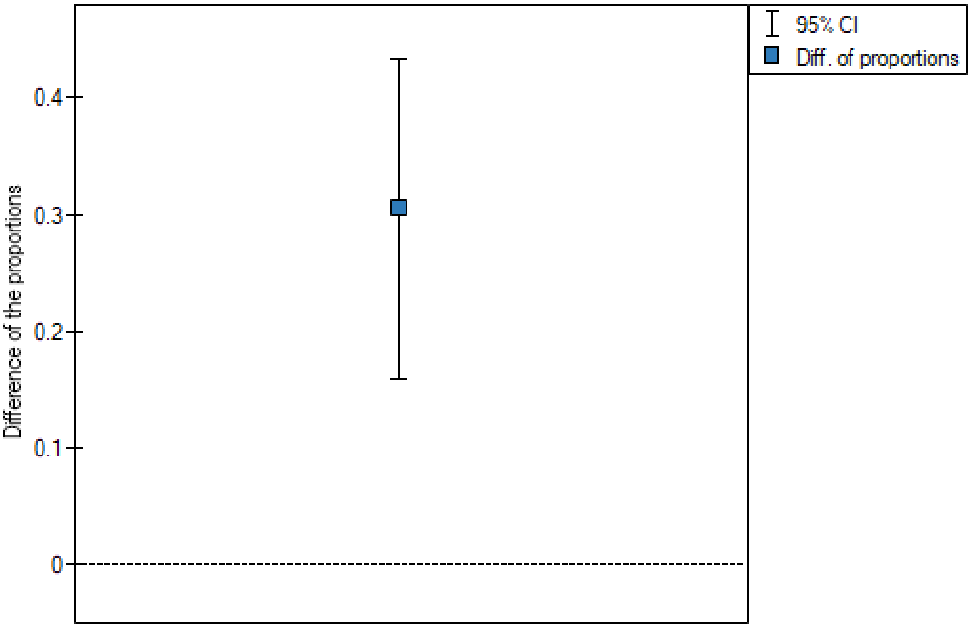
$\mathcal{H}_1$: The proportion of the men who passed the exam is different than the proportion of the women who passed the exam in the analysed population.



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



Note

It is necessary to select the appropriate area (data without headings) before the analysis begins, because usually there are more information in a datasheet. You should also select the option indicating the content of the variable (frequency (numerator) or proportion). The difference between proportions distinguished in the sample is 30.56%, a 95% and the confidence interval for it $(15.90\%, 43.35\%)$ does not contain 0.

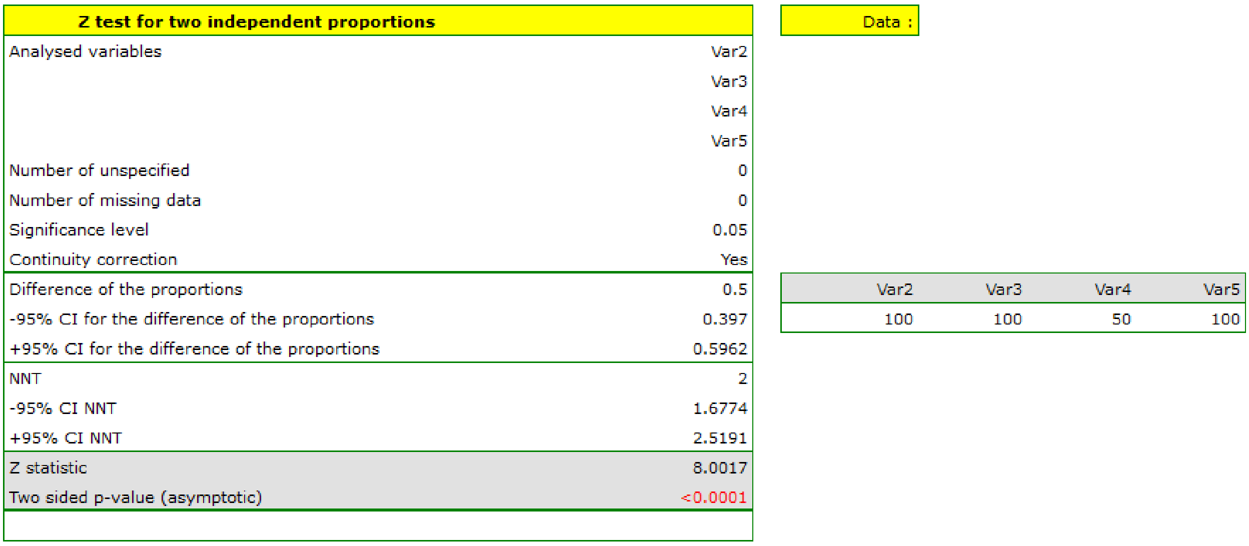
Based on the Z test without the continuity correction as well as on the Z test with the continuity correction (p value < 0.0001), on the significance level α=0.05, the alternative hypothesis can be accepted (similarly to the Fisher exact test, its the mid-p corrections, the χ^2 test and the χ^2 test with the Yate's correction). So, the proportion of men, who passed the exam is different than the proportion of women, who passed the exam in the analysed population. Significantly, the exam was passed more often by women ($\frac{50}{90} = 55.56\%$ out of all the women in the sample who passed the exam) than by men ($\frac{20}{80} = 25.00\%$ out of all the men in the sample who passed the exam).

Example 15.10.

Let us assume that the mortality rate of a disease is 100% without treatment and that therapy lowers the mortality rate to 50% – that is the result of 20 years of study. We want to know how many people have to be treated to prevent 1 death in 20 years. To answer that question, two samples of 100 people were taken from the population of the diseased. In the sample without treatment there are 100 patients of whom we know they will all die without the therapy. In the sample with therapy we also have 100 patients of whom 50 will survive.

Patients – not undergoing therapy		Patients – undergoing therapy	
sample numerator	sample (denominator)	sample numerator	sample (denominator)
100	100	50	100

We will calculate the NNT.



```
Z test for two independent proportions                          Data :
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      Var2   Var3   Var4   Var5
-95% CI for the difference of the proportions      0.397       100    100     50    100
+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
```

The difference between proportions is statistically significant ($p < 0.0001$) but we are interested in the NNT – its value is 2, so the treatment of 2 patients for 20 years will prevent 1 death. The calculated confidence interval value of 95% should be rounded off to a whole number, wherefore the NNT is 2 to 3 patients.

Example 15.11. The value of the certain proportion difference in the study comparing the effectiveness of drug 1 vs drug 2 was: difference (95%CI)=-0.08 (-0.27 do 0.11). This negative proportion difference suggests that drug 1 was less effective than drug 2, so its use put patients at risk. Because the proportion difference is negative, the determined inverse is called the NNH, and because the confidence interval contains infinity NNH(95%CI)= 2.5 (NNH 3.7 to ∞ to NNT 9.1) and goes from NNH to NNT, we should conclude that the result obtained is not statistically significant (Altman (1998)[6]).

15.2.11 The McNemar test, the Bowker test of internal symmetry

Basic assumptions:

- measurement on a nominal scale - any order is not taken into account,
- a dependent model.

The McNemar test

The McNemar test (NcNemar (1947)[116]) is used to verify the hypothesis determining the agreement between the results of the measurements, which were done twice $X^{(1)}$ and $X^{(2)}$ of an X feature (between 2 dependent variables $X^{(1)}$ and $X^{(2)}$). The analysed feature can have only 2 categories (defined here as **(+)** and **(–)**). The McNemar test can be calculated on the basis of raw data or on the basis of a 2×2 contingency table.

Tabela 15.2. 2×2 contingency table for the observed frequencies of dependent variables

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	$O_{11} + O_{21}$	$O_{12} + O_{22}$	$n = O_{11} + O_{12} + O_{21} + O_{22}$

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & O_{12} = O_{21}, \\ \mathcal{H}_1: & O_{12} \neq O_{21}. \end{array}$$

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 a 1 degree of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The Continuity correction for the McNemar test

This correction is a more conservative test than the McNemar test (a null hypothesis is rejected much more rarely than when using the McNemar test). It guarantees the possibility of taking in all the values of real numbers by the test statistic, according to the χ^2 distribution assumption. Some sources give the information that the continuity correction should be used always, but some other ones inform, that only if the frequencies in the table are small.
The test statistic with the continuity correction is defined by:

$$\chi^2 = \frac{(|O_{12} - O_{21}| - 1)^2}{O_{12} + O_{21}}.$$

McNemar's exact test

A common general rule for the asymptotic validity of the McNemar chi-square test is the **Rufibach assumption**, which is that the number of incompatible pairs is greater than 10: $O_{12} + O_{21} \ge 10$ [142] when this condition is not satisfied, then we should base the exact probability values of this test [55]. The exact probability value of the test is based on a binomial distribution and is a conservative test, so the recommended exact value of the mid-p McNemar test is also given in addition to the exact value of the MnNemar test.

Odds ratio of a result change

If the study is carried out twice for the same feature and on the same objects – then, **odds ratio** for the result change (from **(+)** to **(–)** and inversely) is calculated for the table.

The odds for the result change from **(+)** to **(–)** is O_{12}, and the odds for the result change from **(–)** to **(+)** is O_{21}. Odds Ratio (OR) is:

$$OR = \frac{O_{12}}{O_{21}}.$$

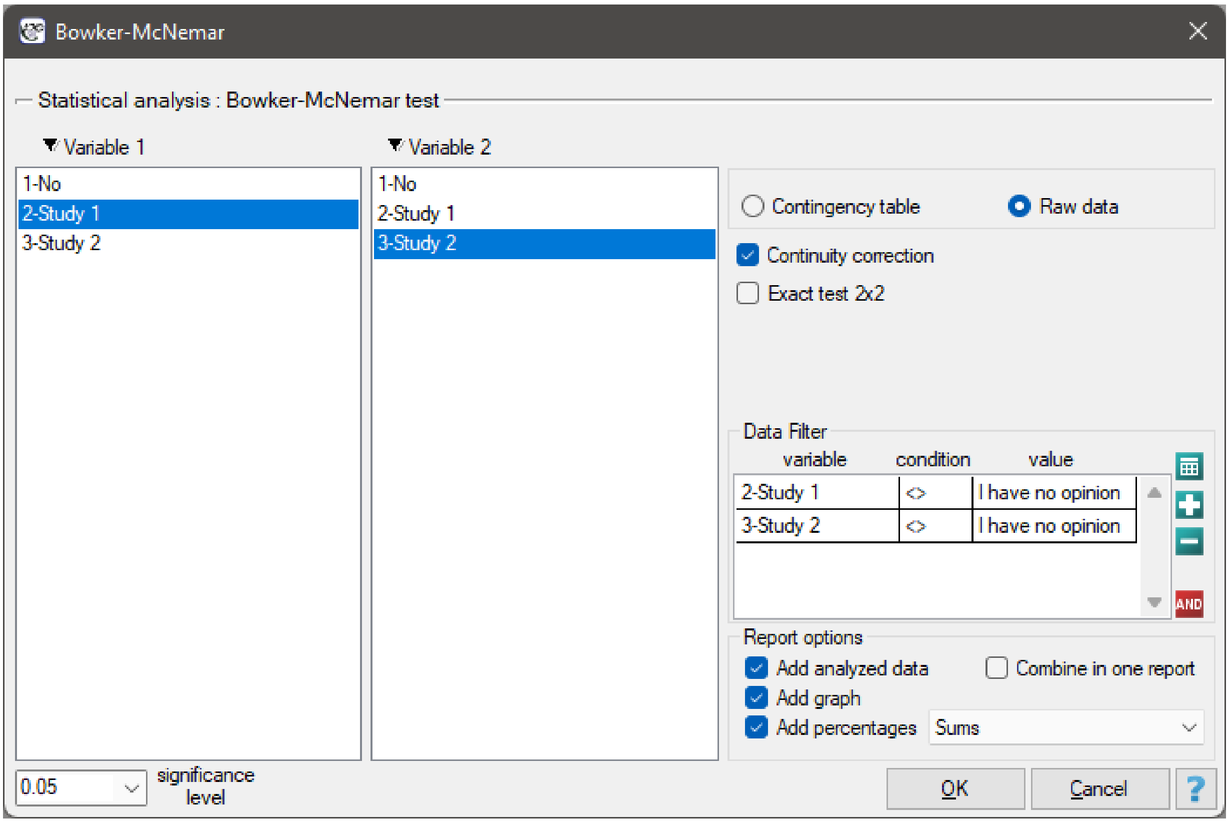
Confidence interval for the odds ratio is calculated on the base of the standard error:

$$SE=\sqrt{\frac{1}{O_{12}}+\frac{1}{O_{21}}}.$$

Note

Additionally, for small sample sizes, the exact range of the confidence interval for the Odds Ratio can be determined[100].

The settings window with the Bowker-McNemar test can be opened in Statistics menu → NonParametric tests → Bowker-McNemar or in Wizard.



Statistical analysis : Bowker-McNemar test

Variable 1: 1-No, **2-Study 1**, 3-Study 2

Variable 2: 1-No, 2-Study 1, **3-Study 2**

○ Contingency table ● Raw data

☑ Continuity correction

□ Exact test 2x2

Data Filter

variable	condition	value
2-Study 1	<>	I have no opinion
3-Study 2	<>	I have no opinion

AND

Report options

☑ Add analyzed data □ Combine in one report

☑ Add graph

☑ Add percentages Sums

0.05 significance level

The Bowker test of internal symmetry

The Bowker test of internal symmetry (Bowker (1948)[23]) is an extension of the McNemar test for 2 variables with more than 2 categories ($c>2$). It is used to verify the hypothesis determining the symmetry of 2 results of measurements executed twice $X^{(1)}$ and $X^{(2)}$ of X feature (symmetry of 2 dependent variables $X^{(1)}$ i $X^{(2)}$). An analysed feature may have more than 2 categories. The Bowker test of internal symmetry can be calculated on the basis of either raw data or a $c\times c$ contingency table.

Tabela 15.3. $c\times c$ contingency table for the observed frequencies of dependent variables

Observed frequencies		$X^{(2)}$				
O_{ij}		$X_1^{(2)}$	$X_2^{(2)}$	...	$X_c^{(2)}$	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:

$$\begin{array}{ll}\mathcal{H}_0: & O_{ij}=O_{ji},\\ \mathcal{H}_1: & O_{ij}\neq O_{ji} \text{ for at least one pair } O_{ij},O_{ji},\end{array}$$

where $j \neq i$, $j \in 1,2,...,c$, $i \in 1,2,...,c$, so O_{ij} and O_{ji} are the frequencies of the symmetrical pairs in the $c \times c$ table

The test statistic is defined by:

$$\chi^2=\sum_{i=1}^{c}\sum_{j>i}\frac{(O_{ij}-O_{ji})^2}{O_{ij}+O_{ji}}.$$

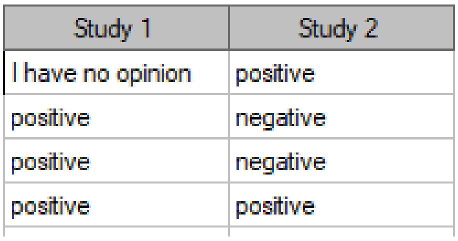
This statistic asymptotically (for large sample size) has the χ^2 distribution with a number of degrees of freedom calculated using the formula: $df=\frac{c(c-1)}{2}$.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

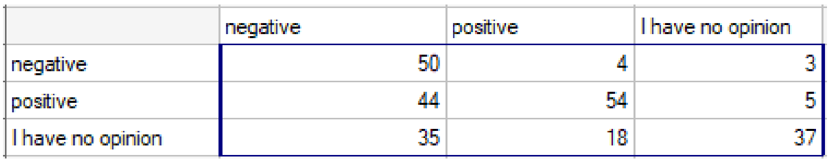
$$\begin{array}{ccl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

Example 15.12. (opinion.pqs file)

Two different surveys were carried out. They were supposed to analyse students' opinions about the particular academic professor. Both the surveys enabled students to give a positive opinion, a negative and a neutral one. Both surveys were carried out on the basis of the same sample of 250 students. But the first one was carried out the day before an exam done by the professor, and the other survey the day after the exam. There are some data below – in a form of raw rows, and all the data – in the form of a contingency table. Check, if both surveys give the similar results.



Study 1	Study 2
I have no opinion	positive
positive	negative
positive	negative
positive	positive



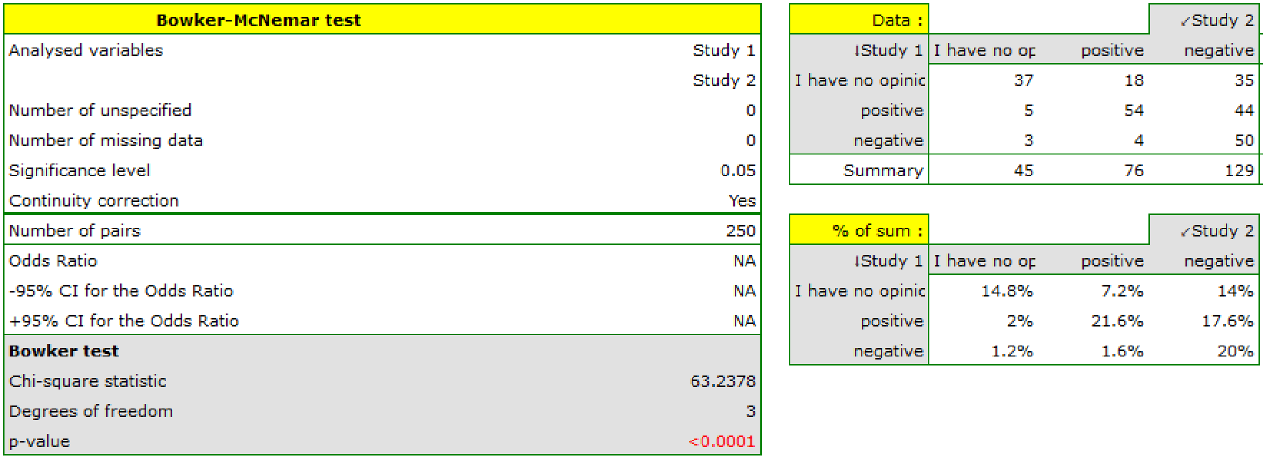
	negative	positive	I have no opinion
negative	50	4	3
positive	44	54	5
I have no opinion	35	18	37

Hypotheses:

$\mathcal{H}_0:$ the number of students, who changed their opinions is exactly the same for each of the possibile symmetric opinion changes,

$\mathcal{H}_1:$ the number of students, who changed their opinions is different for at least one of the possibile symmetric opinion changes,

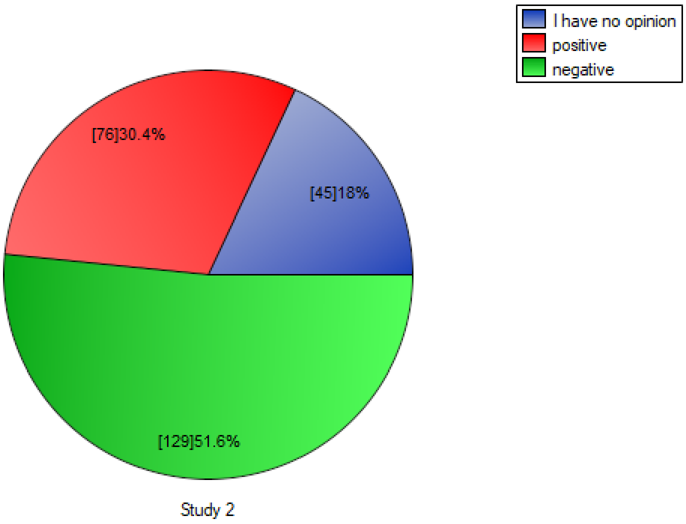
where, for example, changing the opinion from positive to negative one is symmetrical to changing the opinion from negative to positive one.

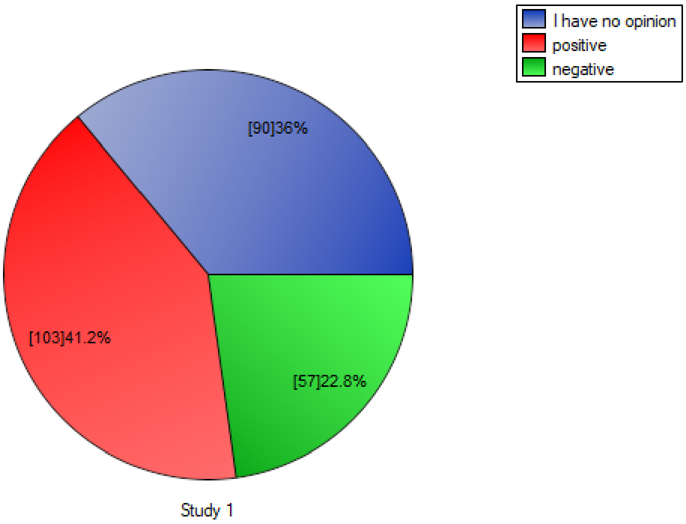


Bowker-McNemar test	
Analysed variables	Study 1
	Study 2
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Number of pairs	250
Odds Ratio	NA
-95% CI for the Odds Ratio	NA
+95% CI for the Odds Ratio	NA
Bowker test	
Chi-square statistic	63.2378
Degrees of 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%



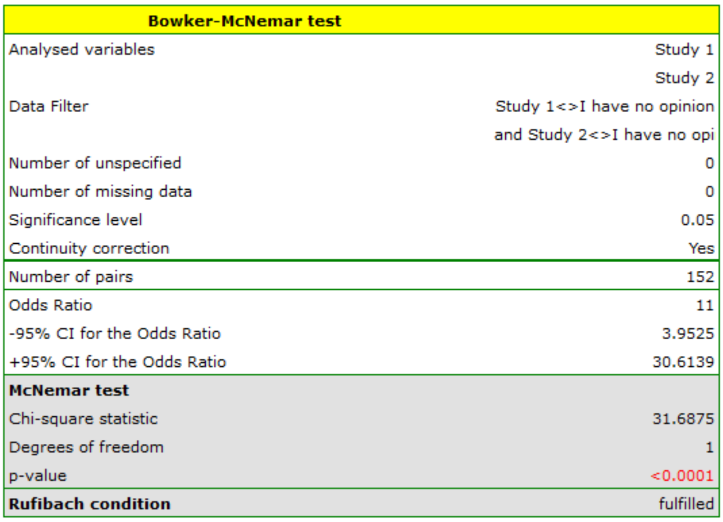


Comparing the p value for the Bowker test (p value < 0.0001) with the significance level $\alpha = 0.05$ it may be assumed that students changed their opinions. Looking at the table you can see that, there were more students who changed their opinions to negative ones after the exam, than those who changed it to positive ones after the exam. There were also students who did not evaluate the professor in the positive way after the exam any more.

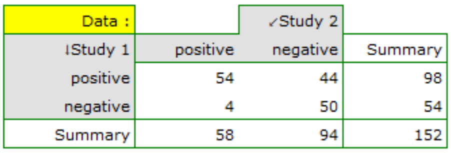
If you limit your analysis only to the people having clear opinions about the professor (positive or negative ones), you can use the McNemar test:

Hypotheses:

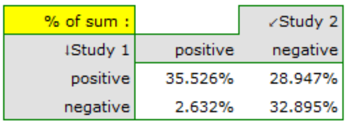
$\mathcal{H}_0$: the number of students, who changed their opinions from negative to positive ones is exactly the same as those, who changed their opinions from positive to negative,
$\mathcal{H}_1$: the number of students, who changed their opinions from negative to positive ones is different from those, who changed their opinions from positive to negative.



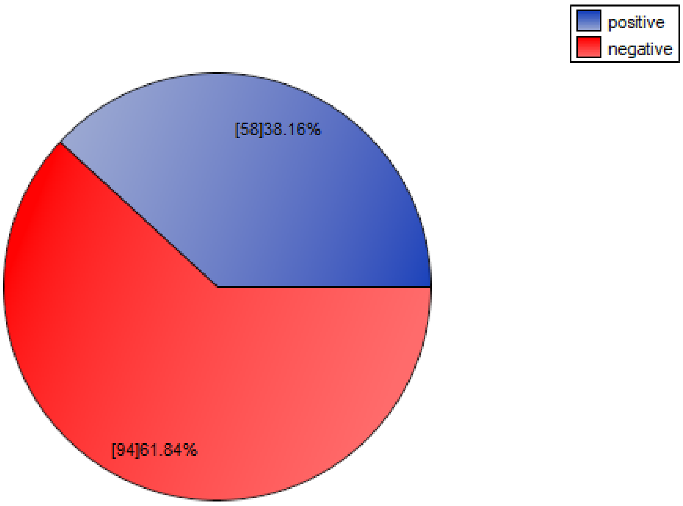
Bowker-McNemar test	
Analysed variables	Study 1
	Study 2
Data Filter	Study 1<>I have no opinion
	and Study 2<>I have no opi
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Continuity correction	Yes
Number of pairs	152
Odds Ratio	11
-95% CI for the Odds Ratio	3.9525
+95% CI for the Odds Ratio	30.6139
McNemar test	
Chi-square statistic	31.6875
Degrees of freedom	1
p-value	<0.0001
Rufibach condition	fulfilled



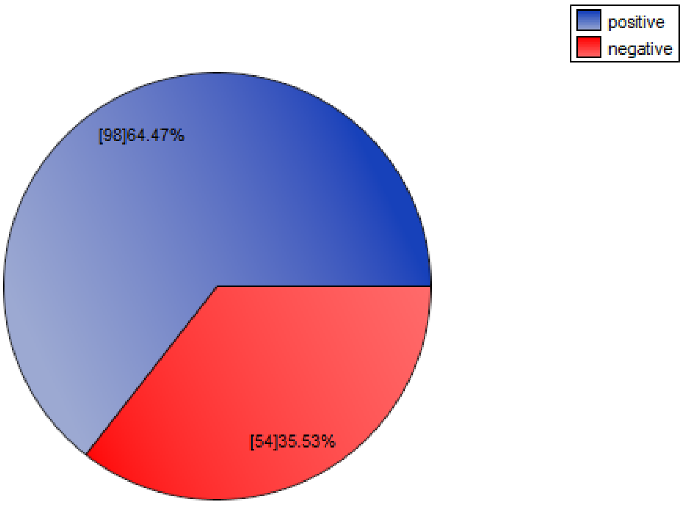
Data :		↙Study 2	
↓Study 1	positive	negative	Summary
positive	54	44	98
negative	4	50	54
Summary	58	94	152



% of sum :		↙Study 2
↓Study 1	positive	negative
positive	35.526%	28.947%
negative	2.632%	32.895%



Study 2



f you compare the p value, calculated for the McNemar test (p value < 0.0001), with the significance level $\alpha = 0.05$, you draw the conclusion that the students changed their opinions. There were much more students, who changed their opinions to negative ones after the exam, than those who changed their opinions to positive ones. The possibility of changing the opinion from positive (before the exam) to negative (after the exam) is eleven $\left(\frac{44}{4}\right)$ times greater than from negative to positive (the chance to change opinion in the opposite direction is: $\left(\frac{4}{44}\right)$).

15.2.12 The Z Test for two dependent proportions

Z Test for two dependent proportions is used in situations similar to the **McNemar's Test**, i.e. when we have 2 dependent groups of measurements ($X^{(1)}$ i $X^{(2)}$), in which we can obtain 2 possible results of the studied feature (**(+)(–)**).

Observed sizes O_{ij}		$X^{(2)}$		
		(+)	**(–)**	**Suma**
$X^{(1)}$	**(+)**	O_{11}	O_{12}	$O_{11}+O_{12}$
	(–)	O_{21}	O_{22}	$O_{21}+O_{22}$
	Sum	$O_{11}+O_{21}$	$O_{12}+O_{22}$	$n=O_{11}+O_{12}+O_{21}+O_{22}$

We can also calculated distinguished proportions for those groups $p_1 = \frac{O_{11}+O_{12}}{n}$ i $p_2 = \frac{O_{11}+O_{21}}{n}$. The test serves the purpose of verifying the hypothesis that the distinguished proportions P_1 and P_2 in the population from which the sample was drawn are equal.

Basic assumptions:

- measurement on the nominal - any order is not taken into account,
- dependent model,
- large sample size.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & P_1 - P_2 = 0,\\ \mathcal{H}_1: & P_1 - P_2 \neq 0,\end{array}$$

where:
P_1, P_2 fractions for the first and the second measurement.

The test statistic has the form presented below:

$$Z = \frac{p_1 - p_2}{\sqrt{O_{21}+O_{12}}} \cdot n,$$

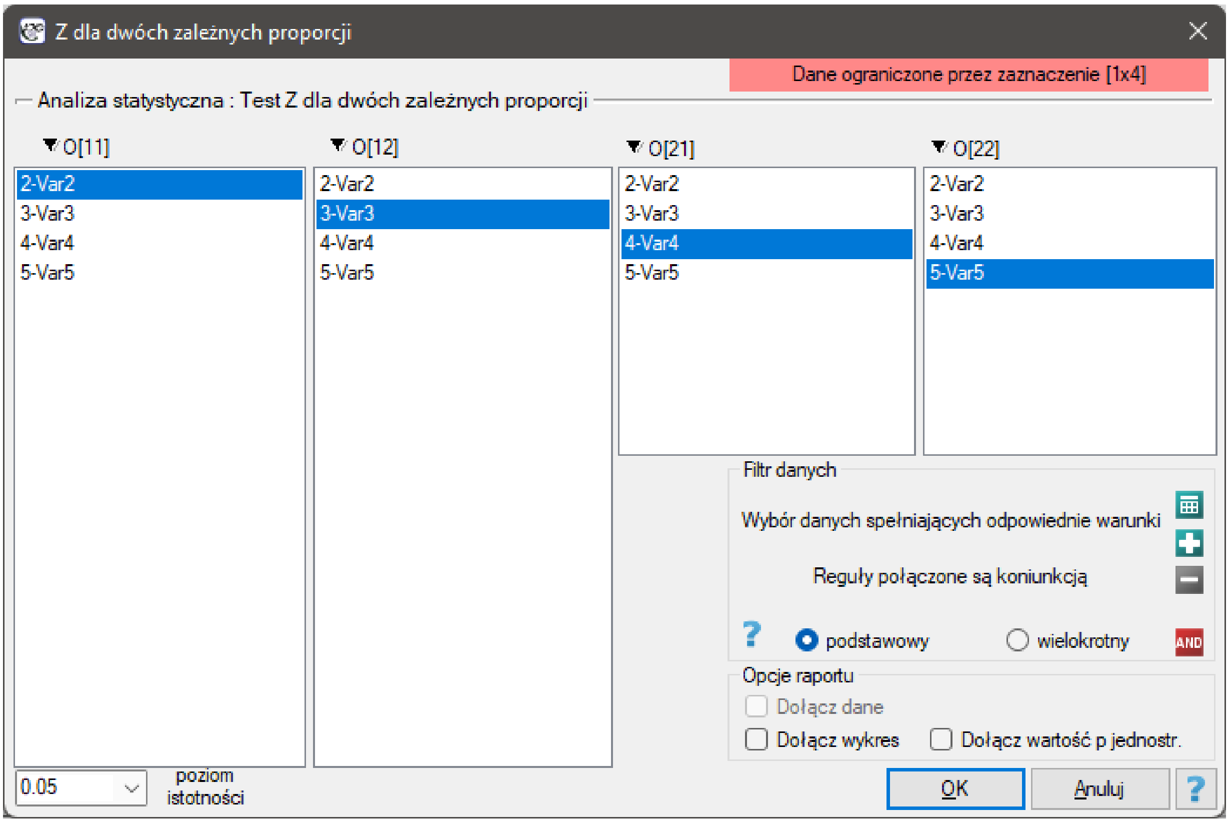
The Z Statistic asymptotically (for the large sample size) has the normal distribution.

On the basis of test statistics, p value is estimated and then compared with the significance level α:

$$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

Note
Confidence interval for the difference of two dependent proportions is estimated on the basis of the Newcombe-Wilson method.

The window with settings for Z-Test for two dependent proportions is accessed via the menu Statistics→Nonparametric tests→Z-Test for two dependent proportions.


Z dla dwóch zależnych proporcji

Dane ograniczone przez zaznaczenie [1x4]

Analiza statystyczna : Test Z dla dwóch zależnych proporcji

O[11]	O[12]	O[21]	O[22]
2-Var2	2-Var2	2-Var2	2-Var2
3-Var3	3-Var3	3-Var3	3-Var3
4-Var4	4-Var4	4-Var4	4-Var4
5-Var5	5-Var5	5-Var5	5-Var5

Selected: O[11] 2-Var2, O[12] 3-Var3, O[21] 4-Var4, O[22] 5-Var5

Filtr danych

Wybór danych spełniających odpowiednie warunki

Reguły połączone są koniunkcją

● podstawowy ○ wielokrotny

Opcje raportu

□ Dołącz dane

□ Dołącz wykres □ Dołącz wartość p jednostr.

0.05 poziom istotności

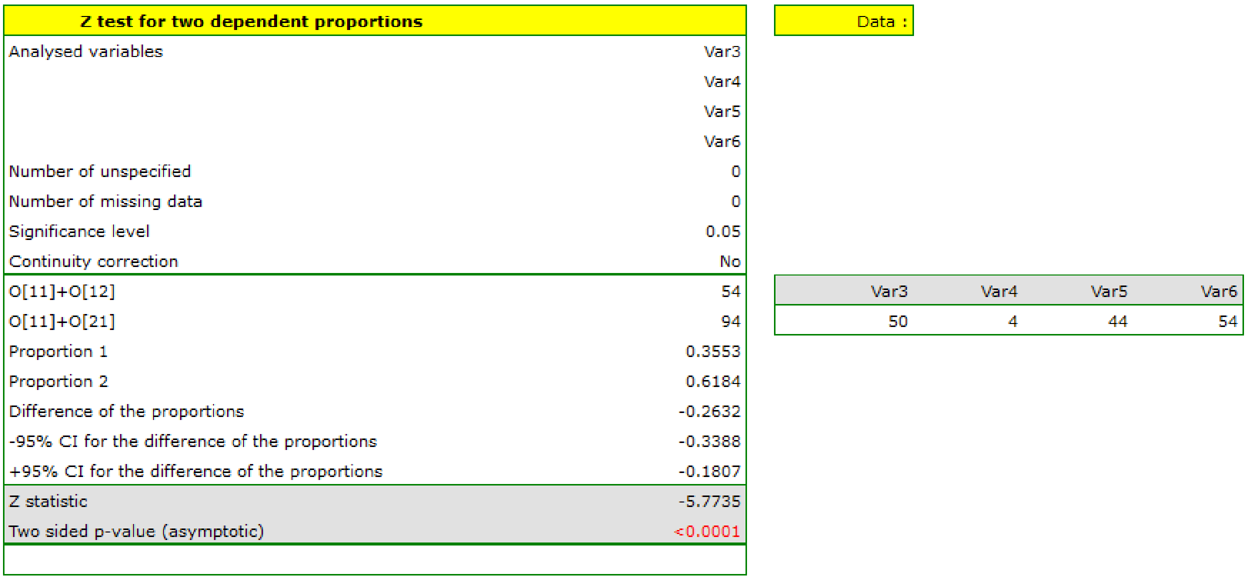
EXAMPLE (15.12) cont. *(opinion.pqs file)*

When we limit the study to people who have a specific opinion about the professor (i.e. those who only have a positive or a negative opinion) we will have 152 such students. The data for calculations are: $O_{11} = 50$, $O_{12} = 4$, $O_{21} = 44$, $O_{22} = 54$. We know that $\frac{50+4}{152} = 35.53\%$ students expressed a negative opinion before the exam. After the exam the percentage was $\frac{50+44}{152} = 61.84\%$.

Hypotheses:

$\mathcal{H}_0$: a lack of a difference between the number of negative evaluations of the professor before and after the exam,

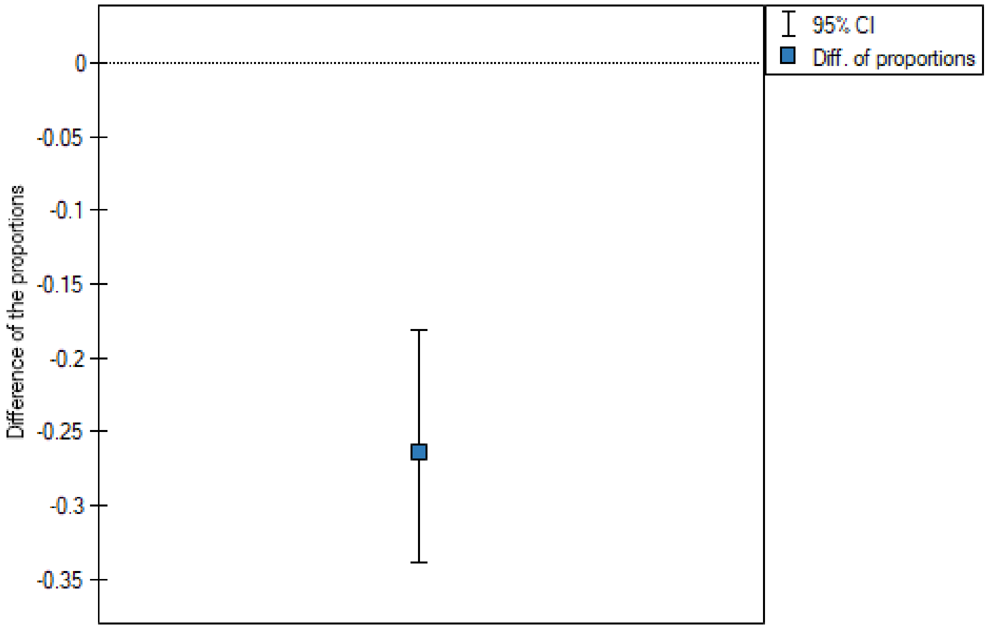
$\mathcal{H}_1$: there is a difference between the number of negative evaluations of the professor before and after the exam.



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 :

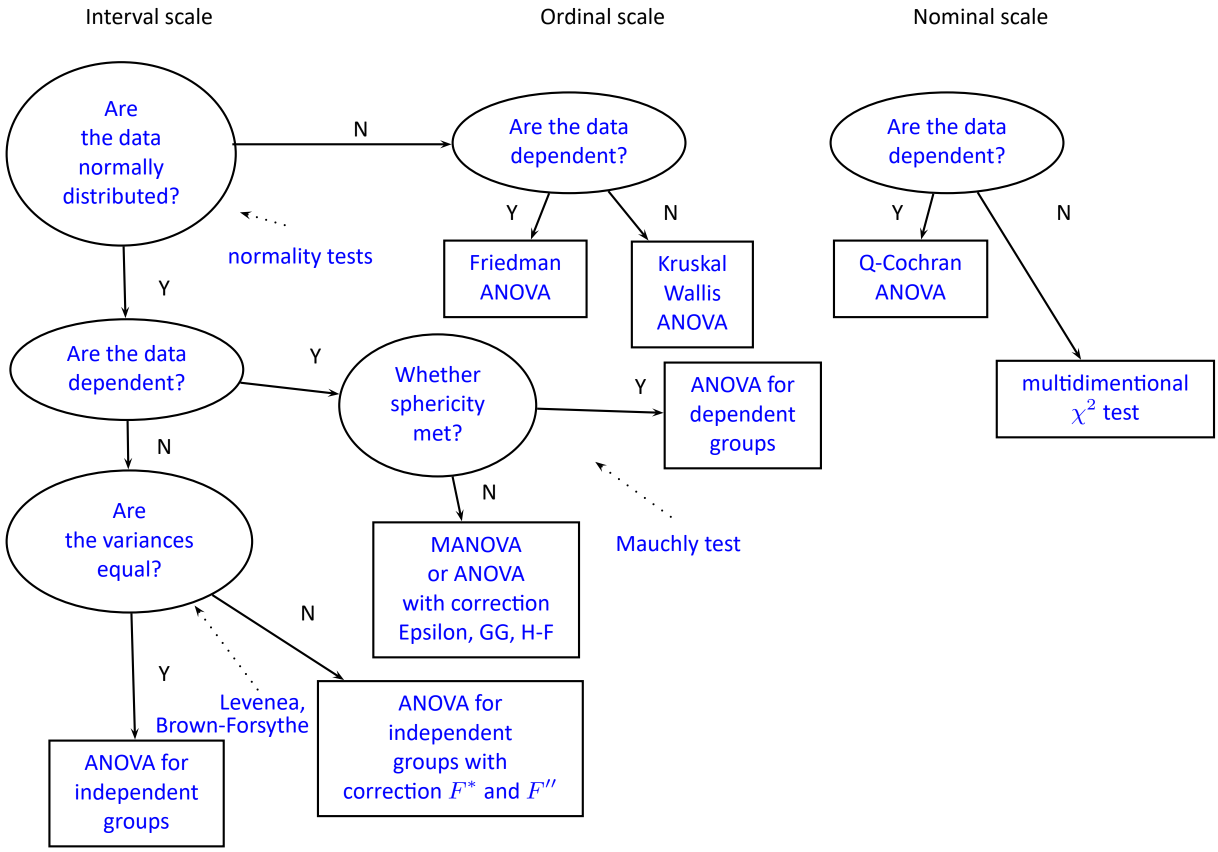
Var3	Var4	Var5	Var6
50	4	44	54



The difference in proportions distinguished in the sample is 26.32%, and the confidence interval of 95% for the sample $(18.07\%, 33.88\%)$ does not contain 0.

On the basis of a Z test (p=0.0001), on the significance level of α=0.05 (similarly to the case of McNemar's test) we accept the alternative hypothesis. Therefore, the proportion of negative evaluations before the exam differs from the proportion of negative evaluations after the exam. Indeed, after the exam there are more negative evaluations of the professor.

16 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.

16.1 PARAMETRIC TESTS

16.1.1 The ANOVA for independent groups

The one-way analysis of variance (ANOVA for independent groups) proposed by Ronald Fisher, is used to verify the hypothesis determining the equality of means of an analysed variable in several ($k \geq 2$) populations.

Basic assumptions:

- measurement on an interval scale,
- normality of distribution of an analysed feature in each population,
- an independent model,
- equality of variances of an analysed variable in all populations.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & \mu_1 = \mu_2 = ... = \mu_k,\\ \mathcal{H}_1: & \text{not all } \mu_j \text{ are equal } (j = 1, 2, ..., k),\end{array}$$

where:
μ_1,μ_2,...,μ_k – means of an analysed variable of each population.

The test statistic is defined by:

$$F = \frac{MS_{BG}}{MS_{WG}},$$

where:
$MS_{BG} = \frac{SS_{BG}}{df_{BG}}$ – mean square between-groups,

$MS_{WG} = \frac{SS_{WG}}{df_{WG}}$ – mean square within-groups,

$SS_{BG} = \sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N}$ – between-groups sum of squares,

$SS_{WG} = SS_T - SS_{BG}$ – within-groups sum of squares,

$SS_T = \left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2\right) - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N}$ – total sum of squares,

$df_{BG} = k - 1$ – between-groups degrees of freedom,
$df_{WG} = df_T - df_{BG}$ – within-groups degrees of freedom,
$df_T = N - 1$ – total degrees of freedom,
$N = \sum_{j=1}^{k} n_j$,
n_j – samples sizes for $(j = 1, 2, ...k)$,
x_{ij} – values of a variable taken from a sample for $(i = 1, 2, ...n_j)$, $(j = 1, 2, ...k)$.

The F statistic has the F Snedecor distribution with df_{BG} and df_{WG} degrees of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

Effect size - partial η^2

This quantity indicates the proportion of explained variance to total variance associated with a factor. Thus, in a one-factor ANOVA model for independent groups, it indicates what proportion of the between-groups variability in outcomes can be attributed to the factor under study determining the independent groups.

$$\eta^2 = \frac{SS_{BG}}{SS_{BG} + SS_{res}}$$

16.1.2 The contrasts and the POST-HOC tests

An analysis of the variance enables you to get information only if there are any significant differences among populations. It does not inform you which populations are different from each other. To gain some more detailed knowledge about the differences in particular parts of our complex structure, you should use **contrasts** (if you do the earlier planned and usually only particular comparisons), or the procedures of multiple comparisons **POST-HOC tests** (when having done the analysis of variance, we look for differences, usually between all the pairs).

The number of all the possible simple comparisons is calculated using the following formula:

$$c = \binom{k}{2} = \frac{k(k-1)}{2}$$

Hypotheses:

The first example - **simple comparisons** (comparison of 2 selected means):

$$\begin{array}{cl}\mathcal{H}_0: & \mu_1 = \mu_2, \\ \mathcal{H}_1: & \mu_1 \neq \mu_2.\end{array}$$

The second example - **complex comparisons** (comparison of combination of selected means):

$$\begin{array}{cl}\mathcal{H}_0: & \mu_1 = \frac{\mu_2+\mu_3}{2}, \\ \mathcal{H}_1: & \mu_1 \neq \frac{\mu_2+\mu_3}{2}.\end{array}$$

If you want to define the selected hypothesis you should ascribe the contrast value c_j, $(j = 1, 2, ...k)$ to each mean. The c_j values are selected, so that their sums of compared sides are the opposite numbers, and their values of means which are not analysed count to 0.

The first example: $c_1 = 1$, $c_2 = -1$, $c_3 = 0, ...c_k = 0.$

The second example: $c_1 = 2$, $c_2 = -1$, $c_3 = -1$, $c_4 = 0$,..., $c_k = 0.$

How to choose the proper hypothesis:

(i) Comparing the differences between the selected means with the **critical difference (CD)** calculated using the proper POST-HOC test:

$$\begin{array}{lcl}\text{if the differences between means} \ge CD & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if the differences between means} < CD & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

(ii) Comparing the p value, designated on the basis of the test statistic of the proper POST-HOC test, with the significance level α:

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The LSD Fisher test

For simple and complex comparisons, equal-size groups as well as unequal-size groups, when the variances are equal.

(i) The value of critical difference is calculated by using the following 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, adequately: 1 and df_{WG}.

(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 are equal.

(i) The value of a critical difference is calculated by using the following 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 are equal.

(i) The value of a critical difference is calculated by using the following 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.

(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 the Lund works (1983)[105]. Other applications or web pages may calculate a little bit different values than PQStat, because they may be based on less precised 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 ANOVA for independent variables, because it is based on the same assumptions, but it captures the alternative hypothesis differently - indicating in it the existence of a trend in the mean values for successive populations. The analysis of the trend in the arrangement of means is based on contrasts Fisher LSD. By building appropriate contrasts, you can study any type of trend such as linear, quadratic, cubic, etc. Below is a table of sample contrast values for selected trends.

		Contrast									
Number of groups	Trends	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
3	line	-1	0	1							
	quadratic	1	-2	1							
4	line	-3	-1	1	3						
	quadratic	1	-1	-1	1						
	cubic	-1	3	-3	1						
5	line	-2	-1	0	1	2					
	quadratic	2	-1	-2	-1	2					
	cubic	-1	2	0	-2	1					
6	line	-5	-3	-1	1	3	5				
	quadratic	5	-1	-4	-4	-1	5				
	cubic	-5	7	4	-4	-7	5				
7	line	-3	-2	-1	0	1	2	3			
	quadratic	5	0	-3	-4	-3	0	5			
	cubic	-1	1	1	0	-1	-1	1			
8	line	-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	line	-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	line	-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

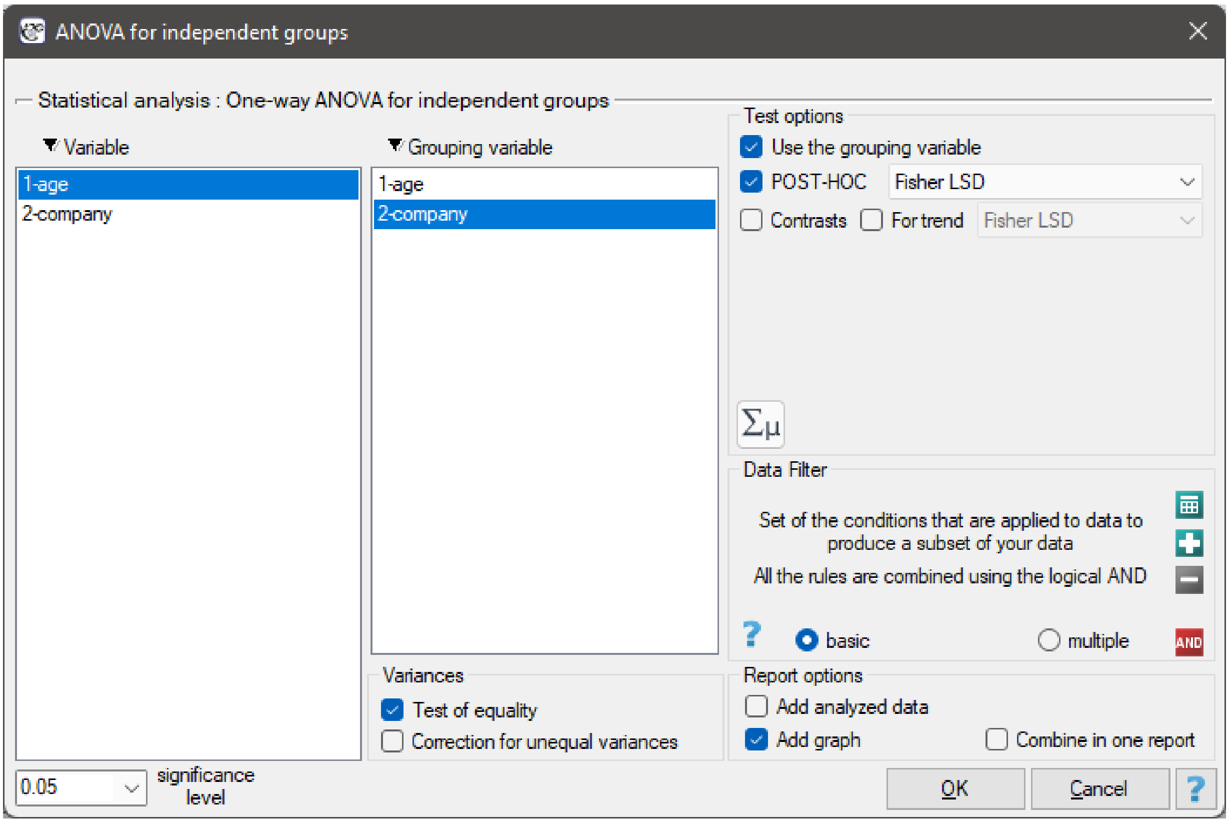
A linear trend, like other trends, can be analyzed by entering the appropriate contrast values. However, if the direction of the linear trend is known, simply use the For trend option and indicate the expected order of the populations by assigning them consecutive natural numbers.

The analysis is performed on the basis of linear contrast, i.e. the groups indicated according to the natural order are assigned appropriate contrast values and the statistics are calculated Fisher LSD .

With the expected direction of the trend known, the alternative hypothesis is one-sided and the one-sided p-value is interpreted. The interpretation of the two-sided p-value means that the researcher does not know (does not assume) the direction of the possible trend.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

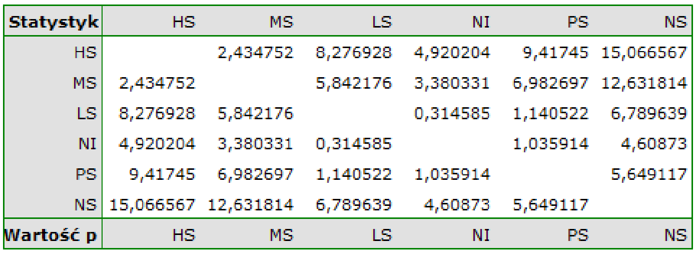
$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the One-way ANOVA for independent groups can be opened in Statistics menu→Parametric tests→ANOVA for independent groups or in Wizard.


ANOVA for independent groups
Statistical analysis : One-way ANOVA for independent groups
Variable
1-age
2-company
Grouping variable
1-age
2-company
Test options
☑ Use the grouping variable
☑ POST-HOC Fisher LSD
□ Contrasts □ For trend Fisher LSD
Σμ
Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple
AND
Variances
☑ Test of equality
□ Correction for unequal variances
Report options
□ Add analyzed data
☑ Add graph □ Combine in one report
0.05 significance level
OK
Cancel
?

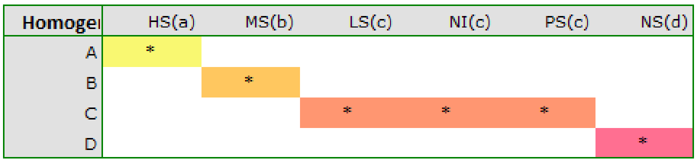
Homogeneous groups.

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 the expiratory parameters for them. In our ANOVA we obtained statistically significant differences in exhalation parameters between the tested groups. In order 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 statistical significance in the form of p:



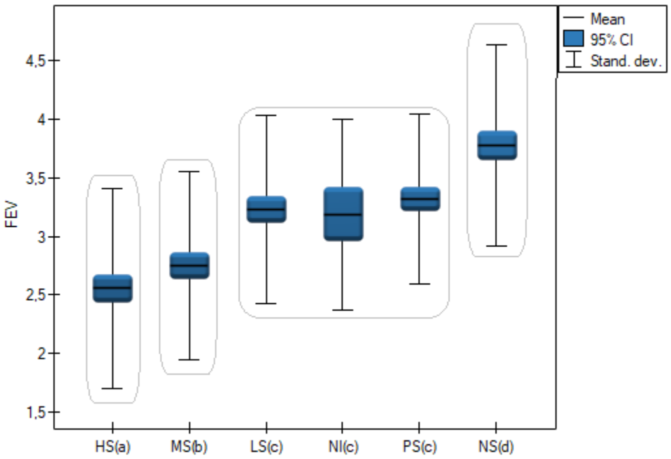
Statystyk	HS	MS	LS	NI	PS	NS
HS		2,434752	8,276928	4,920204	9,41745	15,066567
MS	2,434752		5,842176	3,380331	6,982697	12,631814
LS	8,276928	5,842176		0,314585	1,140522	6,789639
NI	4,920204	3,380331	0,314585		1,035914	4,60873
PS	9,41745	6,982697	1,140522	1,035914		5,649117
NS	15,066567	12,631814	6,789639	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 on the basis of a smaller division, i.e. instead of the six groups we studied originally, further analyses can be conducted on the basis of the four homogeneous groups determined here. The order of groups was determined on the basis of weighted averages calculated for particular homogeneous groups in such a way, that letter A was assigned to the group with the lowest weighted average, and further letters of the alphabet to groups with increasingly higher averages.



Example 16.1. (age ANOVA.pqs file)

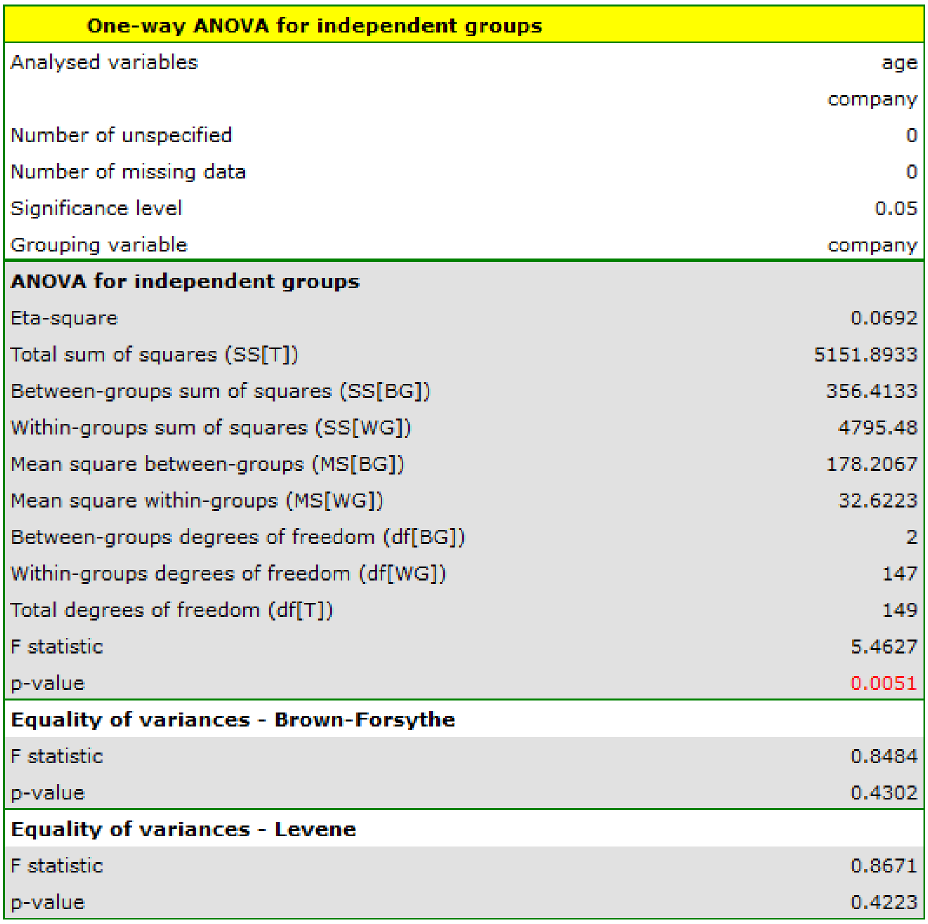
There are 150 persons chosen randomly from the population of workers of 3 different transport companies. From each company there are 50 persons drawn to the sample. Before the experiment begins, you should check if the average age of the workers of these companies is similar, because the next step of the experiment depends on it. The age of each participant is written 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, the normality of the data distribution was confirmed.

The analysis window tested the assumption of equality of variance, obtaining p>0.05 in both tests.

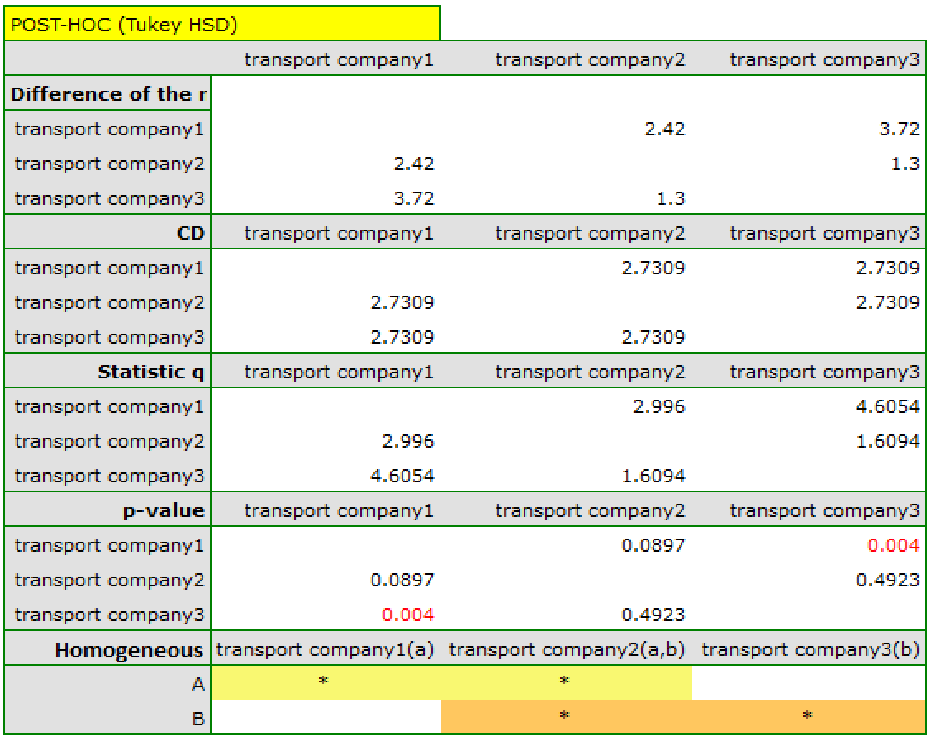
Hypotheses:

$\mathcal{H}_0$: the average age of the workers off all the analysed transport companies is the same,
$\mathcal{H}_1$: at least 2 means are different.



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

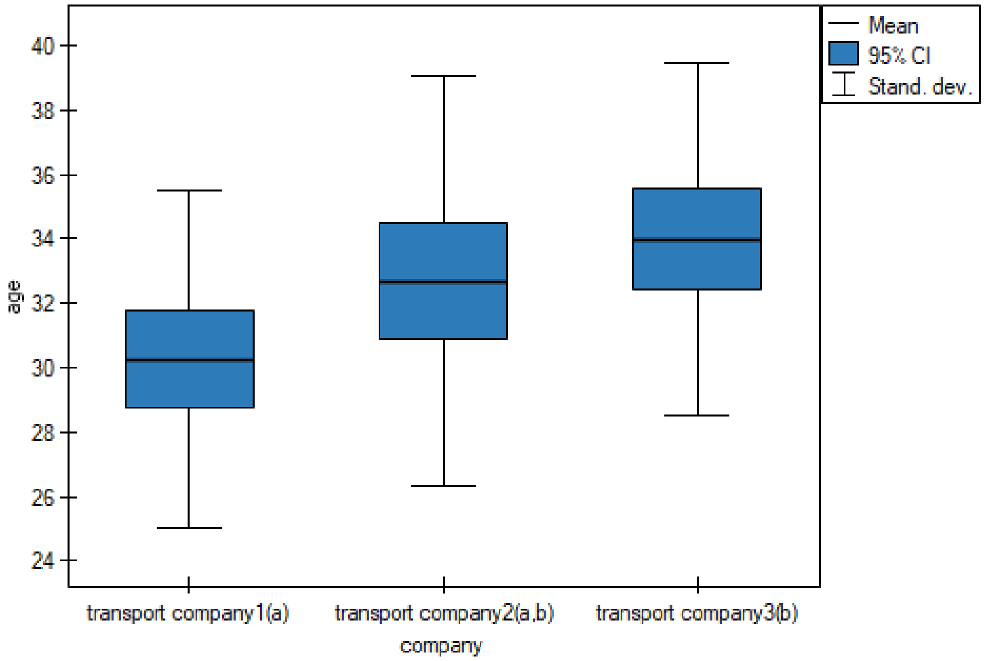
Comparing the p value = 0.005147 of the one-way analysis of variance with the significance level $\alpha = 0.05$, you can draw the conclusion that the average ages of workers of these transport companies is not the same. Based just on the ANOVA result, you do not know precisely which groups differ from others in terms of age. To gain such knowledge, it must be used one of the POST-HOC tests, for example the Tukey test. To do this, you should resume the analysis by clicking Run the recent test and then, in the options window for the test, you should select Tukey HSD and Add graph.



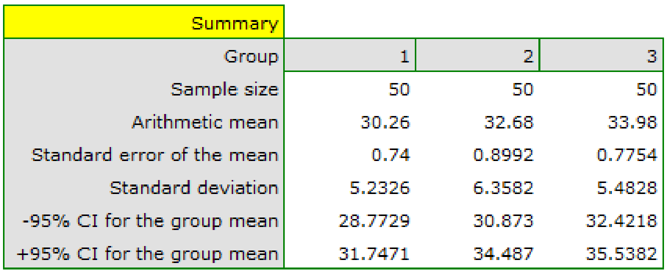
POST-HOC (Tukey HSD)

	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		*	*

The critical difference (CD) calculated for each pair of comparisons is the same (because the groups sizes are equal) and counts to 2.730855. The comparison of the CD value with the value of the mean difference indicates, that there are significant differences only between the mean age of the workers from the first and the third transport company (only if these 2 groups are compared, the CD value is less than the difference of the means). The same conclusion you draw, if you compare the p value of POST-HOC test with the significance level $\alpha = 0.05$. The workers of the first transport company are about 3 years younger (on average) than the workers of the third transport company. Two interlocking homogeneous groups were obtained, which are also marked on the graph.



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

16.1.3 The ANOVA for independent groups with F^* and F'' corrections

F^* (Brown-Forsythe, 1974[32]) and F'' (Welch, 1951[166]) Corrections concern ANOVA for independent groups and are calculated when the assumption of equality of variances is not met.
The test statistic is in the form of:

$$F^* = \frac{SS_{BG}}{\sum_{j=1}^{k}\left(1-\frac{n_j}{n}sd_j^2\right)},$$

$$F'' = \frac{\frac{\sum_{j=1}^{k} w_j(\overline{x}_j-\widetilde{x})}{k-1}}{1+\frac{2(k-2)}{k^2-1}\sum_{j=1}^{k} h_j},$$

where:
sd_j – group standard deviation j,
$w_j = \frac{n_j}{sd_j^2}$ – group weight j,
$\widetilde{x}$ – weighted mean,
$h_j = \frac{\left(1-\frac{w_j}{\sum_{j=1}^{k} w_j}\right)^2}{n_j-1}$.

This statistic is subject toSnedecor's F distribution with $k-1$ and adjusted df_{WG_k} degrees of freedom.

The p value, determined on the basis of the test statistics, is comparde with the significance level α :

$$\begin{array}{ccl} \text{jeżeli } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ adopting } \mathcal{H}_1, \\ \text{jeżeli } p > \alpha & \Longrightarrow & \text{there is no basis to reject } \mathcal{H}_0. \end{array}$$

POST-HOC Tests

Introduction to the **contrasts and POST-HOC tests** was done in chapter 16.1.2 concerning one-way analysis of variance.

T2 Tamhane test

For simple and complex comparisons, equal-size groups as well as unequal-size groups, when the variances differ significantly (Tamhane A. C., 1977[155]).

(i) The value of critical difference is calculated by using the following 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 (statistics) of the Snedecor's F distribution for 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 in the form of:

$$t = \frac{\left(\sum_{j=1}^k c_j\overline{x}_j\right)^2}{\sqrt{\left(\sum_{j=1}^k \frac{c_j^2 sd_j^2}{n_j}\right)}}.$$

This statistic is subject to the t-Student distribution with df_v degrees of freedom, and p value is adjusted by the number of possible simple comparisons.

BF test (Brown-Forsythe)

For simple and complex comparisons, equal-size groups as well as unequal-size groups, when the variances differ significantly (Brown M. B. i Forsythe A. B. (1974)[31]).

(i) The value of critical difference is calculated by using the following 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 (statistics) of the Snedecor's F distribution for a given significance level α as well as $k-1$ and df_v degrees of freedom.

(ii) The test statistic is in the form of:

$$F = \frac{\left(\sum_{j=1}^k c_j\overline{x}_j\right)^2}{(k-1)\left(\sum_{j=1}^k \frac{c_j^2 sd_j^2}{n_j}\right)}.$$

This statistic is subject to Snedecor's F distribution with $k-1$ and df_v degrees of freedom.

GH test (Games-Howell).

For simple and complex comparisons, equal-size groups as well as unequal-size groups, when the variances differ significantly (Games P. A. i Howell J. F. 1976[65]).

(i) The value of critical difference is calculated by using the following formula:

$$CD = \frac{q_{\alpha,k,df_v} \cdot \sqrt{\left(\sum_{j=1}^k \frac{c_j^2 sd_j^2}{n_j}\right)}}{\sqrt{2}},$$

gdzie:
q_{α,k,df_v} - is the critical value (statistics) of the the distribution of the studentised interval for a given significance level α as well as k and df_v degrees of freedom.

(ii) The test statistic is in the form of:

$$q=\sqrt{2}\frac{\sum_{j=1}^{k}c_j\overline{x}_j}{\sqrt{\left(\sum_{j=1}^{k}\frac{c_j^2sd_j^2}{n_j}\right)}}.$$

This statistic follows a studenty distribution with k and df_v degrees of freedom.

Trend test.

The test examining the presence of a trend can be calculated in the same situation as ANOVA for independent groups with correction F^* and F'', because it is based on the same assumptions, however, differently captures the alternative hypothesis - indicating the existence of a trend in the mean values for successive populations. The analysis of the trend of the arrangement of means is based on contrasts (T2 Tamhane). By creating appropriate contrasts you can study any type of trend e.g. linear, quadratic, cubic, etc. A table of sample contrast values for certain trends can be found in the description trend test for Ona-Way ANOVA.

Linear trend

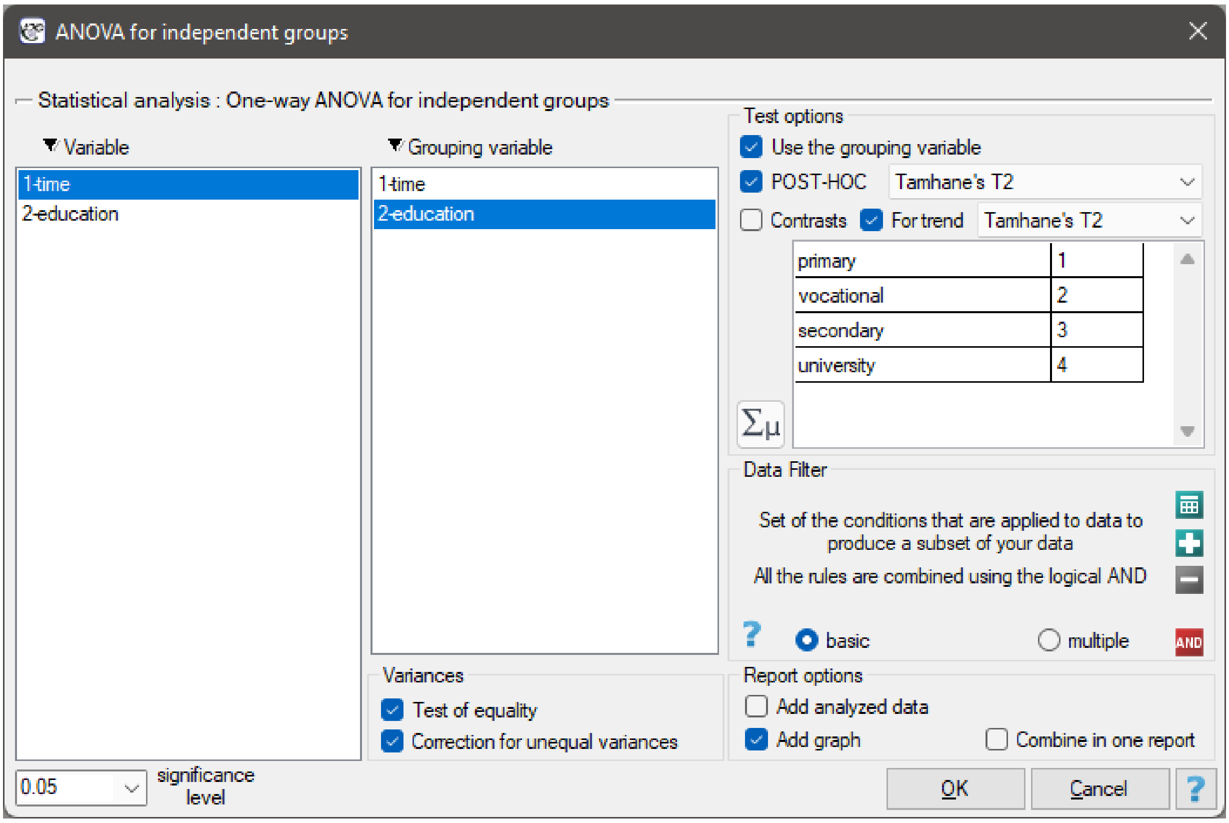
A linear trend, like other trends, can be analyzed by entering the appropriate contrast values. However, if the direction of the linear trend is known, simply use the Linear Trend option and indicate the expected order of the populations by assigning them consecutive natural numbers.

The analysis is performed based on linear contrast, i.e., the groups indicated according to the natural ordering are assigned appropriate contrast values and the T2 Tamhane statistic is calculated.

With the expected direction of the trend being known, the alternative hypothesis is one-sided and the one-sided value of p is subject to interpretation. The interpretation of the two-sided value of p means that the researcher does not know (does not assume) the direction of the possible trend.
The test statistic determined from the test statistic p value is compared with α :

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ adopting } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there are no grounds to reject } \mathcal{H}_0. \end{array}$$

Settings window for the One-way ANOVA for independent groups with F^* and F'' adjustments is opened via menu Statistics→Parametric tests→ANOVA for independent groups or via the Wizard.



ANOVA for independent groups

Statistical analysis : One-way ANOVA for independent groups

Variable: 1-time (selected), 2-education

Grouping variable: 1-time, 2-education (selected)

Test options

☑ Use the grouping variable

☑ POST-HOC Tamhane's T2

□ Contrasts ☑ For trend Tamhane's T2

primary	1
vocational	2
secondary	3
university	4

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Variances

☑ Test of equality

☑ Correction for unequal variances

Report options

□ Add analyzed data

☑ Add graph □ Combine in one report

0.05 significance level

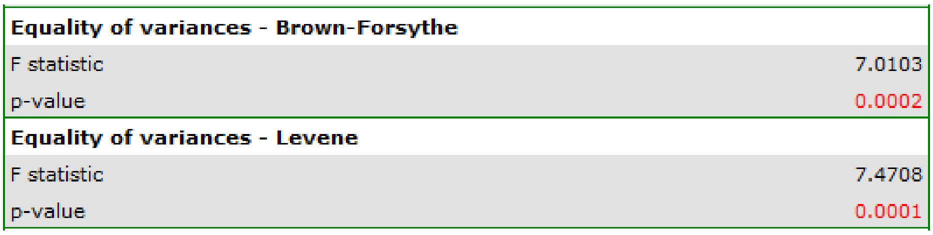
EXAMPLE 16.2. (unemployment.pqs file)

There are many factors that control the time 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) of unemployment are gathered in the file. We want to see if there are differences in average job search time for different education categories.

Hypotheses:

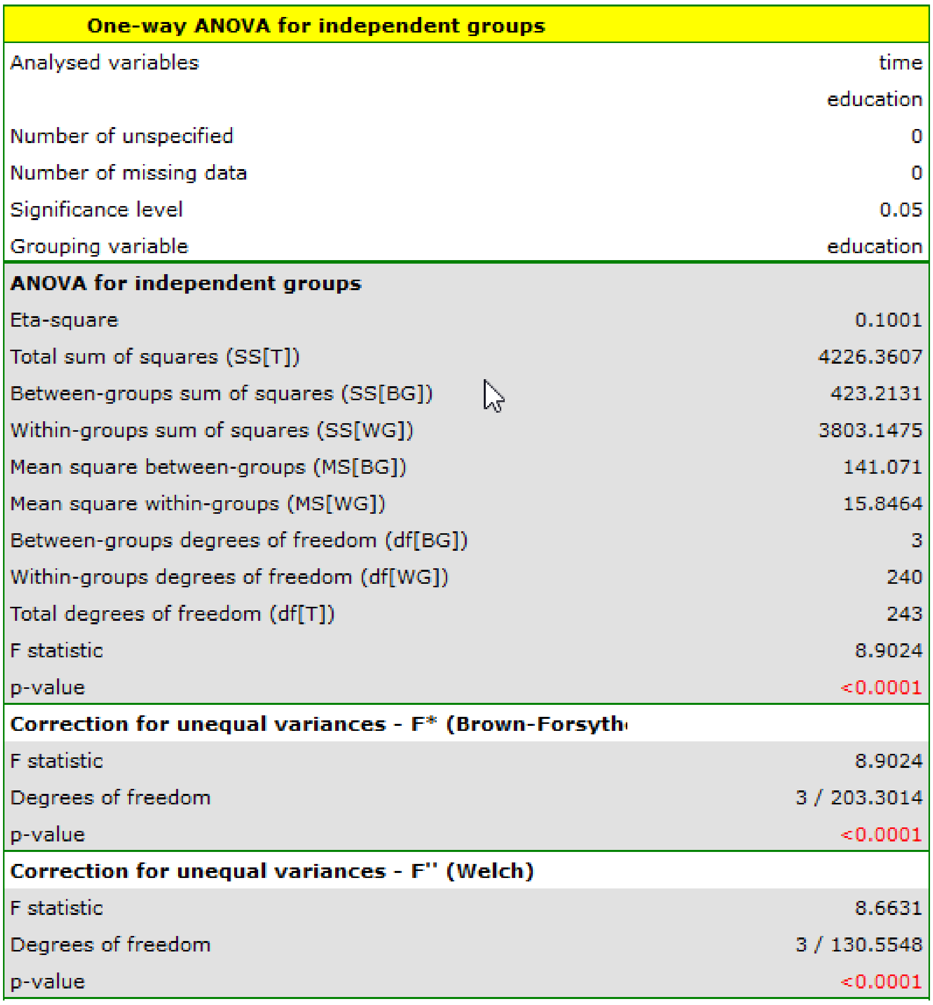
$\mathcal{H}_0$: average job search time is the same
for every category of education,
$\mathcal{H}_1$: at least one education category (one population)
have a different average job search time.

Due to differences in variance between populations (for Levene test $p = 0.0001$ and for Brown-Forsythe test $p = 0.0002$):



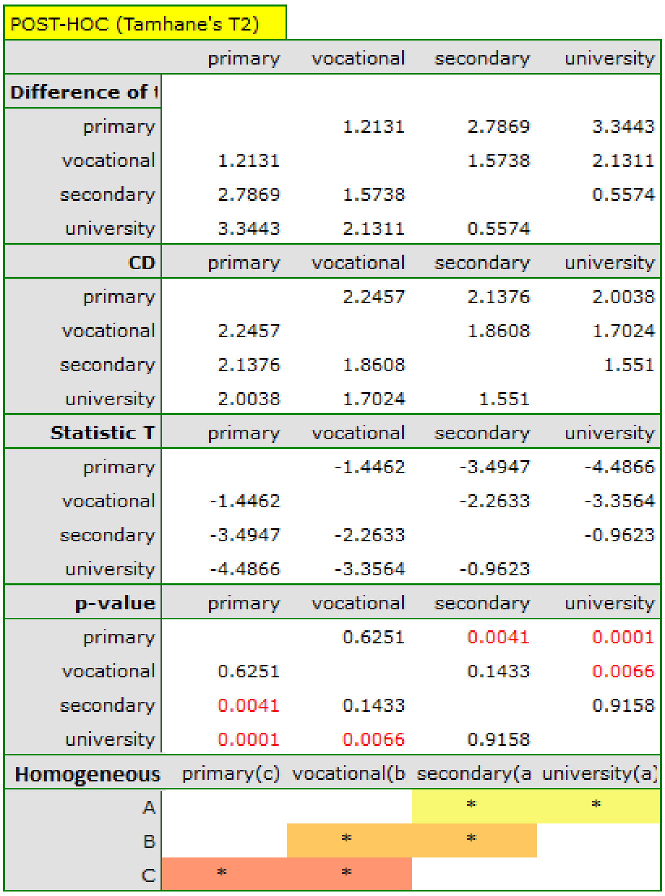
Equality of variances - Brown-Forsythe	
F statistic	7.0103
p-value	0.0002
Equality of variances - Levene	
F statistic	7.4708
p-value	0.0001

the analysis is performed with the correction of various variances enabled. The obtained result of the 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

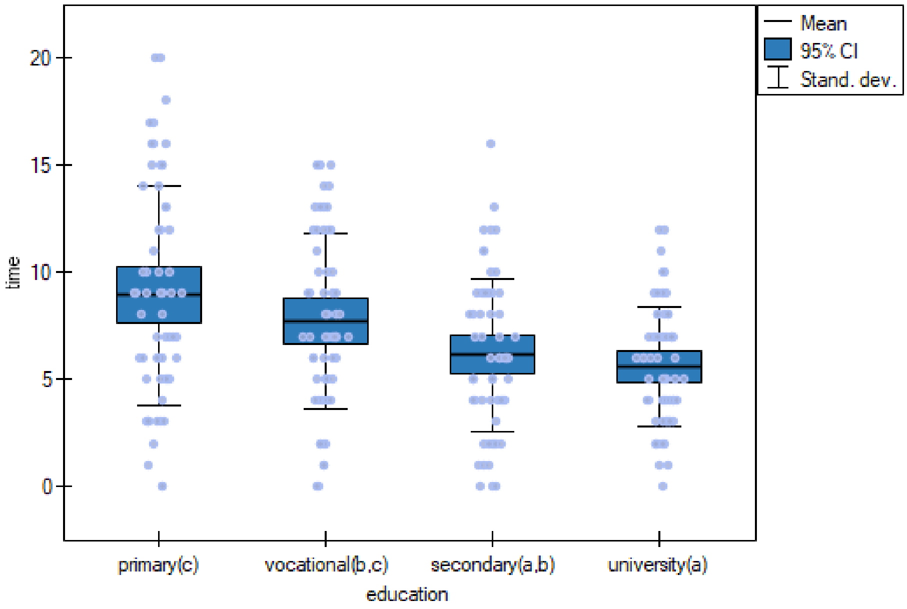
Comparing $p < 0.0001$ (for the F^* test) and $p < 0.0001$ (for the F'' test) with a significance level of $\alpha = 0.05$, we find that the average job search time differs depending on the education one has. By performing one of the POST-HOC tests, designed to compare groups with different variances, we find out which education categories are affected by the differences found:



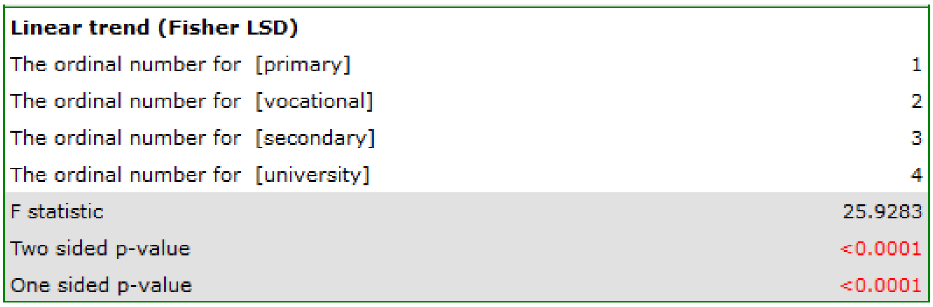
POST-HOC (Tamhane's T2)

	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 least significant difference (LSD) 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 LSD value to the resulting differences in mean values yields the same result as comparing the p value with a significance level of $\alpha = 0.05$. The differences are between primary and higher education, primary and secondary education, and vocational and higher education. The resulting homogeneous groups overlap. In general, however, looking at the graph, we might expect that the more educated a person is, the less time it takes them to find a job.



In order to test the stated hypothesis, it is necessary to perform the trend analysis. To do so, we reopen the analysis with the Run the recent test button and, in the test options window, select: the Tamhane's T2 method, the Contrasts option (and set the appropriate contrast), or the For trend option (and indicate the order of education categories by specifying 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 correlation between education and job search time is known to us, we use a one-sided or two-sided p value. Both of these values are less than the given significance level. The trend we predicted is confirmed, that is, at a significance level of $\alpha = 0.05$ we can say that this trend does indeed exist in the population from which the sample is drawn.

16.1.4 The Brown-Forsythe test and the Levene test

Both tests: the **Levene test** (Levene, 1960 [99]) and the **Brown-Forsythe test** (Brown and Forsythe, 1974 [30]) are used to verify the hypothesis determining the equality of variance of an analysed variable in several ($k >= 2$) populations.

Basic assumptions:

- measurement on an interval scale,
- normality of distribution of an analysed feature in each population,
- an independent model.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2, \\ \mathcal{H}_1: & \text{not all } \sigma_j^2 \text{ are equal } (j = 1, 2, ..., k), \end{array}$$

where:
σ_1^2,σ_2^2,...,σ_k^2 $-$ variances of an analysed variable of each population.

The analysis is based on calculating the absolute deviation of 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 are under the same procedure performed to the analysis of variance for independent groups. Hence, the test statistic is defined by:

$$F = \frac{MS_{BG}}{MS_{WG}},$$

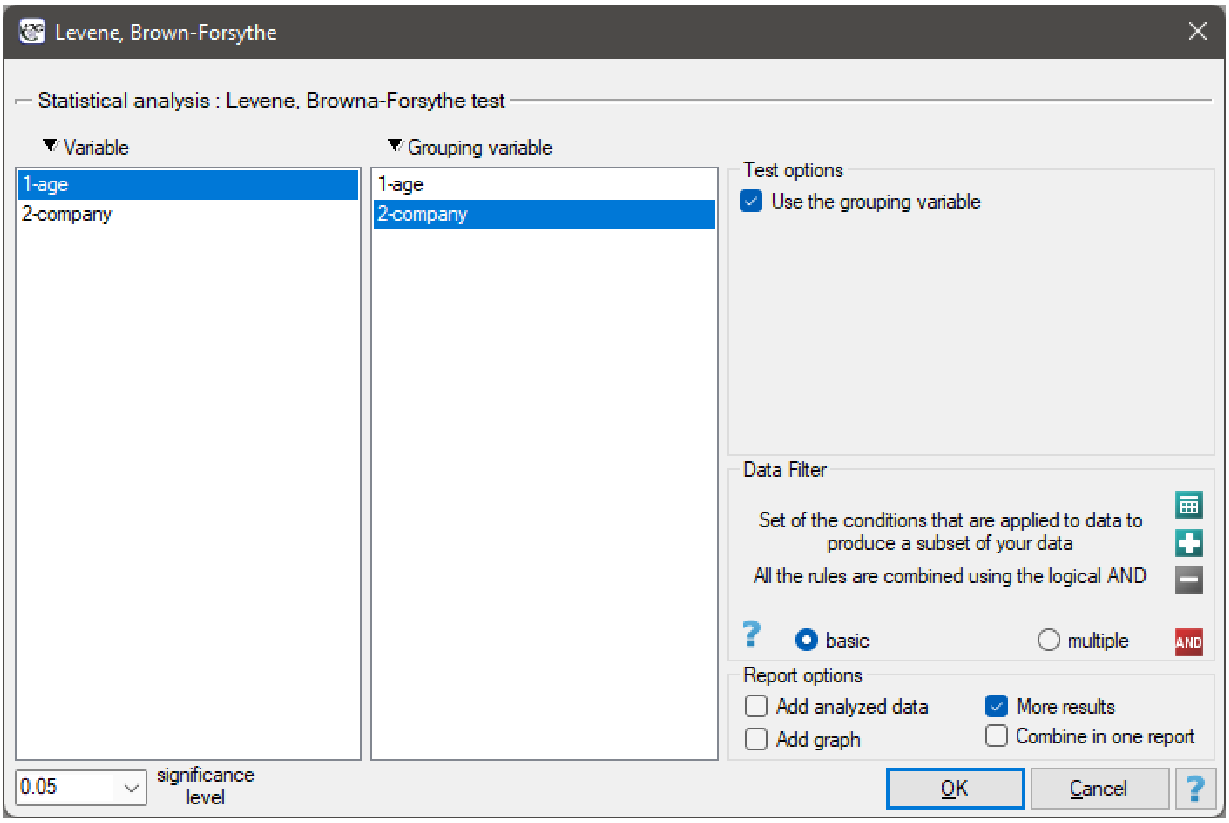
The test statistic has the F Snedecor distribution with df_{BG} and df_{WG} degrees of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note
The Brown-Forsythe test is less sensitive than the Levene test, in terms of an unfulfilled assumption relating to distribution normality.

The settings window with the Levene, Brown-Forsythe tests' can be opened in Statistics menu→Parametric tests→Levene, Brown-Forsythe.



Statistical analysis : Levene, Browna-Forsythe test

Variable: 1-age (selected), 2-company

Grouping variable: 1-age, 2-company (selected)

Test options
☑ Use the grouping variable

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
☐ Add analyzed data ☑ More results
☐ Add graph ☐ Combine in one report

0.05 significance level

16.1.5 The ANOVA for dependent groups

The single-factor repeated-measures analysis of variance (ANOVA for dependent groups) is used when the measurements of an analysed variable are made several times ($k \geq 2$) each time in different conditions (but we need to assume that the variances of the differences between all the pairs of measurements are pretty close to each other).

This test is used to verify the hypothesis determining the equality of means of an analysed variable in several ($k \geq 2$) populations.

Basic assumptions:

- measurement on an interval scale,
- the normal distribution for all variables which are the differences of measurement pairs (or the normal distribution for an analysed variable in each measurement),
- a dependent model.

Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & \mu_1=\mu_2=...=\mu_k,\\ \mathcal{H}_1: & \text{not all } \mu_j \text{ are equal } (j=1,2,...,k),\end{array}$$

where:
μ_1,μ_2,...,μ_k – means for an analysed features, in the following measurements from the examined population.

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,
$SS_{BC}=\displaystyle\sum_{j=1}^{k}\left(\frac{\left(\sum_{i=1}^{n}x_{ij}\right)^2}{n}\right)-\frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n}x_{ij}\right)^2}{N}$ – between-conditions sum of squares,
$SS_{res}=SS_T-SS_{BS}-SS_{BC}$ – residual sum of squares,
$SS_T=\displaystyle\left(\sum_{j=1}^{k}\sum_{i=1}^{n}x_{ij}^2\right)-\frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n}x_{ij}\right)^2}{N}$ – total sum of squares,
$SS_{BS}=\displaystyle\sum_{i=1}^{n}\left(\frac{\left(\sum_{j=1}^{k}x_{ij}\right)^2}{k}\right)-\frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n}x_{ij}\right)^2}{N}$ – between-subjects sum of squares,
$df_{BC}=k-1$ – between-conditions degrees of freedom,
$df_{res}=df_T-df_{BC}-df_{BS}$ – residual degrees of freedom,
$df_T=N-1$ – total degrees of freedom,
$df_{BS}=n-1$ – between-subjects degrees of freedom,
$N=nk$,
n – sample size,
x_{ij} – values of the variable from 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, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lll}\text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

Effect size - partial η^2
This quantity indicates the proportion of explained variance to total variance associated with a factor. Thus, in a repeated measures model, it indicates what proportion of the between-conditions variability in outcomes can be attributed to repeated measurements of the variable.

$$\eta^2=\frac{SS_{BC}}{SS_{BC}+SS_{res}}$$

Testy POST-HOC

Introduction to the **contrasts and the POST-HOC tests** was performed in the 16.1.2 unit, which relates to the one-way analysis of variance.

The LSD Fisher test

For simple and complex comparisons (frequency in particular measurements is always the same).

Hypotheses:
Example - **simple comparisons** (comparison of 2 selected means):

$$\begin{array}{ll}\mathcal{H}_0: & \mu_j = \mu_{j+1},\\ \mathcal{H}_1: & \mu_j \neq \mu_{j+1}.\end{array}$$

(i) The value of the critical difference is calculated by using the following 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, adequately: 1 and df_{res}.

(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.

Note!

For contrasts $SE_{contrast}$ is used instead of $\sqrt{\left(\sum_{j=1}^{k} \frac{c_j^2}{n}\right) MS_{res}}$, and degrees of freedem: df_{BS}.

The Scheffe test

For simple comparisons (frequency in particular measurements is always the same).

(i) The value of the critical difference is calculated by using the following formula:

$$CD = \sqrt{F_{\alpha,df_{BC},df_{res}}} \cdot \sqrt{(k-1)\left(\sum_{j=1}^{k} \frac{c_j^2}{n}\right) MS_{res}},$$

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_{ref} degrees of freedom.

The Tukey test.

For simple comparisons (frequency in particular measurements is always the same).

(i) The value of the critical difference is calculated by using the following 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 statistic of the studentized range distribution in PQStat is based on the Lund works (1983)[105]. Other applications or web pages may calculate a little bit different values than PQStat, because they may be based on less precised or more restrictive algorithms (Copenhaver and Holland (1988), Gleason (1999)).

Test for trend.

The test that examines the existence of a trend can be calculated in the same situation as ANOVA for dependent variables, because it is based on the same assumptions, but it captures the alternative hypothesis differently – indicating in it the existence of a trend of mean values in successive measurements. The analysis of the trend in the arrangement of means is based on contrasts Fisher LSD test. By building appropriate contrasts, you can study any type of trend, e.g. linear, quadratic, cubic, etc. A table of example contrast values for selected trends can be found in the description of the testu dla trendu for ANOVA of independent variables.

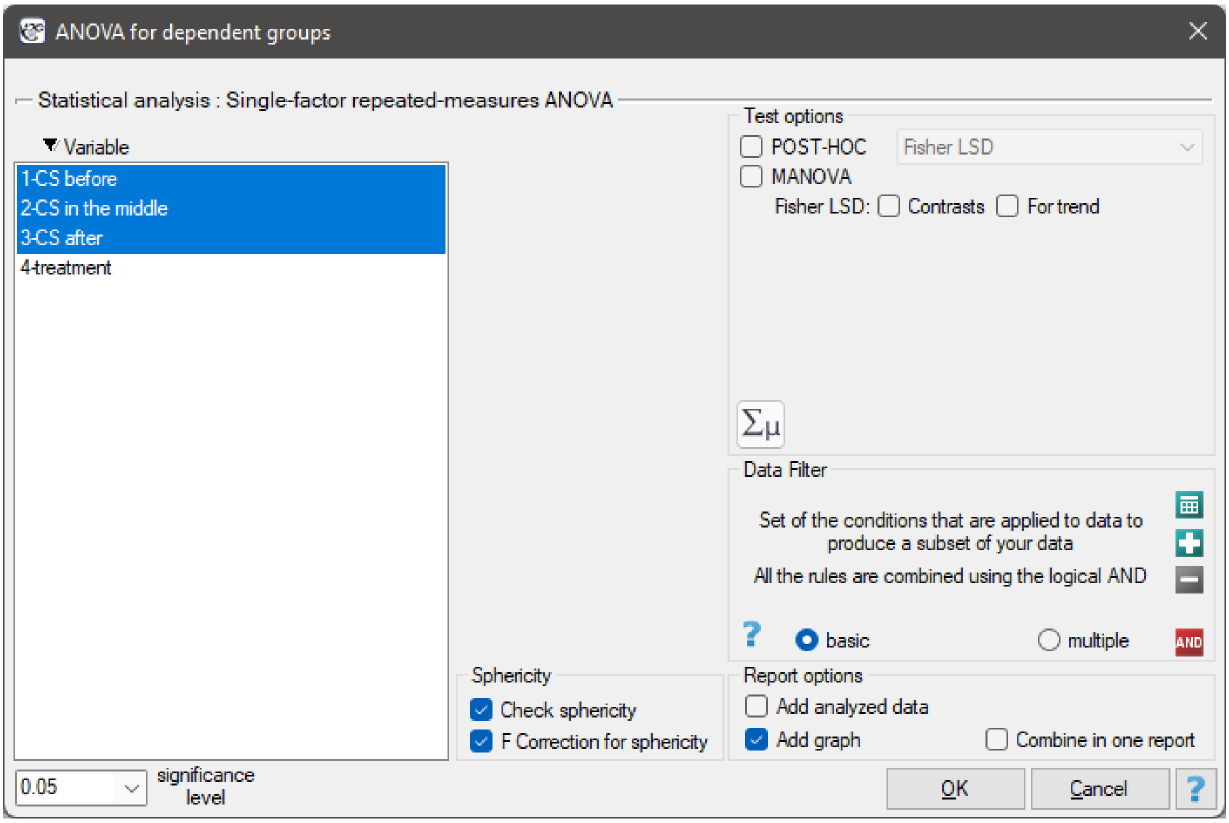
Linear trend
Trend liniowy, tak jak pozostałe trendy, możemy analizować wpisując odpowiednie wartości kontrastów. Jeśli jednak znany jest kierunek trendu liniowego, wystarczy skorzystać z opcji Trend liniowy i wskazać oczekiwaną kolejność populacji przypisując im kolejne liczby naturalne.

A linear trend, like other trends, can be analyzed by entering the appropriate contrast values. However, if the direction of the linear trend is known, simply use the Fisher LSD test option and indicate the expected order of the populations by assigning them consecutive natural numbers.

With the expected direction of the trend known, the alternative hypothesis is one-sided and the one-sided p-values is interpreted. The interpretation of the two-sided p-value means that the researcher does not know (does not assume) the direction of the possible trend.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the Single-factor repeated-measures ANOVA can be opened in Statistics menu→Parametric tests→ANOVA for dependent groups or in Wizard.



ANOVA for dependent groups

Statistical analysis : Single-factor repeated-measures ANOVA

Variable
1-CS before
2-CS in the middle
3-CS after
4-treatment

Test options
□ POST-HOC Fisher LSD
□ MANOVA
Fisher LSD: □ Contrasts □ For trend

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Sphericity
☑ Check sphericity
☑ F Correction for sphericity

Report options
□ Add analyzed data
☑ Add graph □ Combine in one report

0.05 significance level

LOOK AT THE EXAMPLE (16.3)

16.1.6 Mauchly's sphericity

Sphericity assumption is similar but stronger than the assumption of equality of variance. It is met if the variances for the differences between pairs of repeated measurements are the same. Usually, the simpler but more stringent compound symmetry condition is considered in place of the sphericity assumption. This can be done because meeting the compounded symmetry condition entails meeting the sphericity assumption.

Compound symmetry condition assumes, symmetry in the covariance matrix, and therefore equality of variances (elements of the main diagonal of the covariance matrix) and equality of covariances (elements off the main diagonal of the covariance matrix).

Violating the assumption of sphericity or combined symmetry unduly reduces the conservatism of the F-test (makes it easier to reject the null hypothesis).

To check the sphericity assumption, the Mauchly test is used (1940)[115]. Statistical significance ($p \leq \alpha$) here implies a violation of the sphericity assumption.

Basic application conditions:

- measurement on an interval scale,
- multivariate normal distribution or normality of the distribution of each variable tested,
- dependent model.

Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & \sigma_{diff(1)} = \sigma_{diff(2)} = ... = \sigma_{diff(s)}, \\ \mathcal{H}_1: & \text{not all } \sigma_{diff(i)} \text{ are equal } (i = 1, 2, ..., s),\end{array}$$

where:
$\sigma_{diff(i)}$ - population variance of differences between i-th pair of repeated measurements,
s - number of pairs.

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 has the form of:

$$\chi^2 = (f-1)(n-1)\ln W,$$

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.

This statistic has asymptotically (for large sample) χ^2 distribution with $df = \frac{k(k-1)}{2} - 1$ degrees of freedom.

The p value, determined on the basis of test statistics is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ by adopting } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there are no grounds to reject } \mathcal{H}_0.\end{array}$$

A value of $W \approx 1$ is an indication that the sphericity assumption is met. In interpreting the results of this test, however, it is important to note that it is sensitive to violations of the normality assumption of the distribution.

SEE EXAMPLE (16.3) The Epsilon and MANOVA corrections apply to repeated measures ANOVA and are calculated when the sphericity assumption is not met or the variances of the differences between all pairs of measurements are not close to one another.

16.1.7 The ANOVA for dependent groups with Epsilon correction and MANOVA

Correction of non-sphericity

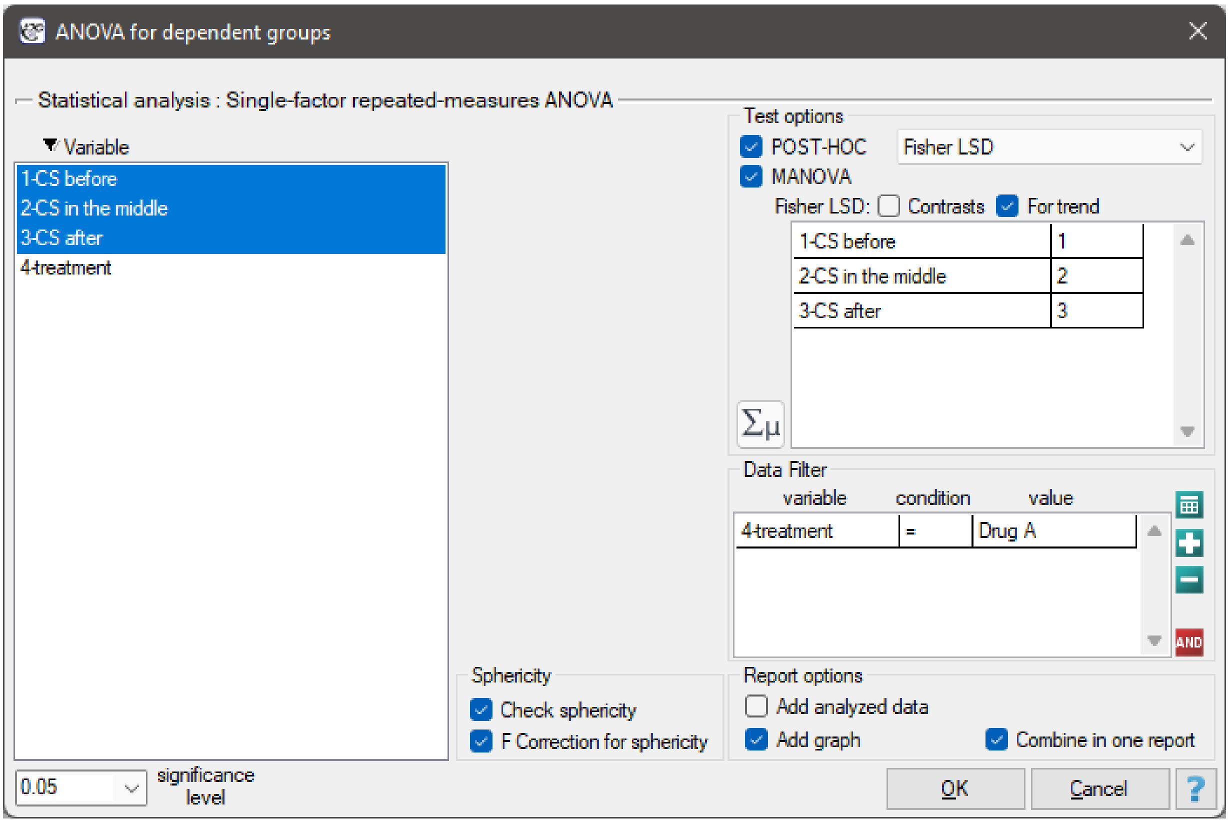
The degree to which sphericity is met is represented by the value of W in the Mauchly test, but also by the values of **Epsilon** (ε) calculated with corrections. $\varepsilon = 1$ indicates strict adherence to the sphericity condition. The smaller the value of Epsilon is compared to 1, the more the sphericity assumption is affected. The lower limit that Epsilon can reach is $\frac{1}{k-1}$.

To minimize the effects of non-sphericity, three corrections can be used to change the number of degrees of freedom when testing from an F distribution. The simplest but weakest is the **Epsilon lower bound correction**. A slightly stronger but also conservative one is the **Greenhouse-Geisser correction** (1959)[69]. The strongest is the **correction by Huynh-Feldt** (1976)[83]. When sphericity is significantly affected, however, it is most appropriate to perform an analysis that does not require this assumption, namely MANOVA.

A multidimensional approach - MANOVA

MANOVA i.e. *multivariate analysis of variance* not assuming sphericity. If this assumption is not met, it is the most efficient method, so it should be chosen as a substitute for analysis of variance for repeated measurements. For a description of this method, see univariate MANOVA. Its use for repeated measures (without the independent groups factor) limits its application to data that are differences of adjacent measurements and provides testing of the same hypothesis as ANOVA for dependent variables.

Settings window for ANOVA for dependent groups with Epsilon correction and MANOVA is opened via menu Statistics→Parametric tests→ANOVA for dependent groups or via Wizard.



ANOVA for dependent groups

Statistical analysis : Single-factor repeated-measures ANOVA

Variable
- 1-CS before
- 2-CS in the middle
- 3-CS after
- 4-treatment

Test options
- ☑ POST-HOC: Fisher LSD
- ☑ MANOVA
- Fisher LSD: □ Contrasts ☑ For trend

1-CS before	1
2-CS in the middle	2
3-CS after	3

Data Filter

variable	condition	value
4-treatment	=	Drug A

Sphericity
- ☑ Check sphericity
- ☑ F Correction for sphericity

Report options
- □ Add analyzed data
- ☑ Add graph
- ☑ Combine in one report

0.05 significance level

Example 16.3. (pressure.pqs file)

The effectiveness of two treatments for hypertension was analyzed. A sample of 56 patients was collected and randomly assigned to two groups: group treated with drug A and 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.

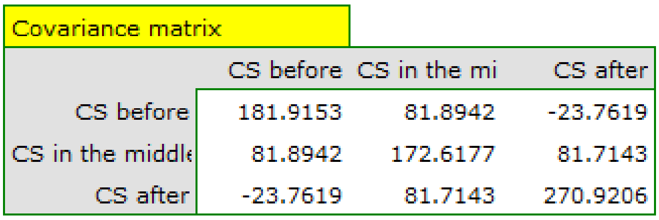
Hypotheses for treated with drug A:

$\mathcal{H}_0$: Mean systolic blood pressure is the same
at any stage of treatment - for those treated with drug A,
$\mathcal{H}_1$: At least one stage of treatment with the drug A
mean systolic blood pressure is different.

The hypotheses for those treated with drug B read similar.

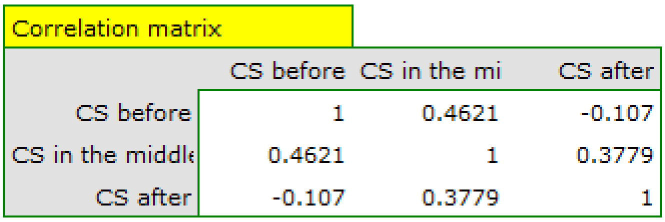
Since the data have a normal distribution, we begin our analysis by testing the assumption of sphericity. We perform the testing for each group separately using a multiple filter.

Failure to meet the sphericity assumption by the group treated with drug B is indicated by both the observed values of the covariance and correlation matrix and the result of the Mauchly test ($W = 0.68$, $p = 0.0063$).



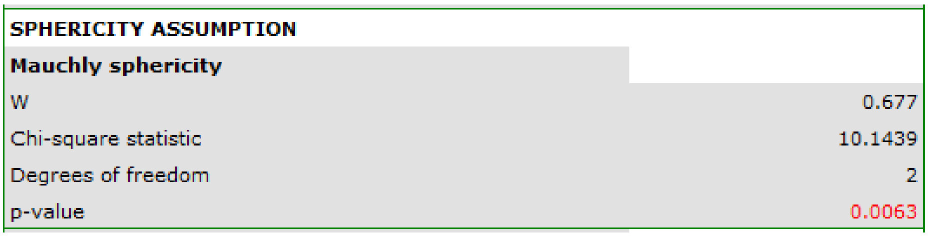
Covariance matrix

	CS before	CS in the mi	CS after
CS before	181.9153	81.8942	-23.7619
CS in the middl	81.8942	172.6177	81.7143
CS after	-23.7619	81.7143	270.9206



Correlation matrix

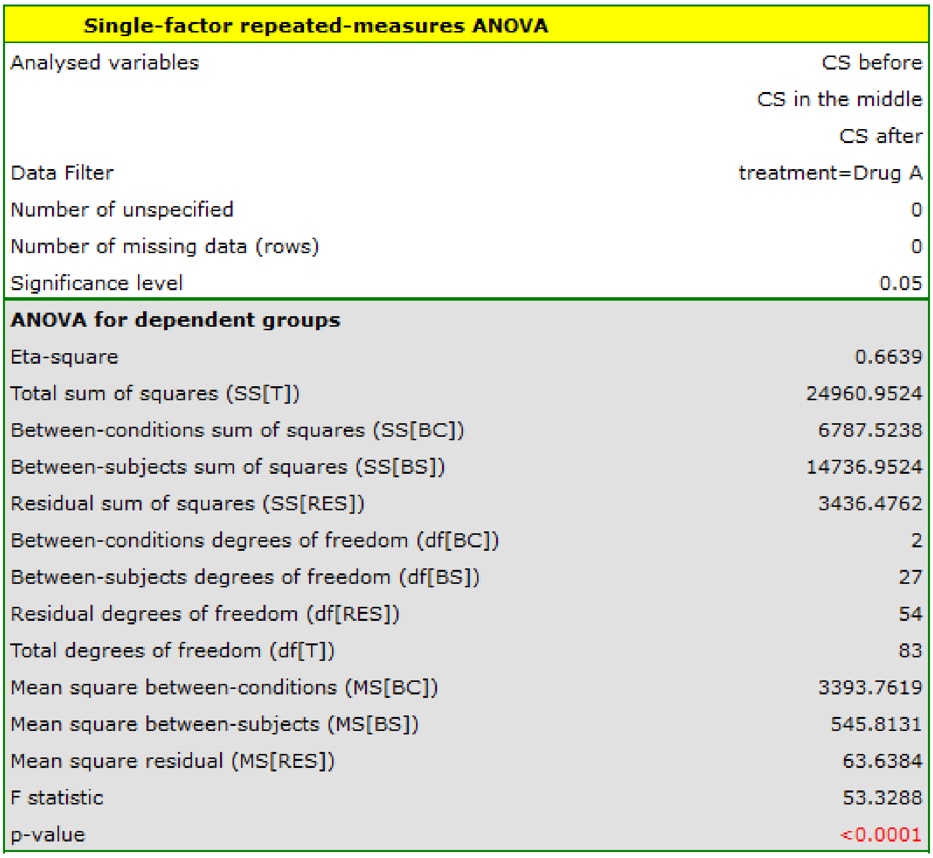
	CS before	CS in the mi	CS after
CS before	1	0.4621	-0.107
CS in the middl	0.4621	1	0.3779
CS after	-0.107	0.3779	1



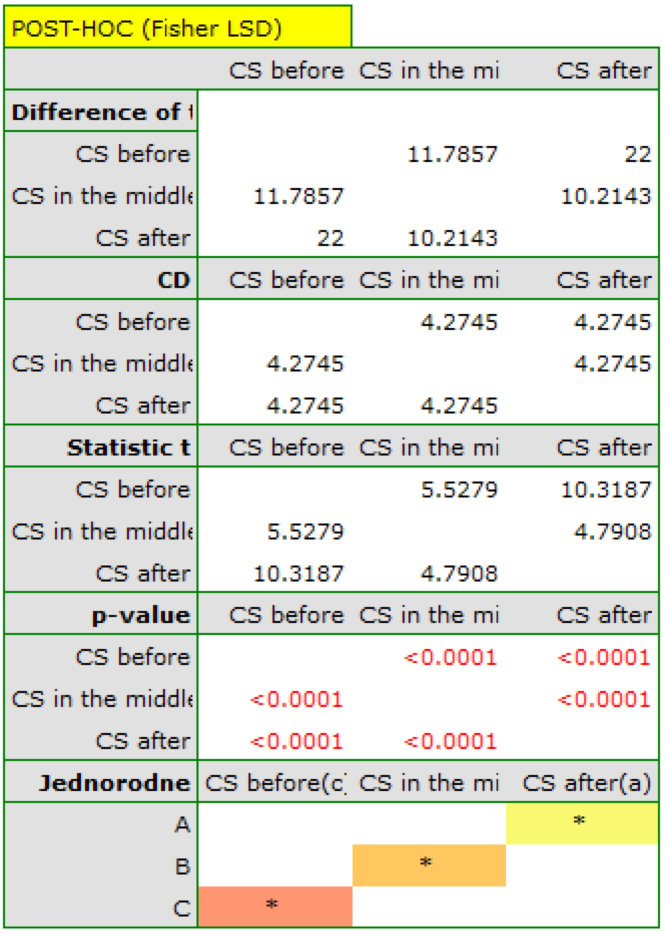
SPHERICITY ASSUMPTION	
Mauchly sphericity	
W	0.677
Chi-square statistic	10.1439
Degrees of freedom	2
p-value	0.0063

We resume our analysis Run the recent test 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 a MANOVA statistic - for those treated with drug B.

Results for those treated with drug A:

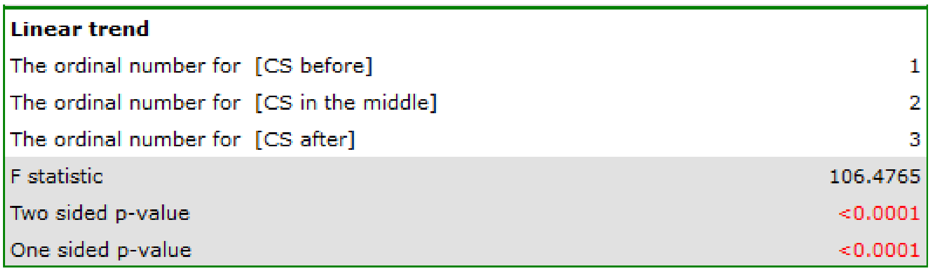


Single-factor repeated-measures ANOVA	
Analysed variables	CS before
	CS in the middle
	CS after
Data Filter	treatment=Drug A
Number of unspecified	0
Number of missing data (rows)	0
Significance level	0.05
ANOVA for dependent groups	
Eta-square	0.6639
Total sum of squares (SS[T])	24960.9524
Between-conditions 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

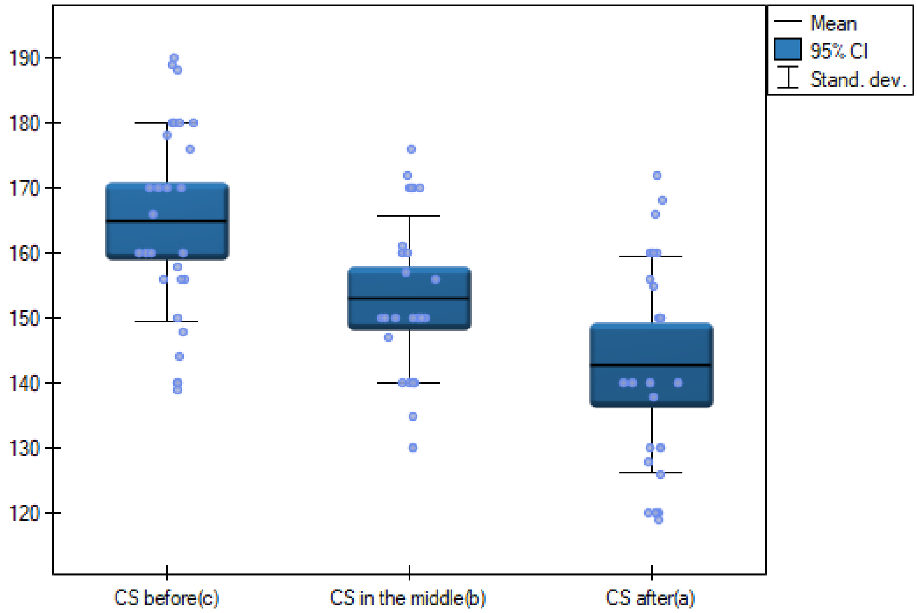


POST-HOC (Fisher LSD)

	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	*		

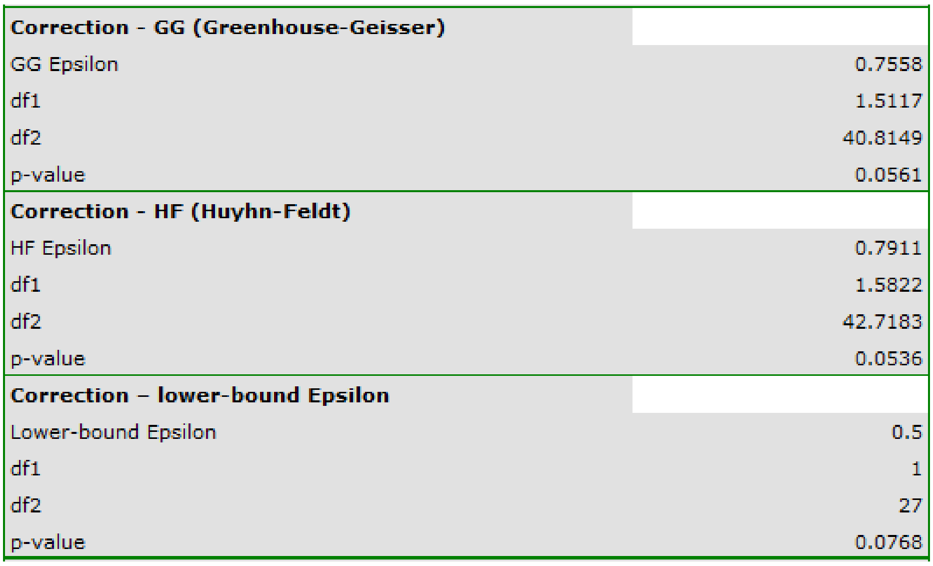


Linear trend	
The ordinal number for [CS before]	1
The ordinal number for [CS in the middle]	2
The ordinal number for [CS after]	3
F statistic	106.4765
Two sided p-value	<0.0001
One sided p-value	<0.0001

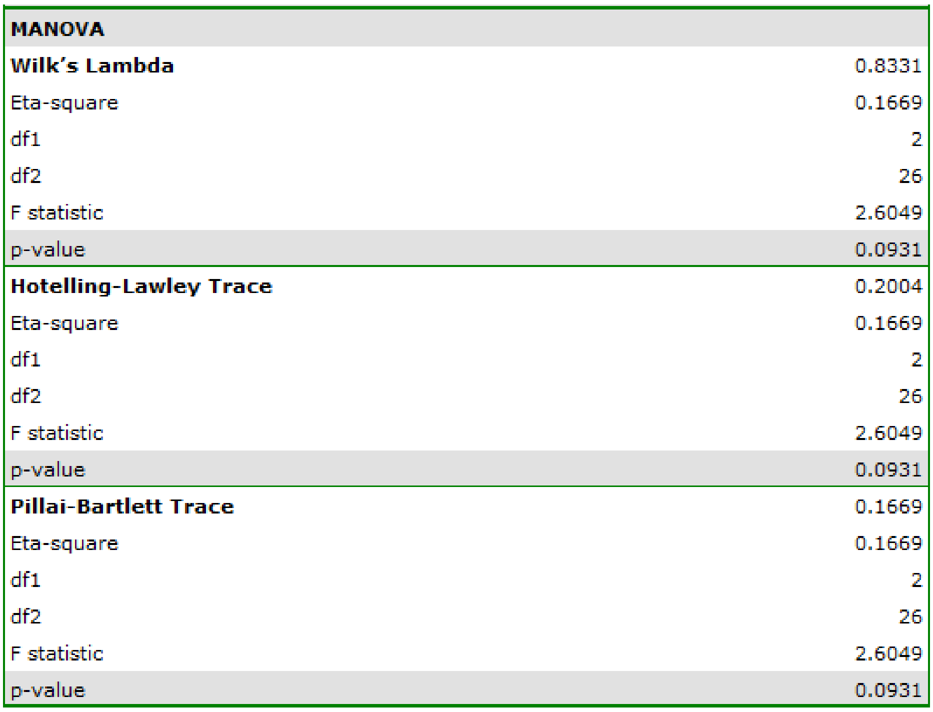


indicate significant (at the level of significance $\alpha = 0.05$) differences between mean systolic blood pressure values (value $p < 0.0001$ for repeated measures ANOVA). More than 66% of the between-conditions variation in outcomes can be explained by the use of drug A ($\eta = 0.66$). The differences apply to all treatment stages compared (POST-HOC score). The decrease in systolic blood pressure due to treatment is also significant ($p < 0.0001$). Thus, we can consider Drug A as an effective drug.

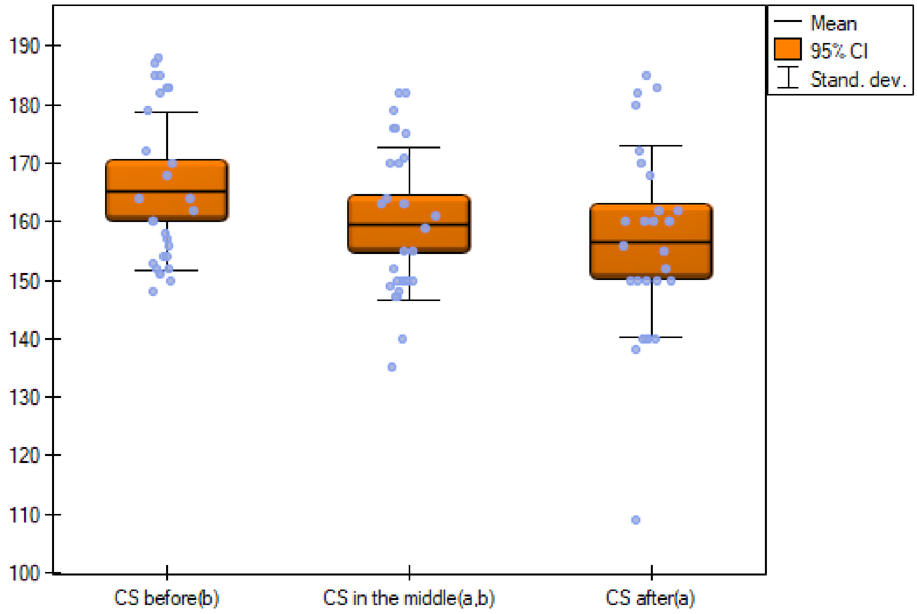
Results for those treated with drug B:



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



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



indicate that there are no significant differences between mean systolic blood pressure values, both when we use epsilon and Lambda Wilks (MANOVA) corrections. As little as 17% of the between-conditions variation in results can be explained by the use of drug B ($\eta = 0.17$).

16.2 NON-PARAMETRIC TESTS

16.2.1 The Kruskal-Wallis ANOVA

The Kruskal-Wallis one-way analysis of variance by ranks (Kruskal 1952 [93]; Kruskal and Wallis 1952 [94]) is an extension of the U-Mann-Whitney test on more than two populations. This test is used to verify the hypothesis that there is no shift in the compared distributions, i.e., most often the insignificant differences between medians of the analysed variable in ($k \geq 2$) populations (but you need to assume, that the variable distributions are similar - comparison of rank variances can be checked using Conover's rank test).

Additional analyses:

- it is possible to test for a trend in the arrangement of the groups under study by performing the Jonckheere-Terpstra test for trend.

Basic assumptions:

- measurement on an ordinal scale or on an interval scale,
- an independent model.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \phi_1=\phi_2=...=\phi_k, \\ \mathcal{H}_1: & \text{not all } \phi_j \text{ are equal } (j=1,2,...,k), \end{array}$$

where:
$\phi_1, \phi_2, ...\phi_k$ distributions of the analysed variable of each population.

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 – samples sizes $(j=1,2,...k)$,
R_{ij} – ranks ascribed to the values of a 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 formula for the test statistic H includes the correction for ties C. This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $C=1$).

The H 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, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The POST-HOC tests

Introduction to **the contrasts and the POST-HOC tests** was performed in the 16.1.2 unit, which relates to the one-way analysis of variance.

The Dunn test

For simple comparisons, equal-size groups as well as unequal-size groups.

The Dunn test (Dunn 1964[51]) includes a correction for tied ranks (Zar 2010[177]) and is a test corrected for multiple testing. The Bonferroni or Sidak correction is most commonly used here, although other, newer corrections are also available, described in more detail in Multiple comparisons.

Example - **simple comparisons** (comparing 2 selected median / mean ranks with each other):

$$\begin{array}{cl}\mathcal{H}_0: & \theta_j = \theta_{j+1},\\ \mathcal{H}_1: & \theta_j \neq \theta_{j+1}.\end{array}$$

(i) The value of critical difference is calculated by using the following formula:

$$CD = Z_{\frac{\alpha}{c}}\sqrt{\frac{N(N+1)}{12}\left(\sum_{j=1}^{k}\frac{c_j^2}{n_j}\right)},$$

where:
$Z_{\frac{\alpha}{c}}$ - is the critical value (statistic) of the normal distribution for a given significance level α corrected on the number of possible simple comparisons c.

(ii) The test statistic is defined by:

$$Z = \frac{\sum_{j=1}^{k} c_j\overline{R}_j}{\sqrt{\frac{N(N+1)}{12}\left(\sum_{j=1}^{k}\frac{c_j^2}{n_j}\right)}},$$

where:
$\overline{R}_j$ – mean of the ranks of the j-th group, for $(j = 1, 2, ...k)$,

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 this correction is not calculated because $\sum(t^3 - t) = 0$).

The test statistic asymptotically (for large sample sizes) has the normal distribution, and the p value is corrected on the number of possible simple comparisons c.

Conover-Inman test

The non-parametric equivalent of Fisher LSD[46], used for simple comparisons of both groups of equal and different sizes.

(i) The value of critical difference is calculated by using the following 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:

$$S^2=\frac{1}{N-1}\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j}R_{ij}^2-N\frac{(N+1)^2}{4}\right)$$

$F_{\alpha,1,N-k}$ is the critical value (statistic) Snedecor's F distribution for a given significance level α and for degrees of freedom respectively: 1 i $N-k$.

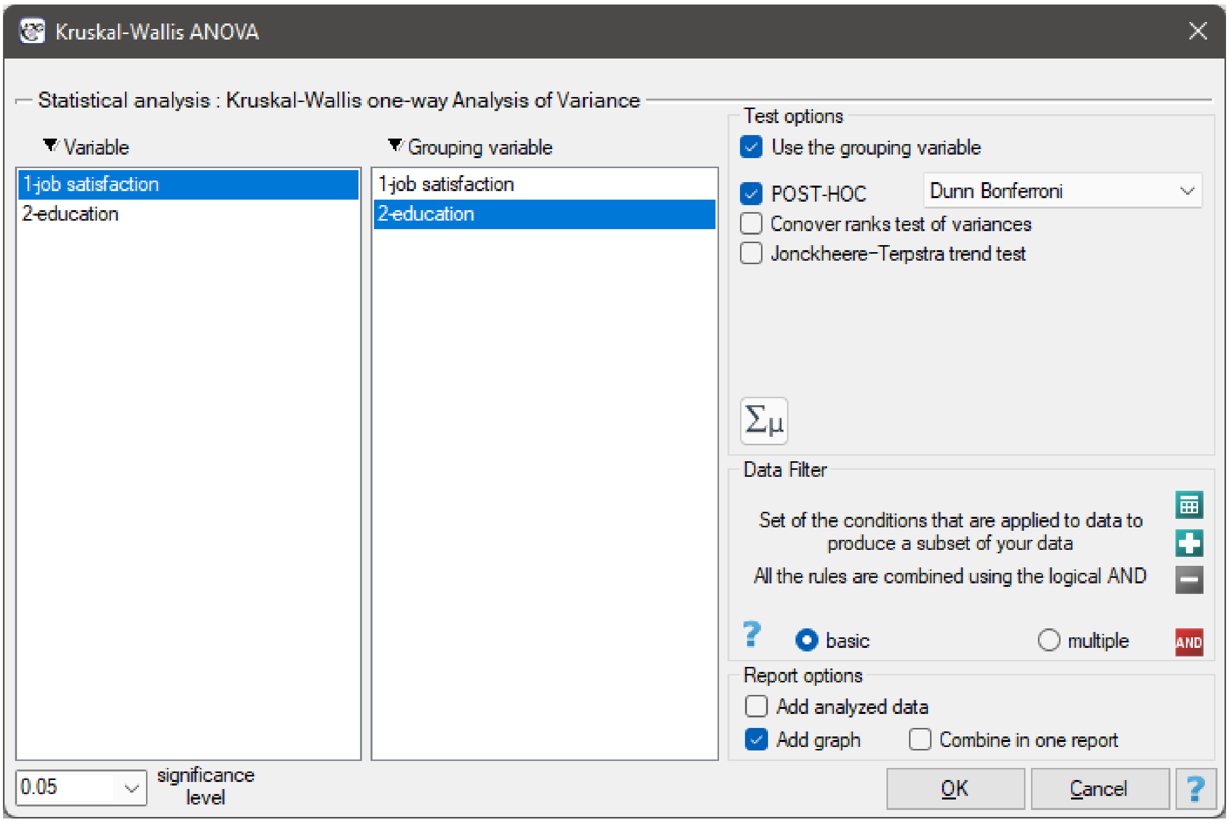
(ii) The test statistic is defined by:

$$t=\frac{\sum_{j=1}^{k}c_j\overline{R}_j}{\sqrt{S^2\frac{N-1-H}{N-k}\sum_{j=1}^{k}\frac{c_j^2}{n_j}}},$$

where:
$\overline{R}_j$ – The mean ranks of the j-th group, for $(j=1,2,...k)$,

This statistic follows a t-Student distribution with $N-k$ degrees of freedom.

The settings window with the Kruskal-Wallis ANOVA can be opened in Statistics menu→NonParametric tests →Kruskal-Wallis ANOVA or in Wizard.



Kruskal-Wallis ANOVA

Statistical analysis : Kruskal-Wallis one-way Analysis of Variance

Variable: 1-job satisfaction (selected), 2-education

Grouping variable: 1-job satisfaction, 2-education (selected)

Test options
☑ Use the grouping variable
☑ POST-HOC Dunn Bonferroni
☐ Conover ranks test of variances
☐ Jonckheere-Terpstra trend test

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
☐ Add analyzed data
☑ Add graph ☐ Combine in one report

0.05 significance level

Example 16.4. (jobSatisfaction.pqs)
A group of 120 people was interviewed, for whom the occupation is their first job obtained after receiving appropriate education. The respondents rated their job satisfaction on a five-point scale, where:

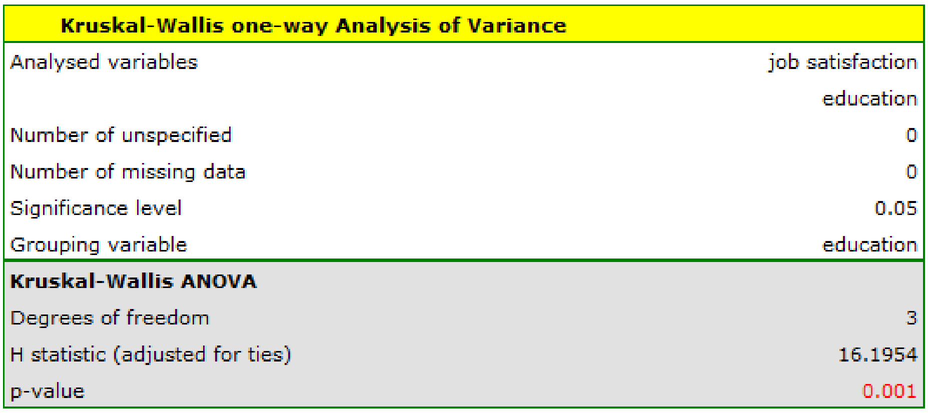
1- unsatisfying job,
2- job giving little satisfaction,

3- job giving an average level of satisfaction,
4- job that gives a fairly high level of satisfaction,
5- job that is very satisfying.

We will test whether the level of reported job satisfaction does not change for each category of education.

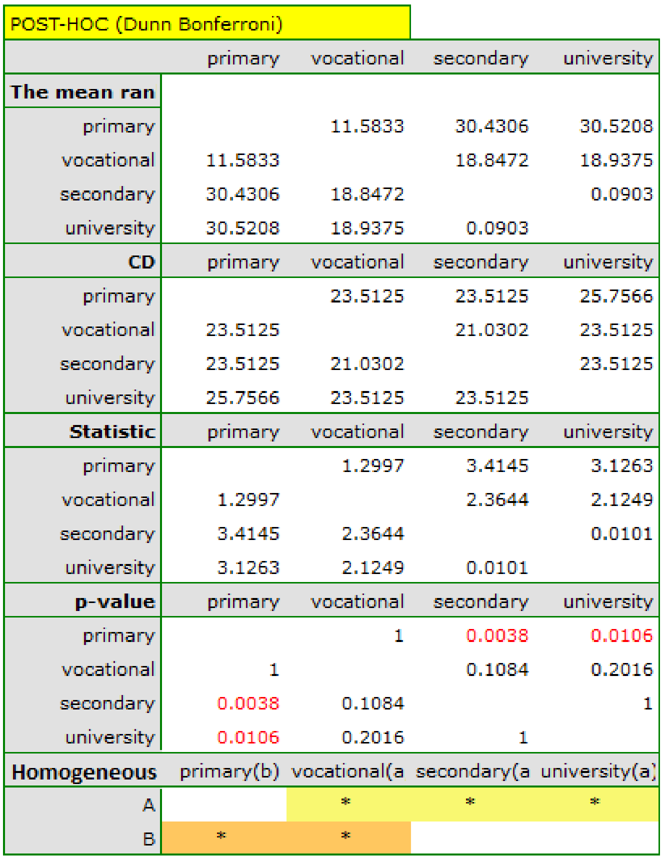
Hypotheses:

$\mathcal{H}_0$: the level of job satisfaction is the same for each education category,
$\mathcal{H}_1$: at least one education category (one population)
has different levels of job satisfaction.



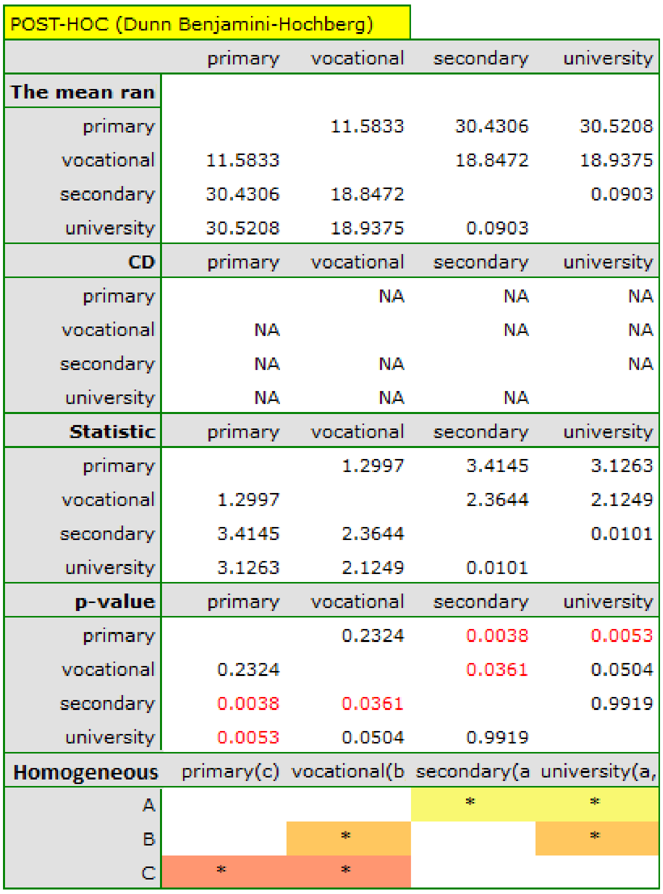
Kruskal-Wallis one-way Analysis of Variance	
Analysed variables	job satisfaction
	education
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	education
Kruskal-Wallis ANOVA	
Degrees of freedom	3
H statistic (adjusted for ties)	16.1954
p-value	0.001

The obtained value of $p = 0.001$ indicates a significant difference in the level of satisfaction between the compared categories of education. Dunn's POST-HOC analysis with Bonferroni's correction shows that significant differences are between those with primary and secondary education and those with primary and tertiary education. Slightly more differences can be confirmed by selecting the stronger POST-HOC Conover-Iman.



POST-HOC (Dunn Bonferroni)

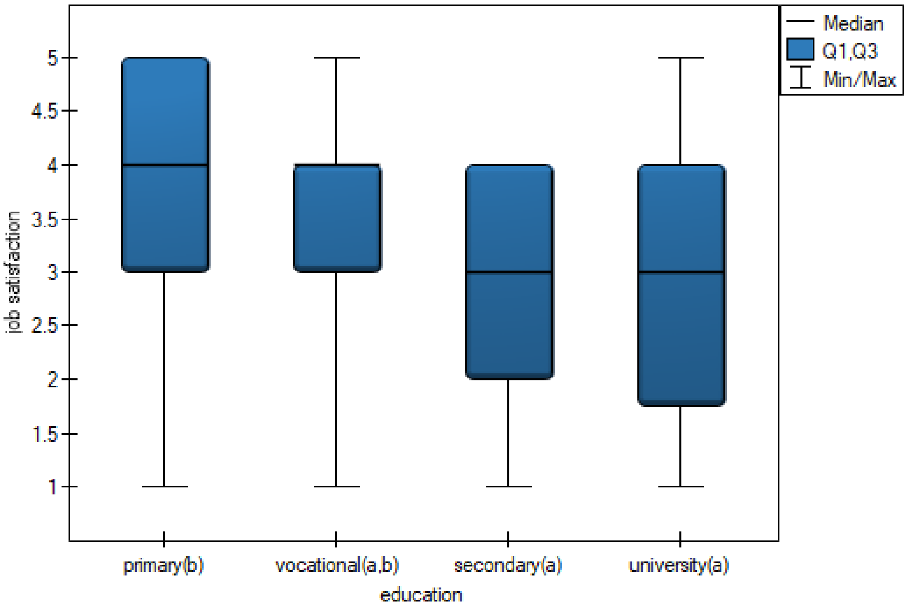
	primary	vocational	secondary	university
The mean ran				
primary		11.5833	30.4306	30.5208
vocational	11.5833		18.8472	18.9375
secondary	30.4306	18.8472		0.0903
university	30.5208	18.9375	0.0903	
CD	primary	vocational	secondary	university
primary		23.5125	23.5125	25.7566
vocational	23.5125		21.0302	23.5125
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	*	*		



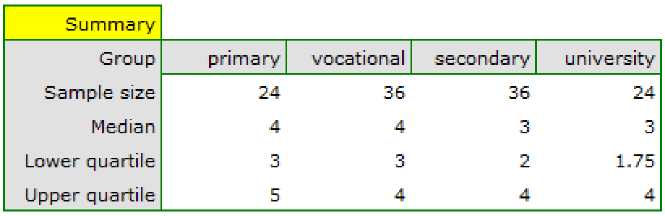
POST-HOC (Dunn Benjamini-Hochberg)

	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	*	*		

In the graph showing medians and quartiles we can see homogeneous groups determined by the POST-HOC test. If we choose to present Dunn's results with Bonferroni correction we can see two homogeneous groups that are not completely distinct, i.e. group (a) - people who rate job satisfaction lower and group (b)- people who rate job satisfaction higher. Vocational education belongs to both of these groups, which means that people with this education evaluate job satisfaction quite differently. The same description of homogeneous groups can be found in the results of the POST-HOC tests.

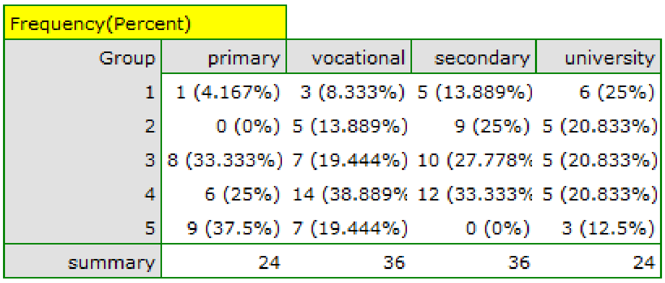


We can provide a detailed description of the data by selecting descriptive statistics in the analysis window Σμ and indicating to add counts and percentages to the description.



Summary

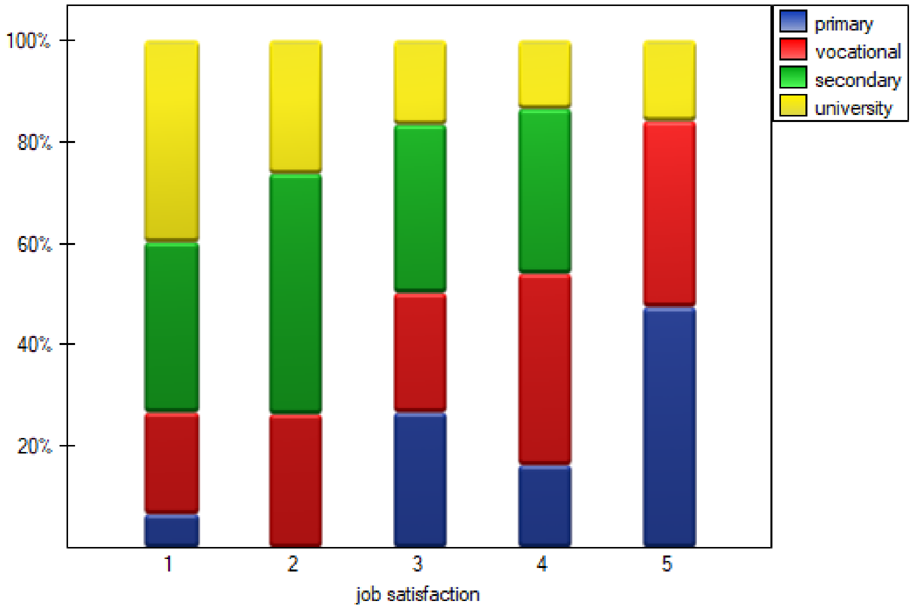
Group	primary	vocational	secondary	university
Sample size	24	36	36	24
Median	4	4	3	3
Lower quartile	3	3	2	1.75
Upper quartile	5	4	4	4



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 plot.



16.2.2 The Jonckheere-Terpstra test for trend

The Jonckheere-Terpstra test for ordered alternatives described independently by Jonckheere (1954) [87] and Terpstra (1952)[158] an be calculated in the same situation as the Kruskal-Wallis ANOVA , as it is based on the same assumptions. The Jonckheere-Terpstra test, however, captures the alternative hypothesis differently - indicating in it the existence of a trend for successive populations.

Hypotheses are simplified to medians:

$$\begin{array}{ll}\mathcal{H}_0: & \theta_1=\theta_2=...=\theta_k,\\ \mathcal{H}_1: & \theta_1\geq\theta_2\geq...\geq\theta_k, \text{ with at least one strict inequality}\end{array}$$

Note

The term: "with at least one strict inequality" written in the alternative hypothesis of this test means that at least the median of one population should be greater than the median of another population in the order specified.

The test statistic has the form:

$$Z=\frac{L-\left[\frac{N^2-\sum_{j=1}^{k}n_j^2}{4}\right]}{SE}$$

where:

$L=$ – sum of l_{ij} values obtained for each pair of compared populations,

l_{ij} – number of results higher than a preset value in the next occurring 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 – umber of cases included in the tied rank,

$N=\sum_{j=1}^{k}n_j$,

n_j – sample sizes for $(j=1,2,...k)$.

Note

To be able to perform a trend analysis, the expected order of the populations must be indicated by assigning consecutive natural numbers.

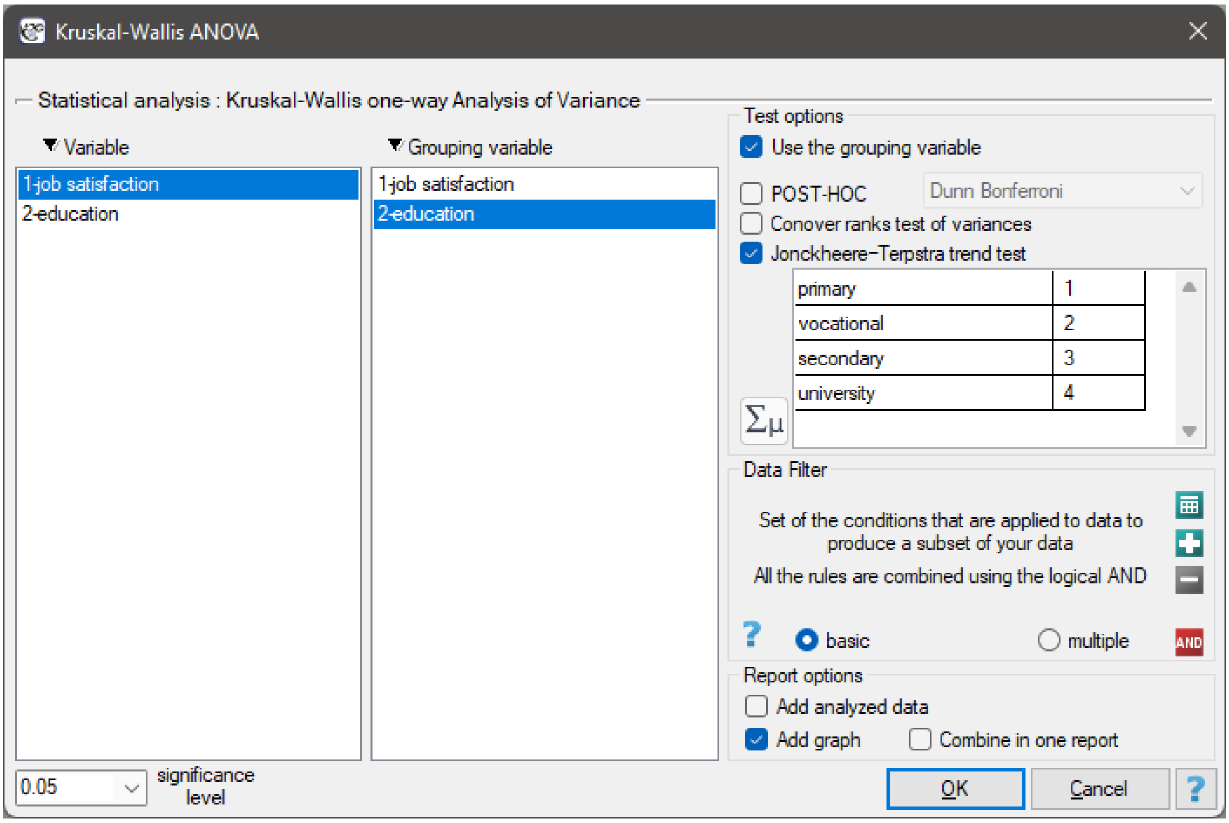
The formula for the test statistic Z includes the correction for ties. This correction is applied when tied ranks are present (when there are no tied ranks the test statistic formula reduces to the original Jonckheere-Terpstra formula without this correction).

The statistic Z has asymptotically (for large samples) normal distribution.

With the expected direction of the trend known, the alternative hypothesis is one-sided and the one-sided p-value is interpreted. The interpretation of the two-sided p-value means that the researcher does not know (does not assume) the direction of the possible trend.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the Jonckheere-Terpstra test for trend can be opened in Statistics menu→NonParametric tests→Kruskal-Wallis ANOVA or in Wizard.



Kruskal-Wallis ANOVA

Statistical analysis : Kruskal-Wallis one-way Analysis of Variance

Variable: 1-job satisfaction (selected), 2-education

Grouping variable: 1-job satisfaction, 2-education (selected)

Test options
☑ Use the grouping variable
□ POST-HOC Dunn Bonferroni
□ Conover ranks test of variances
☑ Jonckheere-Terpstra trend test

primary	1
vocational	2
secondary	3
university	4

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
□ Add analyzed data
☑ Add graph □ Combine in one report

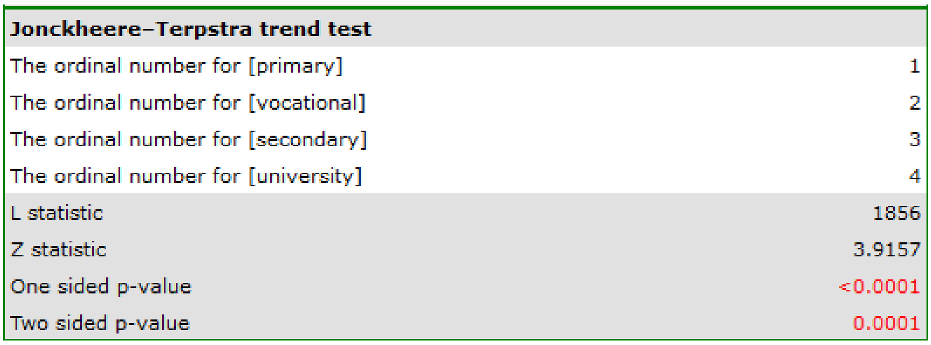
0.05 significance level

Example (16.4) **continued** *(jobSatisfaction.pqs file)* It is suspected that better educated people have high job demands, which may reduce the satisfaction level of the first job, which often does not meet such demands. Therefore, it is worthwhile to conduct a trend analysis.

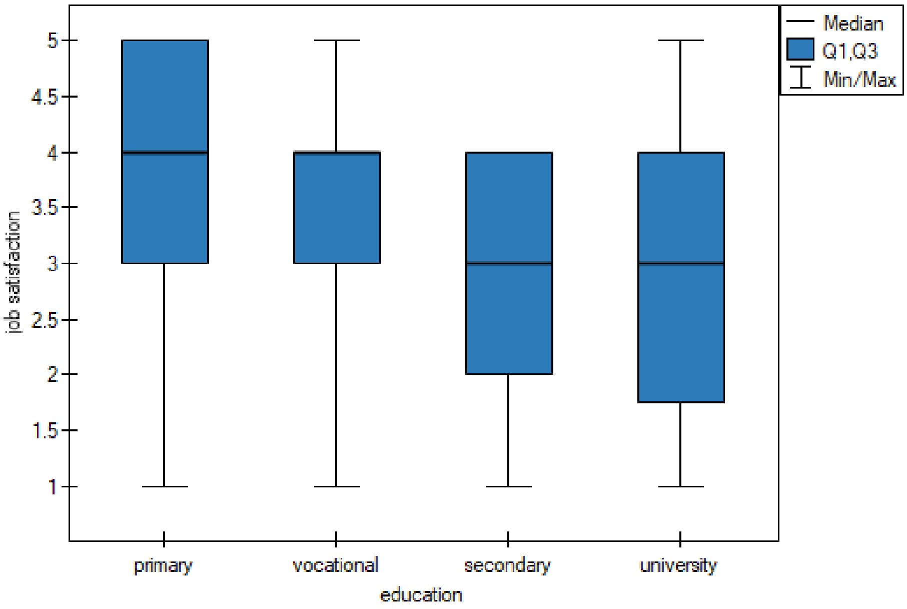
Hypotheses:

$\mathcal{H}_0$: No indicated trend in satisfaction with first job
with increasing education,
$\mathcal{H}_1$: There is an indicated trend in the level of satisfaction with the first job.

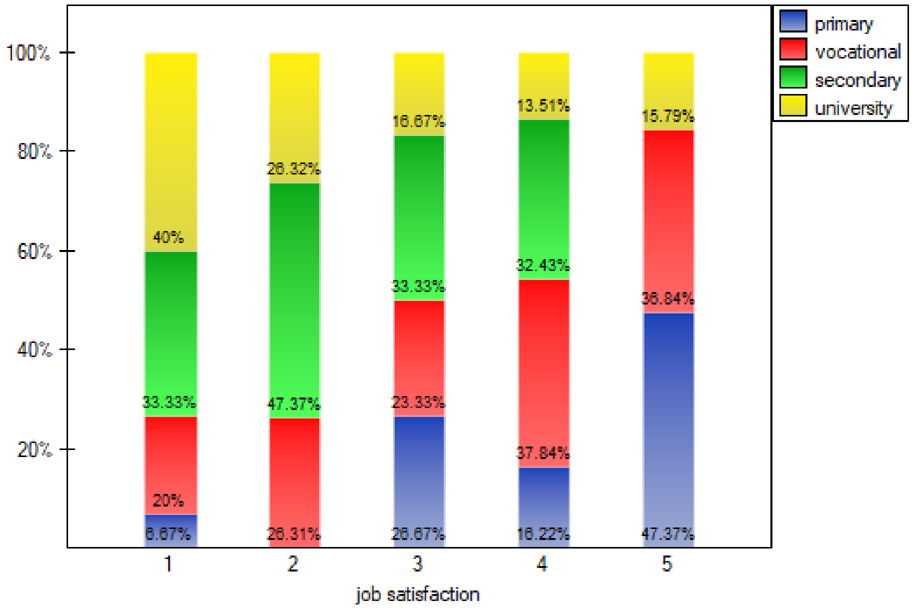
To do this, we resume the analysis with the Run the recent test button, select the Jonckheere-Terpstra trend test option, and assign successive natural numbers to the education categories.



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



The obtained one-sided value $p < 0.0001$ and is less than the set significance level $\alpha = 0.05$, which speaks in favor of a trend actually occurring consistent with the researcher's expectations.
We can also confirm the existence of this trend by showing the percentage distribution of responses obtained.



16.2.3 The Conover ranks test of variance

Conover squared ranks test is used, similarly to Fisher-Snedecor test (for $k = 2$), Levene test and Brown-Forsythe test (for $k >= 2$) to verify the hypothesis of similar variation of the tested variable in several populations. It is the non-parametric counterpart of the tests indicated above, by that it does not assume normality of the data distribution and is based on the ranks[46].However, this test examines variation and therefore distances to the mean, so the basic condition for its use is:

- measurement on an interval scale,

Hypotheses:

$\mathcal{H}_0$: the dispersion of the data in the populations being compared is the same,
$\mathcal{H}_1$: at least two populations differ in the amount of data dispersion.

The test statistic has the form:

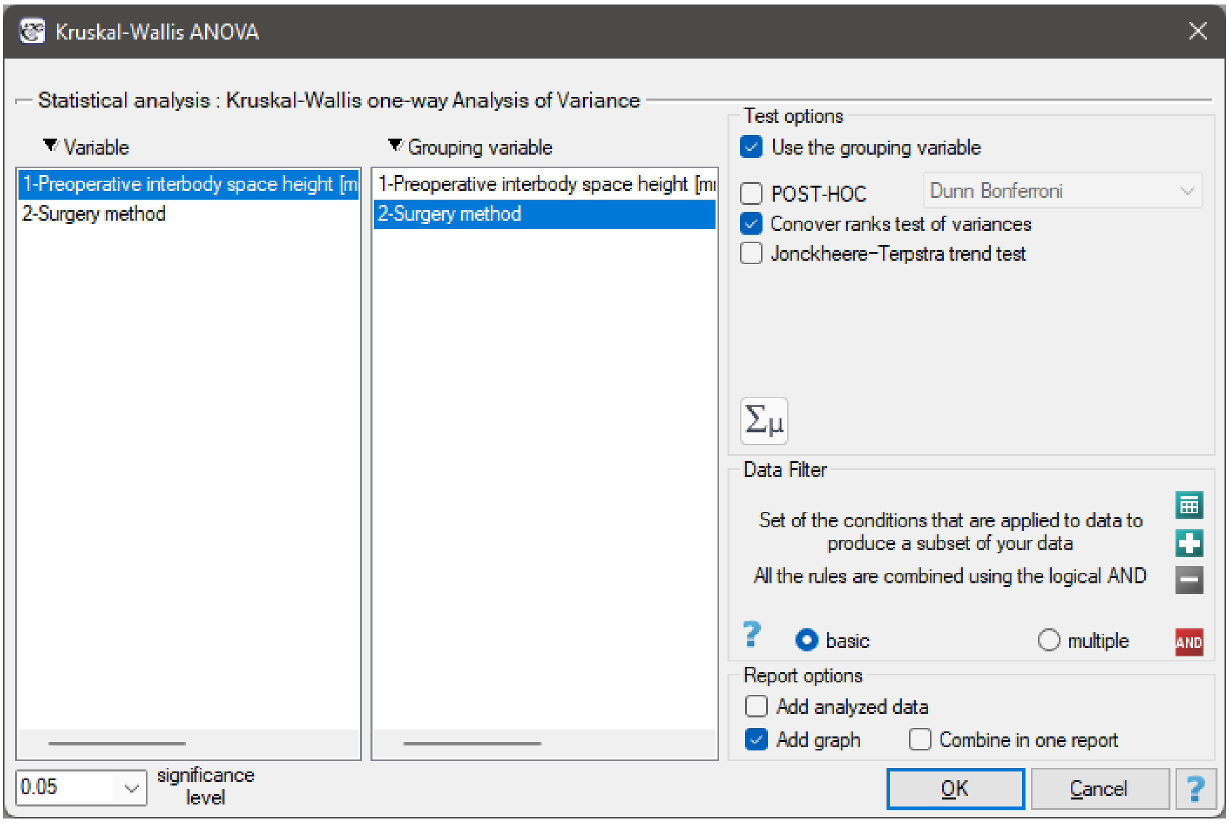
$$\chi^2 = \frac{1}{D^2}\left(\sum_{j=1}^{k}\frac{S_j^2}{n_j} - N\overline{S}^2\right)$$

where:
$N = n_1 + n_2 + ... + n_k$,
n_j – individual group sizes,
S_j –sum of ranks squares in j-th group,
$\overline{S} = \frac{1}{N}\sum_{j=1}^{k}\S_j$ – mean of all ranks squares,
$D^2 = \frac{1}{N-1}\left(\sum_{ji=1}^{N}R_i^4 - N\overline{S}^2\right)$,
R_i – ranks for values representing the distance of the measurement from the mean of a given group.

This statistic has a χ^2distribution with $k-1$ degrees of freedom.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the Conover ranks test of variance can be opened in Statistics menu→NonParametric tests→Kruskal-Wallis ANOVA, option Conover ranks test of variance or textsfStatistics menu→NonParametric tests→Mann-Whitney, option Conover ranks test of variance.



Kruskal-Wallis ANOVA

Statistical analysis : Kruskal-Wallis one-way Analysis of Variance

Variable:
1-Preoperative interbody space height [m
2-Surgery method

Grouping variable:
1-Preoperative interbody space height [mi
2-Surgery method

Test options
☑ Use the grouping variable
☐ POST-HOC Dunn Bonferroni
☑ Conover ranks test of variances
☐ Jonckheere-Terpstra trend test

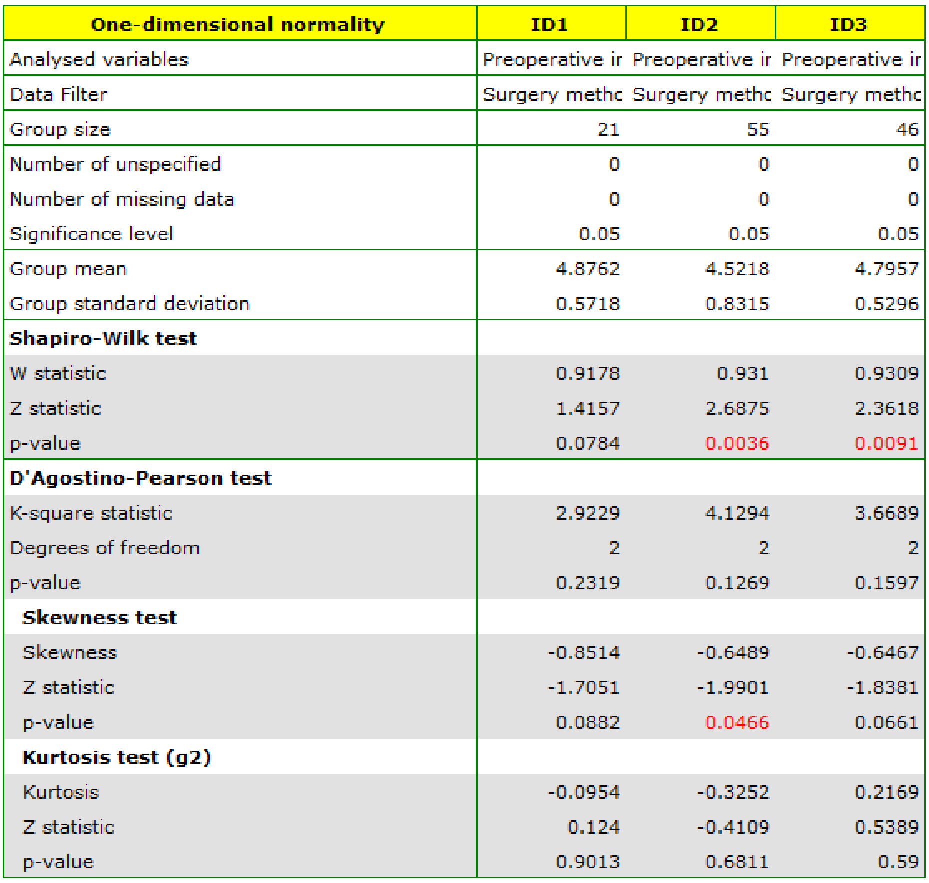
Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
☐ Add analyzed data
☑ Add graph ☐ Combine in one report

0.05 significance level

Example 16.5. (surgeryMethod.pqs file)
Patients have been prepared for spinal surgery. The patients will be operated on by one of three methods. Preliminary allocation of each patient to each type of surgery has been made. At a later stage we intend to compare the condition of the patients after the surgeries, therefore we want the groups of patients to be comparable. They should be similar in terms of the height of the interbody space (WPMT) before surgery. The similarity should concern not only the average values but also the differentiation of the groups.

The distribution of the data was checked



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

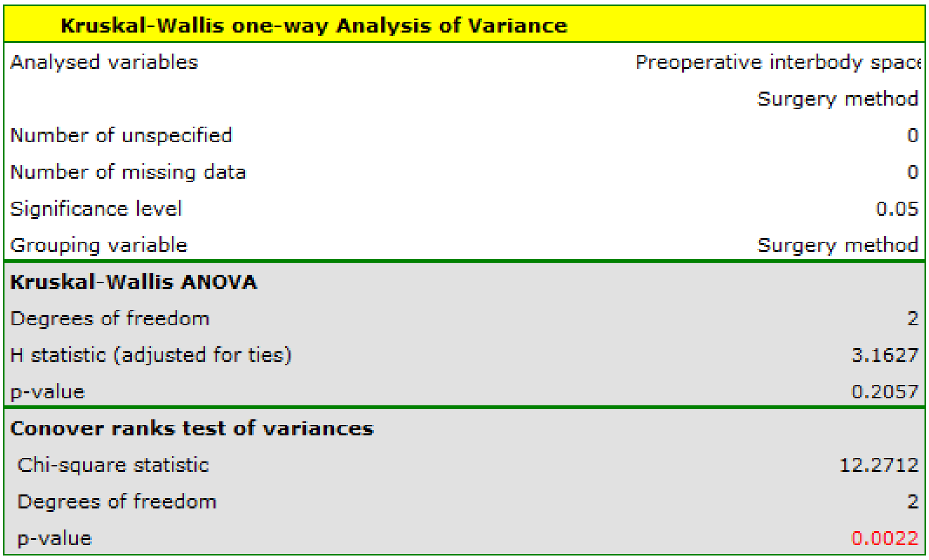
It is found that for the two methods, the WPMT operation exhibits deviations from normality, largely caused by skewness of the data. Further comparative analysis will be conducted using the Kruskal-Wallis test to compare whether the level of WPMT differs between the methods, and the Conover test to indicate whether the spread of WPMT scores is similar in each method.

Hypotheses for Conover's variance test:

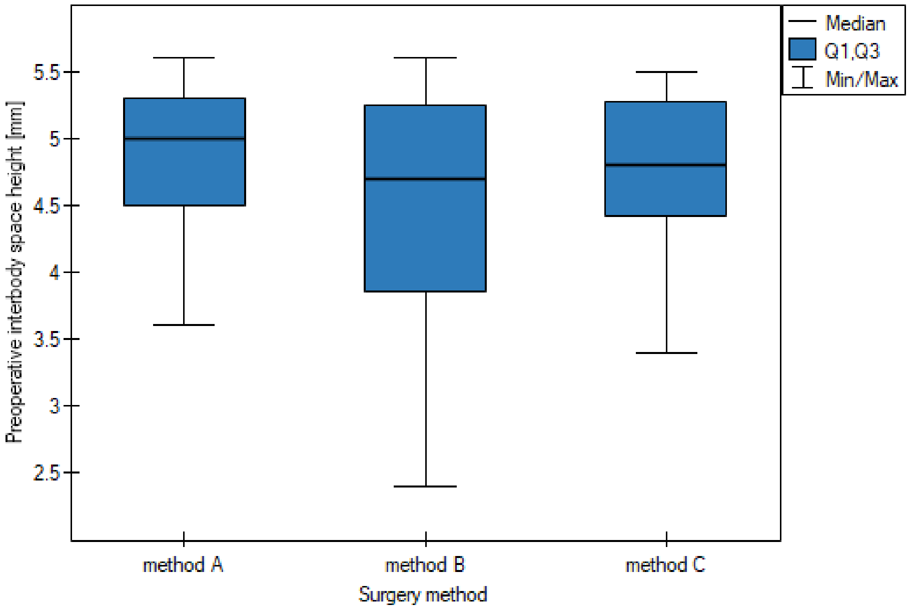
$\mathcal{H}_0$: The diversity (scope) of WPMT is the same for each method of operation,
$\mathcal{H}_1$: WPMT diversity (range) is higher/lower for at least one method of operation.

Hypotheses for Kruskal-Wallis test:

$\mathcal{H}_0$: WPMT level is the same for each operation method,
$\mathcal{H}_1$: WPMT level is higher/lower for at least one method of operation.



Kruskal-Wallis one-way Analysis of Variance	
Analysed variables	Preoperative interbody spac 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



First, the value of Conover's test of variance is interpreted, which indicates statistically significant differences in the ranges of the groups compared (p=0.0022). From the graph, we can conclude that the differences are mainly in group 3. Since differences in WPMT were detected, the interpretation of the result of the Kruskal-Wallis test comparing the level of WPMT for these methods should be cautious, since this test is sensitive to heterogeneity of variance. Although the Kruskal-Wallis test showed no significant differences (p=0.2057), it is recommended that patients with low WPMT (who were mainly assigned to surgery with method B) be more evenly distributed, i.e. to see if they could be offered surgery with method A or C. After reassignment of patients, the analysis should be repeated.

16.2.4 The Friedman ANOVA

The Friedman repeated measures analysis of variance by ranks – the Friedman ANOVA - was described by Friedman (1937)[64]. This test is used when the measurements of an analysed variable are made several times ($k \geq 2$) each time in different conditions. It is also used when we have rankings coming from different sources (form different judges) and concerning a few ($k \geq 2$) objects, but we want to assess the grade of the rankings agreement.

Iman Davenport (1980[84]) has shown that in many cases the Friedman statistic is overly conservative and has made some modification to it. This modification is the non-parametric equivalent of the ANOVA for dependent groups which makes it now recommended for use in place of the traditional Friedman statistic.

Additional analyses:

- It is possible to take missing data into account by using the Accept missing data option, calculating Durbin ANOVA or Skillings-Mack ANOVA;
- it is possible to test the trend in the arrangement of the studied groups by performing Page test for trend.

Basic assumptions:

- measurement on an ordinal scale or on an interval scale,
- a dependent model.

Hypotheses relate to the equality of the sum of ranks for successive measurements (R_j) or are simplified to medians (θ_j)

$$\begin{array}{ll} \mathcal{H}_0: & \theta_1=\theta_2=...=\theta_k, \\ \mathcal{H}_1: & \text{not all } \theta_j \text{ are equal } (j=1,2,...,k), \end{array}$$

where:
$\theta_1, \theta_2, ...\theta_k$ medians for an analysed features, in the following measurements from the examined population.

Two test statistics are determined: the Friedman statistic and the Iman-Davenport modification of this statistic.

The Friedman statistic has the form:

$$T_1=\frac{1}{C}\left(\frac{12}{nk(k+1)}\left(\sum_{j=1}^{k}\left(\sum_{i=1}^{n}R_{ij}\right)^2\right)-3n(k+1)\right),$$

where:
n – sample size,
R_{ij} – ranks ascribed to the following measurements $(j=1,2,...k)$, separately for the analysed objects $(i=1,2,...n)$,
$C=1-\frac{\sum(t^3-t)}{n(k^3-k)}$ – correction for ties,
t – number of cases included in a tie.

The Iman-Davenport modification of the Friedman statistic has the form:

$$T_2=\frac{(n_j-1)T_1}{n_j(k-1)-T_1}$$

The formula for the test statistic T_1 and T_2 includes the correction for ties C. This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $C=1$).

The T_1 statistic has asymptotically ((for large sample sizes) has the χ^2 distribution with $df=k-1$ degrees of freedom.

The statistic T_2 follows the Snedecor's F distribution with $df_1 = k-1$ i $df_2 = (n_j-1)(k-1)$ degrees of freedem.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The POST-HOC tests

Introduction to **the contrasts and the POST-HOC tests** was performed in the 16.1.2 unit, which relates to the one-way analysis of variance.

The Dunn test

For simple comparisons (frequency in particular measurements is always the same).

The Dunn test (Dunn 1964[51]) is a corrected test due to multiple testing. The **Bonferroni** or **Sidak** correction is most commonly used here, although other, newer corrections are also available and are described in more detail in the Multiple Comparisons section.

Hypotheses:

Example - **simple comparisons** (comparison of 2 selected medians):

$$\begin{array}{cl} \mathcal{H}_0: & \theta_j = \theta_{j+1}, \\ \mathcal{H}_1: & \theta_j \neq \theta_{j+1}. \end{array}$$

(i) The value of critical difference is calculated by using the following formula:

$$CD = Z_{\frac{\alpha}{c}} \sqrt{\frac{k(k+1)}{6n}},$$

where:
$Z_{\frac{\alpha}{c}}$ - is the critical value (statistic) of the normal distribution for a given significance level α corrected on the number of possible simple comparisons c.

(ii) The test statistic is defined by:

$$Z = \frac{\sum_{j=1}^{k} c_j R_j}{\sqrt{\frac{k(k+1)}{6n}}},$$

where:
$\overline{R}_j$ – mean of the ranks of the j-th measurement, for $(j = 1, 2, ...k)$,

The test statistic asymptotically (for large sample size) has normal distribution, and the p value is corrected on the number of possible simple comparisons c.

Conover-Inman test

Non-parametric equivalent of Fisher LSD[46], sed for simple comparisons (counts across measurements are always the same).

(i) he value of critical difference is calculated by using the following 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 squares for ranks,

$F_{\alpha,1,df_2}$ to critical value (statistic) Snedecor's F distribution for a given significance level α and for degrees of freedom respectively: 1 and df_2.

(ii) The test statistic is defined by:

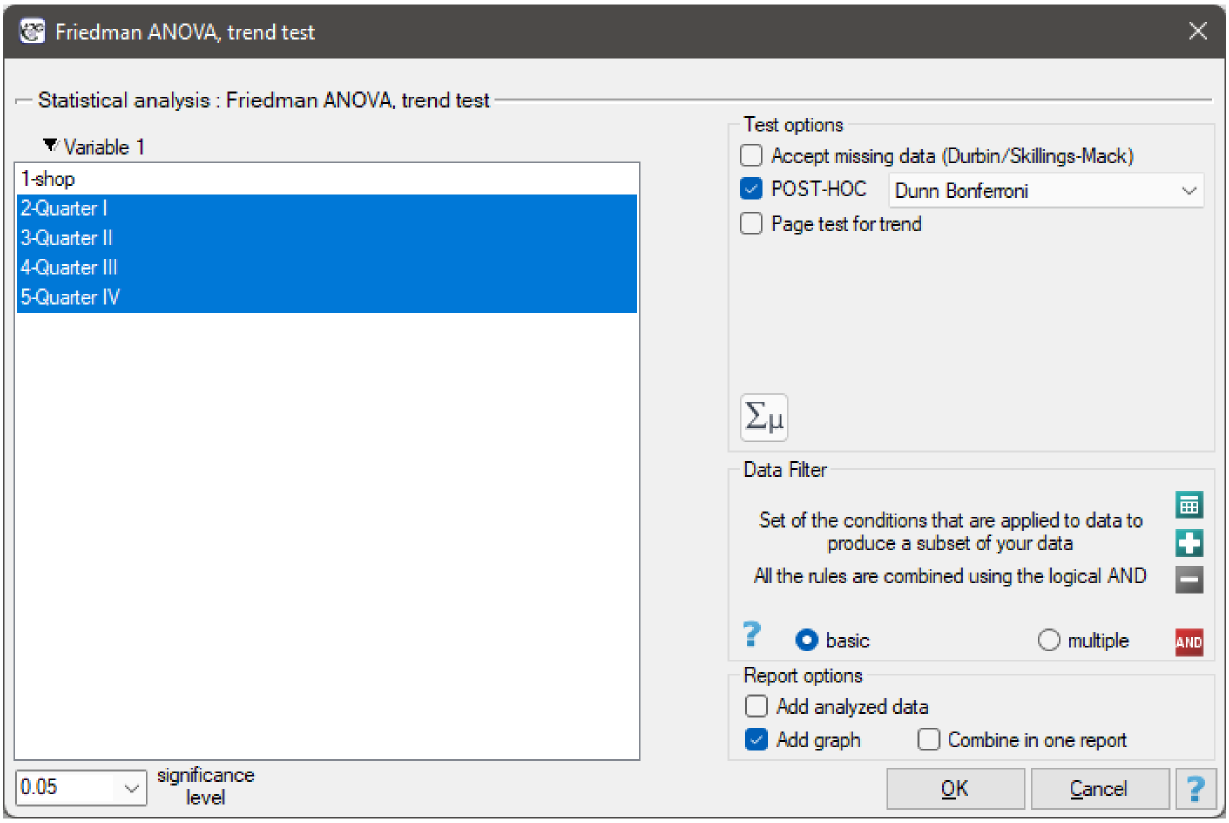
$$t=\frac{\sum_{j=1}^{k}c_jR_j}{\sqrt{\frac{2\left(n_jA-\sum_{j=1}^{t}R_j^k\right)}{(n_j-1)(k-1)}}},$$

where:

R_j – the sum of ranks of jth measurement, for $(j=1,2,...k)$,

The test statistic has$t$-Student distribution with df_2 degrees of freedem.

The settings window with the Friedman ANOVA can be opened in Statistics menu→NonParametric tests →Friedman ANOVA, trend test or in Wizard



Friedman ANOVA, trend test

Statistical analysis : Friedman ANOVA, trend test

Variable 1

1-shop
2-Quarter I
3-Quarter II
4-Quarter III
5-Quarter IV

Test options

□ Accept missing data (Durbin/Skillings-Mack)

☑ POST-HOC Dunn Bonferroni

□ Page test for trend

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data

☑ Add graph □ Combine in one report

0.05 significance level

EXAMPLE 16.6. (chocolate bar.pqs file)

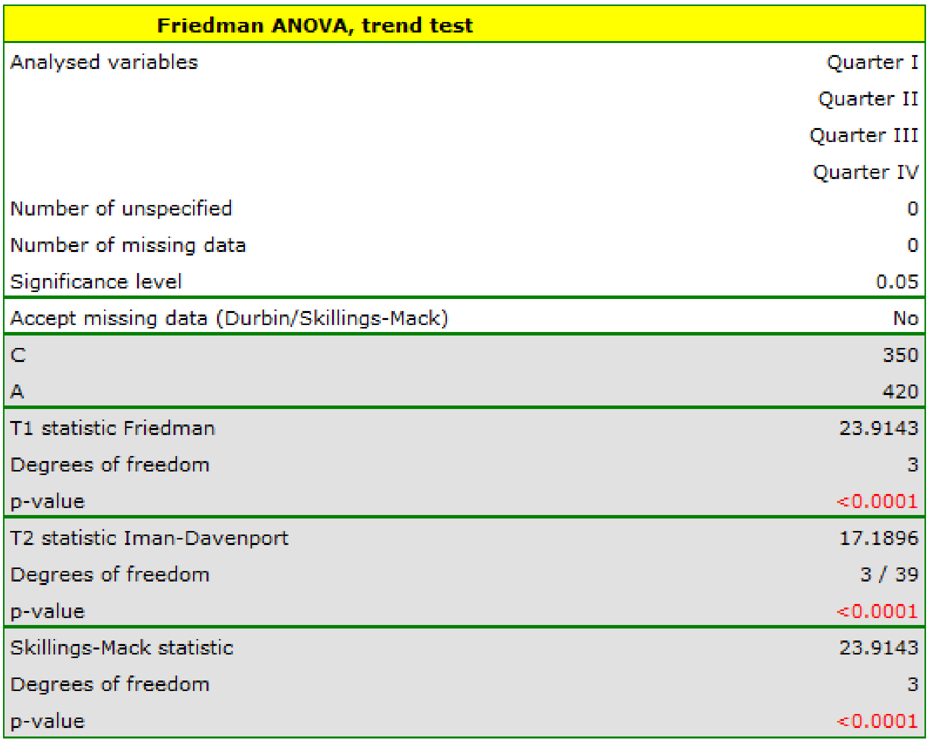
Quarterly sale of some chocolate bar was measured in 14 randomly chosen supermarkets. The study was started in January and finished in December. During the second quarter, the billboard campaign was in full swing. Let's check if the campaign had an influence on the advertised chocolate bar sale.

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

Hypotheses:

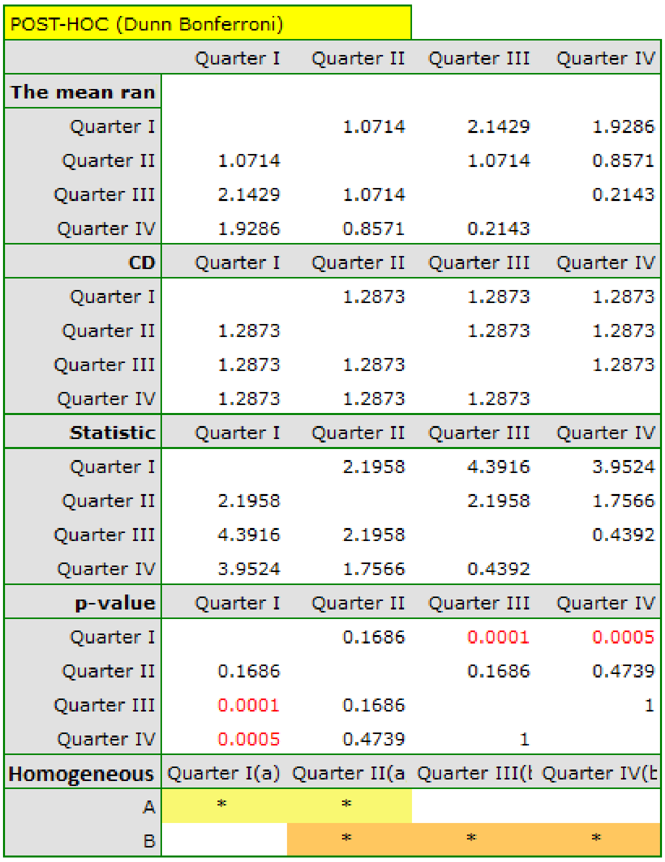
$\mathcal{H}_0$: there is a lack of significant difference in sale values, in the compared quarters, in the population represented by the whole sample,

$\mathcal{H}_1$: the difference in sale values, between at least 2 quarters, is significant, in the population represented by the whole sample.



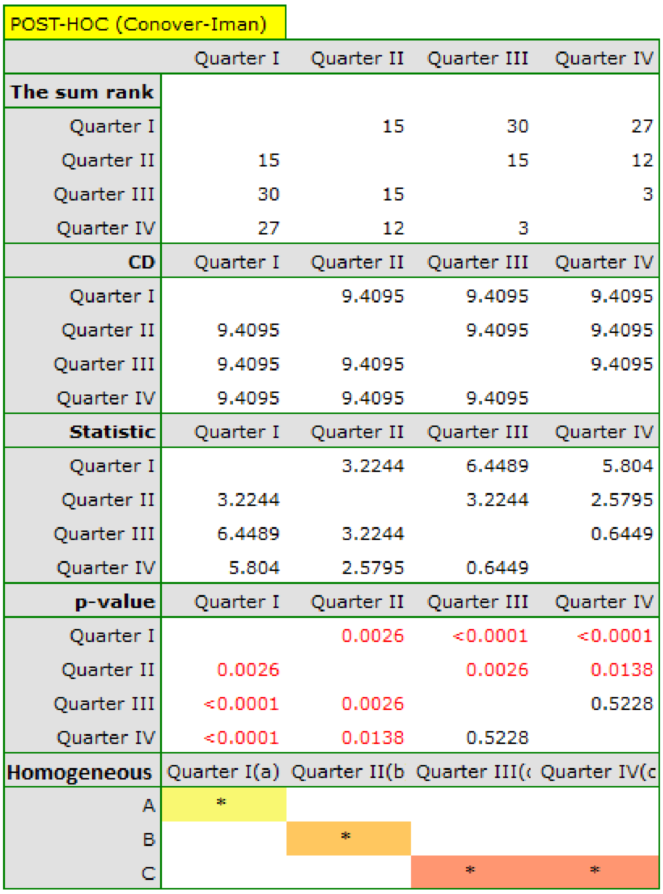
Friedman ANOVA, trend test	
Analysed variables	Quarter I Quarter II Quarter III Quarter IV
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Accept missing data (Durbin/Skillings-Mack)	No
C	350
A	420
T1 statistic Friedman	23.9143
Degrees of freedom	3
p-value	<0.0001
T2 statistic 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

Comparing the p-value of the Friedman test (as well as the p-value of the Iman-Davenport correction of the Friedman test) with a significance level $\alpha = 0.05$, we find that sales of the bar are not the same in each quarter. The POST-HOC Dunn analysis performed with the Bonferroni correction indicates differences in sales volumes pertaining to quarters I and III and I and IV, and an analogous analysis performed with the stronger Conover-Iman test indicates differences between all quarters except quarters III and IV.



POST-HOC (Dunn Bonferroni)

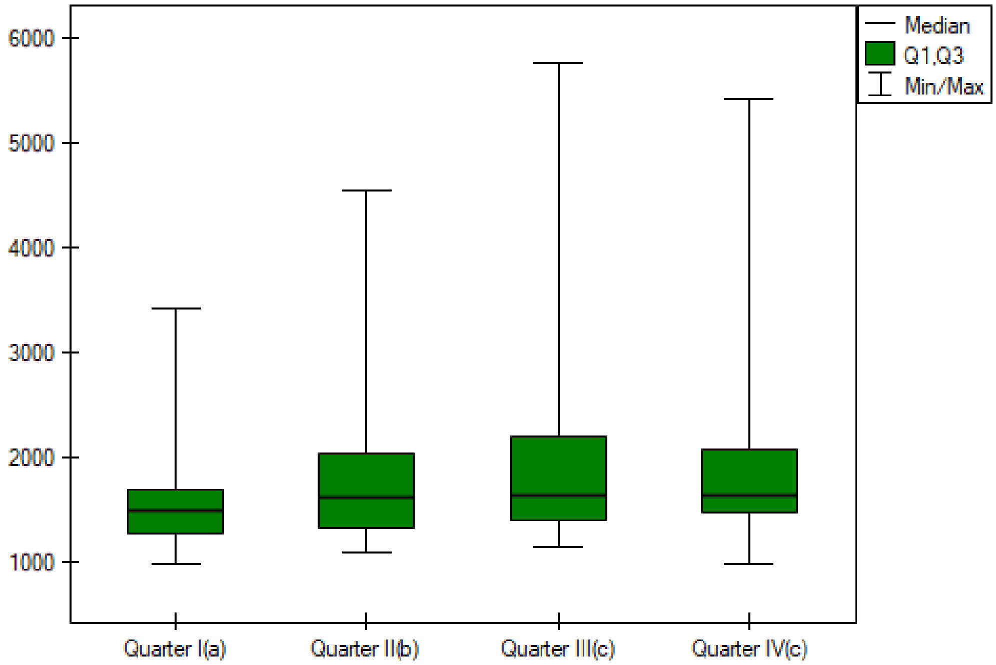
	Quarter I	Quarter II	Quarter III	Quarter IV
The mean ran				
Quarter I		1.0714	2.1429	1.9286
Quarter II	1.0714		1.0714	0.8571
Quarter III	2.1429	1.0714		0.2143
Quarter IV	1.9286	0.8571	0.2143	
CD	Quarter I	Quarter II	Quarter III	Quarter IV
Quarter I		1.2873	1.2873	1.2873
Quarter II	1.2873		1.2873	1.2873
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(	Quarter IV(b
A	*	*		
B		*	*	*



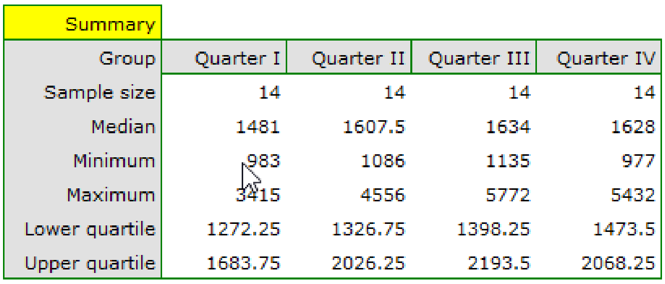
POST-HOC (Conover-Iman)

	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(	Quarter IV(c
A	*			
B		*		
C			*	*

In the graph, we presented homogeneous groups determined by the Conover-Iman test.



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 were described by an ordinal scale with few categories, it would be useful to present it also in numbers and percentages. In our example, this would not be a good method of description.

16.2.5 The Page test for trend

The Page test for ordered alternative described in 1963 by Page E. B. [128] can be computed in the same situation as Friedman's ANOVA, since it is based on the same assumptions. However, Page's test captures the alternative hypothesis differently - indicating that there is a trend in subsequent measurements.

Hypotheses involve equality of the sum of ranks for successive measurements or are simplified to medians:

$$\begin{array}{ll}\mathcal{H}_0: & \theta_1=\theta_2=...=\theta_k,\\ \mathcal{H}_1: & \theta_1\geq\theta_2\geq...\geq\theta_k, \text{ with at least one strict inequality}\end{array}$$

Note
The term: "with at least one strict inequality" written 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

order specified.

The test statistic has the form:

$$Z=\frac{L-\left[\frac{nk(k+1)^2}{4}\right]}{\sqrt{\frac{n(k^3-k)^2}{144(k-1)}}}$$

where:
$L=\sum_{j=1}^{k}R_jc_j$,
R_j – the sum of ranks of jth measurement,
c_j –the weight for j-th measurement informing about the natural order of this measurement among other measurements (weights are consecutive natural numbers).

Note
In order to perform a trend analysis, the expected ordering of measurements must be indicated by assigning consecutive natural numbers to successive measurement groups. These numbers are treated as weights in the analysis c_1, c_2, ..., c_k.

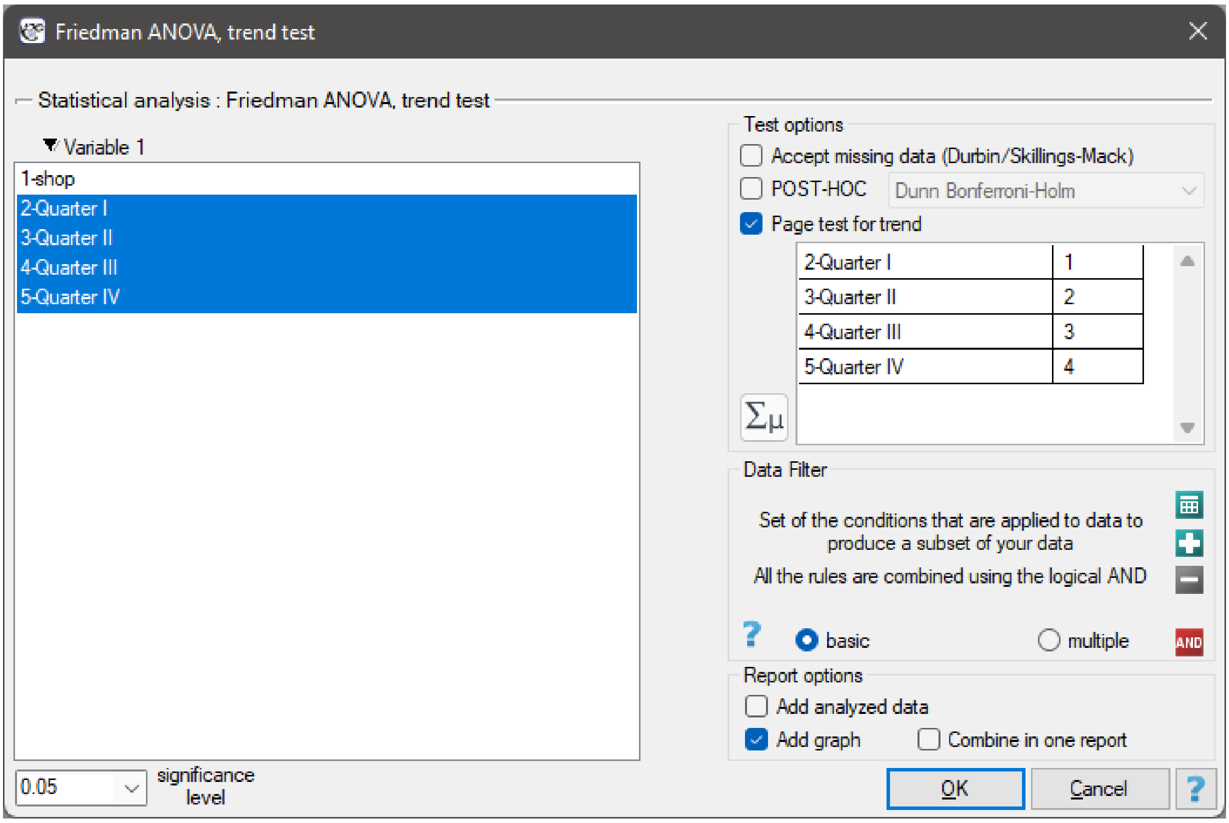
The formula for the test statistic Z does not include a correction for ties, making it somewhat more conservative when tied ranks are present. However, using a correction for tied ranks for this test is not recommended.

The statistic Z has asymptotically (for large sample) 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 a two-sided p-value means that the researcher does not know (does not assume) the direction of the possible trend.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the Page test for trend can be opened in Statistics menu→NonParametric tests →Friedman ANOVA, trend test or in Wizard



Friedman ANOVA, trend test

Statistical analysis : Friedman ANOVA, trend test

Variable 1

- 1-shop
- 2-Quarter I (selected)
- 3-Quarter II (selected)
- 4-Quarter III (selected)
- 5-Quarter IV (selected)

0.05 significance level

Test options

- □ Accept missing data (Durbin/Skillings-Mack)
- □ POST-HOC Dunn Bonferroni-Holm
- ☑ Page test for trend

2-Quarter I	1
3-Quarter II	2
4-Quarter III	3
5-Quarter IV	4

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

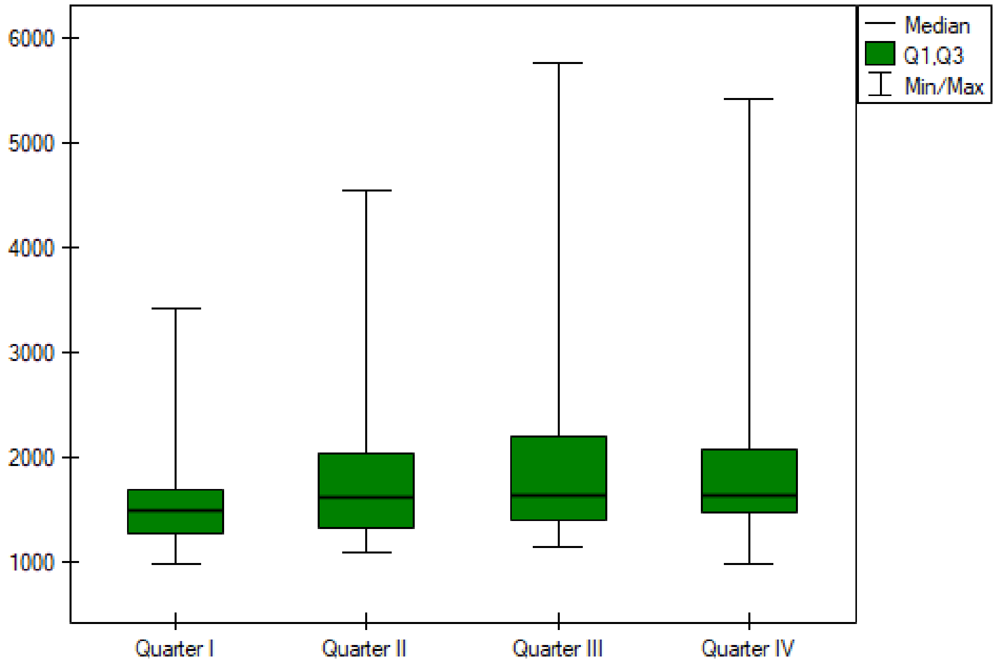
- □ Add analyzed data
- ☑ Add graph
- □ Combine in one report

EXAMPLE (16.6) continued *(chocolate bar.pqs file)* The expected result of the intensive advertising campaign conducted by the company is a steady increase in sales of the offered bar.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \text{no indicated trend in bar sales,} \\ \mathcal{H}_1: & \text{there is an indicated trend in bar sales.} \end{array}$$

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



Comparing a one-sided $p < 0.0001$ with a significance level $\alpha = 0.05$, we find that the campaign produced the expected trend of increased product sales.

16.2.6 The Durbin's ANOVA (missing data)

Durbin's analysis of variance of repeated measurements for ranks was proposed by Durbin (1951)[52]. This test is used when measurements of the variable under study are made several times – a similar situation in which Friedman'sANOVA is used. The original Durbin test and the Friedman test give the same result when we have a complete data set. However, Durbin's test has an advantage – it can also be calculated for an incomplete data set. At the same time, data deficiencies cannot be located arbitrarily, but the data must form a so-called balanced and incomplete block:

- the number of measurements for each object is k ($k \leq t$),
- each measurement is made on r objects ($r \leq b$),
- the number of objects for which the same pair of measurements was taken simultaneously is constant and equal to λ.

  where:
  t – total number of considered measurements,
  b – total number of examined objects.

Basic assumptions:

- measurement on an ordinal scale or on an interval scale,
- a dependent model.

Hypotheses involve equality of the sum of ranks for successive measurements (R_j) or are simplified to medians (θ_j):

$$\begin{array}{cl}\mathcal{H}_0: & \theta_1 = \theta_2 = ... = \theta_k,\\ \mathcal{H}_1: & \text{not all } \theta_j \text{ are equal } (j = 1, 2, ..., k),\end{array}$$

Two test statistics of the following form 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 of the studied objects $(i = 1, 2, ...b)$,
$A = \sum_{i=1}^{b}\sum_{j=1}^{t} R_{ij}^2$ – sum of squared ranks,
$C = \frac{bk(k+1)^2}{4}$ – correction coefficient.

The formula for T_1 and T_2 statistics includes a correction for tied ranks.

For complete data, the T_1 statistic is the same as the Friedman test. It has asymptotically (for large sample sizes) χ^2 distribution with $df = t-1$ degrees of freedom.

The T_2 statistic is the equivalent of Friedman's Iman-Davenport ANOVA adjustment, so it follows Snedecor's F distribution with $df_1 = t-1$ i $df_2 = bk-b-t+1$ degrees of freedom. It is now considered to be more precise than the T_1 statistic and is recommended for use with the T_1 statistic[46].

The p value, designated on the basis of the test statistic, is compared with the significance level α:

if $p \le \alpha$ $\Longrightarrow$ reject $\mathcal{H}_0$ and accept $\mathcal{H}_1$,
if $p > \alpha$ $\Longrightarrow$ there is no reason to reject $\mathcal{H}_0$.

Testy POST-HOC

Introduction to **the contrasts and the POST-HOC tests** was performed in the 16.1.2 unit, which relates to the one-way analysis of variance.

Conover-Inman test

Used for simple comparisons (the counts in each measurement are always the same).

Hypotheses:

Example - **simple comparisons** (comparing 2 selected medians / rank sums between each other):

$$\mathcal{H}_0: \quad \theta_j = \theta_{j+1},$$
$$\mathcal{H}_1: \quad \theta_j \neq \theta_{j+1}.$$

(i) The value of critical difference is calculated by using the following formula:

$$CD = t_{1-\alpha/2, bk-b-t+1}\sqrt{\frac{(A-C)2r}{bk-b-t+1}\left(1-\frac{T_1}{b(k-1)}\right)},$$

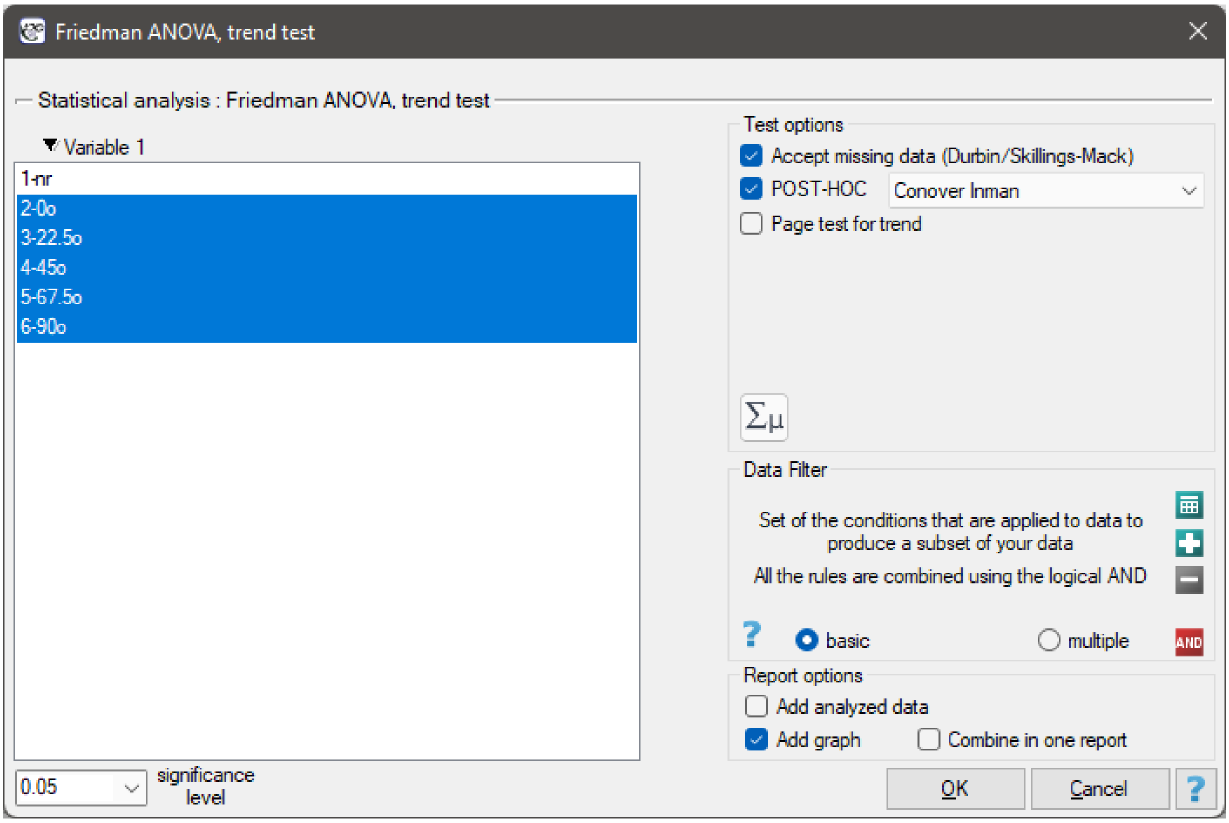
where:
$t_{1-\alpha/2,bk-b-t+1}$ – is the critical value (statistic) of the t-Student distribution for a given significance level α and $df = bk - b - t + 1$ degrees of freedom.

(ii) The test statistic has the form:

$$t = \frac{\sum_{j=1}^{k} c_j R_j}{\sqrt{\frac{(A-C)2r}{bk-b-t+1}\left(1-\frac{T_1}{b(k-1)}\right)}},$$

The test statistic has t-Student distribution with $df = bk - b - t + 1$ degrees of freedom.

The settings window with the Durbin's ANOVA can be opened in Statistics menu→NonParametric tests →Friedman ANOVA, trend test or in Wizard


Friedman ANOVA, trend test

Statistical analysis : Friedman ANOVA, trend test

Variable 1

1-nr
2-0o
3-22.5o
4-45o
5-67.5o
6-90o

Test options
☑ Accept missing data (Durbin/Skillings-Mack)
☑ POST-HOC Conover Inman
☐ Page test for trend

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
☐ Add analyzed data
☑ Add graph ☐ Combine in one report

0.05 significance level

Note
For records with missing data to be taken into account, you must check the Accept missing data option. Empty cells and cells with non-numeric values are treated as missing data. Only records with more than one numeric value will be analyzed.

Example 16.7. (mirror.pqs file)
An experiment was conducted among 20 patients in a psychiatric hospital (Ogilvie 1965)[125]. This experiment involved drawing straight lines according to a presented pattern. The pattern represented 5 lines drawn at different angles ($0^o, 22.5^o, 45^o, 67.5^o, 90^o$) relative to the indicated center. The patients' task was to reproduce the lines while having their hand covered. The time at which the patient drew the line was recorded as the result of the experiment. Ideally, each patient would draw a line from all angles,

but elapsed time and fatigue would have a significant impact on performance. In addition, it is difficult to keep the patient interested and willing to cooperate for an extended period of time. Therefore, the project was planned and conducted in balanced and incomplete blocks. Each of the 20 patients traced a line at two angles (there were five possible angles). Thus, each angle was drawn eight times. The time at which each patient drew a line at a given angle was recorded in the table.

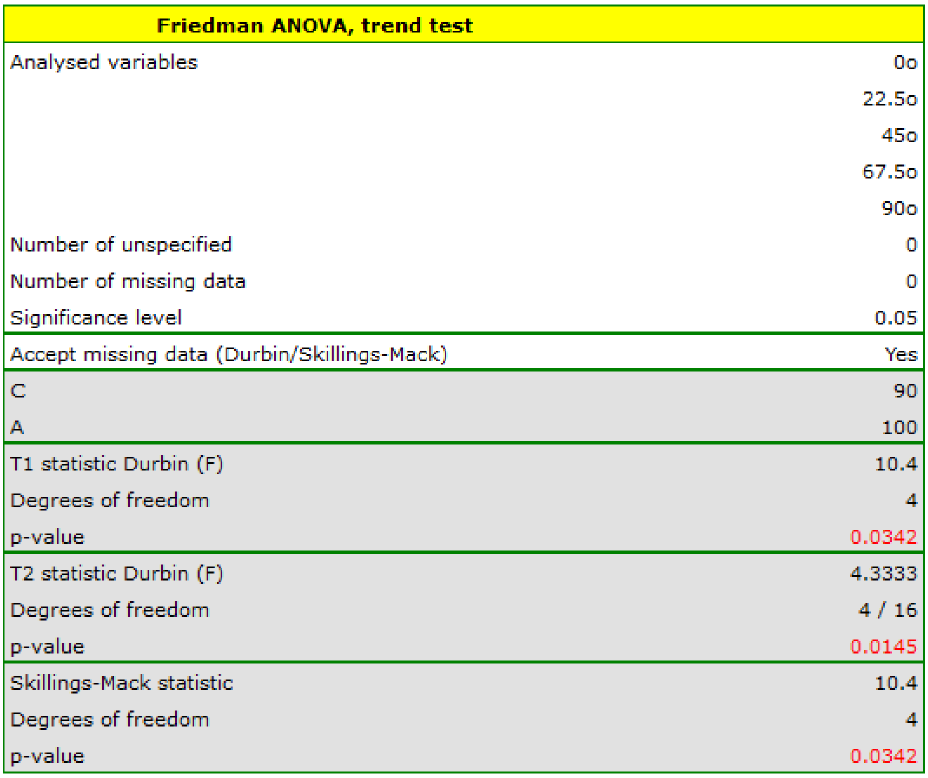
patient number	0^o	22.5^o	45^o	67.5^o	90^o
1	7	15			
2	20		72		
3	8			26	
4	33				36
5	7	16			
6		68	67		
7		33		64	
8		34			12
9	10		96		
10		29	59		
11			17	9	
12			100		15
13	16			32	
14		19		32	
15			36	39	
16				44	54
17	16				38
18		17			12
19			37		11
20				56	6

We want to see if the time taken to draw each line is completely random, or if there are lines that took more or less time to draw.

Hypotheses:

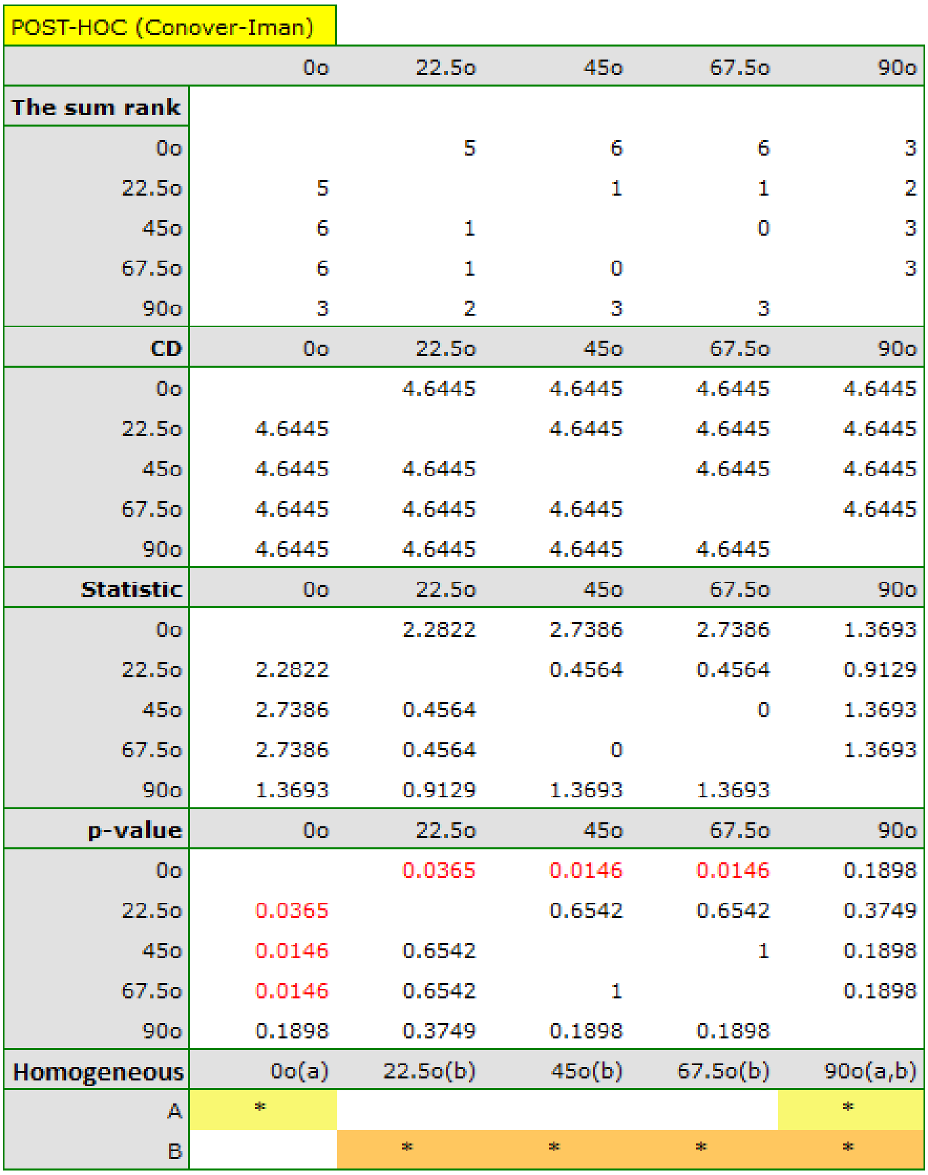
$\mathcal{H}_0$: there is no significant difference between the time taken by patients to draw each line,

$\mathcal{H}_1$: at least one line is drawn in shorter/longer time.



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

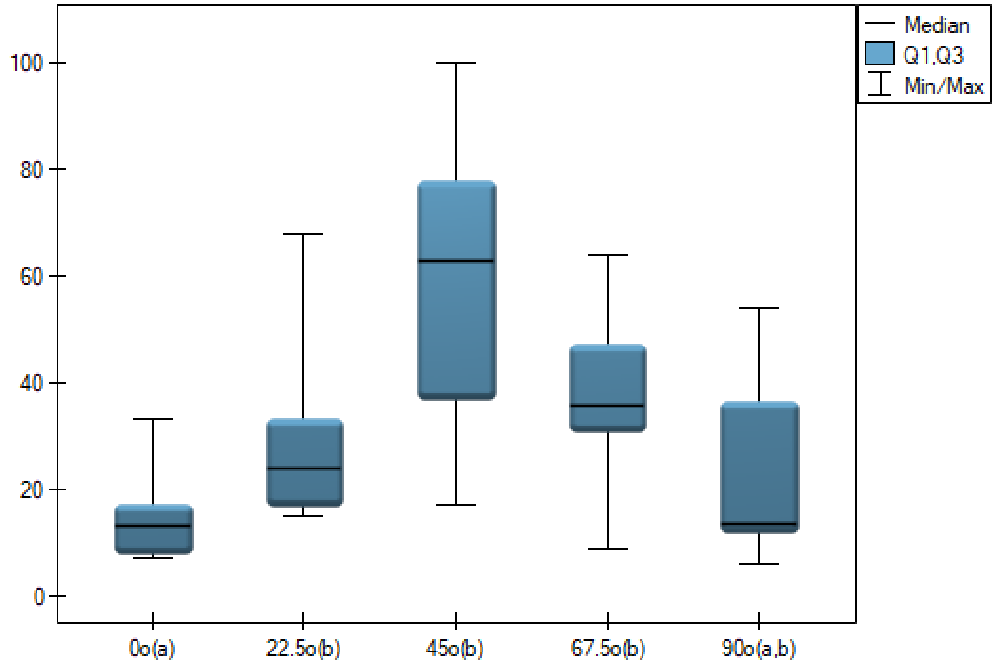
Comparing the $p = 0.0145$ for the T_2 statistic (or the $p = 0.0342$ for the T_1 statistic) with the $\alpha = 0.05$ significance level, we find that the lines are not drawn at the same time. The POST-HOC analysis performed indicates that there is a difference in the time taken to draw the line at angle 0^o. It is drawn faster than the lines at the angle of 22.5^o, 45^o and 67.5^o.



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 homogeneous groups indicated by the post-hoc test.



16.2.7 The Skillings-Mack ANOVA (missing data)

The analysis of variance of repeated measures for Skillings-Mack ranks was proposed by Skillings and Mack in 1981 [152]. t is a test that can be used when there are missing data, but the missing data need not occur in any particular setting. However, each site must have at least two observations. If there are no tied ranks and no gaps are present it is the same as the Friedman's ANOVA, and if data gaps are present in a balanced arrangement it corresponds to the results of Durbin's ANOVA.

Basic assumptions: Basic assumptions:

- measurement on an ordinal scale or on an interval scale,
- a dependent model.

Hypotheses relate to the equality of the sum of ranks for successive measurements (R_j) or are simplified to medians ($theta_j$)

$$\begin{array}{cl}\mathcal{H}_0: & \theta_1=\theta_2=...=\theta_k,\\ \mathcal{H}_1: & \text{nie wszystkie } \theta_j \text{ są sobie równe } (j=1,2,...,k),\end{array}$$

The test statistic has the form:

$$\chi^2 = A\Sigma_0^{-1}A^T$$

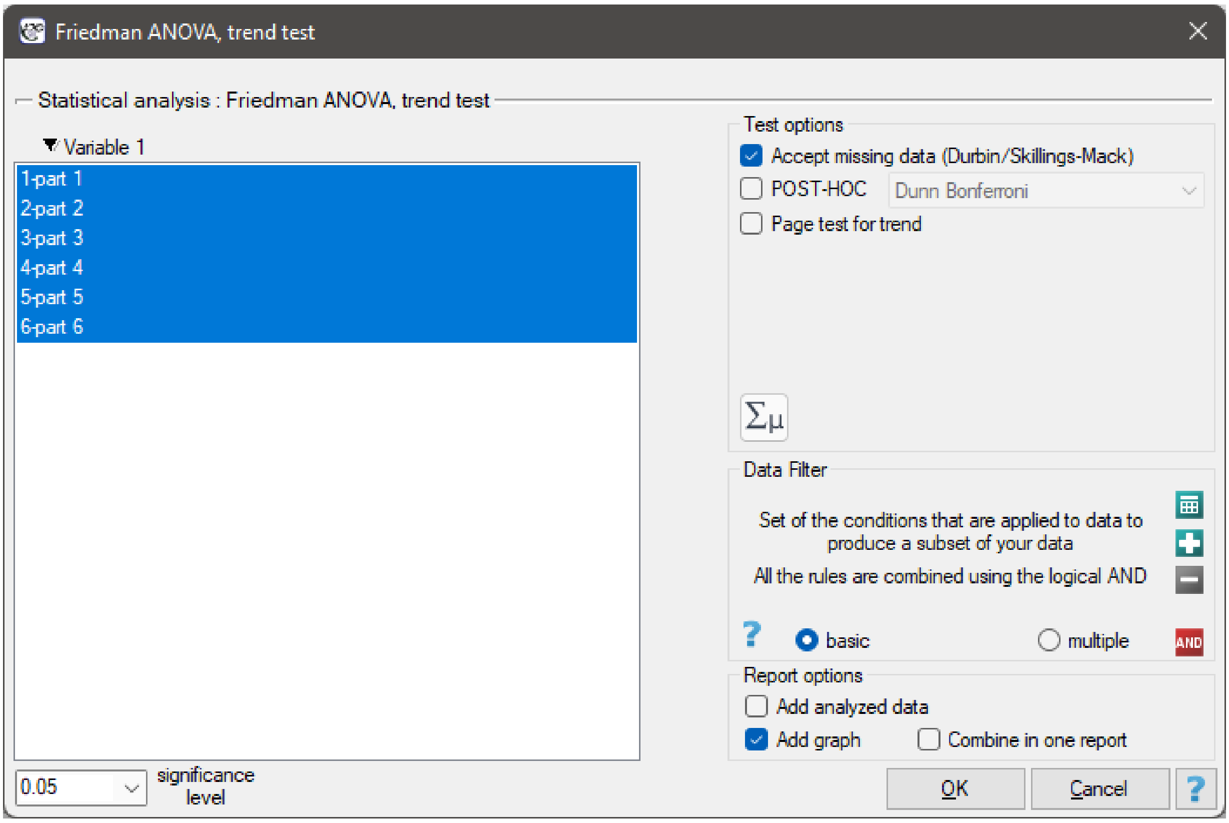
where:
$A = (A_1, A_2, ..., A_{k-1}$
$A_j = \sum_{i=1}^n \sqrt{\frac{12}{s_i+1}}\left(R_{ij} - \frac{s_i+1}{2}\right)$,
s_i – number of observations for i-th object,
R_{ij} – ranks assigned to successive measurements ($j = 1, 2, ...k$), separately for each study object ($i = 1, 2, ...n$), with ranks for missing data equal to the average rank for the object,
Σ_0 – matrix determining the covariances for A at the truth of $\mathcal{H}_0$[152].

When each pair of measurements occurs simultaneously for at least one observation, this statistic has asymptotically (for large sample sizes) the χ^2 distribution with $k-1$ degrees of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}
\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\
\text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.
\end{array}$$

The settings window with the Skillings-Mack ANOVA can be opened in Statistics menu→NonParametric tests →Friedman ANOVA, trend test or in Wizard


Friedman ANOVA, trend test
Statistical analysis : Friedman ANOVA, trend test
Σμ
OK
Cancel
?
AND

Variable 1: 1-part 1, 2-part 2, 3-part 3, 4-part 4, 5-part 5, 6-part 6

Test options
☑ Accept missing data (Durbin/Skillings-Mack)
□ POST-HOC Dunn Bonferroni
□ Page test for trend

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
□ Add analyzed data
☑ Add graph □ Combine in one report

0.05 significance level

Note

For records with missing data to be taken into account, you must check the Accept missing data option. Empty cells and cells with non-numeric values are treated as missing data. Only records containing more than one numeric value will be analyzed.

Example 16.8. (polling.pqs file)

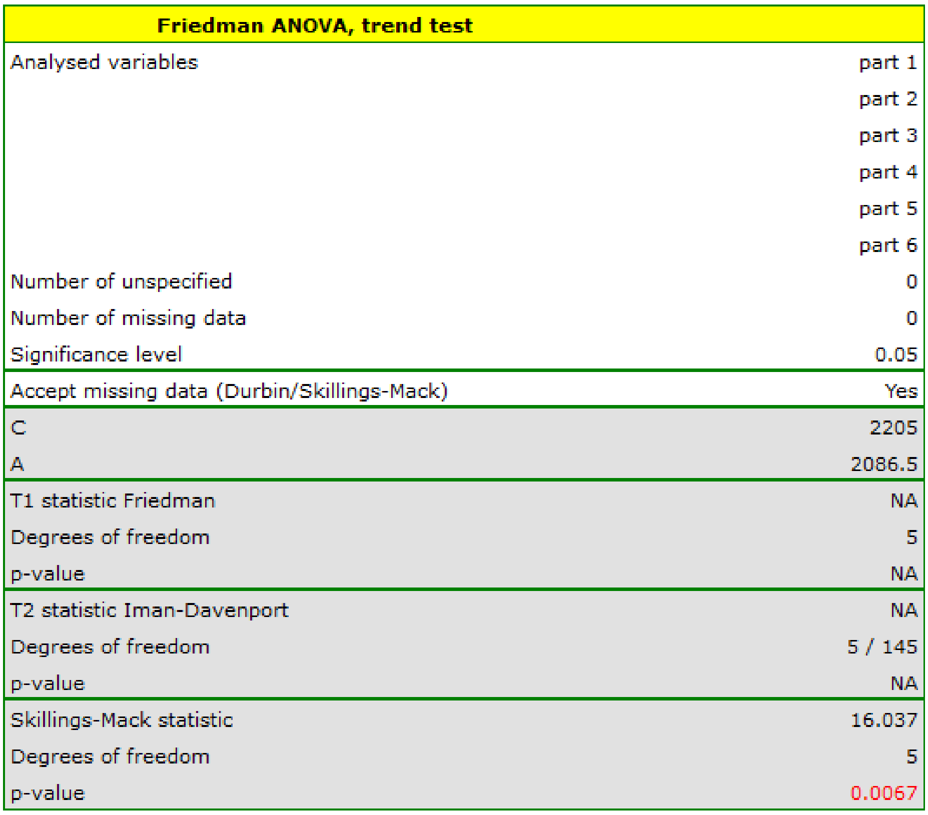
A certain university teacher, wanting to improve the way he conducted his classes, decided to verify his teaching skills. In several randomly selected student groups, during the last class, he asked them to fill in a short anonymous questionnaire. The survey consisted of six questions about how the six specified parts of the material were illustrated. The students could rate it on a 5-point scale, where 1 - the way of presenting the material was completely incomprehensible, 5 - a very clear and interesting way of illustrating the material. The data obtained in this way turned out to be incomplete due to the fact that students did not answer questions about the part of the material they were absent on. In the 30-person group completing the survey, only 15 students provided complete responses. Performing an analysis that does not account for data gaps (in this case, a Friedman analysis) will have limited power by cutting the group size so drastically and will not lead to the detection of significant differences. Data gaps were not planned for and are not present in the balanced block, so this task cannot be performed using Durbin's analysis along with his POST-HOC test.

Hypotheses:

$\mathcal{H}_0$: there is no significant difference in the evaluations of the different parts of the material by studentów,

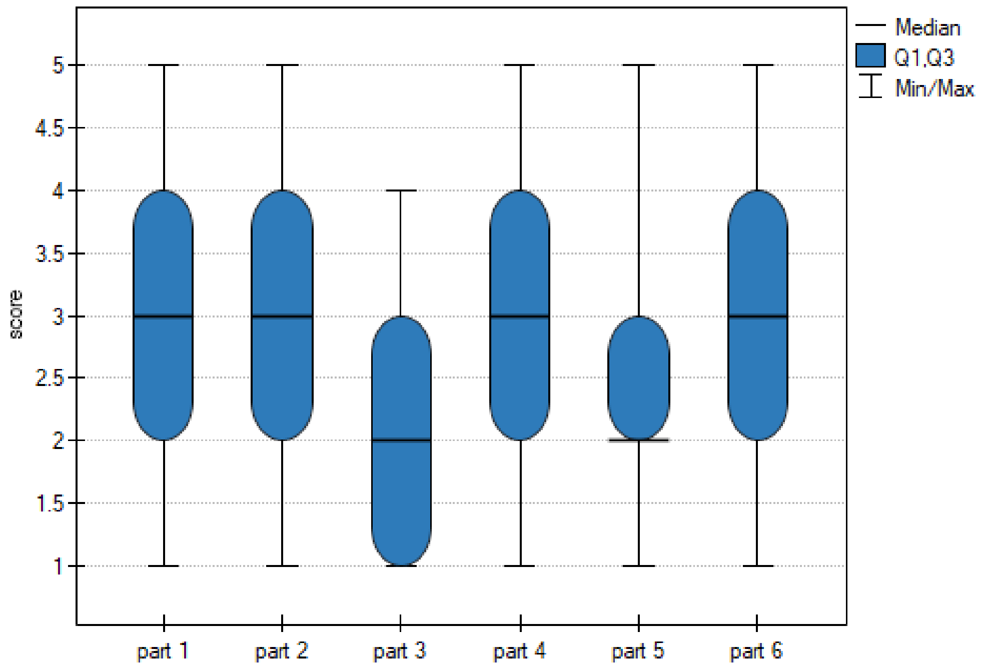
$\mathcal{H}_1$: at least one part of the material is assessed differently by students.

The results of the ANOVA Skillings-Mack analysis are presented in the following report:



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

The p value obtained should be treated with caution due to possible tied ranks. However, for this study, the $p = 0.0067$ is well below the accepted significance level of $\alpha = 0.05$, indicating that significant differences exist. The differences in responses can be observed in the graph; however, there is no POST-HOC analysis available for this test.



16.2.8 The Chi-square test for multidimensional contingency tables

The χ^2 test for multidimensional contingency tables is an extension to the χ^2 test for $(R \times C)$ tables for more than two features.

Basic assumptions:

- measurement on a nominal scale - any order is not taken into account,
- an independent model,
- large expected frequencies (according to the Cochran interpretation (1952)[40].

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & O_{ij...}=E_{ij...} \text{ for all categories,}\\ \mathcal{H}_1: & O_{ij...}\neq E_{ij...} \text{ for at least one category,}\end{array}$$

where:
$O_{ij...}$ and $E_{ij...}$ — observed frequencies in a contingency table and the corresponding expected frequencies.

The test statistic is defined by:

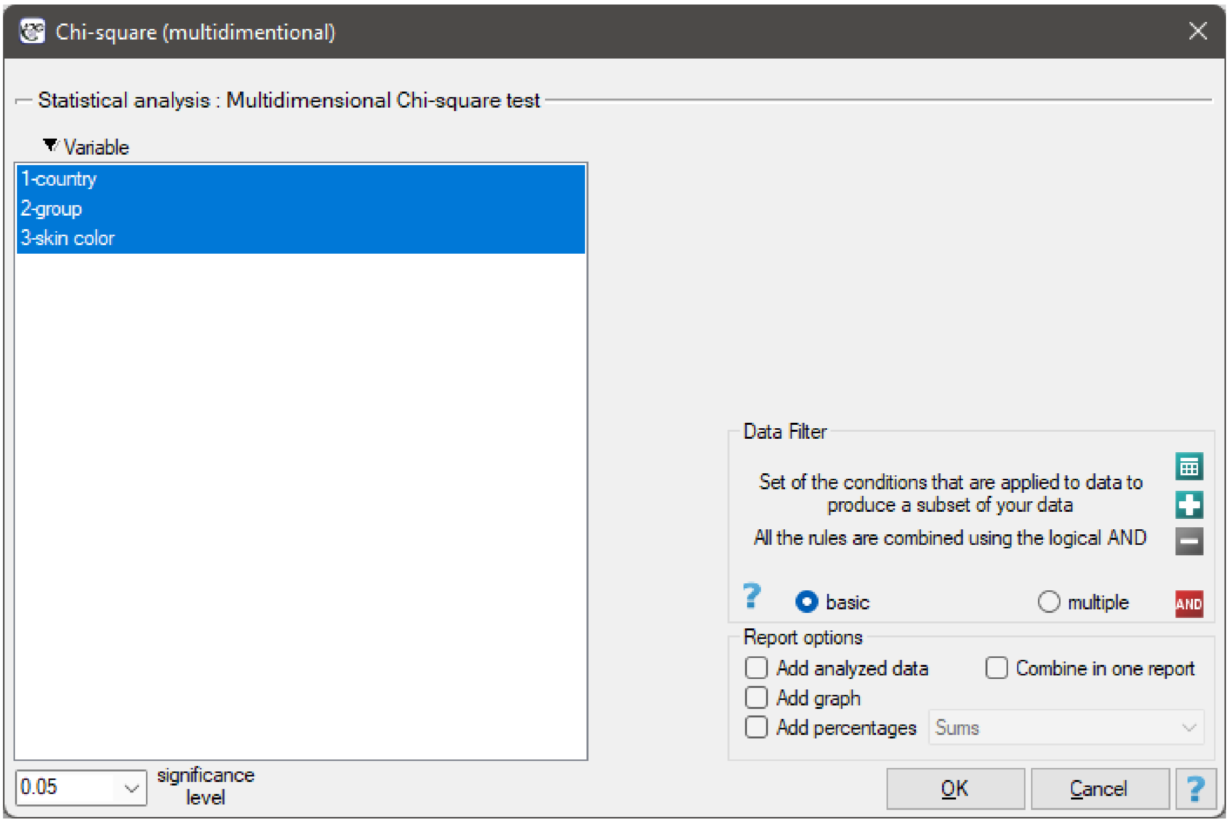
$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\sum\cdots\sum\frac{(O_{ij...}-E_{ij...})^2}{E_{ij...}}.$$

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with a number of degrees of freedom calculated using the formula: $df=(r-l)(c-1)(l-1)+(r-l)(c-1)+(r-1)(l-1)+(c-1)(l-1)$ - for 3-dimensional tables.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the Chi-square (multidimensional) test can be opened in Statistics menu → NonParametric tests (unordered categories)→Chi-square (multidimensional) or in Wizard.


Chi-square (multidimentional)

Statistical analysis : Multidimensional Chi-square test

Variable
1-country
2-group
3-skin color

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
□ Add analyzed data □ Combine in one report
□ Add graph
□ Add percentages Sums

0.05 significance level

Note
This test can be calculated only on the basis of raw data.

16.2.9 The Q-Cochran ANOVA

The Q-Cochran analysis of variance, based on the Q-Cochran test, is described by Cochran (1950)[39]. This test is an extended McNemar test for $k \ge 2$ dependent groups. It is used in hypothesis verification about symmetry between several measurements $X^{(1)}, X^{(2)}, ..., X^{(k)}$ for the X feature. The analysed feature can have only 2 values - for the analysis, there are ascribed to them the numbers: 1 and 0.

Basic assumptions:

- measurement on a nominal scale (dichotomous variables – it means the variables of two categories),
- a dependent model.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \text{all the "incompatible" observed frequencies are equal,} \\ \mathcal{H}_1: & \text{not all the "incompatible" observed frequencies are equal,} \end{array}$$

where:
"incompatible" observed frequencies – the observed frequencies calculated when the value of the analysed feature is different in several measurements.

The test statistic is defined by:

$$Q = \frac{(k-1)\left(kC - T^2\right)}{kT - R}$$

where:
$T = \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}$,
$R = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} x_{ij}\right)^2$,
$C = \sum_{j=1}^{k} \left(\sum_{i=1}^{n} x_{ij}\right)^2$,
x_{ij} – the value of j-th measurement for i-th object (so 0 or 1).

This statistic asymptotically (for large sample size) has the χ^2 distribution with a number of degrees of freedom calculated using the formula: $df = k - 1$.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The POST-HOC tests

Introduction to **the contrasts and the POST-HOC tests** was performed in the 16.1.2 unit, which relates to the one-way analysis of variance.

The Dunn test

For simple comparisons (frequency in particular measurements is always the same).

Hypotheses:

Example - **simple comparisons** (for the difference in proportion in a one chosen pair of measurements):

$$\begin{array}{cl} \mathcal{H}_0: & \text{the chosen "incompatible" observed frequencies are equal,} \\ \mathcal{H}_1: & \text{the chosen "incompatible" observed frequencies are different.} \end{array}$$

(i) The value of critical difference is calculated by using the following 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 a given significance level α corrected on the number of possible simple comparisons c.

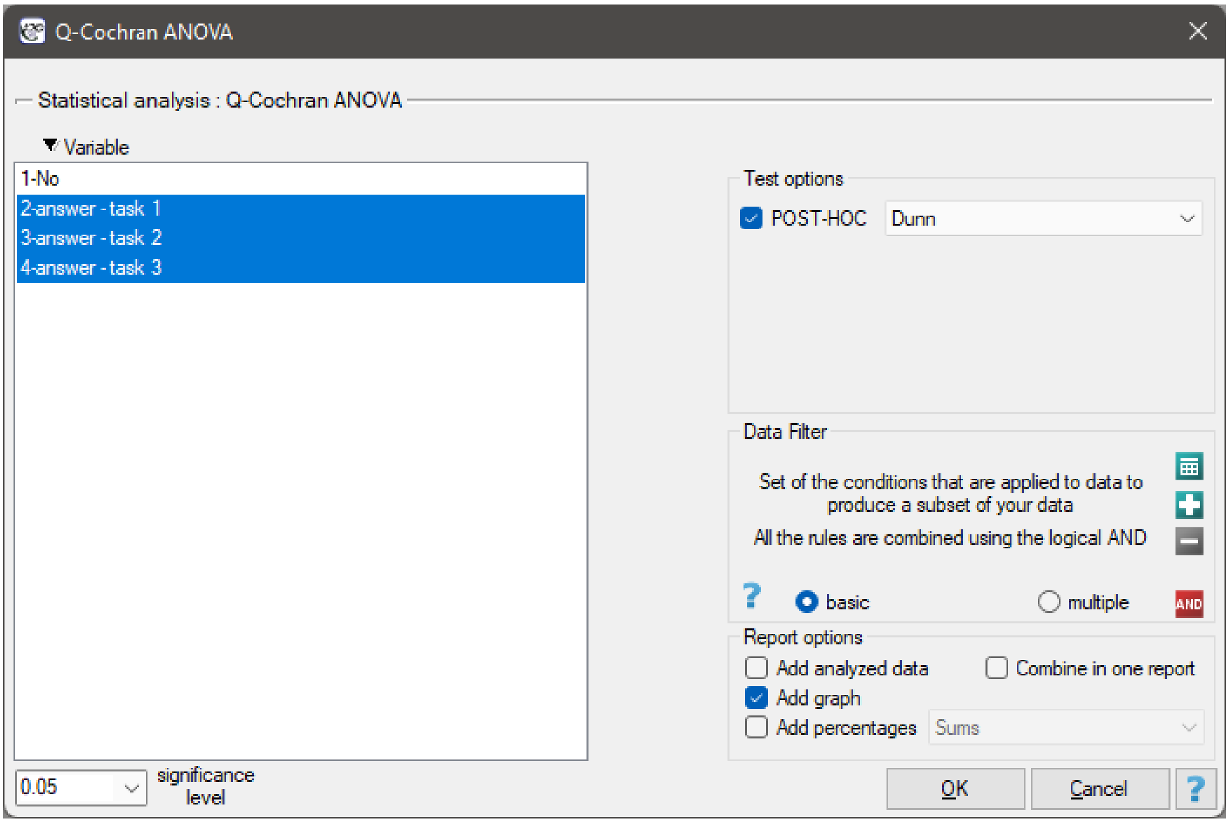
(ii) The test statistic is defined by:

$$Z = \frac{\sum_{j=1}^{k} c_j p_j}{\sqrt{2\frac{kT-R}{n^2 k(k-1)}}},$$

where:
p_j – the proportion j-th measurement $(j = 1, 2, ...k)$,

The test statistic asymptotically (for large sample size) has the normal distribution, and the p value is corrected on the number of possible simple comparisons c.

The settings window with the Cochran Q ANOVA can be opened in Statistics menu→ NonParametric tests→Cochran Q ANOVA or in Wizard.



Q-Cochran ANOVA

Statistical analysis : Q-Cochran ANOVA

Variable: 1-No; 2-answer - task 1 (selected); 3-answer - task 2 (selected); 4-answer - task 3 (selected)

Test options: ☑ POST-HOC Dunn

Data Filter: Set of the conditions that are applied to data to produce a subset of your data. All the rules are combined using the logical AND. ● basic ○ multiple

Report options: □ Add analyzed data □ Combine in one report ☑ Add graph □ Add percentages Sums

0.05 significance level

Note
This test can be calculated only on the basis of raw data.

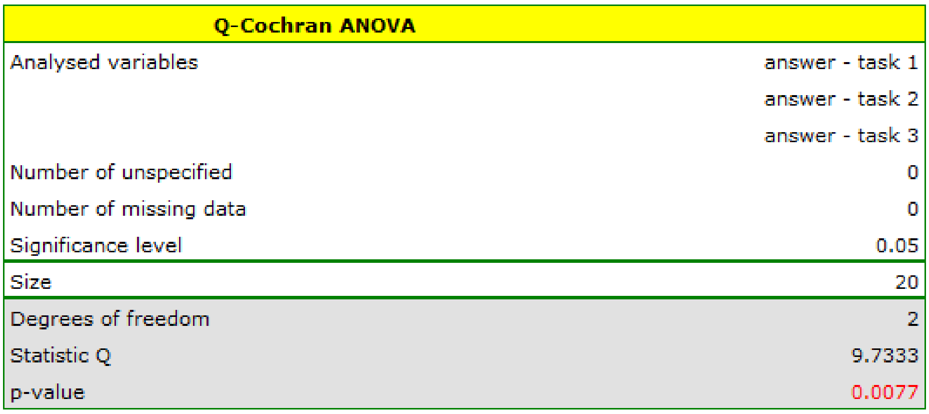
Example 16.9. (test.pqs file)
We want to compare the difficulty of 3 test questions. To do this, we select a sample of 20 people from the analysed population. Every person from the sample answers 3 test questions. Next, we check the correctness of answers (an answer can be correct or wrong). In the table, there are following scores:

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

Hypotheses:

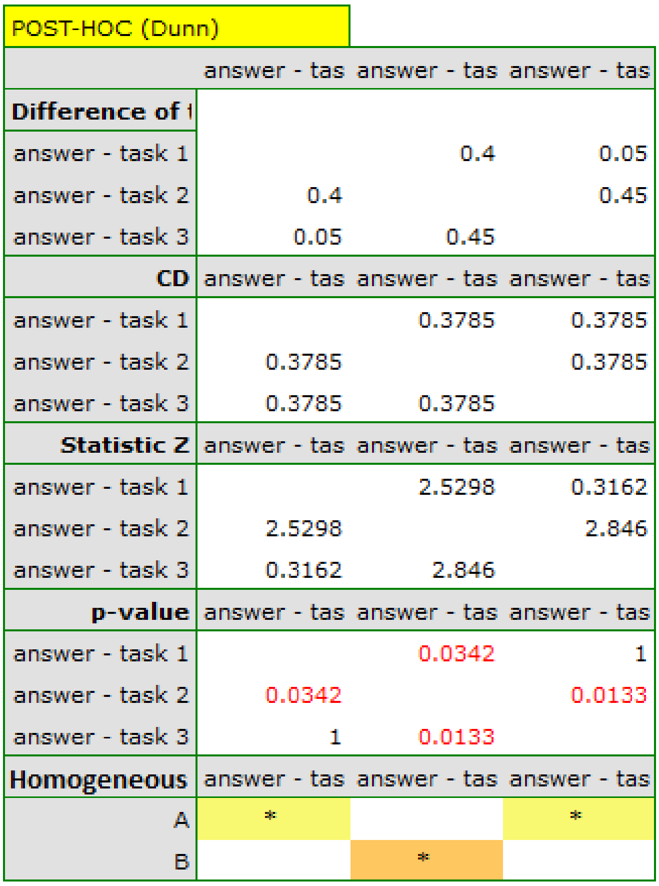
$\mathcal{H}_0$: The individual questions received the same number of correct answers, in the analysed population,

$\mathcal{H}_1$: There are different numbers of correct and wrong answers in individual test questions, in the analysed population.



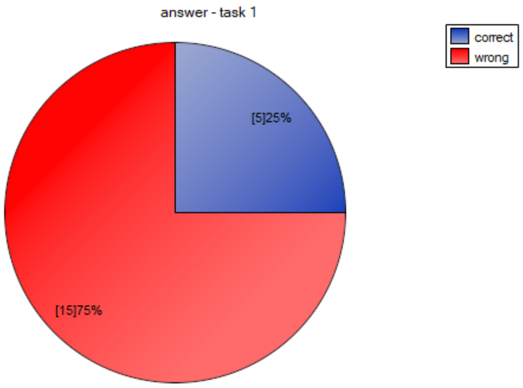
Q-Cochran ANOVA	
Analysed variables	answer - task 1 answer - task 2 answer - task 3
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	20
Degrees of freedom	2
Statistic Q	9.7333
p-value	0.0077

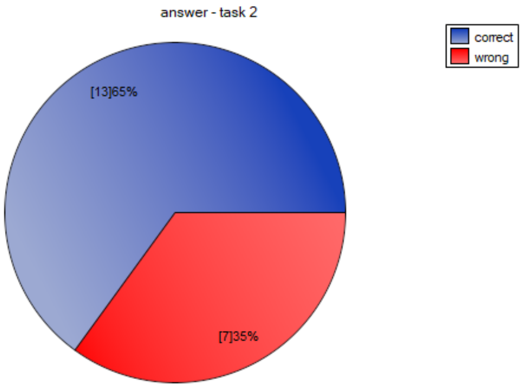
Comparing the p value $p = 0.0077$ with the significance level $\alpha = 0.05$ we conclude that individual test questions have different difficulty levels. We resume the analysis to perform POST-HOC test by clicking Run the recent test, and in the test option window, we select POST-HOC Dunn.

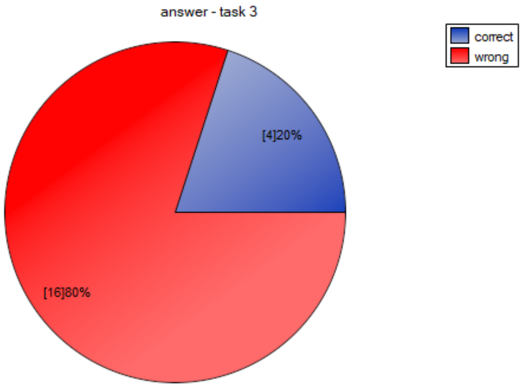


POST-HOC (Dunn)

	answer - tas	answer - tas	answer - tas
Difference of t			
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 carried out POST-HOC analysis indicates that there are differences between the 2-nd and 1-st question and between questions 2-nd and 3-th. The difference is because the second question is easier than the first and the third ones (the number of correct answers the first question is higher).

17 Multicomparisons

Simultaneous testing of multiple hypotheses (the so-called **family of hypotheses**) carries the risk of increasing the error α, which is a major problem in the multicomparison field. When the error α increases, it means that the null hypothesis is too often rejected when it is true. That is, we too often indicate that there are differences, when in fact there are none. To protect against an increase in α one strategy is to adjust (decrease) the level of α or to adjust (increase) the p-values of the tests accordingly. The best known correction is the Bonferroni[2] correction, which is also the most conservative. The Sidak (1967)[150] correction is somewhat more liberal. Both corrections have received several sequential improvements that increase their power. The program uses the step-up Holm's (1979)[79] procedure and the step-down Hochberg's (1988[78] procedure. The most powerful among the proposed corrections is Benjamini's (1995)[19] modified Hochberg's procedure, which does not directly control the error α but minimizes the expected percentage of false differences that occur among detected differences.

If we indicate by c the number of hypotheses being tested, then the adjustments to multicomparisons can be described as follows:

- **Bonferroni's correction**
  It involves multiplying each test probability by the total number of tests performed (or dividing the significance level by that number).

  The p-value adjustment:
  $$p_{(Bonferroni,i)} = p_i \cdot c$$
  The significance level α adjustment:
  $$\alpha_{(Bonferroni,i)} = \frac{\alpha_i}{c}$$

- **Sidak's correction**
  This correction is more powerful than the Bonferroni's correction (therefore it is becoming increasingly popular).

  The p-value adjustment:
  $$p_{(Sidak,i)} = 1-(1-p_i)^c$$
  The significance level α adjustment::
  $$\alpha_{(Sidak,i)} = 1-(1-\alpha_i)^{1/c}$$

- **Bonferroni-Holm's correction**
  It involves using Holm's multi-step procedure for the Bonferroni's correction. The procedure begins with sorting consecutive values of p_i $(p_1, p_2, ...p_c)$ in an ascending order. Bonferroni's correction is then applied to each successive value of p_i (with a corresponding reduction in the number of hypotheses left to test). As a result, all hypotheses that are tested after the first statistically insignificant value of p_i are also insignificant.

  The p-value adjustment:
  $$p_{(Bonferroni,i)} = p_i \cdot (c-i+1)$$

- **Sidak-Holm's correction**
  It involves using Holm's multi-step procedure for Sidak's correction. The procedure begins with sorting consecutive values of $p_i(p_1, p_2, ...p_c)$ in an ascending order. The Sidak correction is then applied to each successive value of p_i (with a corresponding reduction in the number of hypotheses left to test). As a result, all hypotheses that are tested after the first statistically insignificant

value of p_i are also insignificant.

The p-value adjustment:

$$p_{(Sidak,i)} = 1 - (1 - p_i)^{c-i+1}$$

- **Bonferroni-Hochberg's correction**
  It involves the use of the multi-step Hochberg's procedure for the Bonferroni's correction. The procedure begins with sorting consecutive values of p_i $(p_c, p_{c-1}, ..., p_1)$ in a descending order. Bonferroni's correction is then applied to each successive value of p_i (with a corresponding reduction in the number of hypotheses left to test). As a result, all hypotheses that are tested after the first statistically insignificant value of p_i are also insignificant.

  The p-value adjustment:

  $$p_{(Bonferroni,i)} = p_i \cdot (c - i + 1)$$

- **Sidak-Hochberg's correction**
  It involves the use of Hochberg's multi-step procedure for Sidak's correction. The procedure begins with sorting consecutive values of p_i $(p_c, p_{c-1}, ..., p_1)$ in a descending order. A Sidak correction is then applied to each successive value of p_i (with a corresponding reduction in the number of hypotheses left to test). As a result, all hypotheses that are tested after the first statistically insignificant value of p_i are also insignificant.

  Adjustment of the p value:

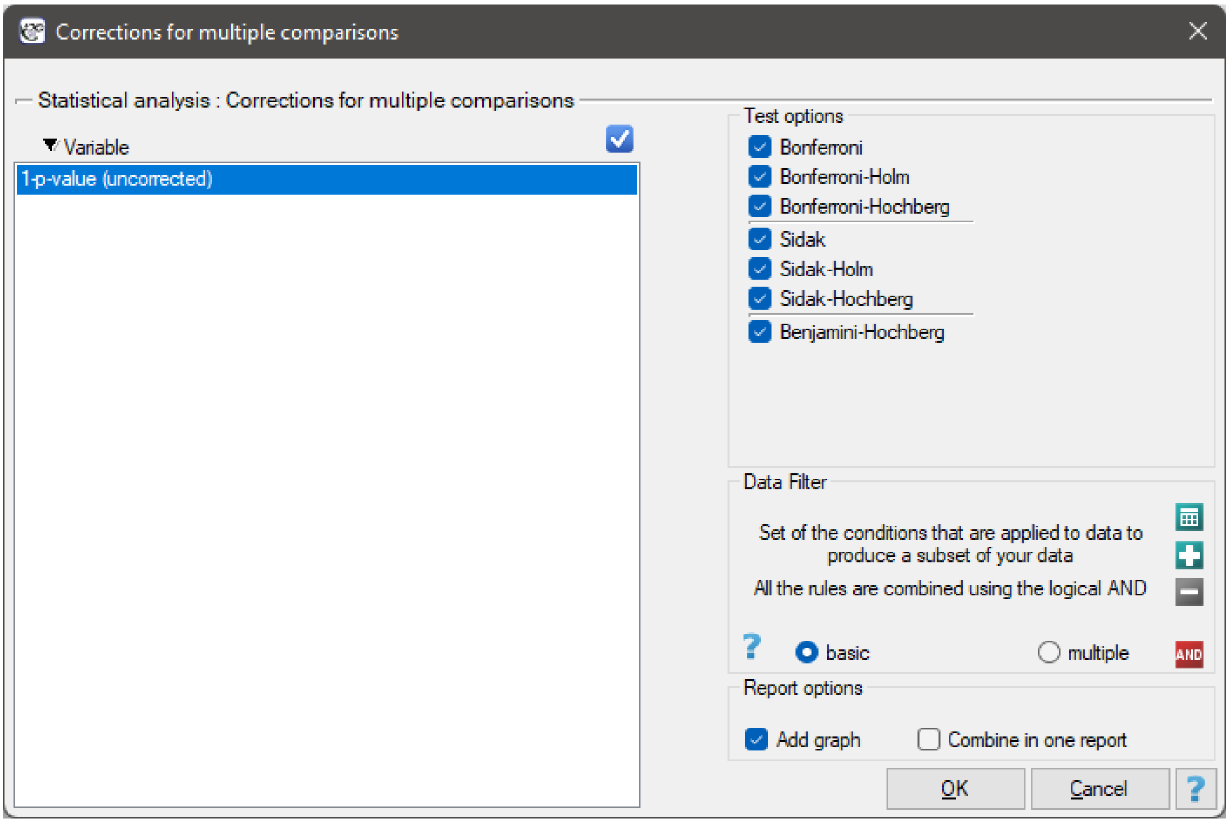
  $$p_{(Sidak,i)} = 1 - (1 - p_i)^{c-i+1}$$

- **Benjamini-Hochberg's correction**
  It involves the use of the multi-step Hochberg procedure for the Benjamini's correction which is a modified version of the Bonferroni's correction. The procedure starts with sorting consecutive values of p_i $(p_c, p_{c-1}, ..., p_1)$ in a descending order. Benjamini's correction is then applied to each successive value of p_i (with a corresponding reduction in the number of hypotheses left to test). As a result, all hypotheses that are tested after the first statistically insignificant value of p_i are also insignificant.

  Adjustment of the p value:

  $$p_{(BH,i)} = p_i \frac{c}{i}$$

To perform a multicomparison correction, consecutive p values are entered into one column of the datasheet.
The window with the multicomparison settings is opened via menu Statistics→Corrections for multiple comaprisons.



Corrections for multiple comparisons

Statistical analysis : Corrections for multiple comparisons

Variable ☑

1-p-value (uncorrected)

Test options

- ☑ Bonferroni
- ☑ Bonferroni-Holm
- ☑ Bonferroni-Hochberg
- ☑ Sidak
- ☑ Sidak-Holm
- ☑ Sidak-Hochberg
- ☑ Benjamini-Hochberg

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

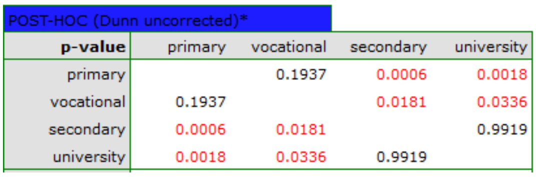
☑ Add graph □ Combine in one report

Note

A family of hypotheses can be defined in many ways. The most common are hypotheses within the POST-HOC procedure, i.e., the performing of multiple tests in a simultaneous comparison of several groups under study. Tests performed as part of a Hotelling-type analysis may also constitute such a family. Families of hypotheses are also found in many geographical analyses. Wherever multiple minor hypotheses are analyzed as part of an overall hypothesis, the multicomparisons correction may be applicable.

Example (16.4) continued *(jobSatisfaction.pqs file)*

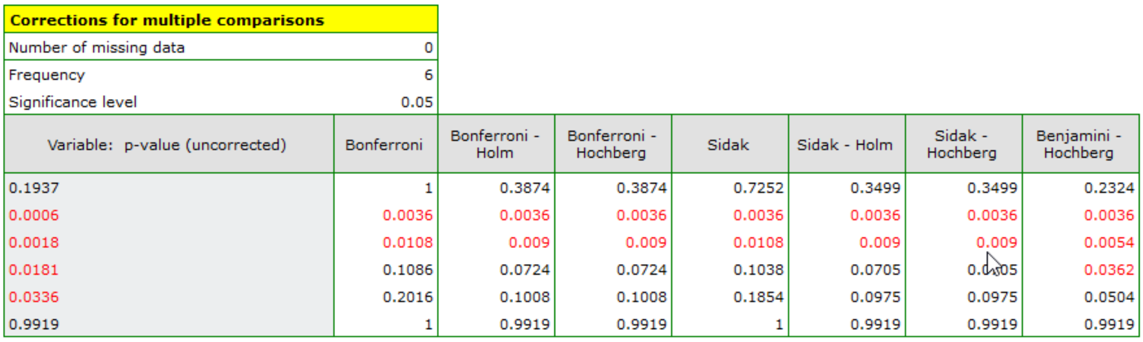
This study tested whether job satisfaction was the same for the four education categories. The family of hypotheses here was made up of hypotheses derived from pairwise comparisons of all groups. To compare all 4 groups, 6 pairs of comparisons were created. In each case, the null hypothesis was that there were no differences in satisfaction among the pairs analyzed. To take advantage of several proposed corrections of multicomparisons, the analysis was conducted using Dunn's uncorrected POST-HOC test.



POST-HOC (Dunn uncorrected)*

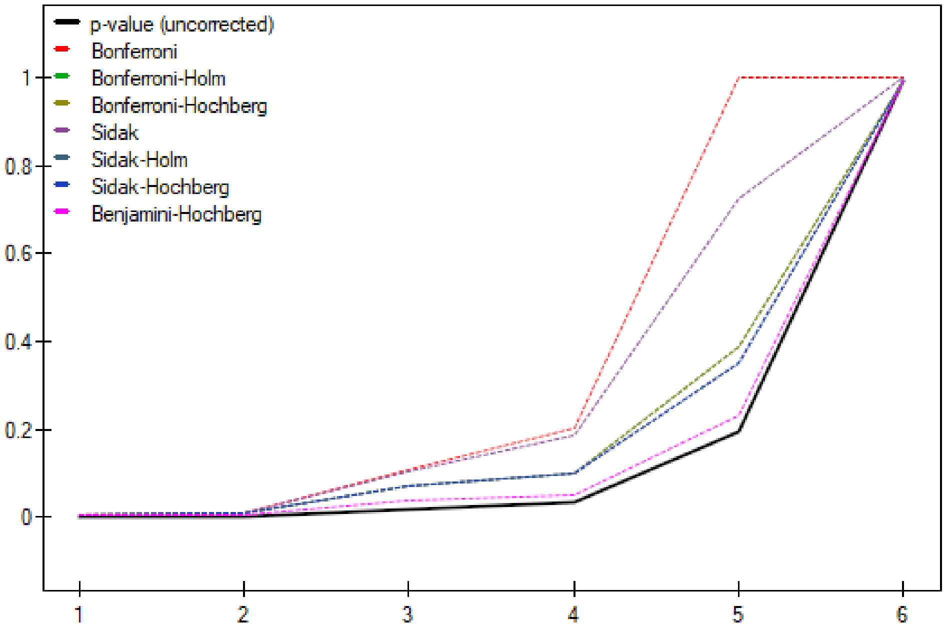
p-value	primary	vocational	secondary	university
primary		0.1937	0.0006	0.0018
vocational	0.1937		0.0181	0.0336
secondary	0.0006	0.0181		0.9919
university	0.0018	0.0336	0.9919	

The resulting p-values were given as data for correction of multicomparisons 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



As a result, differences in job satisfaction were found to be statistically significant for two pairs of comparisons (elementary vs. secondary education and elementary vs. higher education). Using the Benjamini-Hochberg's correction only, differences could be found in three pairs.

OTHER EXAMPLES: **example (??), example (??)**

18 UNIVARIATE MANOVA

Beforehand, we recommend that you read the T-square Hotelling's analysis.

Multivariate analysis of variance, is an extension of one-way ANOVA for independent groups. It is used to verify the hypothesis that the means of the k variables under study are equal across several ($m \geq 2$) populations.

Staying with the ANOVA method involves comparing ($m \geq 2$) populations multiple times (separately for each variable) without taking into account the variables' correlation with each other. A MANOVA-type analysis, on the other hand, examines differences between populations one at a time for multiple variables, taking into account their correlation. In addition, the MANOVA approach is used as an alternative to the ANOVA for dependent groups because it does not require the sphericity assumption to be met.

Basic application conditions:

- measurement on an interval scale,
- A multivariate normal distribution in each population or normality of the distribution of each studied variable in each population,
- independent model,
- equality of the covariance matrix or equality of variances of the examined variables for the compared populations - a condition particularly important in the case of groups of different sizes.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \mu_1=\mu_2=...=\mu_m, \\ \mathcal{H}_1: & \text{not all } \mu_i \text{ are equal,} \end{array}$$

where:
$\mu_i=(\mu_{i1},\mu_{i2},...,\mu_{ik})$ - means of variables in i-th population,
$(i=1,2,...,m)$,
$(j=1,2,...,k)$.

We use several coefficients in MANOVA analyses. The most widely known is the Wilks' Lambda. The Pillai-Bartlett trace is the most conservative, but relatively robust to violations of the MANOVA assumptions and preferred for small sample sizes. The Hotelling-Lawley trace, on the other hand, is the least conservative of the three proposed tests. Work on the development of these techniques was begun by Wilks (1932)[173], Pillai(1955)[131], Lawley(1938)[96], Hotelling(1951)[82], and Roy(1939)[139].

Test statistics are based on Sums of Squares and Cross Products ($SSCP$) matrices. The total matrix $T=SSCP$ is broken down into two matrices, the first of which is related to the hypothesis being tested and is indicated by H (in this case the matrix of between-group sums of squares and mixed products), and the second of which is related to the residuals (errors) and is indicated by E (matrix of within-group sums of squares and mixed products).

Wilks' Lambda

Lambda value is defined as follows:

$$\Lambda=\frac{|E|}{|H+E|}$$

The test statistic is in the form of:

$$F=\frac{1-\Lambda^{\frac{1}{b}}}{\Lambda^{\frac{1}{b}}}\frac{df_2}{df_1}$$

where:
$df_1 = k(m-1)$, $df_2 = ab - c$,
a, b, c - coefficients dependent on the number of variables analyzed and the number of populations compared.

Hotelling-Lawley trail

The Hotelling-Lawley trace is defined as follows:

$$T_0 = trace(HE^{-1})$$

The test statistic is in the form of:

$$F = \frac{T_0^2}{s}\frac{df_2}{df_1}$$

where:
$df_1 = s(2t+s+1)$, $df_2 = 2(su+1)$,
s, t, u - coefficients dependent on the number of variables analyzed and the number of populations compared.

Pillai-Bartlett trail

The Pillai-Bartlett trace is defined as follows:

$$V = trace(H(H+E)^{-1})$$

The test statistic is in the form of:

$$F = \frac{V}{s-V}\frac{df_2}{df_1}$$

where:
$df_1 = s(2t+s+1)$, $df_2 = s(2u+s+1)$,
s, t, u - coefficients dependent on the number of variables analyzed and the number of populations compared.

Each of the test statistics above is subject to Snedecor's F distribution with df_1 and df_2 degrees of freedom.

Designated based on the test statistics value p is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ przyjmując } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no basis to reject } \mathcal{H}_0. \end{array}$$

Effect size - partial η^2

This size indicates the proportion of explained variance to total variance associated with a given factor. In a one-factor MANOVA model for independent groups, it indicates what proportion of the within-group variability in outcomes can be attributed to the factor under study that determines the independent groups.

$$\eta^2 = \frac{F \cdot df_1}{F \cdot df_1 + df_2}$$

Effect size - contrasts, one-dimensional analysis

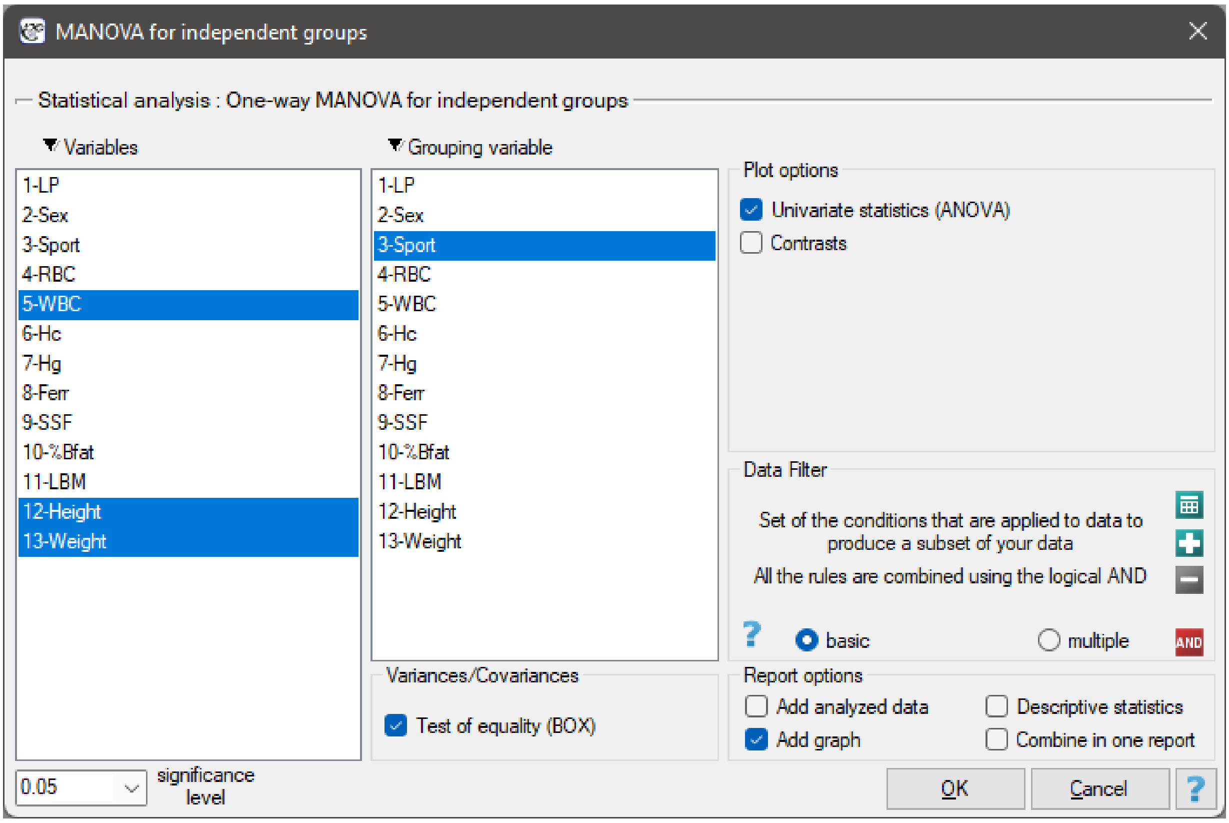
When the analysis performed is to compare selected populations, or a selected set of populations, then we perform a contrasts analysis. This analysis is analogous to the contrasts in one-dimensional analysis but takes into account the interrelatedness of the variables.

For effect sizes, one can also determine **simultaneous confidence intervals** or confidence intervals with Bonferroni correction. When using these intervals, however, it is important to note that they do not take into account associations between variables (which MANOVA takes into account) but only multiple testing.

When looking for variables with differences, we can also use a one-dimensional approach. We then perform the comparisons of the ANOVA for independent groups separately for each variable. Unfortunately, this will not account for intercorrelations, but the p values obtained from the ANOVA can be adjusted in the multiple comparisons section.

Note
The basic principle of MANOVA (as well as Hotelling's tests) is the construction of "multivariate ellipses" of confidence intervals around the centers determined by the means (see example interpretation of Hotelling's test ellipses for a single sample). As a result, using one-dimensional analysis (which does not take into account the interrelationships between variables) we are often unable to obtain identical results.

The settings window for the Single-factor MANOVA for independent groups is opened via menu Statistics→Parametric tests→MANOVA for independent groups.



MANOVA for independent groups

Statistical analysis : One-way MANOVA for independent groups

Variables: 1-LP, 2-Sex, 3-Sport, 4-RBC, **5-WBC**, 6-Hc, 7-Hg, 8-Ferr, 9-SSF, 10-%Bfat, 11-LBM, **12-Height**, **13-Weight**

Grouping variable: 1-LP, 2-Sex, **3-Sport**, 4-RBC, 5-WBC, 6-Hc, 7-Hg, 8-Ferr, 9-SSF, 10-%Bfat, 11-LBM, 12-Height, 13-Weight

Plot options
☑ Univariate statistics (ANOVA)
□ Contrasts

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Variances/Covariances
☑ Test of equality (BOX)

Report options
□ Add analyzed data □ Descriptive statistics
☑ Add graph □ Combine in one report

0.05 significance level

OK Cancel

Example 18.1. (sport.pqs file)
A group of athletes was studied to obtain information on health parameters such as:
WBC - White Blood Count,
Height [cm],
Body weight [kg].

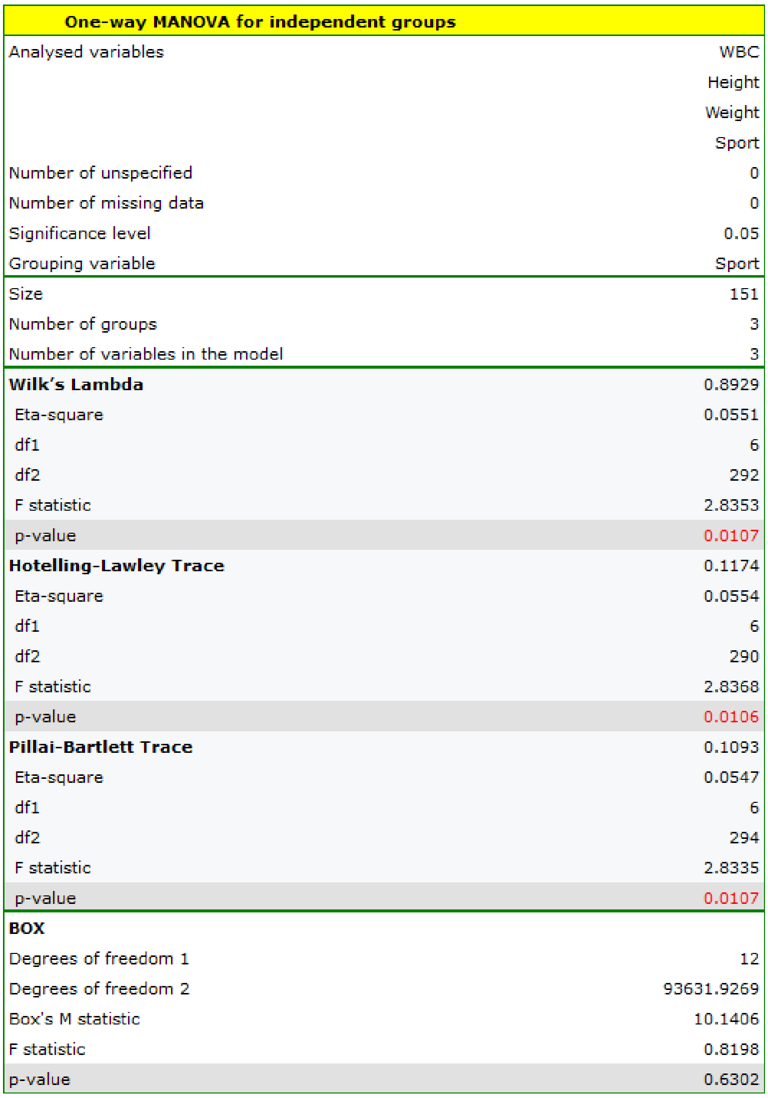
We'd like to know:

1. Whether playing three types of sports professionally: "team games" (such as: basketball, volleyball, etc.) "running" (such as: 100m, 400m, etc.) "aquatic" (like: swimming, rowing, etc.), differ in the levels of these parameters. Whether practicing high effort sports such as: "treadmill" and "aquatic" differ in the levels of these parameters from those practicing "team games"

Re.1) Hypotheses:

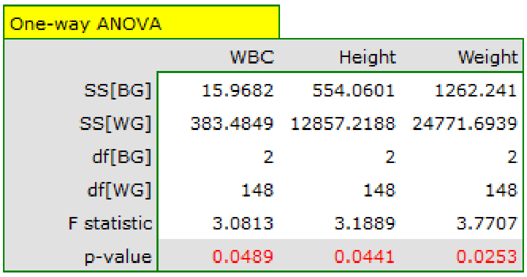
$$\begin{array}{ll}\mathcal{H}_0: & \text{The means of the analyzed parameters are the same}\\ & \text{for athletes participating in specific sports,}\\ \mathcal{H}_1: & \text{at least one parameter has a different mean value}\\ & \text{for the compared populations.}\end{array}$$

The result of Box's test (p=0.6302) allows us to calculate Analyses of the MANOVA type.



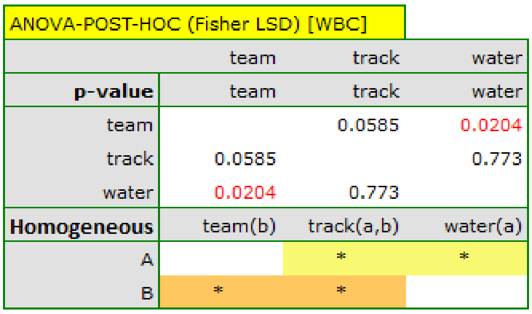
One-way MANOVA for independent groups	
Analysed variables	WBC Height Weight Sport
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	Sport
Size	151
Number of groups	3
Number of variables in the model	3
Wilk's Lambda	0.8929
Eta-square	0.0551
df1	6
df2	292
F statistic	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

The significance of the coefficients: Wilks' Lambda, Hotelling-Lawley trace, and Pillai-Bartlett trace allow us to argue that the study populations of athletes differ on these parameters. To determine the differences we conduct a one-dimensional ANOVA analysis.



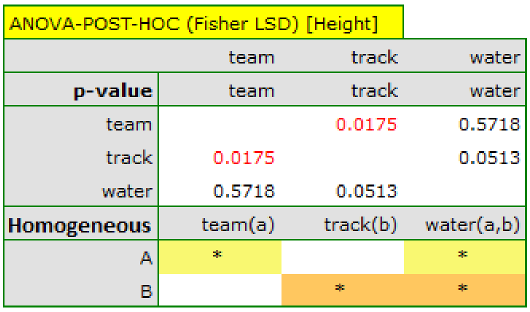
One-way ANOVA

	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



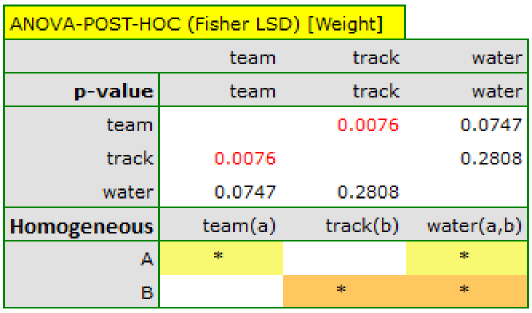
ANOVA-POST-HOC (Fisher LSD) [WBC]

	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]

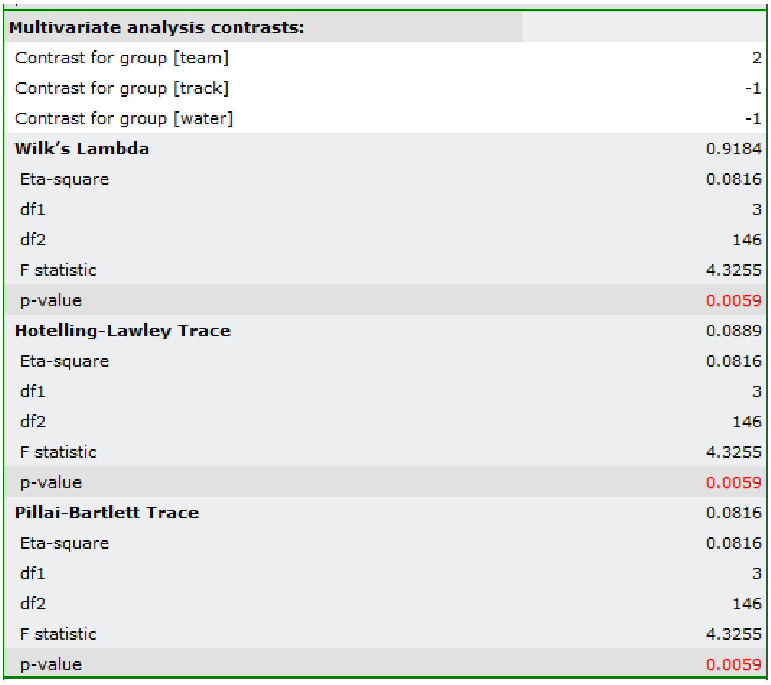
	team	track	water
p-value	team	track	water
team		0.0076	0.0747
track	0.0076		0.2808
water	0.0747	0.2808	
Homogeneous	team(a)	track(b)	water(a,b)
A	*		*
B		*	*

The results should be treated with caution. Although they indicate significant differences in all compared parameters, they yield p-values bordering on statistical significance (for WBC p=0.0489, for height p=0.0441, for weight p=0.0253). Additionally, when interpreting them, it is important to remember that they do not take into account either mutual correlation of parameters or multiple testing. Accounting for multiple testing in this case would require applying one of the p-value adjustments described in the section Multiple comparisons.

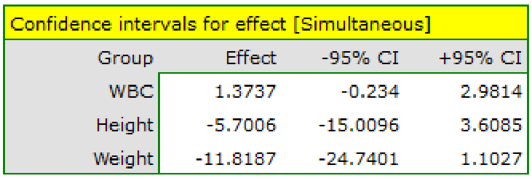
Re.2) Hypotheses:

$\mathcal{H}_0$: means of analyzed parameters for ”team sports”
are not different from the respective means of the athletes in the other two groups
$\mathcal{H}_1$: at least one parameter has a different mean value
for the compared populations.

To check whether the above hypotheses are true, we set an appropriate contrast in the analysis window. As a contrast value we enter 2 for team sports, -1 for treadmill and sports defined as aquatic.

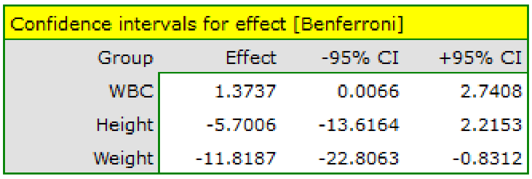


Multivariate analysis contrasts:	
Contrast for group [team]	2
Contrast for group [track]	-1
Contrast for group [water]	-1
Wilk's Lambda	0.9184
Eta-square	0.0816
df1	3
df2	146
F statistic	4.3255
p-value	0.0059
Hotelling-Lawley Trace	0.0889
Eta-square	0.0816
df1	3
df2	146
F statistic	4.3255
p-value	0.0059
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

As a result, the obtained significance of the coefficients: Wilks' Lambda, Hotelling-Lawley trace and Pillai-Bartlett trace (p=0.0059) allows us to argue that athletes practicing high intensity sports differ in these parameters from those practicing team sports. In simultaneous intervals we do not observe these differences, while on the basis of Bonferroni intervals we can state that the difference concerns weight and WBC. WBC values are higher in the team sports group, and weight is significantly lower in this group.

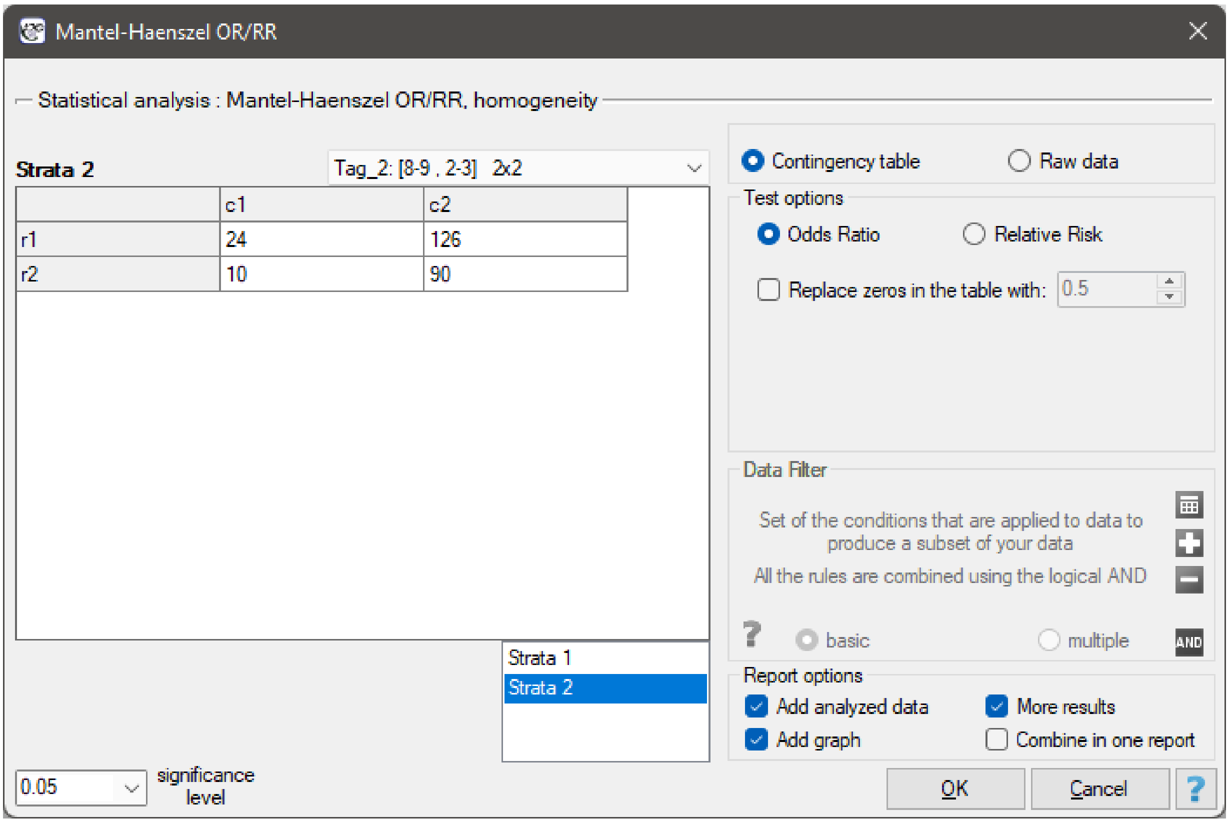
19 ANALYSIS FOR STRATAS

19.1 The Mantel-Haenszel method for several tables

The Mantel-Haenszel method for 2×2 tables proposed by Mantel and Haenszel (1959)[109] then it was extended by Mantel (1963)[110]. A wider review the development of these methods was carried out i.a. by Newman (2001)[123].
This method can be used in analysis 2×2 tables, that occur in several ($w >= 2$) stratas constructed by confounding variable. For the next stratas ($s = 1, ..., w$) the 2×2 contingency tables for observed frequencies are created:

Observed frequencies s-th strata $\left(O_{ij}^{(s)}\right)$		Analysed phenomenon (illness)		
		occurs (case)	not occurs (control)	Total
Risk factor	exposed	$O_{11}^{(s)}$	$O_{12}^{(s)}$	$O_{11}^{(s)} + O_{12}^{(s)}$
	unexposed	$O_{21}^{(s)}$	$O_{22}^{(s)}$	$O_{21}^{(s)} + O_{22}^{(s)}$
	Total	$O_{11}^{(s)} + O_{21}^{(s)}$	$O_{12}^{(s)} + O_{22}^{(s)}$	$n^{(s)} = O_{11}^{(s)} + O_{12}^{(s)} + O_{21}^{(s)} + O_{22}^{(s)}$

The settings window with the Mantel-Haenszel OR/RR can be opened in Statistics menu →Stratified analysis→Mantel-Haenszel OR/RR.


Mantel-Haenszel OR/RR

Statistical analysis : Mantel-Haenszel OR/RR, homogeneity

Strata 2 Tag_2: [8-9 , 2-3] 2x2

	c1	c2
r1	24	126
r2	10	90

Strata 1
Strata 2

0.05 significance level

● Contingency table ○ Raw data

Test options
● Odds Ratio ○ Relative Risk
☐ Replace zeros in the table with: 0.5

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
○ basic ○ multiple AND

Report options
☑ Add analyzed data ☑ More results
☑ Add graph ☐ Combine in one report

OK Cancel ?

19.1.1 The Mantel-Haenszel Odds Ratio

If all tables (created by individual stratas) are homogeneous (the χ^2 test of homogeneity for the OR can check this condition), then, on the basis of these tables, the pooled odds ratio with the confidence interval can be designated. Such odds ratio, is a weighted mean for an odds ratio designated for the individual stratas. The usage of the weighted method, proposed by Mantel and Haenszel allows to include the contribution of the strata weights. Each strata has an influence on the pooled odds ratio (the greater size of the strata, the greater weight and the greater influence on the pooled odds ratio).

Weights for individual stratas are designated according to the following 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:

$R=\sum_{s=1}^{w}\frac{O_{11}^{(s)}\cdot O_{22}^{(s)}}{n^{(s)}}$,

$S=\sum_{s=1}^{w}g^{(s)}$.

The confidence interval for $logOR_{MH}$ is designated on the basis of the standard error (RGB – Robins-Breslow-Greenland[134][135]) calculated according to the following formula:

$$SE_{MH}=\sqrt{\frac{T}{2R^2}+\frac{U+Y}{2RS}+\frac{W}{2S^2}},$$

where:

$T=\sum_{s=1}^{w}T^{(s)}$, $\quad T^{(s)}=\frac{O_{11}^{(s)}\cdot O_{22}^{(s)}\cdot\left(O_{11}^{(s)}+O_{22}^{(s)}\right)}{\left(n^{(s)}\right)^2}$,

$U=\sum_{s=1}^{w}U^{(s)}$, $\quad U^{(s)}=\frac{O_{21}^{(s)}\cdot O_{12}^{(s)}\cdot\left(O_{11}^{(s)}+O_{22}^{(s)}\right)}{\left(n^{(s)}\right)^2}$,

$Y=\sum_{s=1}^{w}Y^{(s)}$, $\quad Y^{(s)}=\frac{O_{11}^{(s)}\cdot O_{22}^{(s)}\cdot\left(O_{21}^{(s)}+O_{12}^{(s)}\right)}{\left(n^{(s)}\right)^2}$,

$W=\sum_{s=1}^{w}W^{(s)}$, $\quad W^{(s)}=\frac{O_{21}^{(s)}\cdot O_{12}^{(s)}\cdot\left(O_{21}^{(s)}+O_{12}^{(s)}\right)}{\left(n^{(s)}\right)^2}$.

The Mantel-Haenszel χ^2 test for the OR_{MH}

The Mantel-Haenszel Chi-square test for the OR_{MH} is used in the hypothesis verification about the significance of designated odds ratio (OR_{MH}). It should be calculated for large frequencies, i.e. when both conditions of the so-called ”rule 5” are satisfied:

- min$(O_{11}^{(s)}+O_{12}^{(s)},O_{11}^{(s)}+O_{21}^{(s)})-\sum_{s=1}^{w}E_{11}^{(s)}\geq 5$ for all the stratas $s=1,2,...,w$,
- max$(0,O_{11}^{(s)}-O_{22}^{(s)})\geq 5$ for all the stratas $s=1,2,...,w$.

When there are zero values in the table, a continuity adjustment (increasing the counts by a value of 0.5) is applied to both the observed counts and the expected counts.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & OR_{MH}=1,\\ \mathcal{H}_1: & OR_{MH}\neq 1.\end{array}$$

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:
$E_{11}^{(s)}=\frac{\left(O_{11}^{(s)}+O_{21}^{(s)}\right)\left(O_{11}^{(s)}+O_{12}^{(s)}\right)}{n^{(s)}}$ are the expected frequencies in the first contingency table cell, for the individual stratas $s=1,2,...,w$,
$V=\sum_{s=1}^{w}V^{(s)}$,
$V^{(s)}=\frac{\left(O_{11}^{(s)}+O_{12}^{(s)}\right)\left(O_{21}^{(s)}+O_{22}^{(s)}\right)\left(O_{11}^{(s)}+O_{21}^{(s)}\right)\left(O_{12}^{(s)}+O_{22}^{(s)}\right)}{\left(n^{(s)}\right)^2\left(n^{(s)}-1\right)}.$

This statistic asymptotically (for large frequencies) has the χ^2 distribution with 1 degree of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The χ^2 test of homogeneity for the OR

The Chi-square test of homogeneity for the OR is used in the hypothesis verification that the variable, creating stratas, is the modifying effect, i.e. it influences on the designated odds ratio in the manner that, the odds ratios are significant different for individual stratas.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & OR_{MH}=OR^{(s)}, \text{ for all the stratas } s=1,2,...,w,\\ \mathcal{H}_1: & OR_{MH}\neq OR^{(s)}, \text{ for at least one strata.} \end{array}$$

The test statistic (Breslow-Day (1980)[25], Tarone (1985)[26][157]) 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 solution to 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 frequencies) has the χ^2 distribution with the number of degrees of freedom calculated using the formula: $df=w-1$.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Example 19.1. (leptospirosis.pqs file)

The following table presents hypothetical poll results, conducted among inhabitants of a city and village (the village is treated as a risk factor) in West India. The poll aim was to detect risk factors of leptospirosis[20]. The occurrence of leptospirosis antibodies is a indirect evidence about infection.

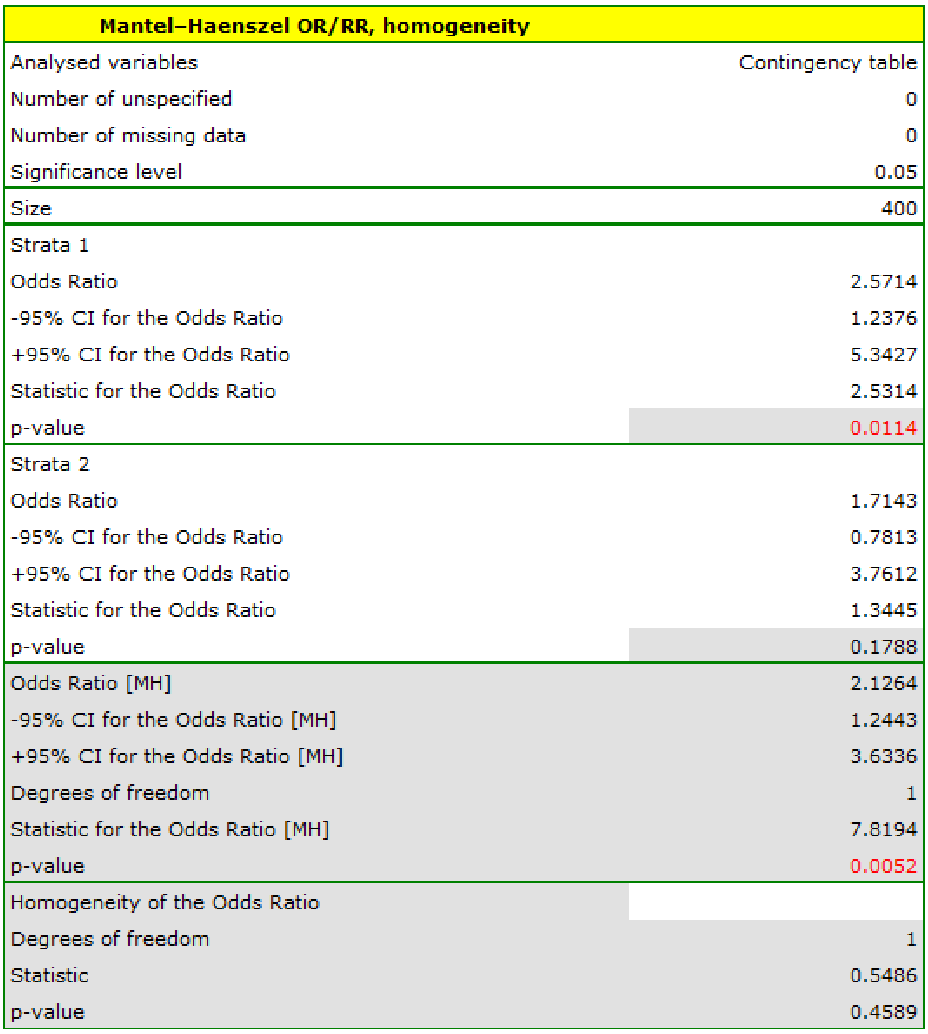
Observed frequencies O_{ij}		leptospirosis antibodies	
		occur	not occur
place of residence	rural	60	140
	urban	60	140

The odds of the occurrence of leptospirosis antibodies, among inhabitants of the city and the village, is the same (OR=1). Let's include gender in the analysis and check what odds will be then. The sample has to be divided into 2 stratas, because of gender (they are marked in a file as a saved selection):

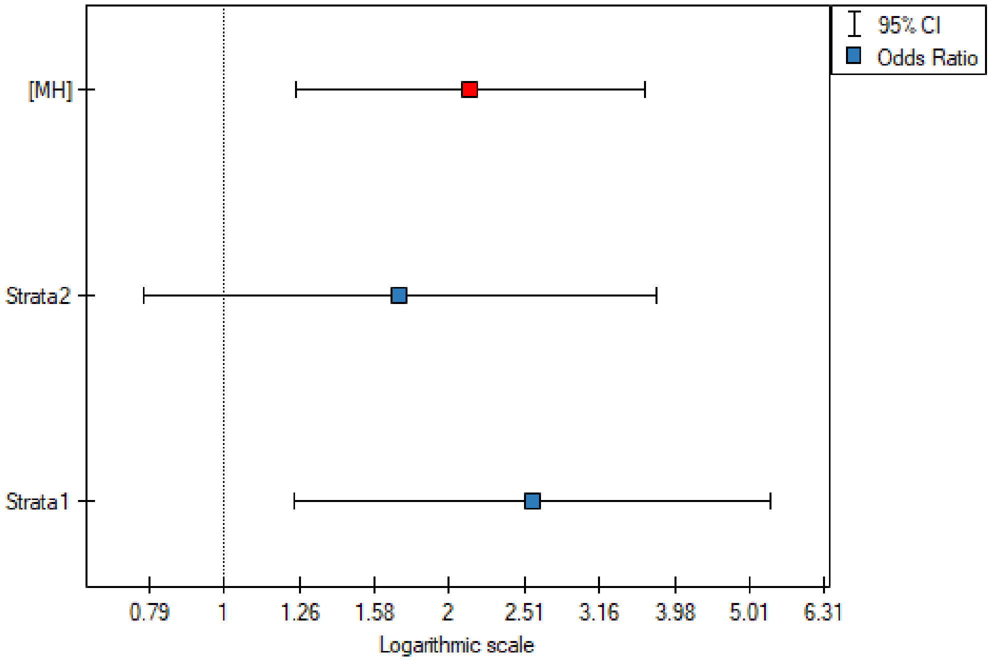
Observed frequencies for men		leptospirosis antibodies	
		occur	not occur
place of residence	rural	36	14
	urban	50	50

Observed frequencies for women		leptospirosis antibodies	
		occur	not occur
place of residence	rural	24	126
	urban	10	90

Gender is associated with both factors (the occurrence of leptospirosis anibodies and the residence in West India). This is a significant factor. Its ignorance can lead to errors in 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



The odds of the occurrence of leptospirosis antibodies is larger among village inhabitants, both among women (OR[95%CI]=2.57[1.24, 5.34]) and men (OR[95%CI]=1.71[0.78, 3.76]). The tables are homogeneous (p=0.4589). Thus, we can use the calculated odds ratio, which is mutual for both tables (OR_{MH}[95%CI]=2.13[1.24, 3.65]). Finally, the obtained result indicates that the odds of the occurrence of leptospirosis antibodies is significantly greater among village inhabitants (p=0.0052).

19.1.2 The Mantel-Haenszel Relative Risk

If all tables (created by individual stratas) are homogeneous (the χ^2 test of homogeneity for the RR), can check this condition), then, on the basis of these tables, the pooled relative risk with the confidence interval can be designated. Such relative risk is a weighted mean for a relative risk designated for the individual stratas. The usage of the weighted method, proposed by Mantel and Haenszel allows to include the contribution of the strata weights. Each strata of the input has an influence on the pooled relative risk construction (the greater size of the strata, the greater weight and the greater influence on the pooled relative risk).

Weights for individual stratas are designated according to the following 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 $logRR_{MH}$ is designated on the basis of the standard error calculated according to the following 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 Manel-Hanszel χ^2 test for the RR_{MH}

The Mantel-Haenszel Chi-square test for the RR_{MH} is used in the hypothesis verification about the significance of designated relative risk (RR_{MH}). It should be calculated for large frequencies, in a contingency table.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & RR_{MH} = 1, \\ \mathcal{H}_1: & RR_{MH} \neq 1. \end{array}$$

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:
$E_{11}^{(s)}=\frac{\left(O_{11}^{(s)}+O_{21}^{(s)}\right)\left(O_{11}^{(s)}+O_{12}^{(s)}\right)}{n^{(s)}}$ are the expected frequencies in the first contingency table cell, for individual stratas $s=1,2,...,w$.

This statistic asymptotically (for large frequencies) has the χ^2 distribution with 1 degree of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

The χ^2 test of homogeneity for the RR

The Chi-square test of homogeneity for the RR is used in the hypothesis verification that the variable creating stratas, is the modifying effect, i.e. it influences on the designated relative risk in the manner that, the relative risks are significant different for individual stratas.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & RR_{MH}=RR^{(s)}, \text{ for all the stratas } s=1,2,...,w,\\ \mathcal{H}_1: & RR_{MH}\neq RR^{(s)}, \text{ for at least one strata.}\end{array}$$

The test statistic, using weighted least squares method, is defined by:

$$\chi^2=\sum_{s=1}^{w}v^{(s)}\left(\ln(RR^{(s)})-\ln(RR_{MH})\right)^2$$

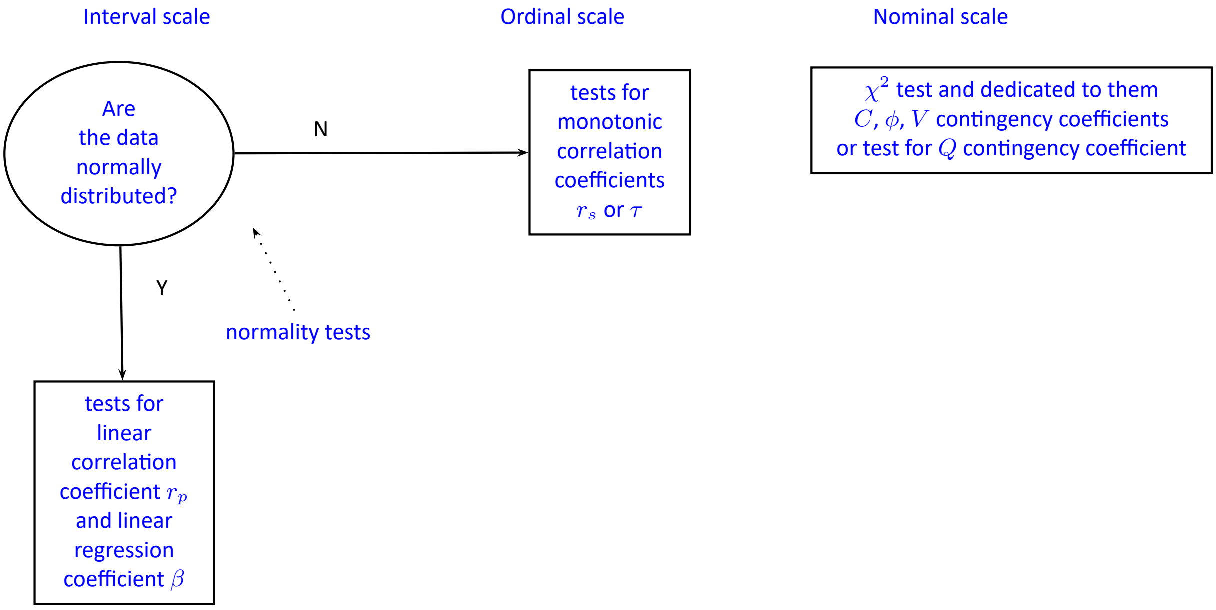
where:
$$v^{(s)}=\left(\frac{O_{12}^{(s)}}{O_{11}^{(s)}\left(O_{11}^{(s)}+O_{12}^{(s)}\right)}+\frac{O_{22}^{(s)}}{O_{21}^{(s)}\left(O_{21}^{(s)}+O_{22}^{(s)}\right)}\right)^{-1}.$$

This statistic asymptotically (for large frequencies) has the χ^2 distribution with the number of degrees of freedom calculated using the formula: $df=w-1$.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

20 CORRELATION


Interval scale
Ordinal scale
Nominal scale
Are the data normally distributed?
N
Y
tests for monotonic correlation coefficients r_s or τ
χ^2 test and dedicated to them C, ϕ, V contingency coefficients or test for Q contingency coefficient
normality tests
tests for linear correlation coefficient r_p and linear regression coefficient β

The Correlation coefficients are one of the measures of descriptive statistics which represent the level 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 done. Calculation of coefficients is one of the first steps of the correlation analysis. Then the statistic significance of the gained coefficients may be checked using adequate tests.

Note

Note, that the dependence between variables does not always show the cause-and-effect relationship.

20.1 PARAMETRIC TESTS

20.1.1 THE LINEAR CORRELATION COEFFICIENTS

The Pearson product-moment correlation coefficient r_p called also the Pearson's linear correlation coefficient (Pearson (1896,1900)) is used to decribe the strength of linear relations between 2 features. It may be calculated on an interval scale only if the distribution of the analyed features is a normal one.

$$r_p = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}},$$

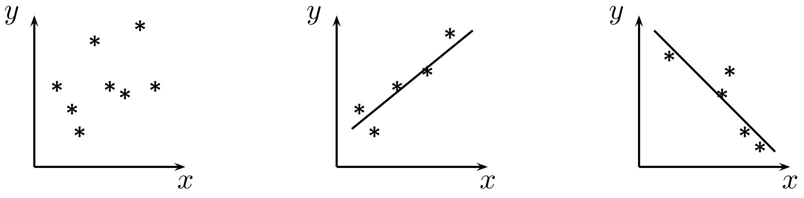
where:
x_i, y_i - the following values of the feature X and Y,
$\overline{x}, \overline{y}$ - means values of features: X and Y,
n - sample size.

Note
R_p – the Pearson product-moment correlation coefficient in a population;
r_p – the Pearson product-moment correlation coefficient in a sample.

The value of $r_p \in < -1; 1 >$, and it should be interpreted the following way:

- $r_p \approx 1$ means a strong positive linear correlation – measurement points are closed to a straight line and when the independent variable increases, the dependent variable increases too;
- $r_p \approx -1$ means a strong negative linear correlation – measurement points are closed to a straight line, but when the independent variable increases, the dependent variable decreases;
- if the correlation coefficient is equal to the value or very closed to zero, there is no linear dependence between the analysed features (but there might exist another relation - a not linear one).

Wykres 20.1. Graphic interpretation of r_p.



$r_p \approx 0$ $r_p \approx 1$ $r_p \approx -1$

If one out of the 2 analysed features is constant (it does not matter if the other feature is changed), the features are not dependent from each other. In that situation r_p can not be calculated.

Note
You are not allowed to calculate the correlation coefficient if: there are outliers in a sample (they may make that the value and the sign of the coefficient would be completely wrong), if the sample is clearly heterogeneous, or if the analysed relation takes obviously the other shape than linear.

The coefficient of determination: r_p^2 – reflects the percentage of a dependent variable a variability which is explained by variability of an independent variable.

A created model shows a linear relationship:

$$y = \beta x + \alpha.$$

β **and** α **coefficients of linear regression equation** can be calculated using formulas:

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \alpha = \overline{y} - \beta\overline{x}.$$

20.1.2 The Pearson correlation coefficient significance

The test of significance for Pearson product-moment correlation coefficient is used to verify the hypothesis determining the lack of linear correlation between an analysed features of a population and it is based on the Pearson's linear correlation coefficient calculated for the sample. The closer to 0 the value of r_p is, the weaker dependence joins the analysed features.

Basic assumptions:

- measurement on the interval scale,
- normality of distribution of residuals or an analysed features in a population.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & R_p = 0, \\ \mathcal{H}_1: & R_p \neq 0. \end{array}$$

The test statistic is defined by:

$$t = \frac{r_p}{SE},$$

where $SE = \sqrt{\frac{1 - r_p^2}{n-2}}$.

The value of the test statistic can not 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, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

20.1.3 The slope coefficient significance

The test of significance for the coefficient of linear regression equation

This test is used to verify the hypothesis determining the lack of a linear dependence between an analysed features and is based on the slope coefficient (also called an effect), calculated for the sample. The closer to 0 the value of β is, the weaker dependence presents the fitted line.

Basic assumptions:

- measurement on the interval scale,
- normality of distribution of residuals or an analysed features in a population.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & \beta=0,\\ \mathcal{H}_1: & \beta\neq 0.\end{array}$$

The test statistic is defined by:

$$t=\frac{\beta}{SE}$$

where:
$SE=\frac{s_{yx}}{sd_x\sqrt{n-1}}$,
$s_{yx}=sd_y\sqrt{\frac{n-1}{n-2}(1-r^2)}$,
sd_x, sd_y – standard deviation of the value of features: X and Y.

The value of the test statistic can not 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, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl}\text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

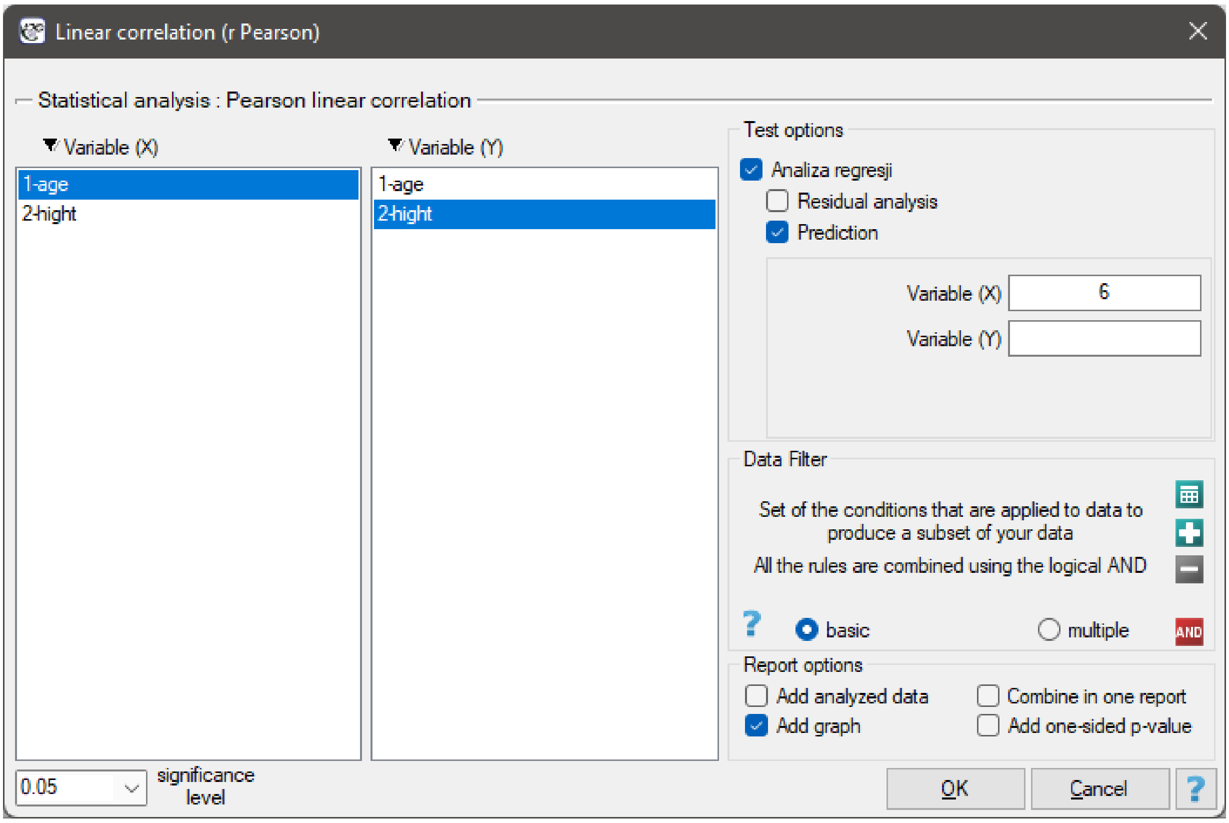
Prediction is used to predict the value of a one variable (mainly a dependent variable y_0) on the basis of a value of an another variable (mainly an independent variable x_0). The accuracies of a calculated value are defined by prediction intervals calculated for it.

- **Interpolation** is used to predict the value of a variable, which occurs inside the area for which the regression model was done. Interpolation is mainly a safe procedure - it is assumed only the continuity of the function of analysed variables.

- **Extrapolation** is used to predict the value of variable, which occurs outside the area for which the regression model was done. As opposed to interpolation, extrapolation is often risky and is performed only not far away from the area, where the regression model was created. Similarly to the interpolation, it is assumed the continuity of the function of analysed variables.

Analysis of model residuals - explanation in the Multiple Linear Regression module.

The settings window with the Pearson's linear correlation can be opened in Statistics menu→Parametric tests→linear correlation (r-Pearson) or in Wizard.



Linear correlation (r Pearson)

Statistical analysis : Pearson linear correlation

Variable (X)	Variable (Y)
1-age	1-age
2-hight	2-hight

Selected: Variable (X): 1-age; Variable (Y): 2-hight

Test options

☑ Analiza regresji

□ Residual analysis

☑ Prediction

Variable (X) 6

Variable (Y)

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

☑ Add graph □ Add one-sided p-value

0.05 significance level

Example 20.1. (age-height.pqs file)

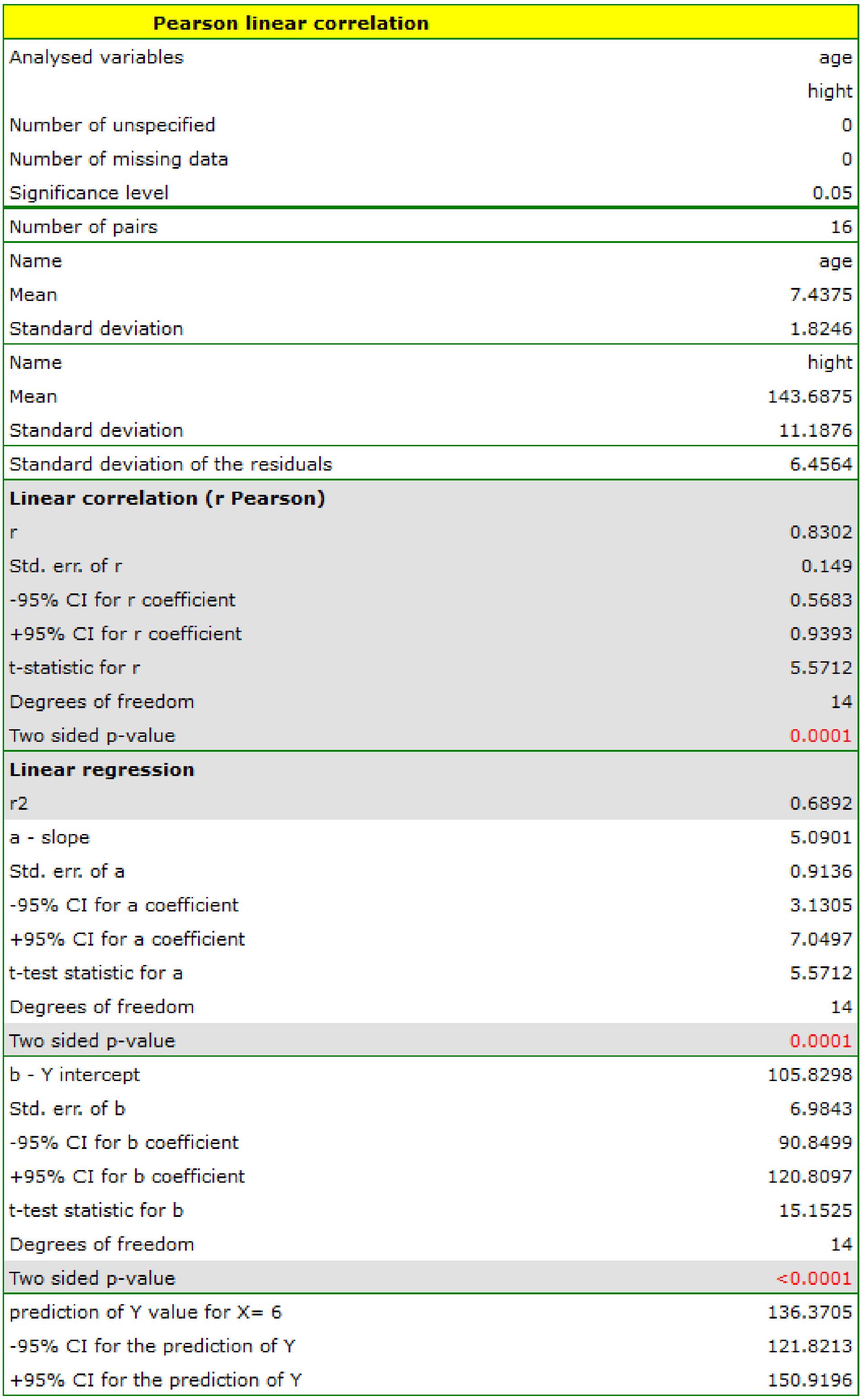
Among some students of a ballet school, the dependence between age and height was analysed. The sample consists of 16 children and the following results of these features (related to the children) were written down:

(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).

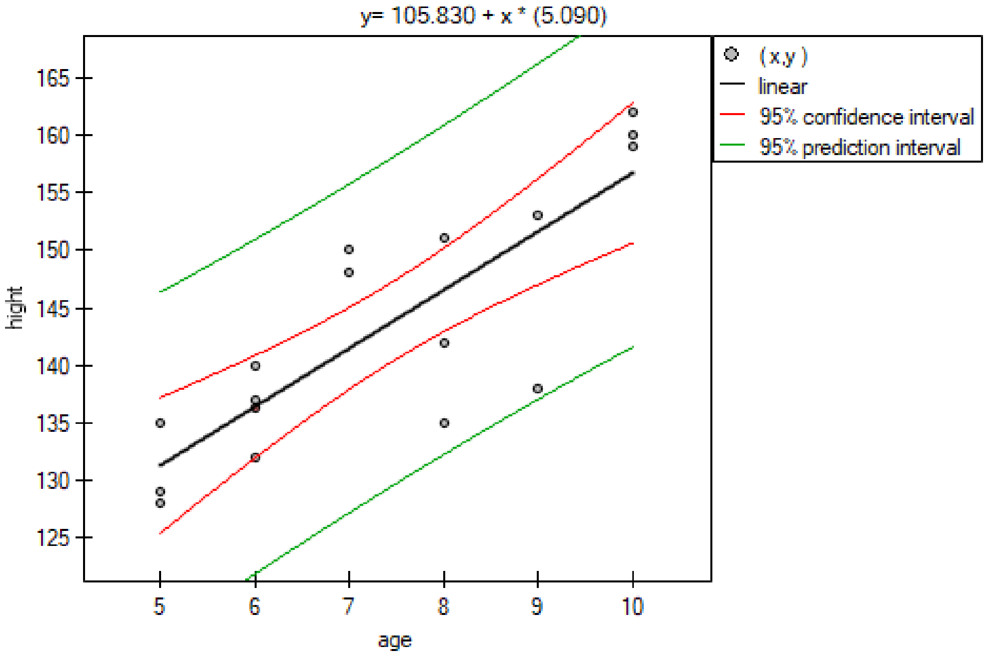
Hypotheses:

$\mathcal{H}_0$: there is no linear dependence between age and height
for the population of children who attend to the analysed school,

$\mathcal{H}_1$: there is a linear dependence between age and height
for the population of children who attend to the analysed school.



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



Comparing the p value < 0.0001 with the significance level $\alpha = 0.05$, we draw the conclusion, that there is a linear dependence between age and height in the population of children attening to the analysed school. This dependence is directly proportional, it means that the children grow up as they are getting older.

The Pearson product-moment correlation coefficient, so the strength of the linear relation between age and height counts to r_p=0.83. Coefficient of determination $r_p^2 = 0.69$ means that about 69% variability of height is explained by the changing of age.

From the regression equation:

$$height = 5.09 \cdot age + 105.83$$

it is possible to calculate the predicted value for a child, for example: in the age of 6. The predicted height of such child is 136.37cm.

20.1.4 Comparison of correlation coefficients

The test for checking the equality of the Pearson product-moment correlation coefficients, which come from 2 independent populations

This test is used to verify the hypothesis determining the equality of 2 Pearson's linear correlation coefficients (R_{p_1}, R_{p_2}).

Basic assumptions:

- r_{p_1} and r_{p_2} come from 2 samples which are chosen randomly from independent populations,
- r_{p_1} and r_{p_2} describe the strength of dependence of the same features: X and Y,
- sizes of both samples (n_1 and n_2) are known.

Hypotheses:

$$\begin{array}{cl}
\mathcal{H}_0: & R_{p_1} = R_{p_2}, \\
\mathcal{H}_1: & R_{p_1} \neq R_{p_2}.
\end{array}$$

The test statistic is defined by:

$$t=\frac{z_{r_{p1}}-z_{r_{p2}}}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}},$$

where:

$$z_{r_{p1}}=\frac{1}{2}\ln\left(\frac{1+r_{p1}}{1-r_{p1}}\right),$$

$$z_{r_{p2}}=\frac{1}{2}\ln\left(\frac{1+r_{p2}}{1-r_{p2}}\right).$$

The test statistic has the t-Student distribution with n_1+n_2-4 degrees of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

20.1.5 Comparison of the slope of regression lines

The test for checking the equality of the coefficients of linear regression equation, which come from 2 independent populations

This test is used to verify the hypothesis determining the equality of 2 coefficients of the linear regression equation β_1 and β_2 in analysed populations.

Basic assumptions:

- β_1 and β_2 come from 2 samples which are chosen randomly from independent populations,
- β_1 and β_2 describe the strength of dependence of the same features: X and Y,
- both sample sizes (n_1 and n_2) are known,
- standard deviations for the values of both features in both samples (sd_{x_1}, sd_{y_1} and sd_{x_2}, sd_{y_2}) are known,
- the Pearson product-moment correlation coefficients of both samples (r_{p_1} and r_{p_2}) are known.

Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & \beta_1=\beta_2,\\ \mathcal{H}_1: & \beta_1\neq\beta_2.\end{array}$$

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:

$$s_{yx_1}=sd_{y_1}\sqrt{\frac{n_1-1}{n_1-2}(1-r^2_{p_1})},$$

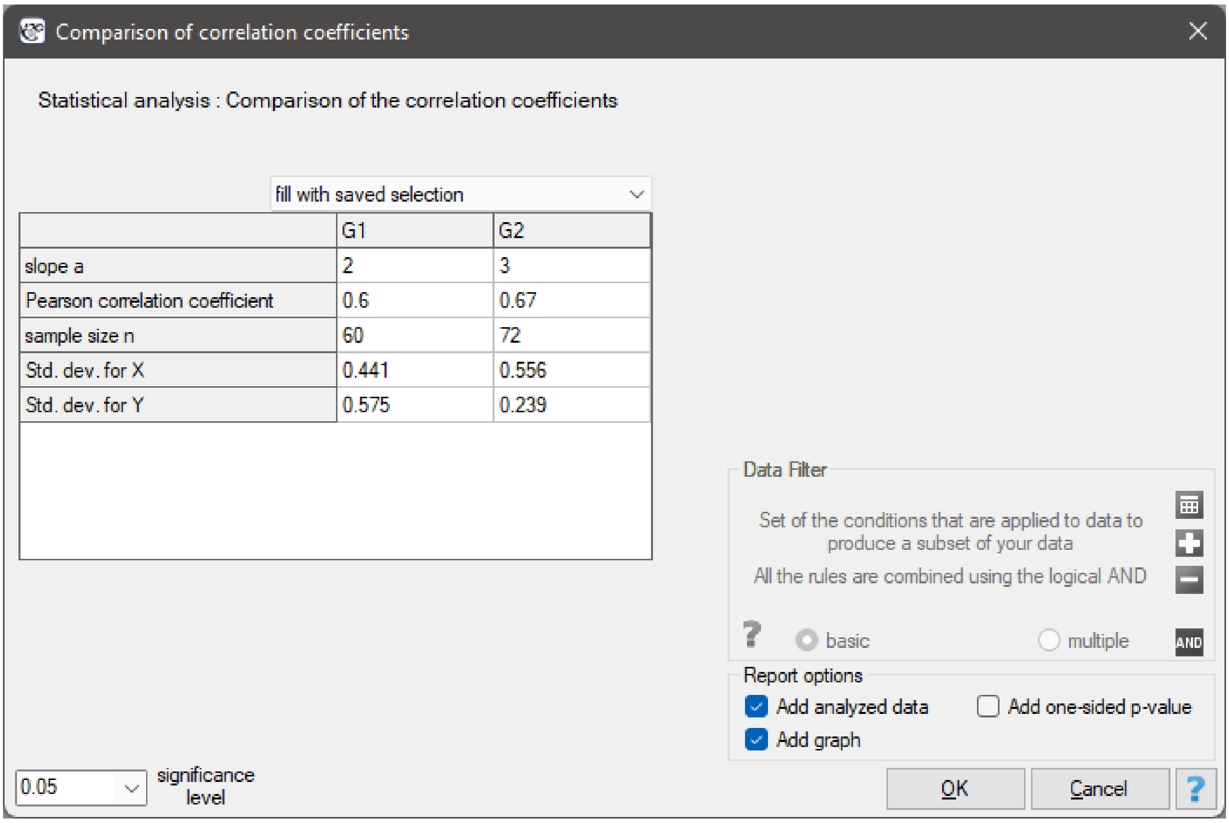
$$s_{yx_2}=sd_{y_2}\sqrt{\frac{n_2-1}{n_2-2}(1-r^2_{p_2})}.$$

The test statistic has the t-Student distribution with n_1+n_2-4 degrees of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the comparison of correlation coefficients can be opened in Statistics menu → Parametric tests → Comparison of correlation coefficients.



Comparison of correlation coefficients

Statistical analysis : Comparison of the correlation coefficients

fill with saved selection

	G1	G2
slope a	2	3
Pearson correlation coefficient	0.6	0.67
sample size n	60	72
Std. dev. for X	0.441	0.556
Std. dev. for Y	0.575	0.239

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

☑ Add analyzed data ☐ Add one-sided p-value

☑ Add graph

0.05 significance level

20.2 NON-PARAMETRIC TESTS

20.2.1 THE MONOTONIC CORRELATION COEFFICIENTS

The monotonic correlation may be described as monotonically increasing or monotonically decreasing. The relation between 2 features is presented by the monotonic increasing if the increasing of the one feature accompanies with the increasing of the other one. The relation between 2 features is presented by the monotonic decreasing if the increasing of the one feature accompanies with the decreasing of the other one.

The Spearman's rank-order correlation coefficient r_s is used to describe the strength of monotonic relations between 2 features: X and Y. It may be calculated on an ordinal scale or an interval one. The value of the Spearman's rank correlation coefficient should be calculated using the following formula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where:
$d_i = R_{x_i} - R_{y_i}$ – difference of ranks for the feature X and Y,
n number of d_i.

This formula is modified when there are ties:

$$r_s = \frac{\Sigma_X + \Sigma_Y - \sum_{i=1}^{n} d_i^2}{2\sqrt{\Sigma_X \Sigma_Y}},$$

where:

$\Sigma_X = \frac{n^3-n-T_X}{12}$, $\Sigma_Y = \frac{n^3-n-T_Y}{12}$,
$T_X = \sum_{i=1}^{s}(t_{i_{(X)}}^3 - t_{i_{(X)}})$, $T_Y = \sum_{i=1}^{s}(t_{i_{(Y)}}^3 - t_{i_{(Y)}})$,
t – number of cases included in tie.

This correction is used, when ties occur. If there are no ties, the correction is not calculated, because the correction is reduced to the formula describing the above equation.

Note
R_s – the Spearman's rank correlation coefficient in a population;
r_s – the Spearman's rank correlation coefficient in a sample.

The value of $r_s \in < -1; 1 >$, and it should be interpreted the following way:

- $r_s \approx 1$ means a strong positive monotonic correlation (increasing) – when the independent variable increases, the dependent variable increases too;
- $r_s \approx -1$ means a strong negative monotonic correlation (decreasing) – when the independent variable increases, the dependent variable decreases;
- if the Spearman's correlation coefficient is of the value equal or very close to zero, there is no monotonic dependence between the analysed features (but there might exist another relation - a non monotonic one, for example a sinusoidal relation).

The Kendall's $\tilde{\tau}$ correlation coefficient (Kendall (1938)[89]) is used to describe the strength of monotonic relations between features . It may be calculated on an ordinal scale or interval one. The value of the Kendall's $\tilde{\tau}$ correlation coefficient should be calculated using the following formula:

$$\tilde{\tau} = \frac{2(n_C - n_D)}{\sqrt{n(n-1)-T_X}\sqrt{n(n-1)-T_Y}},$$

where:

n_C – number of pairs of observations, for which the values of the ranks for the X feature as well as Y feature are changed in the same direction (the number of agreed pairs),

n_D – number of pairs of observations, for which the values of the ranks for the X feature are changed in the different direction than for the Y feature (the number of disagreed pairs),

$T_X = \sum_{i=1}^{s}(t_{i_{(X)}}^2 - t_{i_{(X)}})$, $T_Y = \sum_{i=1}^{s}(t_{i_{(Y)}}^2 - t_{i_{(Y)}})$,

t – number of cases included in a tie.

The formula for the $\tilde{\tau}$ correlation coefficient includes the correction for ties. This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $T_X = 0$ i $T_Y = 0$) .

Note
τ – the Kendall's correlation coefficient in a population;
$\tilde{\tau}$ – the Kendall's correlation coefficient in a sample.

The value of $\tilde{\tau} \in < -1; 1 >$, and it should be interpreted the following way:

- $\tilde{\tau} \approx 1$ means a strong agreement of the sequence of ranks (the increasing monotonic correlation) – when the independent variable increases, the dependent variable increases too;
- $\tilde{\tau} \approx -1$ means a strong disagreement of the sequence of ranks (the decreasing monotonic correlation) – when the independent variable increases, the dependent variable decreases;
- if the Kendall's $\tilde{\tau}$ correlation coefficient is of the value equal or very close to zero, there is no monotonic dependence between analysed features (but there might exist another relation - a non monotonic one, for example a sinusoidal relation).

The Spearman's r_s versus the Kendall's $\tilde{\tau}$

- for an interval scale with a normality of the distribution, the r_s gives the results which are close to r_p, but $\tilde{\tau}$ may be totally different from r_p,
- the $\tilde{\tau}$ value is less or equal to r_p value,
- the $\tilde{\tau}$ is an unbiased estimator of the population parameter τ, while the r_s is a biased estimator of the population parameter R_s.

20.2.2 Significance Test for Spearman

The test of significance for the Spearman's rank-order correlation coefficient is used to verify the hypothesis determining the lack of monotonic correlation between analysed features of the population and it is based on the Spearman's rank-order correlation coefficient calculated for the sample. The closer to 0 the value of r_s is, the weaker dependence joins the analysed features.

Basic assumptions:

- measurement on an ordinal scale or on an interval scale.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & R_s = 0, \\ \mathcal{H}_1: & R_s \neq 0. \end{array}$$

The test statistic is defined by:

$$t = \frac{r_s}{SE},$$

where $SE = \sqrt{\frac{1-r_s^2}{n-2}}$.

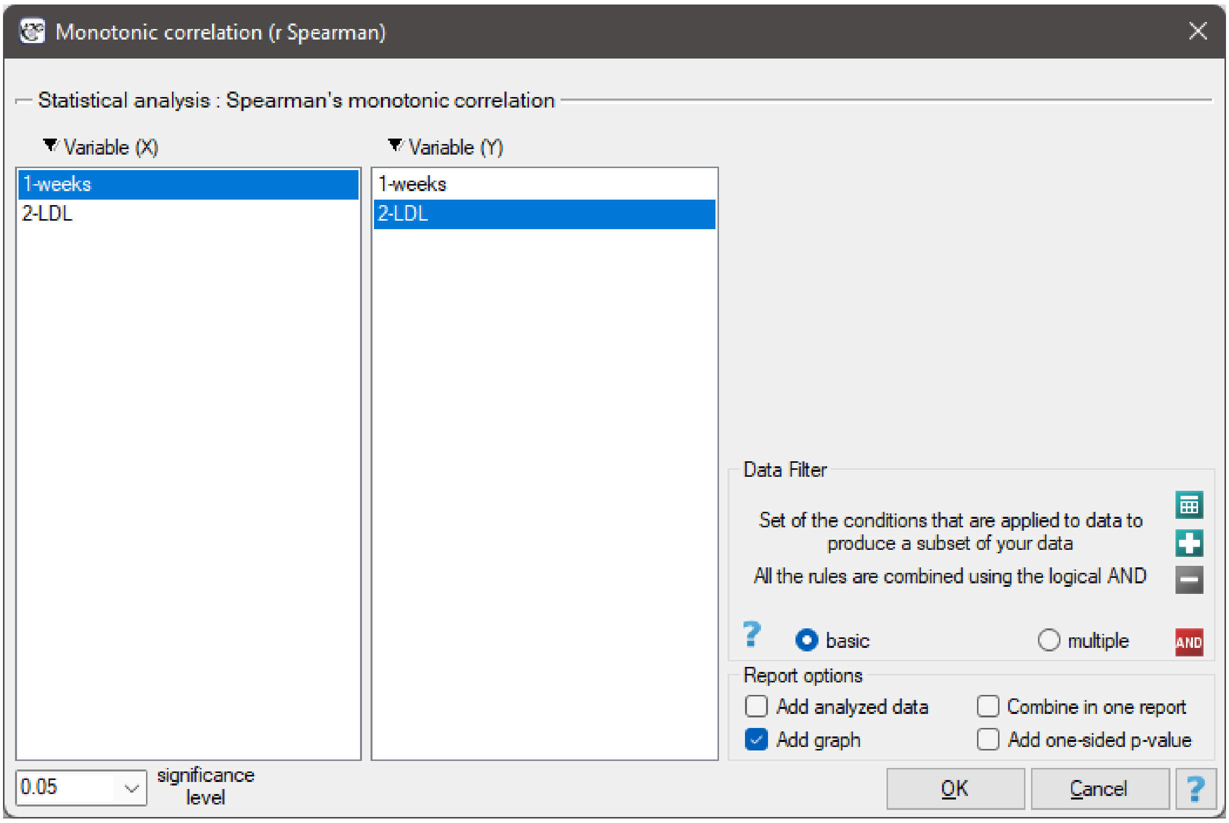
The value of the test statistic can not be calculated when $r_s = 1$ lub $r_s = -1$ or when $n < 3$.

The test statistic has the t-Student distribution with $n-2$ degrees of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the Spearman's monotonic correlation can be opened in Statistics menu → NonParametric tests→monotonic correlation (r-Spearman) or in Wizard.



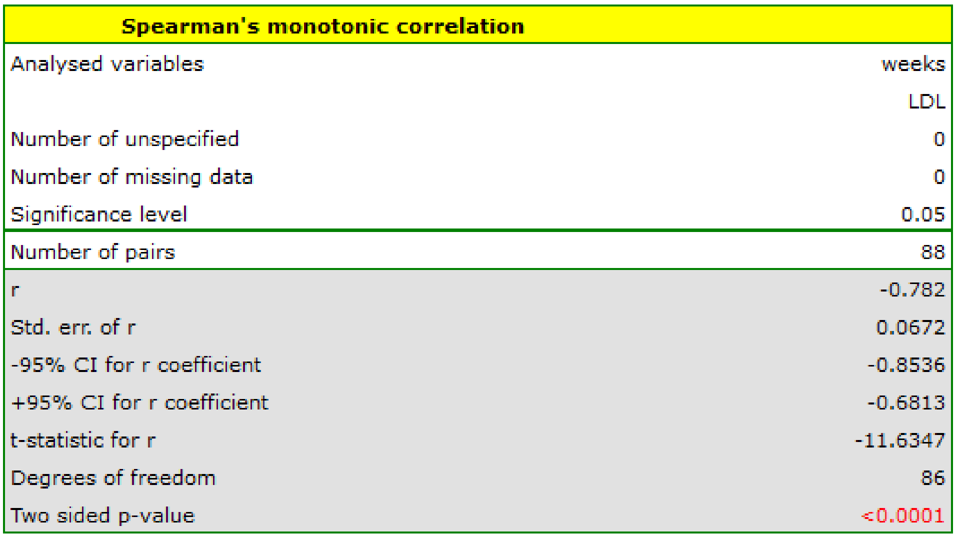
Monotonic correlation (r Spearman)
Statistical analysis : Spearman's monotonic correlation
Variable (X)
1-weeks
2-LDL
Variable (Y)
1-weeks
2-LDL
Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic
○ multiple
AND
Report options
☐ Add analyzed data
☐ Combine in one report
☑ Add graph
☐ Add one-sided p-value
0.05
significance level
OK
Cancel

Example 20.2. (LDL weeks.pqs file)

The effectiveness of a new therapy designed to lower cholesterol levels in the LDL fraction was studied. 88 people at different stages of the treatment were examined. We will test whether LDL cholesterol levels decrease and stabilize with the duration of the treatment (time in weeks).

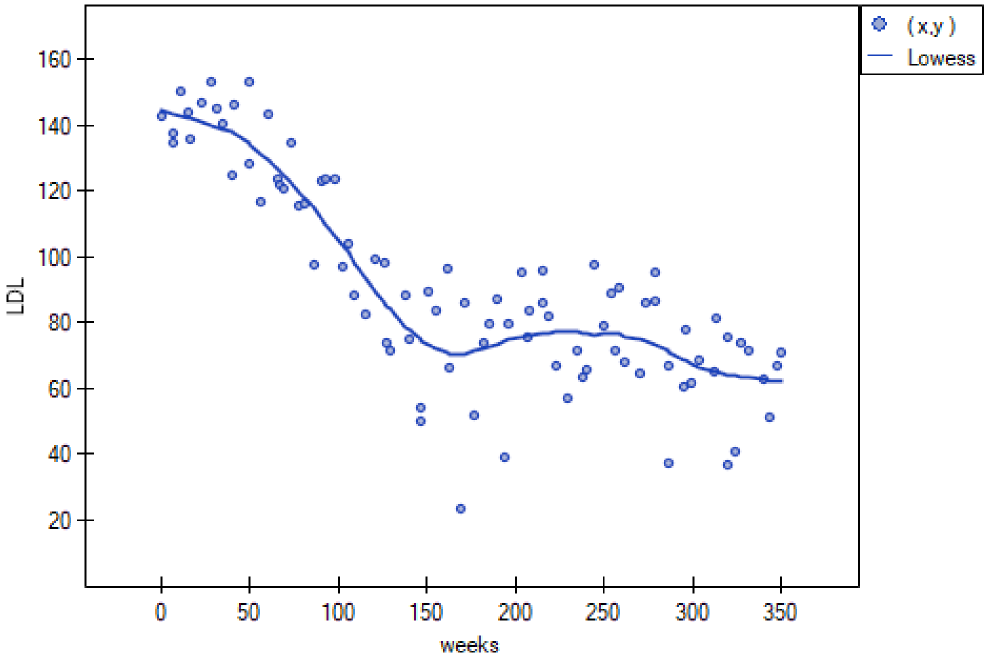
Hypotheses:

$\mathcal{H}_0$: In the population, there is no monotonic relationship between treatment time and LDL levels,
$\mathcal{H}_1$: In the population, there is a monotonic relationship between treatment time and LDL levels.



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

Comparing p<0.0001 with a significance level $\alpha = 0.05$ we find that there is a statistically significant monotonic relationship between treatment time and LDL levels. This relationship is initially decreasing and begins to stabilize after 150 weeks. The Spearman's monotonic correlation coefficient and therefore the strength of the monotonic relationship for this relationship is quite high at r_s=-0.78. The graph was plotted by curve fitting through local LOWESS linear smoothing techniques.



20.2.3 Significance Test for Kendall's tau

The test of significance for the Kendall's $\tilde{\tau}$ correlation coefficient is used to verify the hypothesis determining the lack of monotonic correlation between analysed features of population. It is based on the Kendall's tau correlation coefficient calculated for the sample. The closer to 0 the value of $\tilde{\tau}$ is, the weaker dependence joins the analysed features.

Basic assumptions:

– measurement on an ordinal scale or on an interval scale.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & \tau = 0,\\ \mathcal{H}_1: & \tau \neq 0.\end{array}$$

The test statistic is defined by:

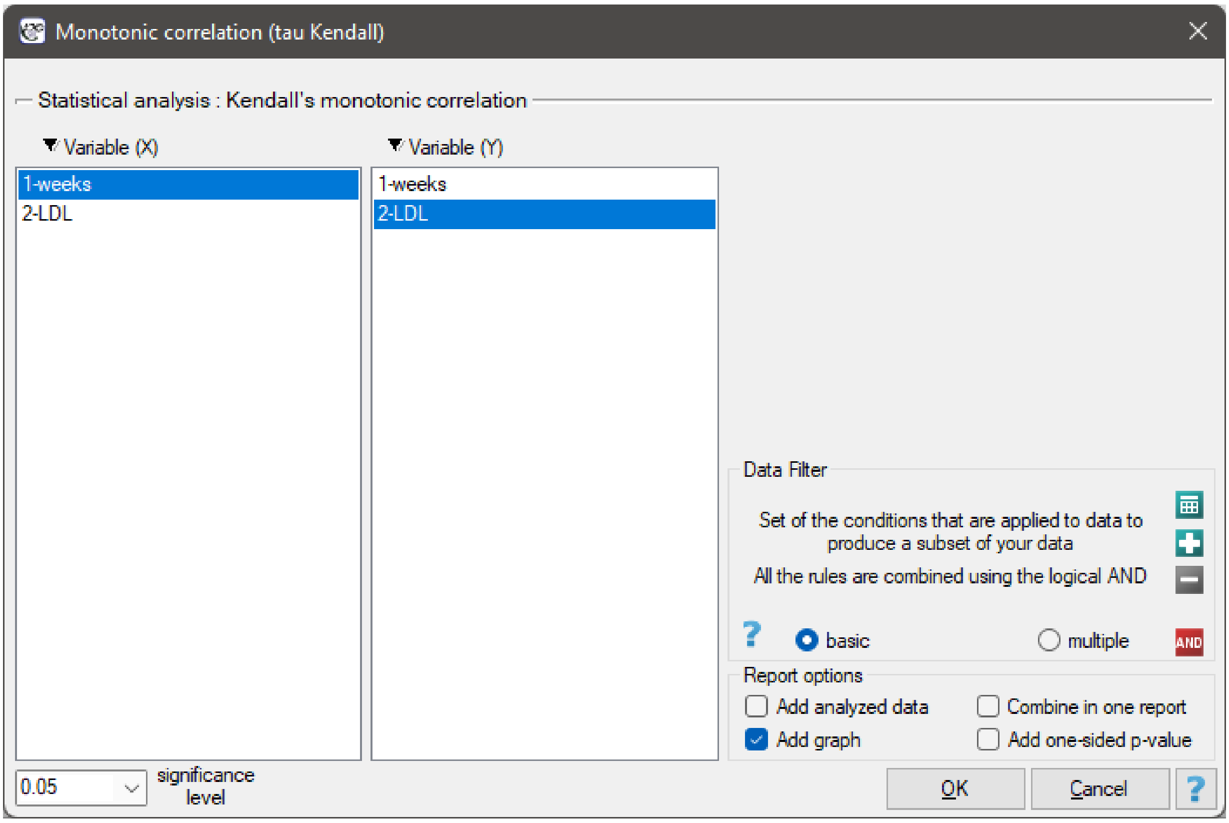
$$Z = \frac{3\tilde{\tau}\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}.$$

The test statistic asymptotically (for a large sample size) has the normal distribution.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

The settings window with the Kendall's monotonic correlation can be opened in Statistics menu → NonParametric tests→monotonic correlation (tau-Kendall) or in Wizard.



Monotonic correlation (tau Kendall)

Statistical analysis : Kendall's monotonic correlation

Variable (X): 1-weeks (selected), 2-LDL

Variable (Y): 1-weeks, 2-LDL (selected)

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

☑ Add graph □ Add one-sided p-value

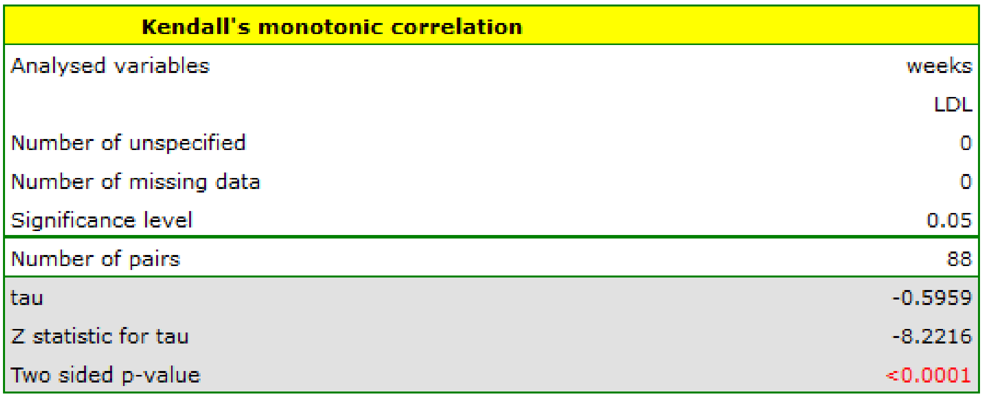
0.05 significance level

Example (20.2) c.d. *(LDL weeks.pqs file)*

Hypotheses:

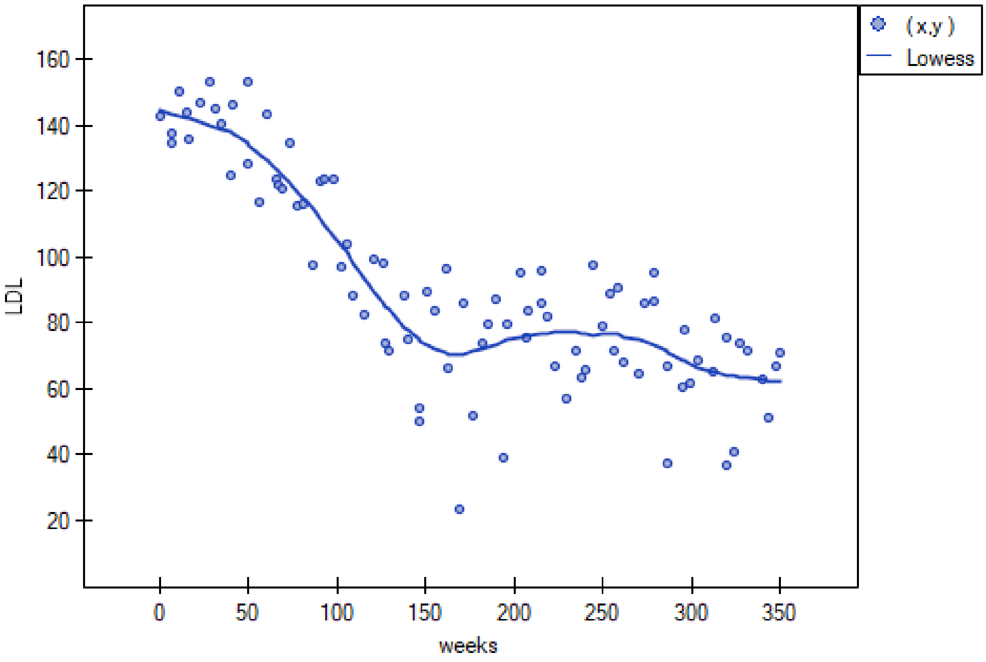
$\mathcal{H}_0:$ In the population, there is no monotonic relationship between treatment time and LDL levels,

$\mathcal{H}_1:$ In the population, there is a monotonic relationship between treatment time and LDL levels.



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

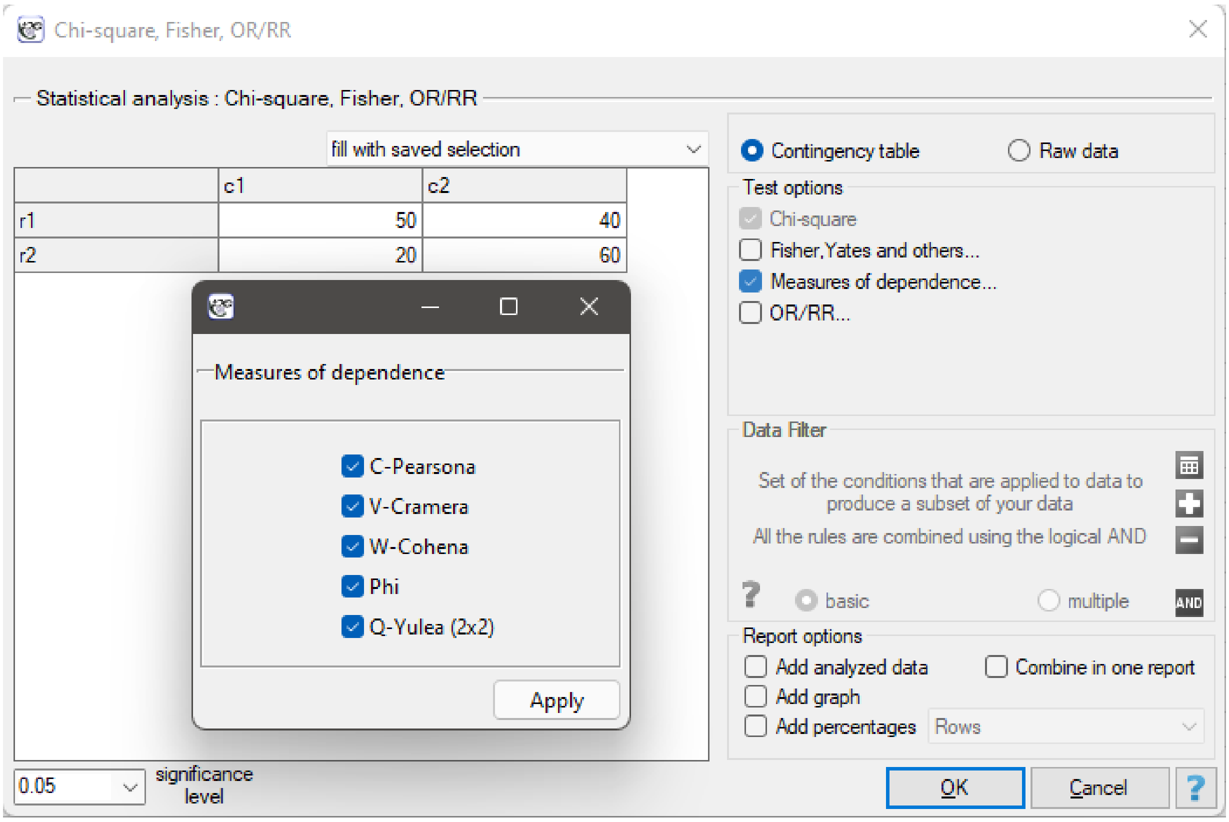
Comparing p<0.0001 with a significance level $\alpha = 0.05$ we find that there is a statistically significant monotonic relationship between treatment time and LDL levels. This relationship is initially decreasing and begins to stabilize after 150 weeks. The Kendall's monotonic correlation coefficient, and therefore the strength of the monotonic relationship for this relationship is quite high at $\tilde{\tau}$=-0.60. The graph was plotted by curve fitting through local LOWESS linear smoothing techniques.



20.2.4 CONTINGENCY TABLES COEFFICIENTS AND THEIR STATISTICAL SIGNIFICANCE

The contingency coefficients are calculated for the raw data or the data gathered in a contingency table (look at the table (10.1)).

The settings window with the measures of correlation can be opened in Statistics menu → NonParametric tests → Ch-square, Fisher, OR/RR option Measures of dependence... or in Wizard.



Chi-square, Fisher, OR/RR

Statistical analysis : Chi-square, Fisher, OR/RR

fill with saved selection

	c1	c2
r1	50	40
r2	20	60

Measures of dependence

☑ C-Pearsona
☑ V-Cramera
☑ W-Cohena
☑ Phi
☑ Q-Yulea (2x2)

Apply

● Contingency table ○ Raw data

Test options

☑ Chi-square
□ Fisher,Yates and others...
☑ Measures of dependence...
□ OR/RR...

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

○ basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report
□ Add graph
□ Add percentages Rows

0.05 significance level

OK Cancel

The Yule's Q contingency coefficient

The Yule's Q contingency coefficient (Yule, 1900[176]) is a measure of correlation, which can be calculated for 2×2 contingency tables.

$$Q = \frac{O_{11}O_{22} - O_{12}O_{21}}{O_{11}O_{22} + O_{12}O_{21}},$$

where:
$O_{11}, O_{12}, O_{21}, O_{22}$ - observed frequencies in a contingency table.

The Q coefficient value is included in a range of $< -1; 1 >$. The closer to 0 the value of the Q is, the weaker dependence joins the analysed features, and the closer to -1 or +1, the stronger dependence joins the analysed features. There is one disadvantage of this coefficient. It is not much resistant to small observed frequencies (if one of them is 0, the coefficient might wrongly indicate the total dependence of features).

The statistic significance of the Yule's Q coefficient is defined by the Z test.
Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & Q = 0, \\ \mathcal{H}_1: & Q \neq 0. \end{array}$$

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 a large sample size) has the normal distribution.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The ϕ contingency coefficient

The Phi contingency coefficient is a measure of correlation, which can be calculated for 2×2 contingency tables.

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

where:
χ^2 – value of the χ^2 test statistic,
n – total frequency in a contingency table.

The ϕ coefficient value is included in a range of $< 0; 1 >$. The closer to 0 the value of ϕ is, the weaker dependence joins the analysed features, and the closer to 1, the stronger dependence joins the analysed features.

The ϕ contingency coefficient is considered as **statistically significant**, if the p-value calculated on the basis of the χ^2 test (designated for this table) is equal to or less than the significance level α.

The Cramer's V contingency coefficient

The Cramer's V contingency coefficient (Cramer, 1946[48]), is an extension of the ϕ coefficient on $r \times c$ contingency tables.

$$V = \sqrt{\frac{\chi^2}{n(w'-1)}},$$

where:
χ^2 – value of the χ^2 test statistic,
n – total frequency in a contingency table,
w' – the smaller the value out of r and c.

The V coefficient value is included in a range of $< 0; 1 >$. The closer to 0 the value of V is, the weaker dependence joins the analysed features, and the closer to 1, the stronger dependence joins the analysed features. The V coefficient value depends also on the table size, so you should not use this coefficient to compare different sizes of contingency tables.

The V contingency coefficient is considered as **statistically significant**, if the p-value calculated on the basis of the χ^2 test (designated for this table) is equal to or less than the significance level α.

W-Cohen contingency coefficient

The W-Cohen contingency coefficient (Cohen (1988)[45]), is a modification of the V-Cramer coefficient and is computable for $r \times c$ tables.

$$W = \sqrt{\frac{\chi^2}{n(w'-1)}}\sqrt{w'-1},$$

where:
χ^2 – value of the χ^2 test statistic,
n – total frequency in a contingency table,
w' – the smaller the value out of r and c.

The W coefficient value is included in a range of $< 0; \max W >$, where $\max W = \sqrt{w' - 1}$ (for tables where at least one variable contains only two categories, the value of the coefficient W is in the range $< 0; 1 >$). The closer to 0 the value of W is, the weaker dependence joins the analysed features, and the closer to $\max W$, the stronger dependence joins the analysed features. The W coefficient value depends also on the table size, so you should not use this coefficient to compare different sizes of contingency tables.

The W contingency coefficient is considered as **statistically significant**, if the p-value calculated on the basis of the χ^2 test (designated for this table) is equal to or less than the significance level α.

The Pearson's C contingency coefficient

The Pearson's C contingency coefficient is a measure of correlation, which can be calculated for $r \times c$ contingency tables.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

where:
χ^2 – value of the χ^2 test statistic,
n – total frequency in a contingency table.

The C coefficient value is included in a range of $< 0; 1)$. The closer to 0 the value of C is, the weaker dependence joins the analysed features, and the farther from 0, the stronger dependence joins the analysed features. The C coefficient value depends also on the table size (the bigger table, the closer to 1 C value can be), that is why it should be calculated the top limit, which the C coefficient may gain – for the particular table size:

$$C_{max} = \sqrt{\frac{w' - 1}{w}},$$

where:
w' – the smaller value out of r and c.

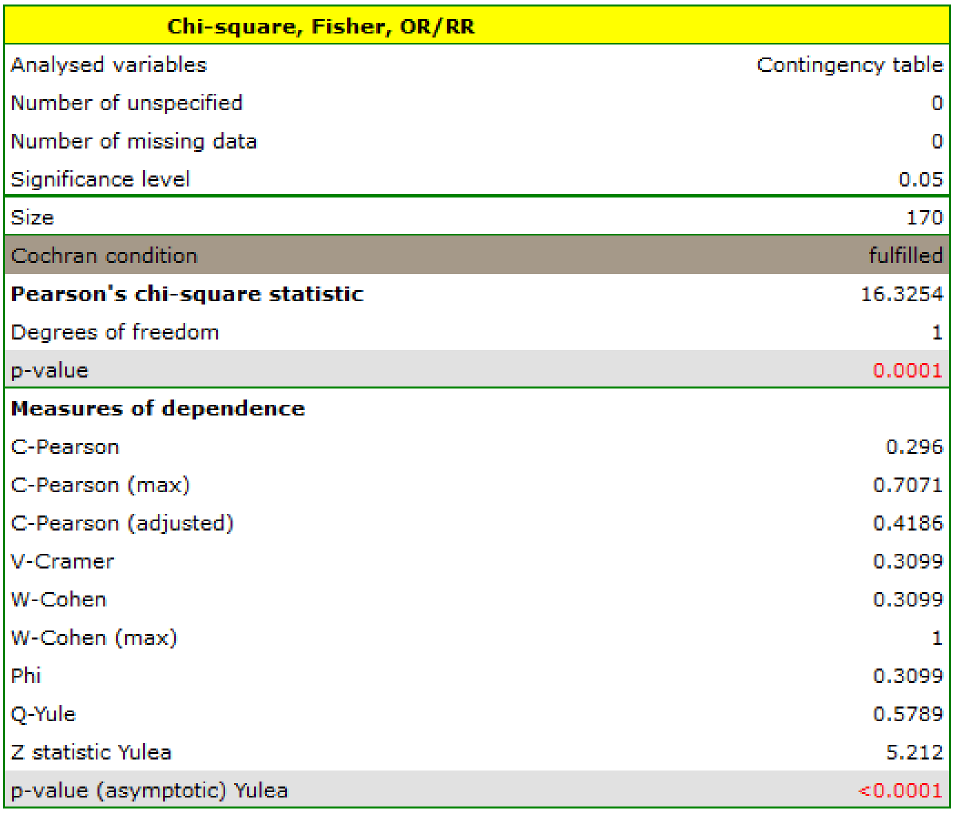
An uncomfortable consequence of dependence of C value on a table size is the lack of possibility of comparison the C coefficient value calculated for the various sizes of contingency tables. A little bit better measure is a contingency coefficient adjusted for the table size (C_{adj}):

$$C_{adj} = \frac{C}{C_{max}}.$$

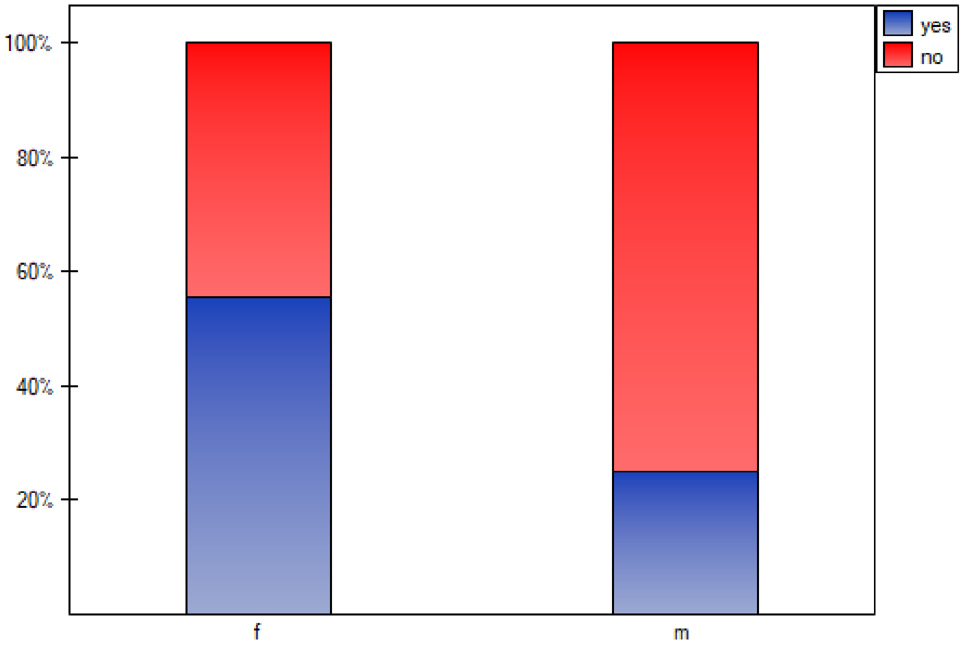
The C contingency coefficient is considered as **statistically significant**, if the p-value calculated on the basis of the χ^2 test (designated for this table) is equal to or less than significance level α.

Example 20.3. (sex-exam.pqs file)
There is a sample of 170 persons ($n = 170$), who have 2 features analysed (X=sex, Y=passing the exam). Each of these features occurs in 2 categories (X_1=f, X_2=m, Y_1=yes, Y_2=no). Basing on the sample, we would like to get to know, if there is any dependence between sex and passing the exam in an analysed population. The data distribution is presented in a 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



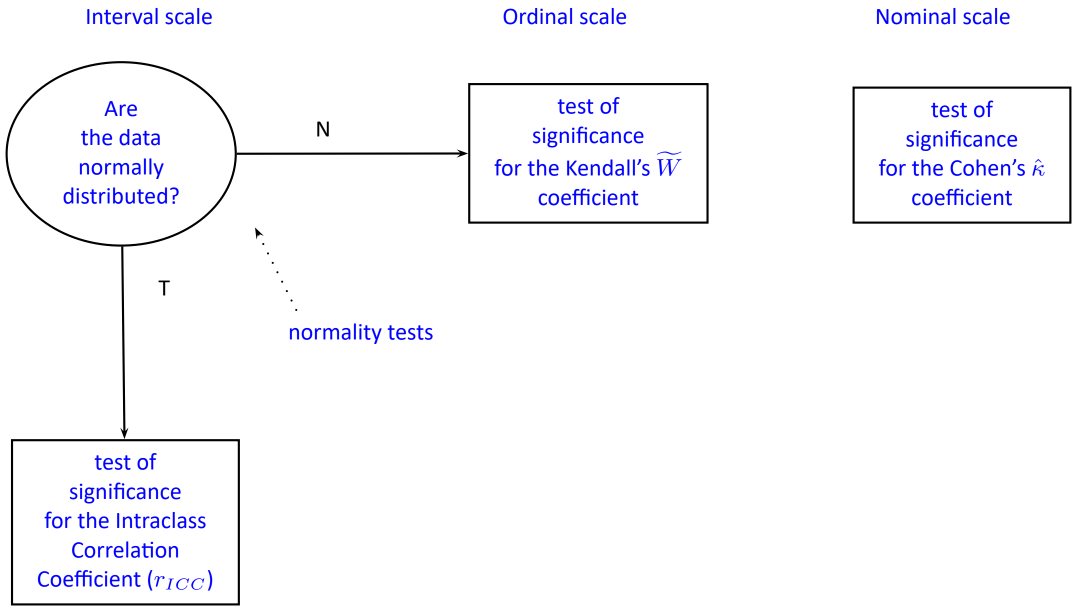
The test statistic value is $\chi^2 = 16.33$ and the p value calculated for it: $p < 0.0001$. The result indicates that there is a statistically significant dependence between sex and passing the exam in the analysed population.

Coefficient values, which are based on the χ^2 test, so the strength of the correlation between analysed features are:

C_{adj}-Pearson = 0.42.
V-Cramer = ϕ = W-Cohen = 0.31

The Q-Yule = 0.58, and the p value of the Z test (similarly to χ^2 test) indicates the statistically significant dependence between the analysed features.

21 AGREEMENT ANALYSIS


Interval scale
Ordinal scale
Nominal scale
Are the data normally distributed?
N
T
normality tests
test of significance for the Kendall's $\widetilde{W}$ coefficient
test of significance for the Cohen's $\hat{\kappa}$ coefficient
test of significance for the Intraclass Correlation Coefficient (r_{ICC})

21.1 PARAMETRIC TESTS

21.1.1 The Intraclass Correlation Coefficient and a test to examine its significance

The intraclass correlation coefficient is used when the measurement of variables is done by a few ”raters” ($k \geq 2$). It measures the strength of **interrater reliability** $-$ the degree of its assessment concordance.

Since it can be determined in several different situations, there are several variations depending on the model and the type of concordance. Depending on the variability present in the data, we can distinguish between 2 main research models and 2 types of concordance.

Model 1 For each of the n randomly selected judged objects, a set of k judges is randomly selected from the population of judges. Whereby for each object a different set of k judges can be drawn.

The ICC coefficient is then determined by the random model ANOVA for independent groups. The question of the reliability of a single judge's ratings is answered by ICC(1,1) given by the formula:

$$ICC(1,1)=\frac{MS_{WG}-MS_{BG}}{MS_{WG}+(k-1)MS_{BG}}.$$

To estimate the reliability of scores that are the average of the judges' ratings (for k judges), determine ICC(1,k) given by the formula:

$$ICC(1,k)=\frac{MS_{WG}-MS_{BG}}{MS_{WG}},$$

where:
MS_{WG} – mean of squares within groups,
MS_{BG} – mean of squares between objects.

Model 2 A set of k judges is randomly selected from a population of judges and each judge evaluates all n random objects. The ICC coefficient is then determined in a random model ANOVA for dependent groups.

Depending on the type of concordance we are looking for, we can estimate: absolute agreement, i.e., if the judges agree absolutely, they give exactly the same ratings, e.g., perfectly concordant will be such ratings given by a pair of judges (2,2), (5,5), (8,8); or consistency, i.e., the judges may use different ranges of values but beyond this shift there should be no differences to keep the verdict consistent, e.g., perfectly consistent will be such ratings given by a pair of judges (2,5), (5,8), (8,11).

Absolute agreement

The question about the reliability of a single judge's ratings is answered by ICC(2,1) given by the formula:

$$ICC(2,1)=\frac{MS_{BS}-MS_{res}}{MS_{BS}+(k-1)MS_{res}+\frac{k}{n}(MS_{BC}-MS_{res})}.$$

To estimate the reliability of scores that are the average of the judges' ratings (for k judges), determine ICC(2,k) given by the formula:

$$ICC(2,k)=\frac{MS_{BS}-MS_{res}}{MS_{BS}+(MS_{BC}-MS_{res})/n},$$

where:
MS_{BC} – the mean of squares between judges,
MS_{BS} – the mean of squares between objects,
MS_{res} – mean of squares for the residuals.

Consistency

The question about the reliability of a single judge's ratings is answered by ICC(2,1) given by the formula:

$$ICC(2,1) = \frac{MS_{BS} - MS_{res}}{MS_{BS} + (k-1)MS_{res}},$$

To estimate the reliability of scores that are the average of the judges' ratings (for k judges), determine ICC(2,k) given by the formula:

$$ICC(2,k) = \frac{MS_{BS} - MS_{res}}{MS_{BS}},$$

where:
MS_{BS} – the mean of squares between objects,
MS_{res} – mean of squares for the residuals.

Note

Sometimes, there is a need to consider **model 3** [149], i.e., a set of k judges is selected and each judge evaluates all n random objects. The concordance score applies only to these particular k judges. The ICC coefficient is then determined in a mixed model (since the randomness only affects the objects and not the judges). Since we are ignoring the variability concerning the judges, we are examining consistency (rather than absolute agreement) and the coefficients from the second model may apply: ICC(2,1) and ICC (2,k), since they are the same as the coefficients ICC(3,1) and ICC (3,k) desired in this case under the assumption of no interaction between objects and judges.

Note

We interpret the value $ICC \in < -1; 1 >$ as follows:

- $ICC \approx 1$ it is an strong concordance of objects assessment made by judgess; it is especially reflected in a high-variance between objects (a significant means difference between n objects) and a low-variance between judges assessments (a small means difference of assessments designated by k judges);
- $ICC \approx -1$ a negative intraclass coefficient is treated in the same ways as $r_{ICC} \approx 0$;
- $ICC \approx 0$ denotes a lack of concordance in the judges' evaluations of individual objects, as reflected by low variance between objects (a small difference in means between n objects) and high variance between judges' evaluations (a significant difference in mean scores determined for k judges).

F-test for testing the significance of intraclass correlation coefficient

Basic assumptions:

- measurement on an interval scale,
- the normal distribution for all variables which are the differences of measurement pairs (or the normal distribution for an analysed variable in each measurement),
- for model 1 - independent model, for model 2 / 3 - dependent model.

Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & ICC = 0\\ \mathcal{H}_1: & ICC \neq 0 \quad (ICC = 1)\end{array}$$

The test statistic has the form:

$$F = \frac{MS_{BS}}{MS_{res}} \text{ – in the dependent model,}$$

or

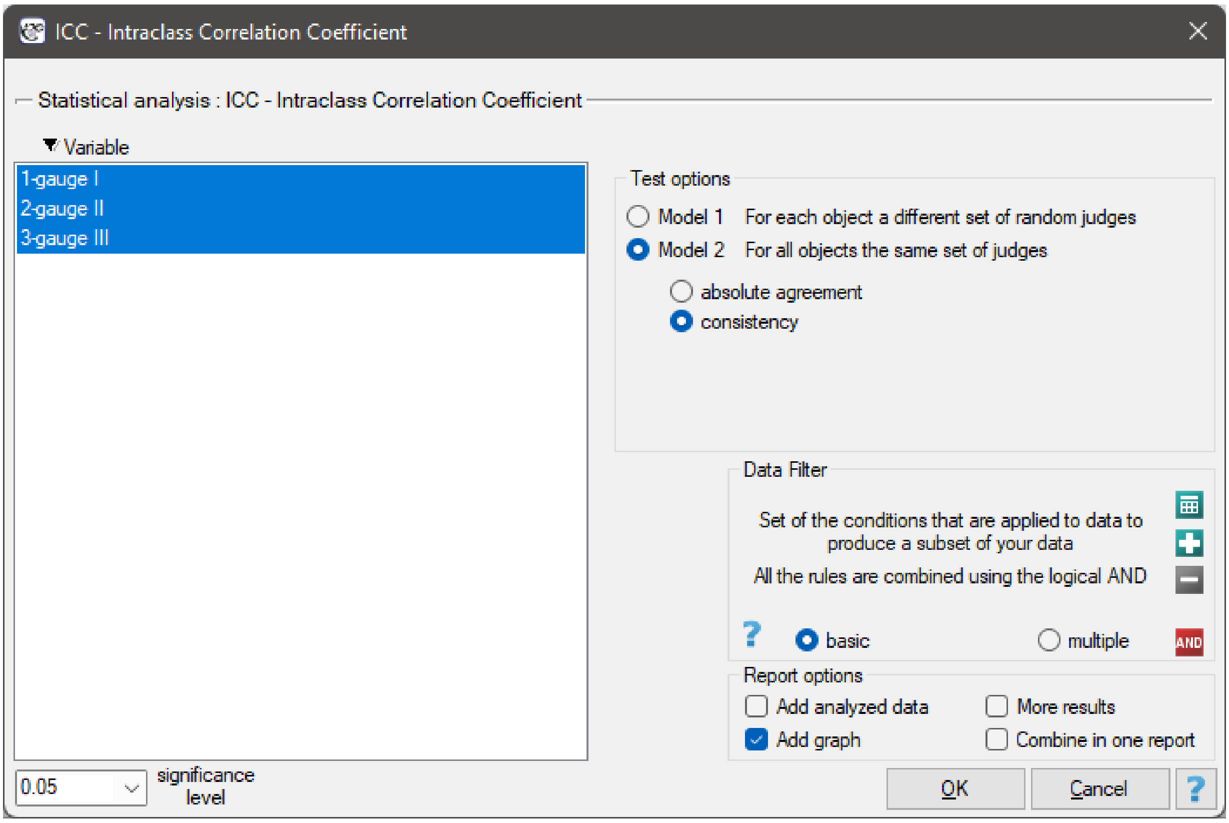
$$F = \frac{MS_{WG}}{MS_{BG}} \text{ – in the independent model.}$$

This statistic has the F Snedecor distribution with the number of degrees of freedom defined in the model.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

The settings window with the ICC – Intraclass Correlation Coefficient can be opened in Statistics menu→Parametric tests→ICC – Intraclass Correlation Coefficient or in Wizard.



ICC - Intraclass Correlation Coefficient

Statistical analysis : ICC - Intraclass Correlation Coefficient

Variable
- 1-gauge I
- 2-gauge II
- 3-gauge III

Test options
○ Model 1 For each object a different set of random judges
● Model 2 For all objects the same set of judges
○ absolute agreement
● consistency

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

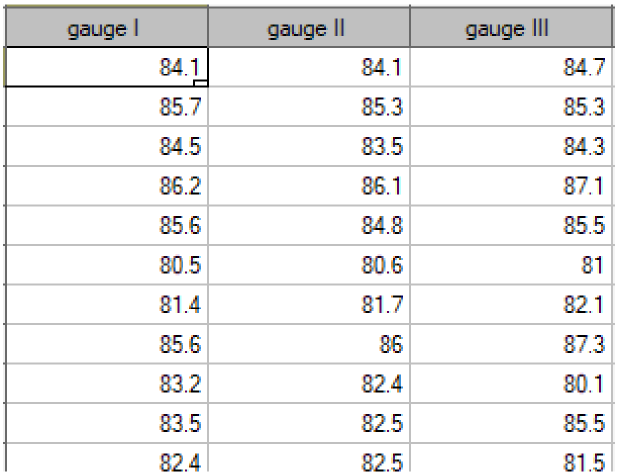
Report options
□ Add analyzed data □ More results
☑ Add graph □ Combine in one report

0.05 significance level

OK Cancel

Example 21.1. (sound intensity.pqs file)

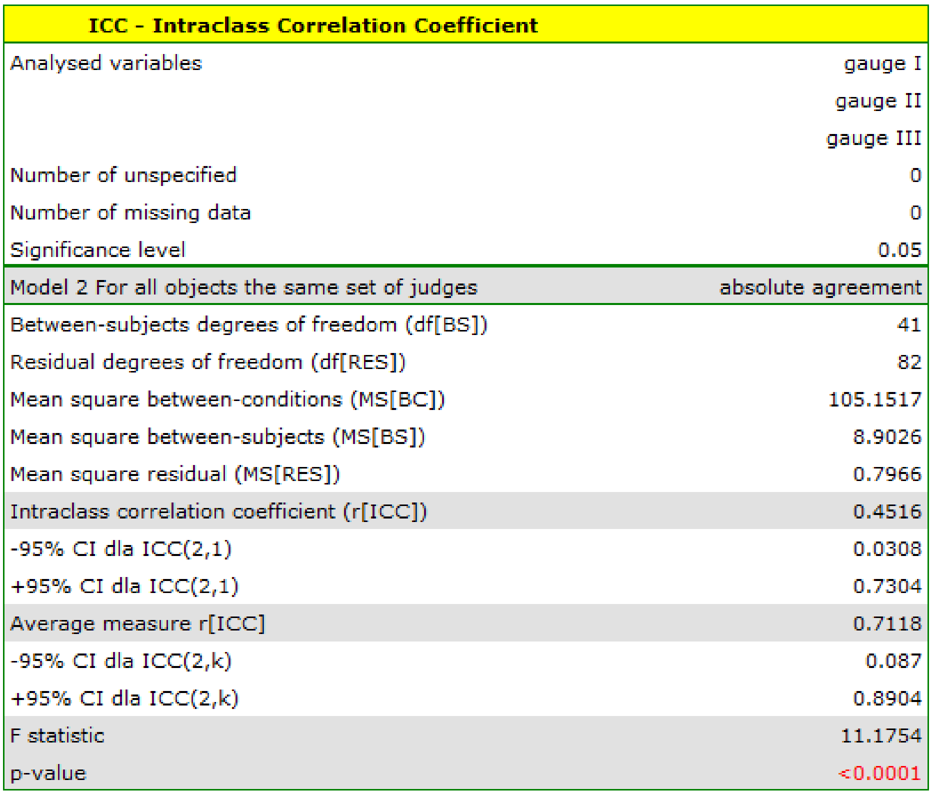
In order to effectively care for the hearing of workers in the workplace, it is first necessary to reliably estimate the sound intensity in the various areas where people are present. One company decided to conduct an experiment before choosing a sound intensity meter (sonograph). Measurements of sound intensity were made at 42 randomly selected measurement points in the plant using 3 drawn analog sonographs and 3 randomly selected digital sonographs. A part of collected measurements is presented 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

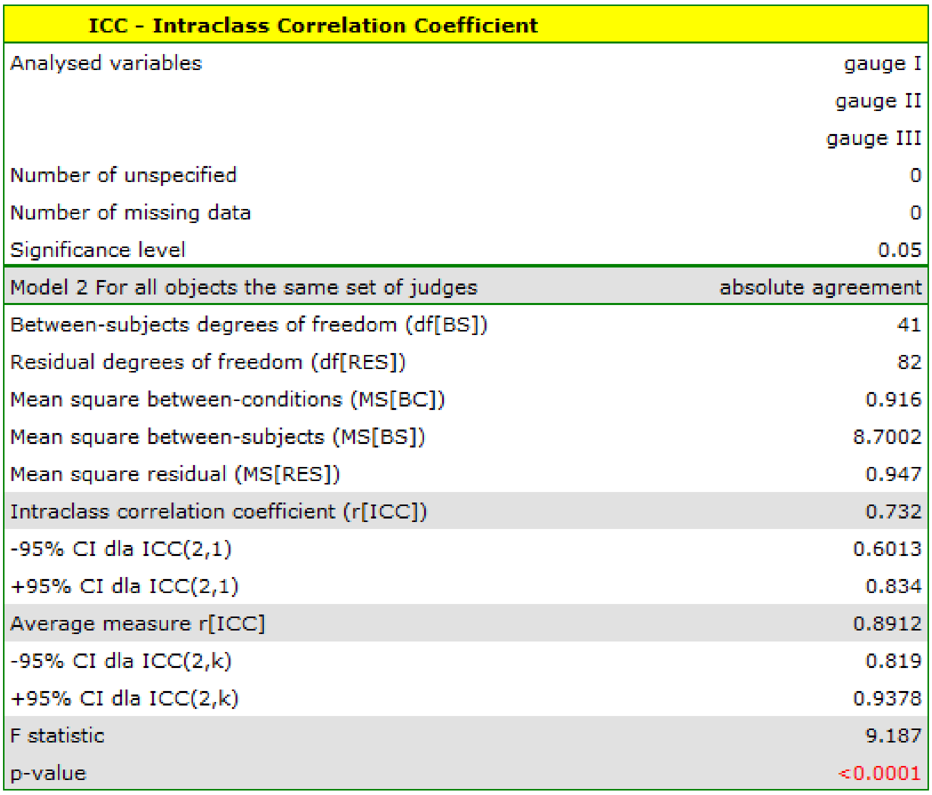
To find out which type of instrument (analog or digital) will better accomplish the task at hand, the ICC in model 2 should be determined by examining the absolute agreement. The type of meter with the higher ICC will have more reliable measurements and will therefore be used in the future.

The analysis performed for the analog meters shows significant consistency of the measurements ($p < 0.0001$). The reliability of the measurement made by the analog meter is $ICC(2,1) = 0.45$, while the reliability of the measurement that is the average of the measurements made by the 3 analog meters is slightly higher and is $ICC(2,k) = 0.71$. However, the lower limit of the 95 percent confidence interval for these coefficients is disturbingly low.



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

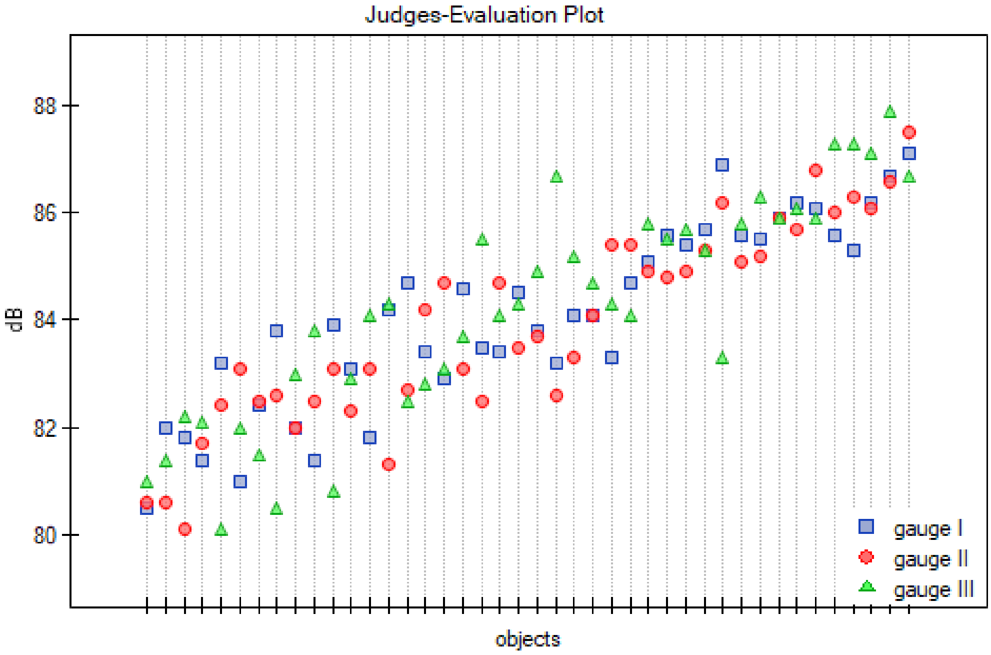
A similar analysis performed for digital meters produced better results. The model is again statistically significant, but the ICC coefficients and their confidence intervals are much higher than for analog meters, so the absolute agreement obtained is higher $ICC(2,1) = 0.73$, $ICC(2,k) = 0.89$.



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

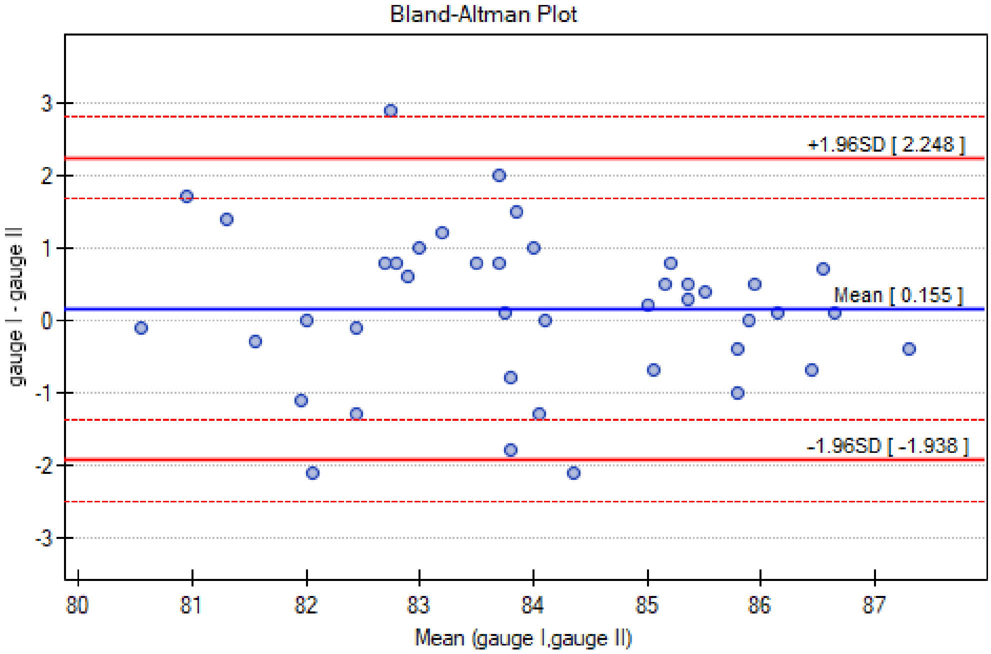
Therefore, eventually digital meters will be used in the workplace.

The agreement of the results obtained for the digital meters is shown in a dot plot, where each measurement point is described by the sound intensity value obtained for each meter.



By presenting a graph for the previously sorted data according to the average value of the sound intensity, one can check whether the degree of agreement increases or decreases as the sound intensity increases. In the case of our data, a slightly higher correspondence (closeness of positions of points on the graph) is observed at high sound intensities.

Similarly, the consistency of the results obtained can be observed in the Blanda-Altmana graphs[8][22] constructed separately for each pair of meters. The graph for Meter I and Meter II is shown below.



Here, too, we observe higher agreement (points are concentrated near the horizontal axis y=0) for higher sound intensity values.

Note
If the researcher was not concerned with estimating the actual sound level at the worksite, but wanted to identify where the sound level was higher than at other sites or to see if the sound level varied over time, then Model 2, which tests consistency, would be a sufficient model.

21.2 NON-PARAMETRIC TESTS

21.2.1 The Kendall's concordance coefficient and a test to examine its significance

The Kendall's $\widetilde{W}$ coefficient of concordance is described in the works of Kendall, Babington-Smith (1939)[90] and Wallis (1939)[164]. It is used when the result comes from different sources (from different raters) and concerns a few ($k \geq 2$) objects. However, the assessment concordance is necessary. Is often used in measuring the **interrater reliability** strength – the degree of (raters) assessment concordance.

The Kendall's coefficient of concordance is calculated on an ordinal scale or a interval scale. Its value is calculated according to the following formula:

$$\widetilde{W} = \frac{12U - 3n^2k(k+1)^2}{n^2k(k^2-1) - nC},$$

where:
n – number of different assessments sets (the number of raters),
k – number of ranked objects,
$U = \sum_{j=1}^{k}\left(\sum_{i=1}^{n} R_{ij}\right)^2$,
R_{ij} – ranks ascribed to the following objects $(j = 1, 2, ...k)$, independently for each rater

$(i = 1, 2, ...n)$,
$C = \sum(t^3 - t)$ – a correction for ties,
t – number of cases incorporated into tie.

The coefficient's formula includes C – the correction for ties. This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $C = 0$).

Note
W – the Kendall's coefficient in a population;
$\widetilde{W}$ – the Kendall's coefficient in a sample.

The value of $W \in < 0; 1 >$ and it should be interpreted in the following way:

- $\widetilde{W} \approx 1$ means a strong concordance in raters assessments;
- $\widetilde{W} \approx 0$ means a lack of concordance in raters assessments.

The Kendall's $\widetilde{W}$ coefficient of concordance vs. the Spearman r_s coefficient:

When the values of the Spearman r_s correlation coefficient (for all possible pairs) are calculated, the **average** r_s **coefficient** – marked by $\bar{r}_s$ is a linear function of $\widetilde{W}$ coefficient:

$$\bar{r}_s = \frac{n\widetilde{W} - 1}{n - 1}$$

The Kendall's $\widetilde{W}$ coefficient of concordance vs. 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 chi-square test statistic for the Kendall's coefficient of concordance and the value of the chi-square test statistic for the Friedman ANOVA are the same.

The chi-square test of significance for the Kendall's coefficient of concordance

Basic assumptions:

- measurement on an ordinal scale or on an interval scale.

Hypotheses:

$$\begin{array}{ll} \mathcal{H}_0: & W = 0 \\ \mathcal{H}_1: & W \neq 0 \end{array}$$

The test statistic is defined by:

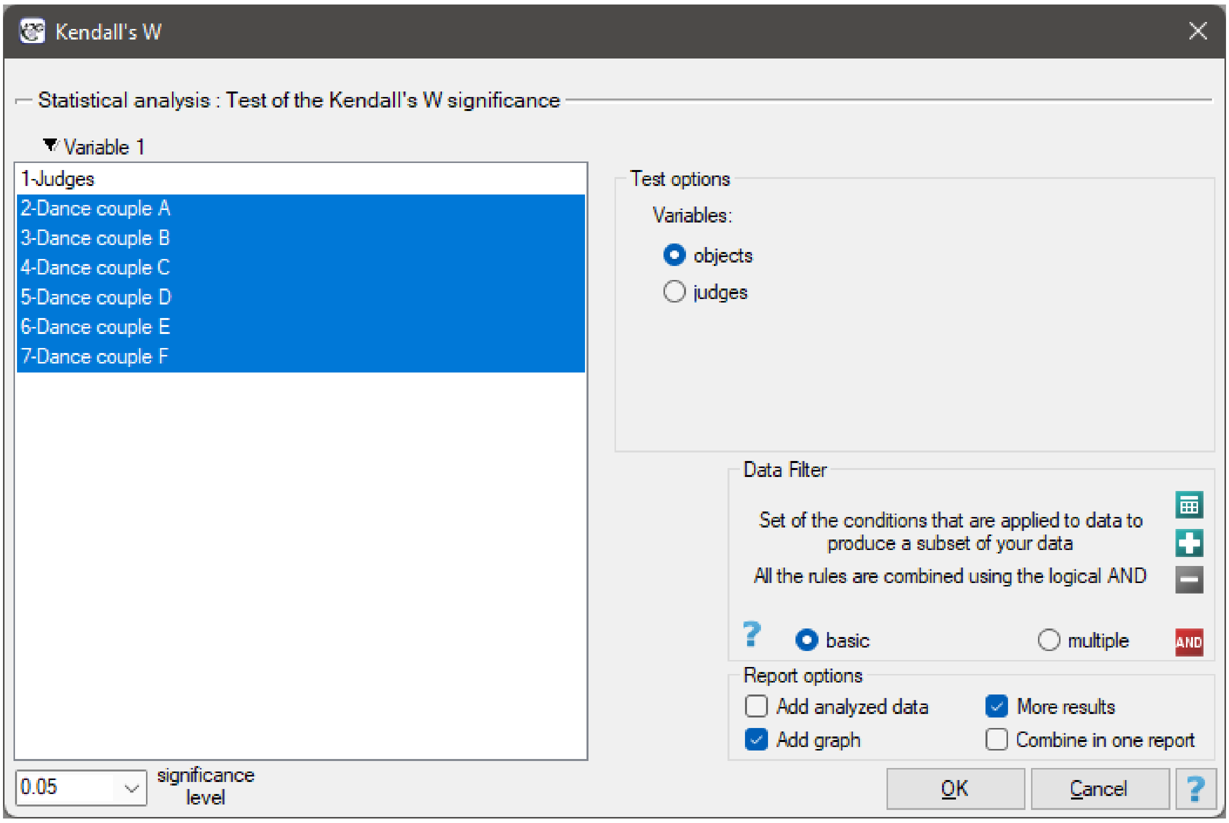
$$\chi^2 = n(k - 1)\widetilde{W}$$

This statistic asymptotically (for large sample sizes) has the χ^2 distribution with the degrees of freedom calculated according to the following formula: $df = k - 1$.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the test of the Kendall's W significance can be opened in Statistics menu →NonParametric tests→Kendall's W or in Wizard.



Kendall's W

Statistical analysis : Test of the Kendall's W significance

Variable 1

1-Judges
2-Dance couple A
3-Dance couple B
4-Dance couple C
5-Dance couple D
6-Dance couple E
7-Dance couple F

Test options

Variables:

● objects
○ judges

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data ☑ More results
☑ Add graph □ Combine in one report

0.05 significance level

Example 21.2. (judges.pqs file)

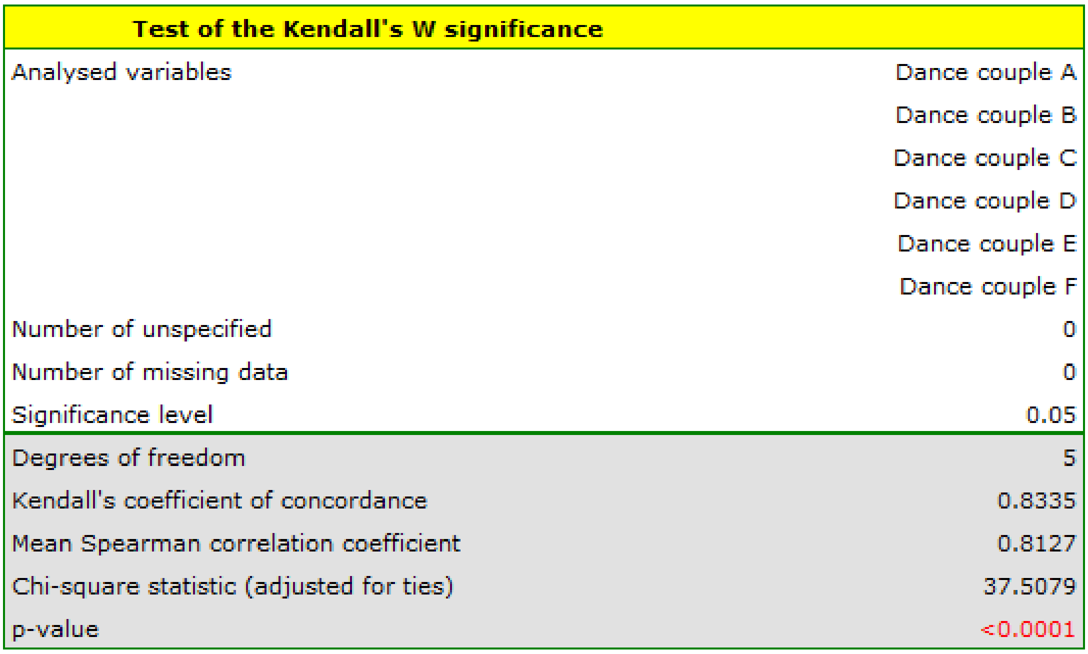
In the 6.0 system, dancing pairs grades are assessed by 9 judges. The judges point for example an artistic expression. They asses dancing pairs without comparing each of them and without placing them in the particular ”podium place” (they create a ranking). Let’s check if the judges assessments are concordant.

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

Hypotheses:

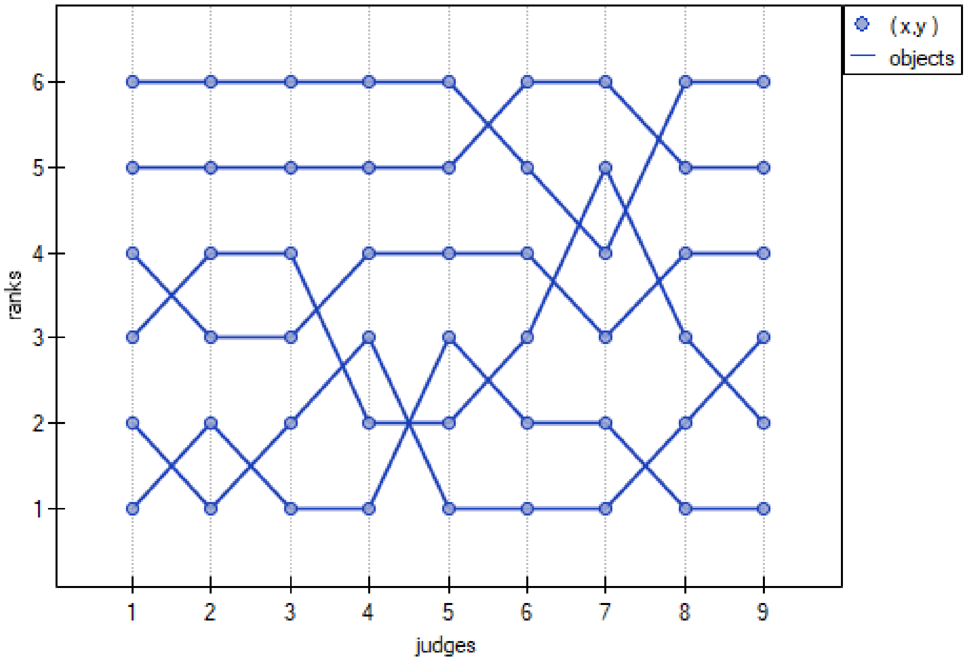
$\mathcal{H}_0$: a lack of concordance between 9 judges assessments, in the population represented by the sample,

$\mathcal{H}_1$: the 9 judges assessments in the population represented by the sample are concordant.



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

Comparing the $p < 0.0001$ with the significance level $\alpha = 0.05$, we have stated that the judges assessments are statistically concordant. The concordance strength is high: $\widetilde{W} = 0.83$, similarly the average Spearman's rank-order correlation coefficient: $\bar{r}_s = 0.81$. This result can be presented in the graph, where the X-axis represents the successive judges. Then the more intersection of the lines we can see (the lines should be parallel to the X axis, if the concordance is perfect), the less there is the concordance of rateres evaluations.



21.2.2 The Cohen's Kappa coefficient and the test examining its significance

The Cohen's Kappa coefficient (Cohen J. (1960)[43]) defines the agreement level of two-times measurements of the same variable in different conditions. Measurement of the same variable can be performed by 2 different observers (reproducibility) or by a one observer twice (recurrence). The $\hat{\kappa}$ coefficient is calculated for categorial dependent variables and its value is included in a range from -1 to 1. A 1 value means a full agreement, 0 value means agreement on the same level which would occur for data spread in a contingency table randomly. The level between 0 and -1 is practically not used. The negative $\hat{\kappa}$ value means an agreement on the level which is lower than agreement which occurred for the randomly spread data in a contingency table. The $\hat{\kappa}$ coefficient can be calculated on the basis of raw data or a $c \times c$ contingency table.

Unweighted Kappa (i.e., Cohen's Kappa) or weighted Kappa can be determined as needed. The assigned weights (w_{ij}) refer to individual cells of the contingency table, on the diagonal they are 1 and off the diagonal they belong to the range $< 0; 1)$.

Unweighted Kappa

It is calculated for data, the categories of which cannot be ordered, e.g. data comes from patients, who are divided according to the type of disease which was diagnosed, and these diseases cannot be ordered, e.g. pneumonia (1), bronchitis (2) and other (3). In such a situation, one can check the concordance of the diagnoses given by the two doctors by using unweighted Kappa, or Cohen's Kappa. Discordance of pairs $(1),(3)$ and $(1),(2)$ will be treated equivalently, so the weights off the diagonal of the weight matrix will be zeroed.

Weighted Kappa

In situations where data categories can be sorted, e.g., data comes from patients who are divided by the lesion grade into: no lesion (1), benign lesion (2), suspected cancer (3), cancer (4), one can build the concordance of the ratings given by the two radiologists taking into account the possibility of sorting. The ratings of $(1),(4)$ than $(1),(2)$ may then be considered as more discordant pairs of ratings. For this to be the case, so that the order of the categories affects the compatibility score, the weighted Kappa should be determined.

The assigned weights can be in linear or quadratic form.

- **Linear weights** (Cicchetti, 1971[36]) – calculated according to the formula:

$$w_{ij} = 1 - \frac{|i-j|}{c-1}.$$

The greater the distance from the diagonal of the matrix the smaller the weight, with the weights decreasing proportionally. Example weights for matrices of size 5x5 are shown in the table:

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

- **Square weights** (Cohen, 1968[44]) – calculated according to the formula:

$$w_{ij} = 1 - \frac{(i-j)^2}{(c-1)^2}.$$

The greater the distance from the diagonal of the matrix, the smaller the weight, with weights decreasing more slowly at closer distances from the diagonal and more rapidly at farther distances. Example weights for matrices of size 5x5 are shown in the table:

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

Quadratic scales are of greater interest because of the practical interpretation of the Kappa coefficient, which in this case is the same as the intraclass correlation coefficient [60].

To determine the Kappa coefficient compliance, the data are presented in the form of a table of observed counts O_{ij} (15.3), and this table is transformed into a probability contingency table $p_{ij} = O_{ij}/n$.

The Kappa coefficient ($\hat{\kappa}$) is expressed by the formula:

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e},$$

where:
$P_o = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} p_{ij}$,
$P_e = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} p_{i.} p_{.i}$,
$p_{i.}$, $p_{.i}$ - total sums of columns and rows of the probability contingency table.

Note
$\hat{\kappa}$ denotes the concordance coefficient in the sample, while κ in the population.

The standard error for Kappa is expressed by the formula:

$$SE_{\hat{\kappa}} = \frac{1}{(1 - P_e)\sqrt{n}} \sqrt{\sum_{i=1}^{c} \sum_{j=1}^{c} p_{i.} p_{.j} [w_{ij} - (\overline{w}_{i.} + (\overline{w}_{.j})]^2 - P_e^2}$$

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 of significance for the Cohen's Kappa ($\hat{\kappa}$) (Fleiss,2003[61]) is used to verify the hypothesis informing us about the agreement of the results of two-times measurements $X^{(1)}$ and $X^{(2)}$ features X and it is based on the $\hat{\kappa}$ coefficient calculated for the sample.

Basic assumptions:

- measurement on a nominal scale (unweighted Kappa) or on a nominal scale (weighted Kappa).

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \kappa = 0, \\ \mathcal{H}_1: & \kappa \neq 0. \end{array}$$

The test statistic is defined by:

$$Z = \frac{\hat{\kappa}}{SE_{\kappa_{distr}}},$$

Where:

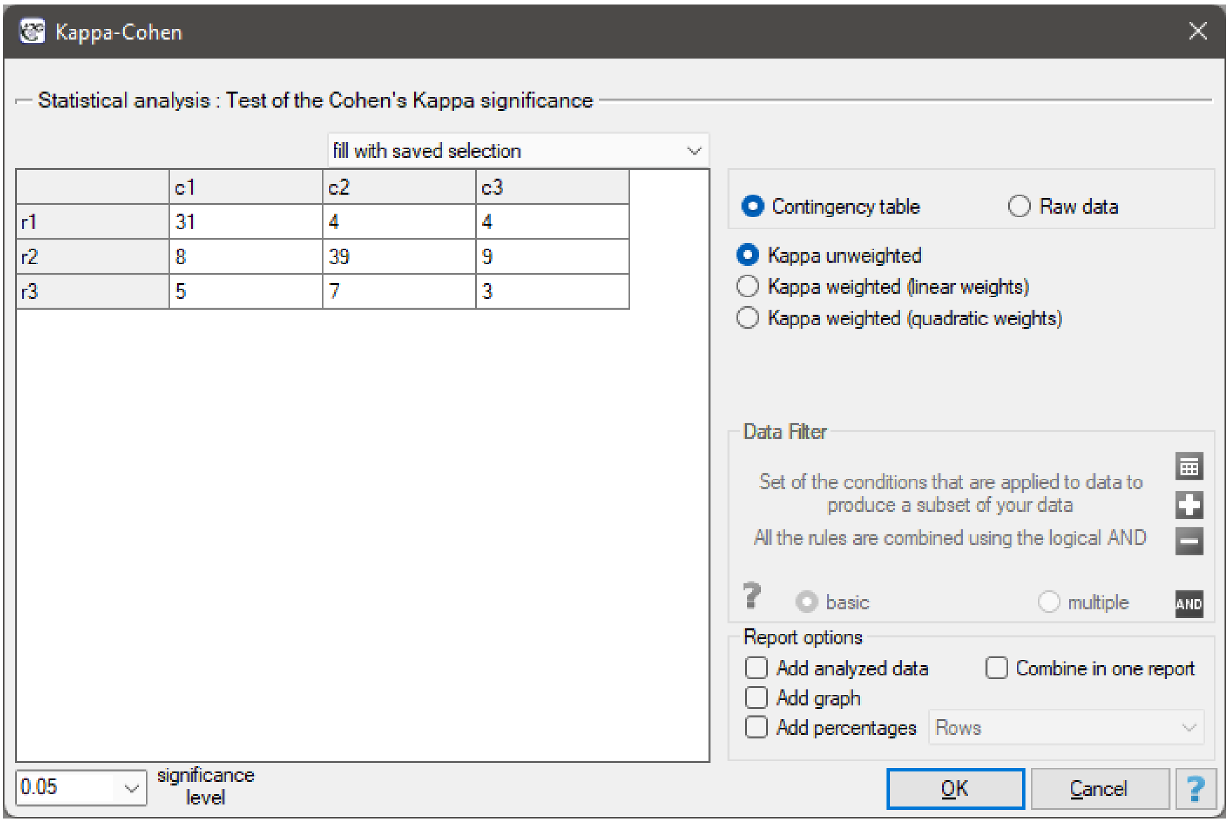
$$SE_{\kappa_{distr}} = \frac{1}{(1 - P_e)\sqrt{n}} \sqrt{\sum_{i=1}^{c} \sum_{j=1}^{c} p_{ij} [w_{ij} - (\overline{w}_{i.} + \overline{w}_{.j})(1 - \hat{\kappa})]^2 - [\hat{\kappa} - P_e(1 - \hat{\kappa})]^2}.$$

The Z statistic asymptotically (for a large sample size) has the normal distribution.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The settings window with the test of Cohen's Kappa significance can be opened in Statistics menu → NonParametric tests → Kappa-Cohen or in Wizard.



Kappa-Cohen

Statistical analysis : Test of the Cohen's Kappa significance

fill with saved selection

	c1	c2	c3
r1	31	4	4
r2	8	39	9
r3	5	7	3

● Contingency table ○ Raw data

● Kappa unweighted
○ Kappa weighted (linear weights)
○ Kappa weighted (quadratic weights)

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

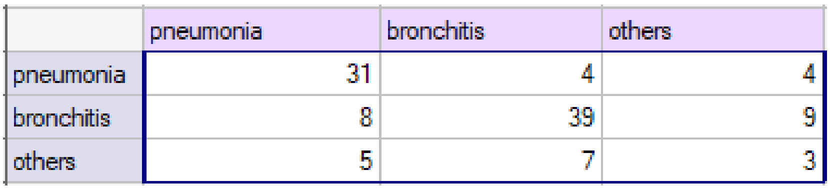
○ basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report
□ Add graph
□ Add percentages Rows

0.05 significance level

Example 21.3. (diagnosis.pqs file)
You want to analyse the compatibility of a diagnosis made by 2 doctors. To do this, you need to draw 110 patients (children) from a population. The doctors treat patients in a neighbouring doctors' offices. Each patient is examined first by the doctor A and then by the doctor B. Both diagnoses, made by the doctors, are shown in the table below.



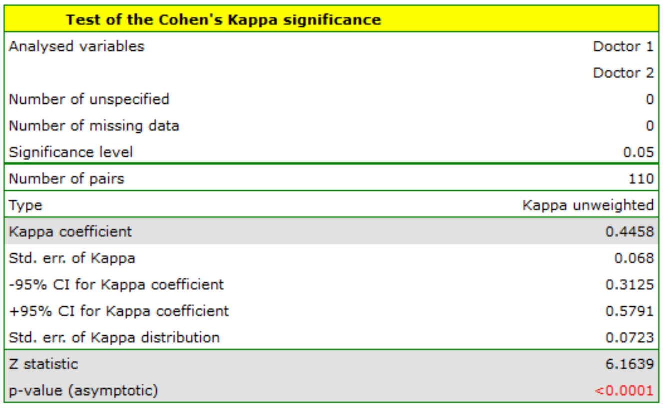
	pneumonia	bronchitis	others
pneumonia	31	4	4
bronchitis	8	39	9
others	5	7	3

Hypotheses:

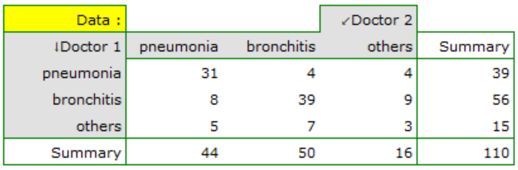
$$\begin{array}{cl} \mathcal{H}_0: & \kappa = 0, \\ \mathcal{H}_1: & \kappa \neq 0. \end{array}$$

We could analyse the agreement of the diagnoses using just the percentage of the compatible values. In this example, the compatible diagnoses were made for 73 patients (31+39+3=73) which is 66.36%

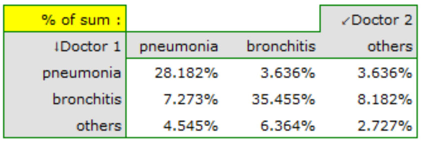
of the analysed group. The kappa coefficient introduces the correction of a chance agreement (it takes into account the agreement occurring by chance).



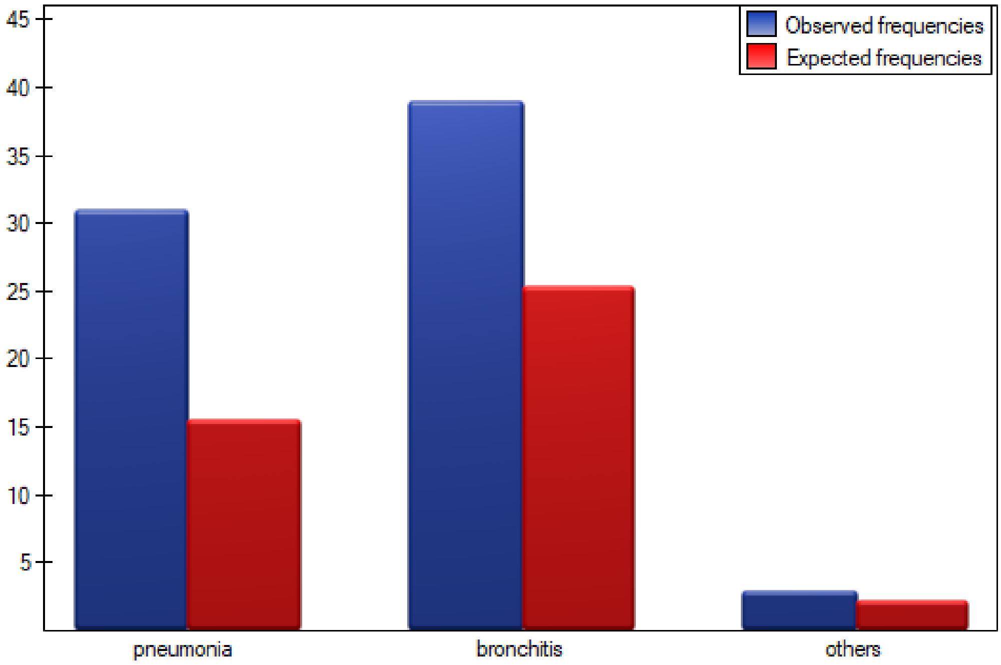
Test of the Cohen's Kappa significance	
Analysed variables	Doctor 1
	Doctor 2
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of pairs	110
Type	Kappa unweighted
Kappa coefficient	0.4458
Std. err. of Kappa	0.068
-95% CI for Kappa coefficient	0.3125
+95% CI for Kappa coefficient	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



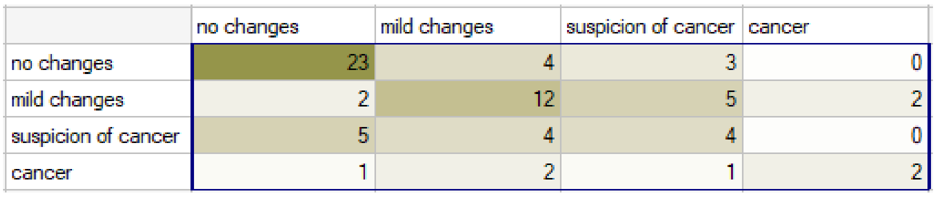
% of sum :			✓Doctor 2
↓Doctor 1	pneumonia	bronchitis	others
pneumonia	28.182%	3.636%	3.636%
bronchitis	7.273%	35.455%	8.182%
others	4.545%	6.364%	2.727%



The agreement with a chance adjustment $\hat{\kappa} = 44,58$% is smaller than the one which is not adjusted for the chances of an agreement.

The p value < 0.0001. Such result proves an agreement between these 2 doctors' opinions, on the significance level $\alpha = 0.05$,.

Example 21.4. (radiology.pqs file)
Radiological imaging assessed liver damage in the following categories: no changes (1), mild changes (2), suspicion of cancer (3), cancer (4). The evaluation was done by two independent radiologists based on a group of 70 patients. We want to check the concordance of the diagnosis.

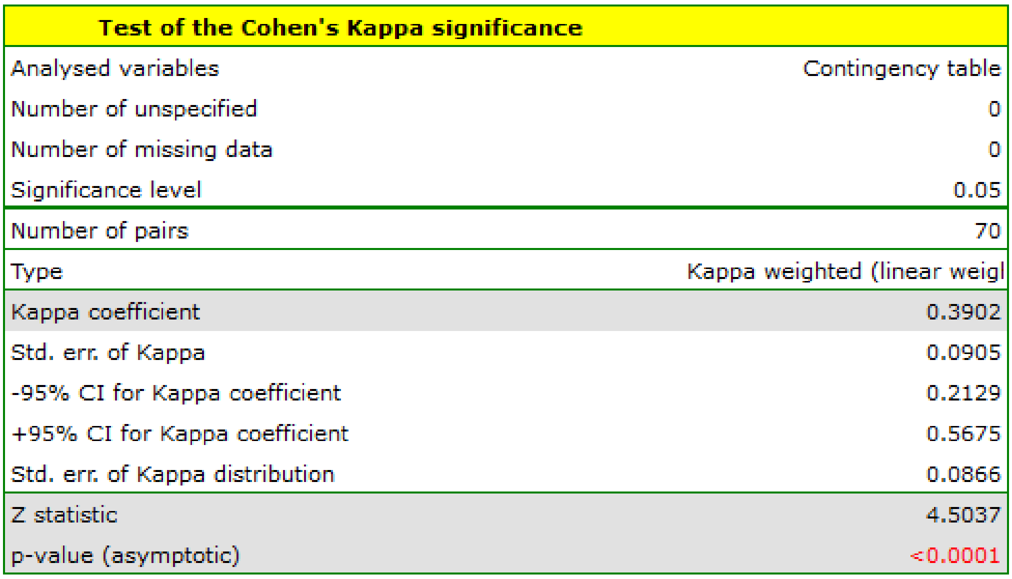


	no changes	mild changes	suspicion of cancer	cancer
no changes	23	4	3	0
mild changes	2	12	5	2
suspicion of cancer	5	4	4	0
cancer	1	2	1	2

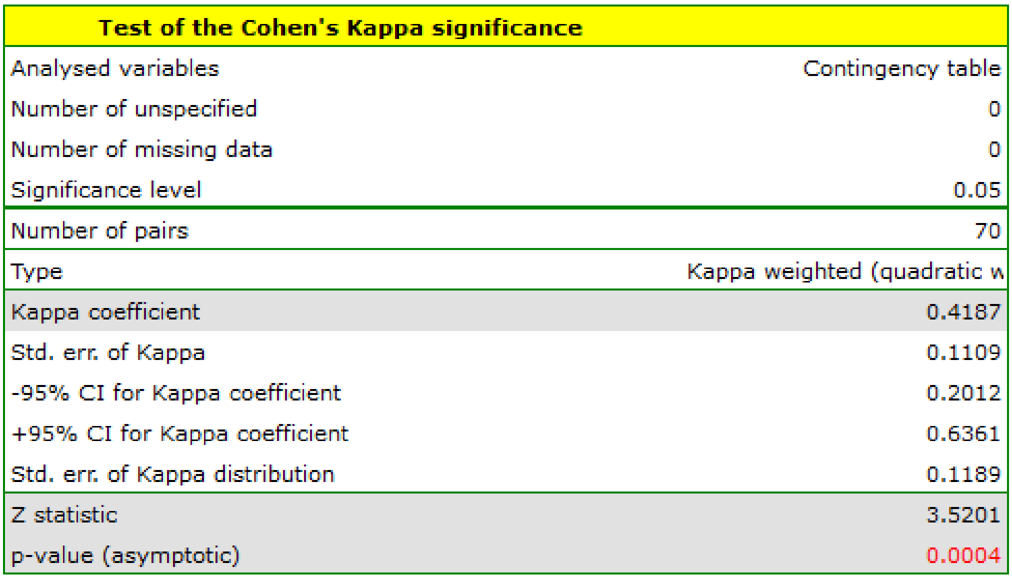
Hypotheses:

$$\begin{array}{ll} \mathcal{H}_0: & \kappa = 0, \\ \mathcal{H}_1: & \kappa \neq 0. \end{array}$$

Because the diagnosis is issued on an ordinal scale, an appropriate measure of concordance 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

Because the data are mainly concentrated on the main diagonal of the matrix and in close proximity to it, the coefficient weighted by the linear weights is lower ($\hat{\kappa} = 0.39$) than the coefficient determined for the quadratic weights ($\hat{\kappa} = 0.42$). In both situations, this is a statistically significant result (at the $\alpha = 0.05$ significance level), $p < 0.0001$.

If there was a large disagreement in the ratings concerning the two extreme cases and the pair: (no change and cancer) located in the upper right corner of the table occurred far more often, e.g., 15 times, then such a large disagreement would be more apparent when using quadratic weights (the Kappa coefficient would drop dramatically) than when using linear weights.

21.2.3 The Kappa Fleiss coefficient and a test to examine its significance

This coefficient determines the concordance of measurements conducted by a few judges (Fleiss, 1971[59]) and is an extension of Cohen's Kappa coefficient, which allows testing the concordance of only two judges. With that said, it should be noted that each of n randomly selected objects can be judged by a different random set of k judges. 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 the judges assign the test objects. Thus, each row in the table gives x_{ij}, which is the number of judges making the judgments specified in that column.

The Kappa coefficient ($\hat{\kappa}$) is then expressed by the formula:

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e},$$

where:
$P_o = \frac{1}{kn(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{c} x_{ij} - kn,$
$P_e = \sum_{i=1}^{c} q_j^2,$
$q_j = \frac{1}{km} \sum_{i=1}^{n} x_{ij}.$

A value of $\hat{\kappa} = 1$ indicates full agreement among judges, while $\hat{\kappa} = 0$ indicates the concordance that would arise if the judges' opinions were given at random. Negative values of Kappa, on the other hand, indicate concordance less than that at random.

For a coefficient of $\hat{\kappa}$ the standard error SE can be determined, which allows statistical significance to be tested and asymptotic confidence intervals to be determined.

Z test for significance of Fleiss' Kappa coefficient ($\hat{\kappa}$) (Fleiss, 2003[61]) is used to test the hypothesis that the ratings of several judges are consistent and is based on the coefficient $\hat{\kappa}$ calculated for the sample.

Basic assumptions:

- measurement on a nominal scale – possible category ordering is not taken into account.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \kappa = 0, \\ \mathcal{H}_1: & \kappa \neq 0. \end{array}$$

The test statistic has the form:

$$Z = \frac{\hat{\kappa}}{SE},$$

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

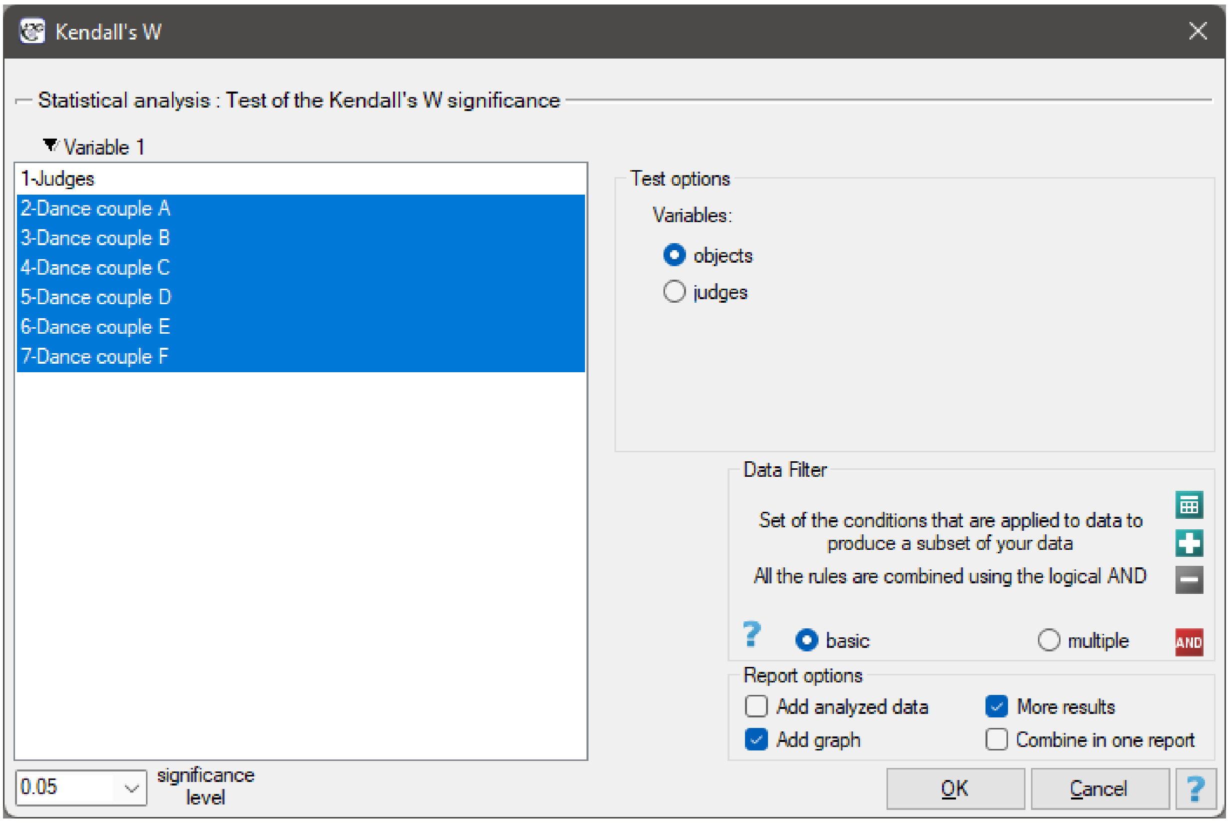
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note
The determination of Fleiss's Kappa coefficient is conceptually similar to the Mantel-Haenszel method. The determined Kappa is a general measure that summarizes the concordance of all judge ratings and can be determined as the Kappa formed from individual layers, which are specific judge ratings (Fleiss, 2003[61]). Therefore, as a summary of each layer, the judges' concordance (Kappa coefficient) can be

determined summarizing each possible rating separately.

The settings window with the test of the Fleiss's Kappa significance can be opened in Statistics menu →NonParametric tests→Fleiss Kappa.



Kendall's W

Statistical analysis : Test of the Kendall's W significance

Variable 1

1-Judges
2-Dance couple A
3-Dance couple B
4-Dance couple C
5-Dance couple D
6-Dance couple E
7-Dance couple F

Test options

Variables:

● objects
○ judges

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

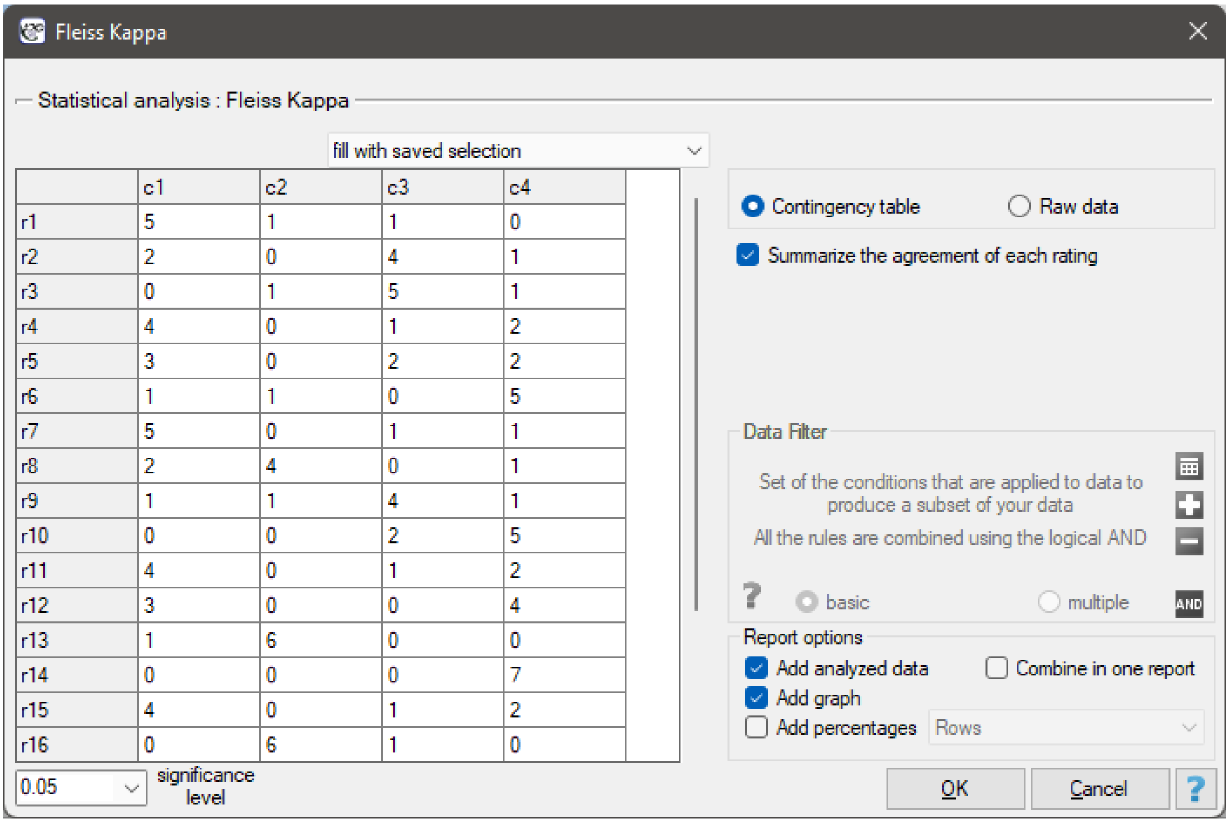
All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data ☑ More results
☑ Add graph □ Combine in one report

0.05 significance level



Fleiss Kappa

Statistical analysis : Fleiss Kappa

fill with saved selection

	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

● Contingency table ○ Raw data

☑ Summarize the agreement of each rating

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

○ basic ○ multiple

Report options

☑ Add analyzed data ☐ Combine in one report

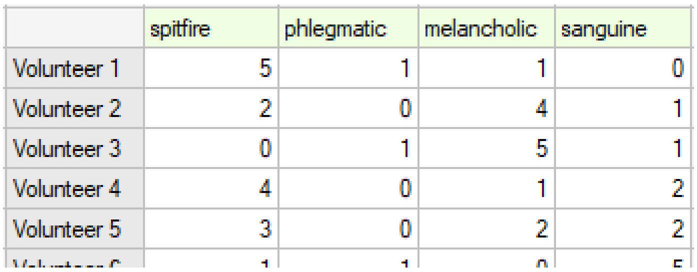
☑ Add graph

☐ Add percentages Rows

0.05 significance level

Example 21.5. (temperament.pqs file)

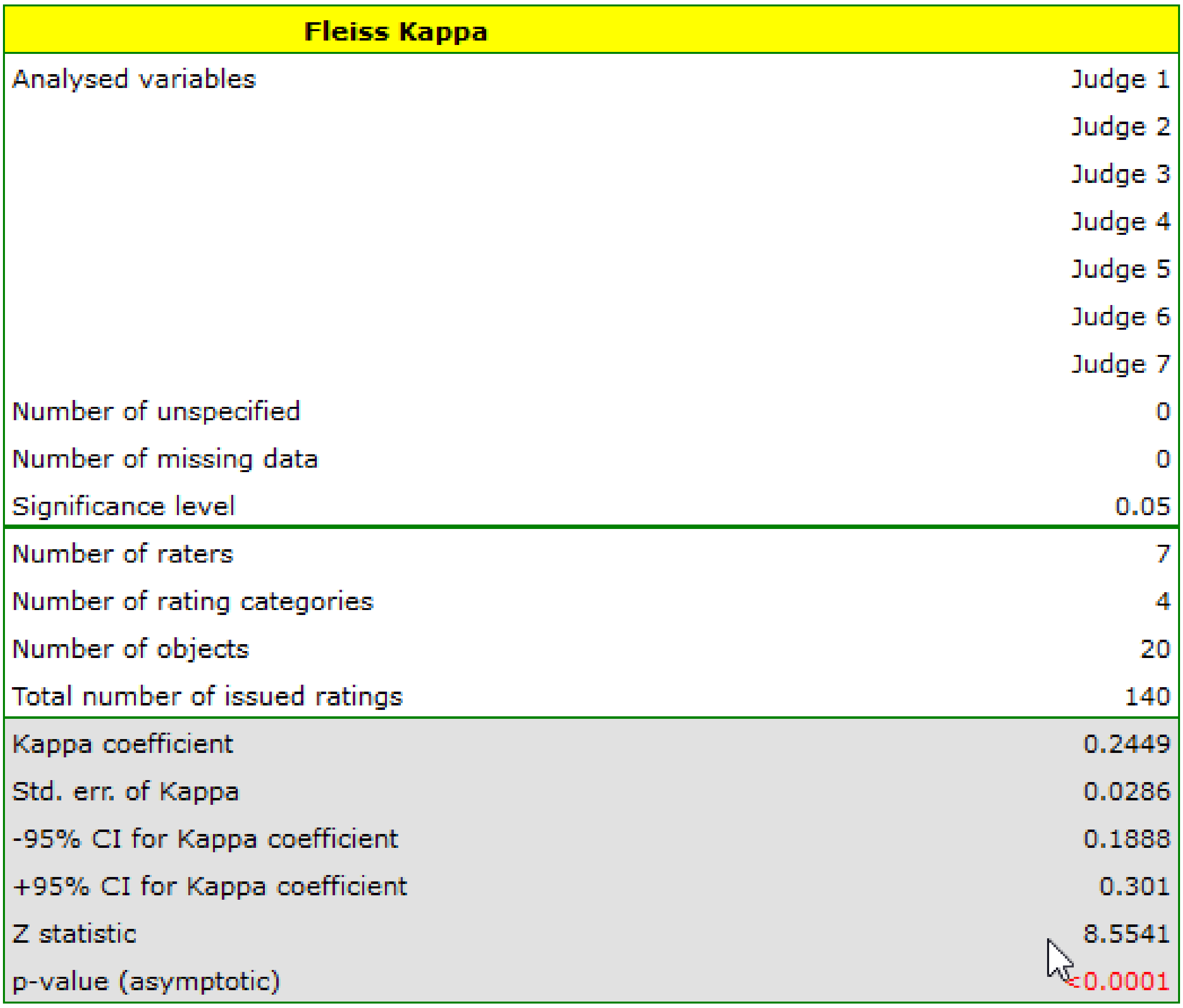
20 volunteers take part in a game to determine their personality type. Each volunteer has a rating given by 7 different observers (usually people from their close circle or family). Each observer has been introduced to the basic traits describing temperament in each personality type: choleric, phlegmatic, melancholic, sanguine. We examine observers' concordance in assigning personality types. An excerpt of the data is shown in the table below.



	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

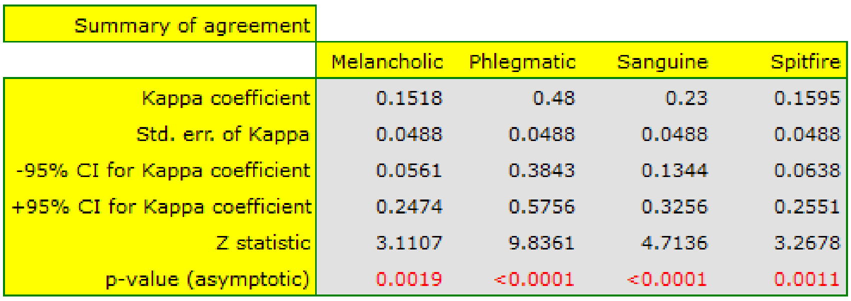
Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & \kappa = 0,\\ \mathcal{H}_1: & \kappa \neq 0.\end{array}$$

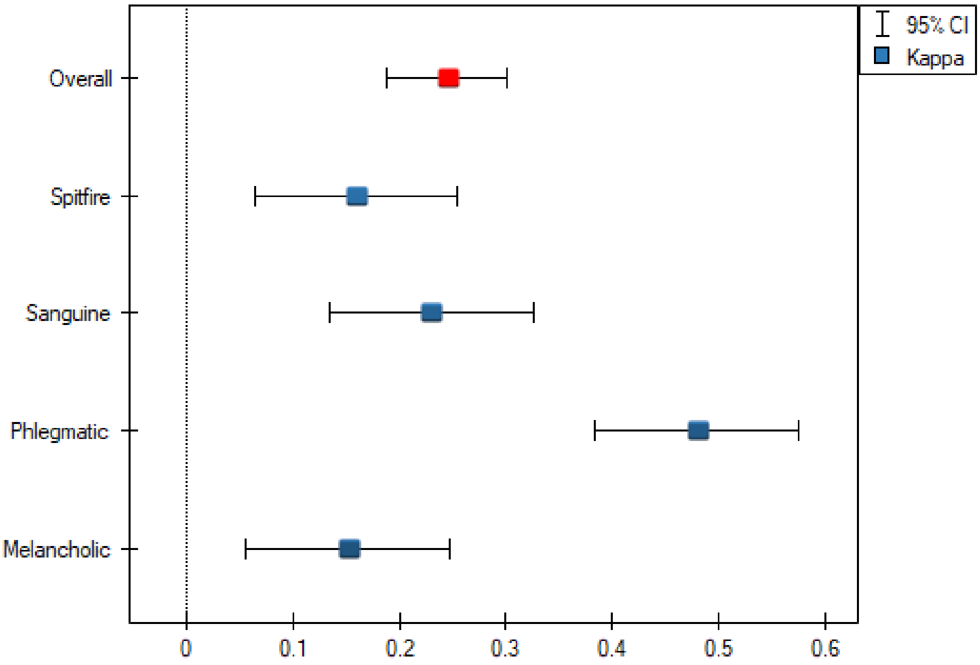


Fleiss Kappa	
Analysed variables	Judge 1 Judge 2 Judge 3 Judge 4 Judge 5 Judge 6 Judge 7
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Number of raters	7
Number of rating categories	4
Number of objects	20
Total number of issued ratings	140
Kappa coefficient	0.2449
Std. err. 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

We observe an unimpressive Kappa coefficient = 0.24, but statistically significant (p<0.0001), indicating non-random agreement between judges' ratings. The significant concordance applies to each grade, as evidenced by the concordance summary report for each stratum (for each grade) and the graph showing the individual Kappa coefficients and Kappa summarizing the total.

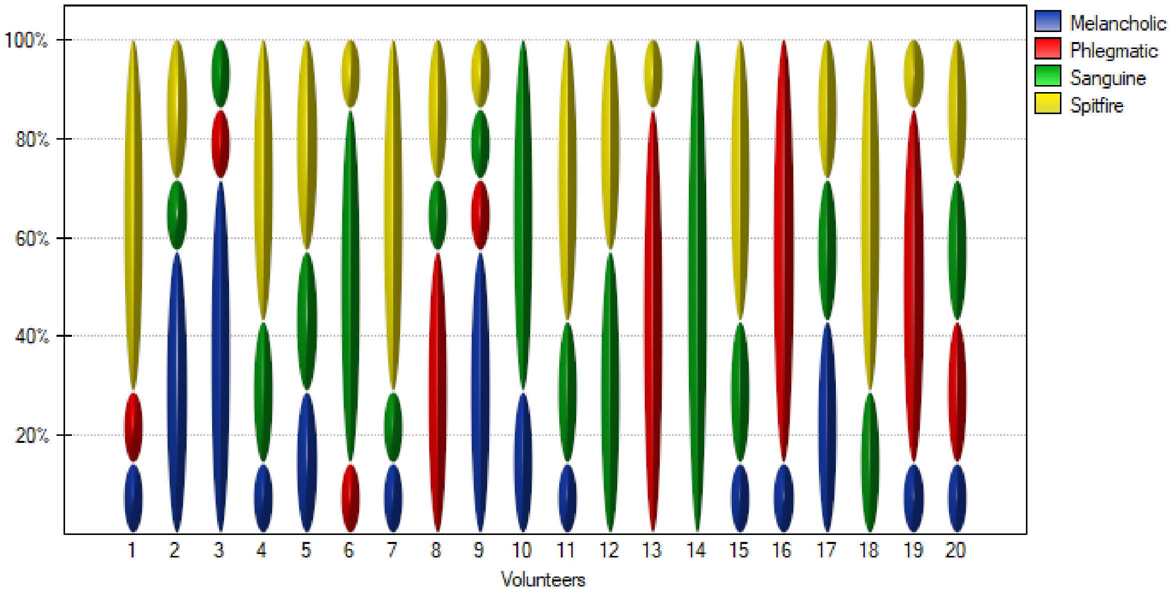


Summary of agreement	Melancholic	Phlegmatic	Sanguine	Spitfire
Kappa coefficient	0.1518	0.48	0.23	0.1595
Std. err. of Kappa	0.0488	0.0488	0.0488	0.0488
-95% CI for Kappa coefficient	0.0561	0.3843	0.1344	0.0638
+95% CI for Kappa coefficient	0.2474	0.5756	0.3256	0.2551
Z statistic	3.1107	9.8361	4.7136	3.2678
p-value (asymptotic)	0.0019	<0.0001	<0.0001	0.0011



It may be interesting to note that the highest concordance is for the evaluation of phlegmatics (Kappa=0.48).

With a small number of people observed, it is also useful to make a graph showing how observers rated each person.



In this case, only person no 14 received an unambiguous personality type rating – sanguine. Person no. 13 and 16 were assessed as phlegmatic by 6 observers (out of 7 possible). In the case of the remaining persons, there was slightly less agreement in the ratings. The most difficult to define personality type seems to be characteristic of the last person, who received the most diverse set of ratings.

22 DIAGNOSTIC TESTS

22.1 EVALUATION OF DIAGNOSTIC TEST

Suppose that using a diagnostic test we calculate the occurrence of a particular feature (most often disease) and know the gold-standard, so we know that the feature really occurs among the examined people. On the basis of these 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
FP – false positive
FN – false negative
TN – true negative

For such a table we can calculate the following measurements.

- **Sensitivity and specificity of diagnostic test**

  Every diagnostic test, in some cases, can obtain results different than actual results, for example a diagnostic test, basing on the obtained parameters, classifies a patient to the group of people suffering from a particular disease, or to the group of healthy people. In reality, the number of people approved for the above groups by the test may differ from the number of people genuinely ill and genuinely healthy.

  There are two evaluation measurements of the test accuracy. They are:

  Sensitivity – describes the ability to detect people genuinely ill (having a particular feature). If we examine a group of ill people, the sensitivity provides us with the information what percentage of them have a positive test result.

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

  Confidence interval is built on the basis of the Clopper-Pearson method for a single proportion.

  Specificity – describes the ability to detect people genuinely healthy (without a particular feature). If we examine a group of genuinely healthy people, the specificity provides us with the information about the percentage of people having a negative test result.

$$\text{specificity} = \frac{TN}{FP + TN}$$

  Confidence interval is built on the basis of the Clopper-Pearson method for a single proportion.

- **Positive predictive values, negative predictive values and prevalence rate**

Positive predictive value (PPV) – the probability, that a person having a positive test result suffered from a disease. If the examined person obtains a positive test result, the PPV informs them how they can be sure, that they suffer from a particular disease.

$$PPV = \frac{TP}{TP + FP}$$

Confidence interval is built on the basis of the Clopper-Pearson method for a single proportion.

Negative predictive value (hypertargetNPVNPV) – the probability that a person having a negative test result did not suffer from any disease. If the examined person obtains a negative test result, the NPV informs them how they can be sure that they do not suffer from a particular disease.

$$NPV = \frac{TN}{FN + TN}$$

Confidence interval is built on the basis of the Clopper-Pearson method for a single proportion.

Positive and negative predictive values depend on the prevalence rate.

Prevalence – probability of disease in the population for which the diagnostic test was conducted.

$$\text{prevalence} = \frac{TP + FN}{n}$$

Confidence interval is built on the basis of the Clopper-Pearson method for a single proportion.

- **Likelihood ratio of positive test and likelihood ratio of negative test**

Likelihood ratio of positive test (LR_+) – this measurement enables the comparison of some test results matching to the gold-standard. It does not depend on the prevalence of the disease. It is the ratio of two odds: the odds that a person from the group of ill people will obtain a positive test result, and the same effect will be observed among healthy people.

$$LR_+ = \frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{TP\,(TP + FN)}{FP\,(FP + TN)}$$

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 negative test (LR_-) – it is the ratio of two odds: the odds that a person from the group of ill people will obtain a negative test result, and the same effect will be observed among healthy people.

$$LR_- = \frac{1 - \text{sensitivity}}{\text{specificity}} = \frac{FN\,(TP + FN)}{TN\,(FP + TN)}$$

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**

Accuracy (Acc) – the probability of a correct diagnose using a diagnostic test. If the examined person obtains a positive or a negative test result, the Acc informs how they can be sure about the definitive diagnosis.

$$Acc = \frac{TP + TN}{n}$$

Confidence interval is built on the basis of the Clopper-Pearson method for a single proportion.

- **Diagnostic Odds Ratio**

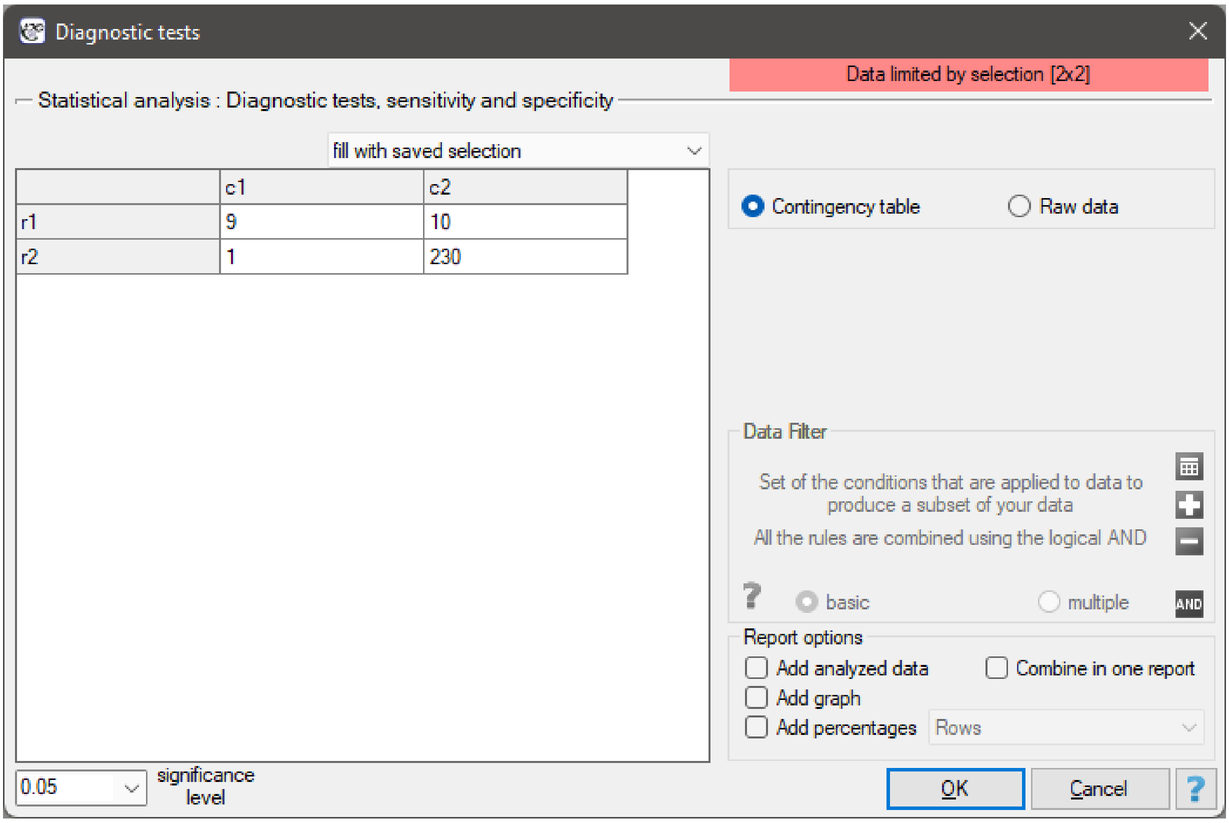
Diagnostic Odds Ratio — is the ratio of two chances: the chance of a positive test result from a diseased person to the chance of a positive test result from a healthy person.

$$DOR = \frac{TP/FN}{FP/TN}$$

Confidence interval for DOR is built on the basis of the standard error:

$$SE = \sqrt{\frac{1}{TP} + \frac{1}{FN} + \frac{1}{FP} + \frac{1}{TN}}.$$

The settings window with the diagnostic tests can be opened in Advanced stistics menu →Diagnostic tests → Diagnostic tests



Diagnostic tests

Data limited by selection [2x2]

Statistical analysis : Diagnostic tests, sensitivity and specificity

fill with saved selection

	c1	c2
r1	9	10
r2	1	230

● Contingency table ○ Raw data

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

○ basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

□ Add graph

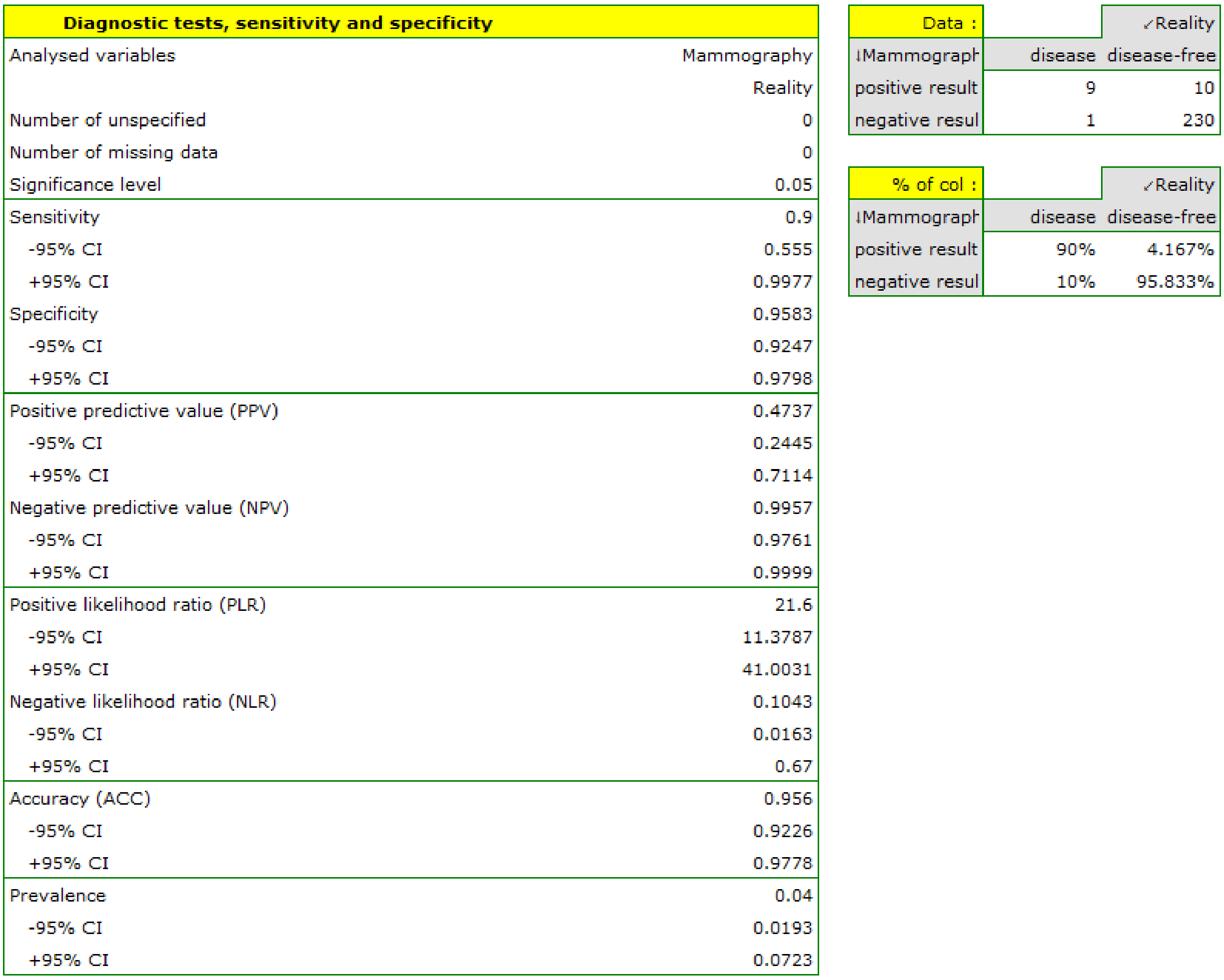
□ Add percentages Rows

0.05 significance level

Example 22.1. (mammography.pqs file)
Mammography is one of the most popular screening tests which enables the detection of breast cancer. The following study has been carried out on the group of 250 people, so-called ”asymptomatic” women at the age from 40 to 50. Mammography can detect an outbreak of cancer smaller than 5 mm and enables to note the change which is not a nodule yet but a change in the structure of tissues.

Observed frequencies		Reality (histopatology)		
		disease **(+)**	disease free **(–)**	**Total**
mammography	positive result **(+)**	9	10	19
	negative result **(–)**	1	230	231
	Total	10	240	250

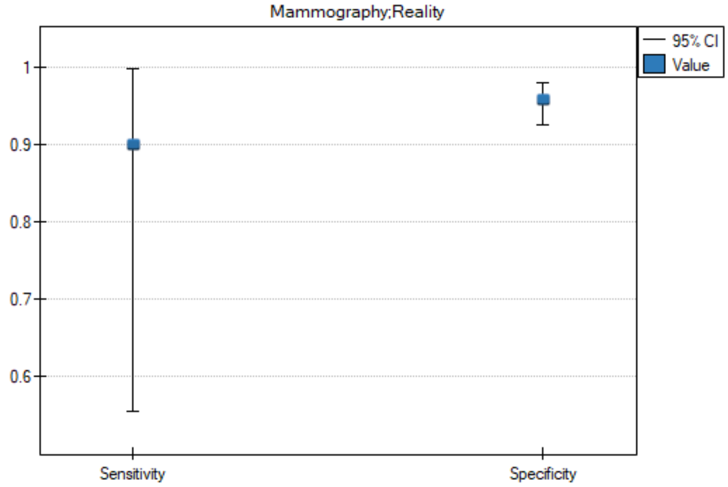
We will calculate the values enabling the assessment of the performed diagnostic test.



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%



- 90% of women suffering from breast cancer have been correctly defined, so they have obtained the positive result of mammography;
- 95.83% of healthy women (not suffering from breast cancer) have been correctly defined, so they have obtained the negative result of mammography;
- 4 out of 100 examined women suffer from breast cancer;
- A woman who have obtained a positive mammography result can be 47.37% sure that she suffers from breast cancer;
- A women who have obtained a negative test result can be 99.57% sure that she does not suffer from breast cancer;
- The probability that the positive mammography result will be obtained by a woman genuinely suffering from cancer is 21.60 times greater than the probability that the positive mammography result will be obtained by a healthy woman (not suffering from breast cancer);
- The probability that the negative mammography result will be obtained by a woman genuinely suffering from breast cancer is 10.43% of the probability that the negative mammography result will be obtained by a healthy woman (not suffering from breast cancer);
- A woman undergoing mammography (regardless of age) can be 96.50% sure of the definitive diagnosis;
- The chance of a positive test result in a woman who actually has breast cancer is 207 times greater than the chance of such a result in a healthy woman.

22.2 The ROC CURVE

The diagnostic test is used for differentiating objects with a given feature (marked as **(+)**, e.g. ill people) from objects without the feature (marked as **(–)**, e.g. healthy people). For the diagnostic test to be considered valuable, it should yield a relatively small number of wrong classifications. If the test is based on a dichotomous variable then the proper tool for the evaluation of the quality of the test is the analysis of a 2×2 contingency table of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. Most frequently, though, diagnostic tests are based on continuous variables or ordered categorical variables. In such a situation the proper means of evaluating the capability of the test for differentiating **(+)** and **(–)** are ROC (Receiver Operating Characteristic) curves.

It is frequently observed that the greater the value of the diagnostic variable, the greater the odds of occurrence of the studied phenomenon, or the other way round: the smaller the value of the diagnostic variable, the smaller the odds of occurrence of the studied phenomenon. Then, with the use of ROC curves, the choice of the optimum cut-off is made, i.e. the choice of a certain value of the diagnostic variable which best separates the studied statistical population into two groups: **(+)** in which the given phenomenon occurs and **(–)** in which the given phenomenon does not occur.

When, on the basis of the studies of the same objects, two or more ROC curves are constructed, one can compare the curves with regard to the quality of classification.

Let us assume that we have at our disposal a sample of n elements, in which each object has one of the k values of the diagnostic variable. Each of the received values of the diagnostic variable $x_1, x_2, ...x_k$ becomes the cut-off x_{cat}.

If the diagnostic variable is:

- **stimulant** (the growth of its value makes the odds of occurrence of the studied phenomenon greater), then values greater than or equal to the cut-off ($x_i >= x_{cat}$) are classified in group **(+)**;
- **destimulant** (the growth of its value makes the odds of occurrence of the studied phenomenon smaller), then values smaller than or equal to the cut-off ($x_i >= x_{cat}$) are classified in group **(+)**;

For each of the k cut-offs we define true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.

stimulant		Reality	
		(+)	**(–)**
diagnostic variable	$x_i >= x_{cat}$ **(+)**	TP	FP
	$x_i < x_{cat}$ **(–)**	FN	TN

destimulant		Reality	
		(+)	**(–)**
diagnostic variable	$x_i <= x_{cat}$ **(+)**	TP	FP
	$x_i > x_{cat}$ **(–)**	FN	TN

On the basis of those values each cut-off x_{cat} can be further described by means of sensitivity and specificity, positive predictive values(PPV), negative predictive values (NPV), positive result likelihood ratio (LR_+), negative result likelihood ratio (LR_-), and accuracy (Acc).

Note

The PQStat program computes the prevalence coefficient on the basis of the sample. The computed prevalence coefficient will reflect the occurrence of the studied phenomenon (illness) in the population in the case of screening of a large sample representing the population. If only people with suspected illness are directed to medical examinations, then the computed prevalence coefficient for them can be much higher than the prevalence coefficient for the population.

Because both the positive and negative predictive value depend on the prevalence coefficient, when the coefficient for the population is known a priori, we can use it to compute, for each cut-off x_{cat}, corrected predictive values according to Bayes's formulas:

$$PPV_{revised} = \frac{\text{Sensitivity} \cdot P_{apriori}}{\text{Sensitivity} \cdot P_{apriori} + (1 - \text{Specificity}) \cdot (1 - P_{apriori})}$$

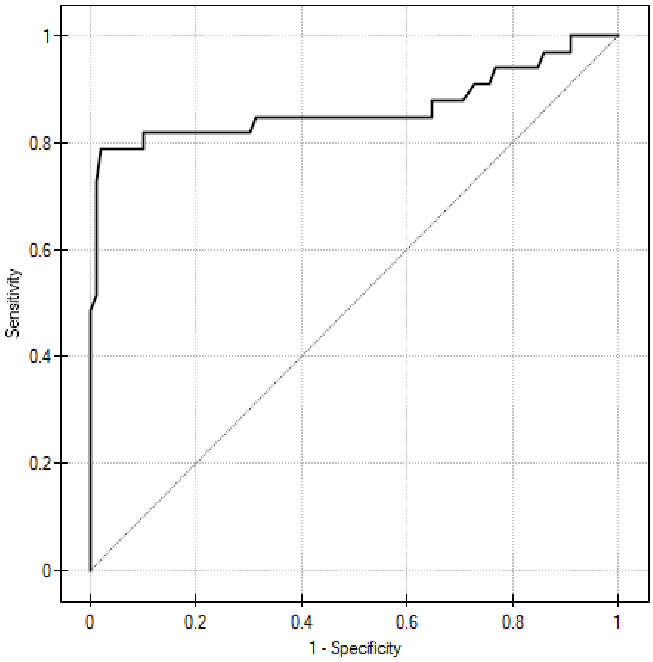
$$NPV_{revised} = \frac{\text{Specificity} \cdot (1 - P_{apriori})}{\text{Specificity} \cdot (1 - P_{apriori}) + (1 - \text{Sensitivity}) \cdot P_{apriori}}$$

where:

$P_{apriori}$ - the prevalence coefficient put in by the user, the so-called pre-test probability of disease

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}

The ROC curve is created on the basis of the calculated values of sensitivity and specificity. On the abscissa axis the x=1-specificity is placed, and on the ordinate axis y=sensitivity. The points obtained in that manner are linked. The constructed curve, especially the area under the curve, presents the classification quality of the analyzed diagnostic variable. When the ROC curve coincides with the diagonal $y = x$, then the decision made on the basis of the diagnostic variable is as good as the random distribution of studied objects into group **(+)** and group **(–)**.



AUC (*area under curve*) – the size of the area under the ROC curve falls within $< 0; 1 >$. The greater the field the more exact the classification of the objects in group **(+)** and group **(–)** on the basis of the analyzed diagnostic variable. Therefore, that diagnostic variable can be even more useful as a classifier. The area AUC, error SE_{AUC} and confidence interval for AUC are calculated on the basis of:

- nonparametric **DeLong** method (DeLong E.R. et al. 1988[50], Hanley J.A. i Hajian-Tilaki K.O. 1997[**?**]) - **recommended**,
- nonparametric **Hanley-McNeil** method (Hanley J.A. i McNeil M.D. 1982[73]),
- **Hanley-McNeil** method which presumes double negative exponential distribution (Hanley J.A. i McNeil M.D. 1982[73]) - computed only when groups **(+)** and **(–)** are equinumerous.

For the classification to be better than random distribution of objects into to classes, the area under the ROC curve should be significantly larger than the area under the line $y = x$, i.e. than 0.5.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & AUC = 0.5, \\ \mathcal{H}_1: & AUC \neq 0.5. \end{array}$$

The test statistics has the form presented below:

$$Z = \frac{AUC - 0.5}{SE_{0.5}},$$

where:
$SE_{0.5} = \sqrt{\frac{n_{(+)}+n_{(-)}+1}{12n_{(+)}n_{(-)}}}$,
$n_{(+)}$ – size of the sample **(+)** in which the given phenomenon occurs,
$n_{(-)}$ – size of the sample **(–)**, in which the given phenomenon does not occur.

The Z statistic asymptotically (for large sample sizes) has the normal distribution.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lll}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

In addition, when we assume that the diagnostic parameter forms a high field (AUC), we can select the optimal cut-off point.

22.2.1 Selection of optimum cut-off

The point which is looked for is a certain value of the diagnostic variable, which provides the optimum separation of the studied population into to groups: **(+)** in which the given phenomenon occurs and **(–)** in which the given phenomenon does not occur. The selection of the optimum cut-off is not easy because it requires specialist knowledge about the topic of the study. For example, different cut-offs will be required in, on the one hand, a test used for screening of a large group of people, e.g. for a mammography study, and, on the other hand, in invasive studies conducted for the purpose of confirming an earlier suspicion, e.g. in histopathology. With the help of an advanced mathematical apparatus we can find a cut-off which will be the most useful from the perspective of mathematics.

PQStat allows you to select the optimal cut-off point by:

- **Tangent method (cost index)** – calculated based on sensitivity, specificity, cost of erroneous decisions and prevalence.

  Errors which can be made when classifying the studied objects as belonging to group **(+)** and group **(–)** are false positive results (FP) and false negative results (FN). If committing those errors is equally costly (ethical, financial, and other costs), then in the field Cost FP and in the field Cost FN we enter the same positive value – usually 1. However, if we come to the conclusion that one type of error is encumbered with a greater cost than the other one, then we will assign appropriately greater weight to it.

  The optimum cut-off value is calculated on the basis of sensitivity, specificity, and with the help of value m – slope of the tangent line to the ROC curve. The slope angle m is defined in relation to two values: the costs of wrong decisions and the prevalence coefficient. Normally the costs of wrong decisions have the value 1 and the prevalence coefficient is estimated from the sample. Knowing, a priori, the prevalence coefficient ($P_{apriori}$) and the costs of wrong decisions, the user can influence the value m and, consequently, the search for an optimum cut-off. As a result, the optimum cut-off is determined to be such a value of the diagnostic variable for which the formula:

  $$\text{Sensitivity} - m \cdot (1 - \text{Specificity})$$

  reaches the minimum (Zweig M.H. 1993[178]).

  The optimum cut-off point of the diagnostic variable, selected as described above, will finally be marked on the ROC curve.

- **Costs graph** – presents the calculated values of an wrong diagnosis together with their costs. The values are computed according to the formula:

  $$cost = cost_{FP} \cdot FP + cost_{FN} \cdot FN$$

  The point marked on the graph is the minimum of the function presented above.

- **Youden's Index** – Conceptually, it is the maximum distance between the line that is the diagonal of a square of side 1 and the point of the ROC curve [175]. This index is calculated from the formula:

  $$d = \text{Sensitivity} + \text{Specificity} - 1$$

The optimal cut-off point of the diagnostic variable thus selected will eventually be marked on the ROC curve plot.

- **Distance from the top left corner** – Conceptually, it is the minimum distance between the upper left corner of a square of side 1 (i.e., the place where sensitivity and specificity can be highest) and the point of the ROC curve. This index is calculated from the formula:

$$d = \sqrt{(1 - \text{Sensitivity})^2 + (1 - \text{Sspecificity})^2}$$

  The optimal cut-off point of the diagnostic variable thus selected will eventually be marked on the ROC curve plot.

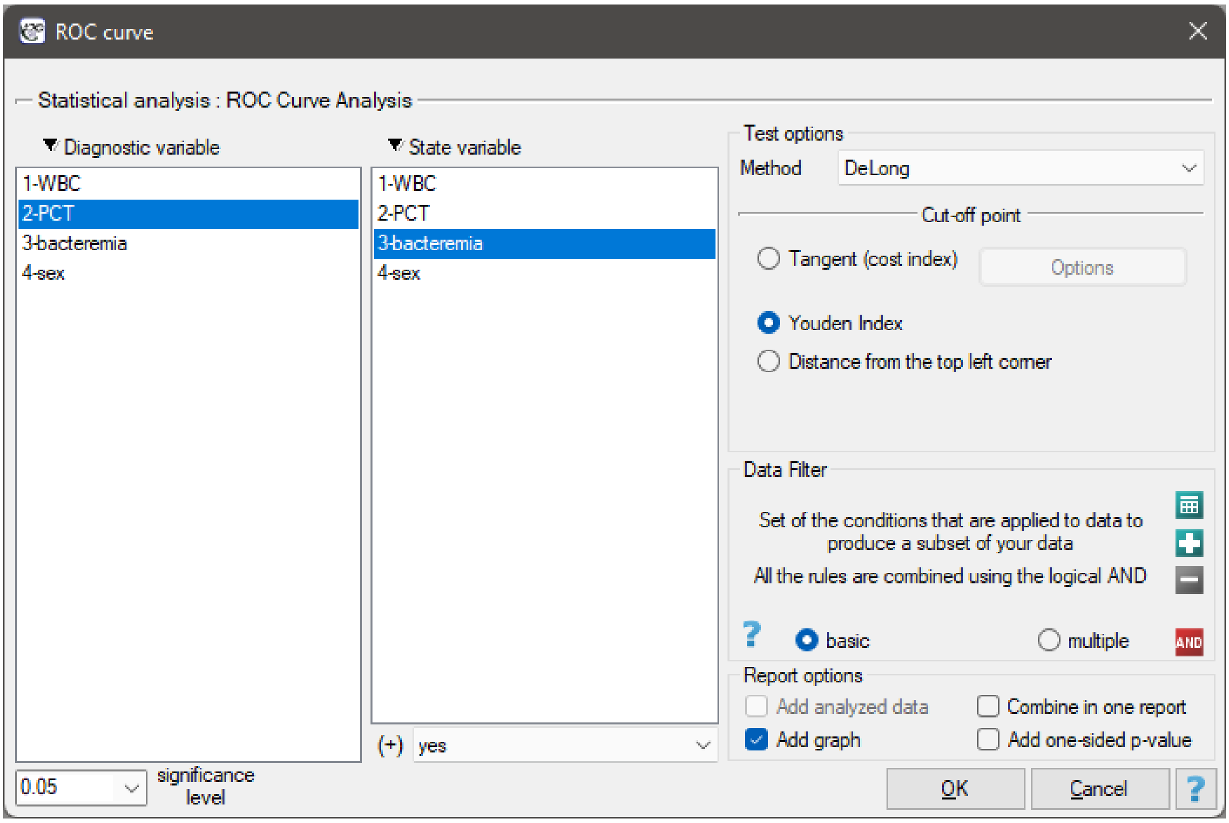
- **Costs graph** – presents the calculated values of an wrong diagnosis together with their costs. The values are computed according to the formula:

$$cost = cost_{FP} \cdot FP + cost_{FN} \cdot FN$$

  The point marked on the graph is the minimum of the function presented above.

- **Sensitivity and specificity intersection graph** – allows the localization of the point in which the value of sensitivity and specificity is simultaneously the greatest.

The window with settings for ROC analysis is accessed via the menu Advanced statistics → Diagnostic tests→ROC curve.



ROC curve

Statistical analysis : ROC Curve Analysis

Diagnostic variable: 1-WBC, **2-PCT**, 3-bacteremia, 4-sex

State variable: 1-WBC, 2-PCT, **3-bacteremia**, 4-sex

(+) yes

0.05 significance level

Test options

Method: DeLong

Cut-off point

○ Tangent (cost index) Options

● Youden Index

○ Distance from the top left corner

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

☑ Add graph □ Add one-sided p-value

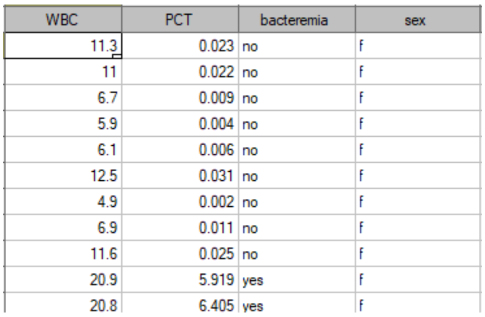
EXAMPLE 22.2. (file bacteriemia.pqs)

Persistent high fever in an infant or a small child without clearly diagnosed reasons is a premise for testing for bacteremia. The most useful and reliable parameters for screening and monitoring bacterial infections are the following indicators:

WBC – the number of white blood cells

PCT – procalcitonin.

It is assumed that in a healthy infant or a small child WBC should not exceed 15 thousand/μl and PCT should be lower than 0.5 ng/ml.
The sample values of those indicators for 136 children of up to 3 years old with persistent fever $> 39^0C$ is presented in the table fragment below:



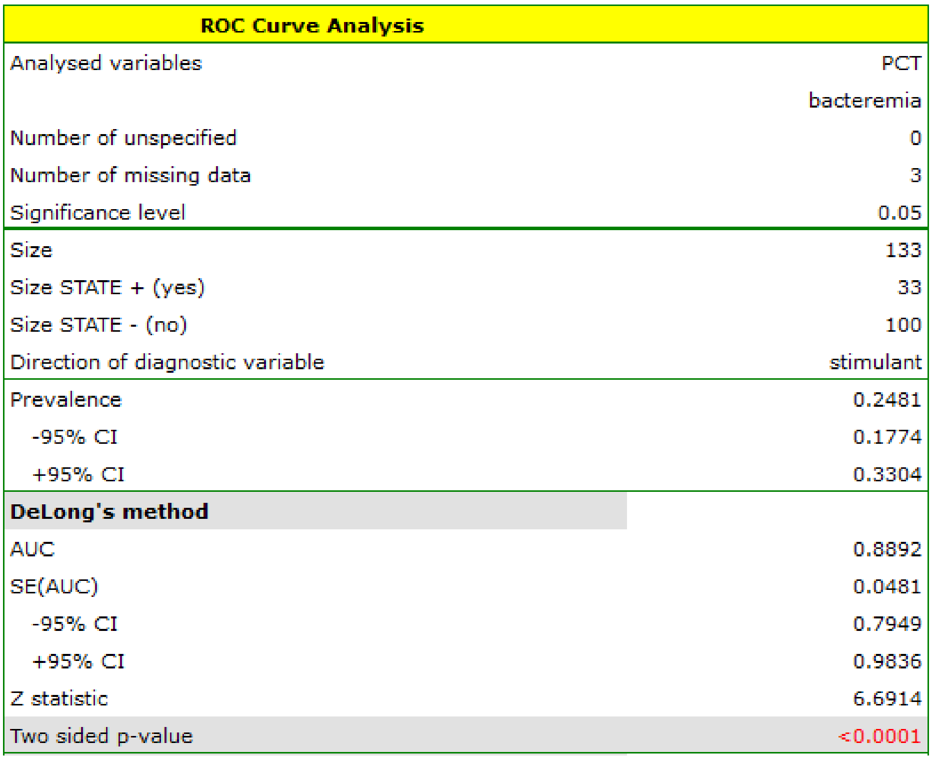
WBC	PCT	bacteremia	sex
11.3	0.023	no	f
11	0.022	no	f
6.7	0.009	no	f
5.9	0.004	no	f
6.1	0.006	no	f
12.5	0.031	no	f
4.9	0.002	no	f
6.9	0.011	no	f
11.6	0.025	no	f
20.9	5.919	yes	f
20.8	6.405	yes	f

One method of analyzing the PCT indicator is transforming it into a dichotomous variable by selecting a cut-off (e.g. x_{cat}=0.5 ng/ml) above which the study is considered to be ”positive”. The level of adequacy of such a division will be indicated by the value of sensitivity and specificity. We want to use a more complex approach, that is, calculate the sensitivity and specificity not only for one value but for each PCT value obtained in the sample - which means constructing a ROC curve. On the basis of the information obtained in that manner we want to check if the PTC indicator is indeed useful for diagnosing bacteremia. If so, then we want to check what is the optimal cut-off above which we can consider the study to be ”positive” – detecting bacteremia.

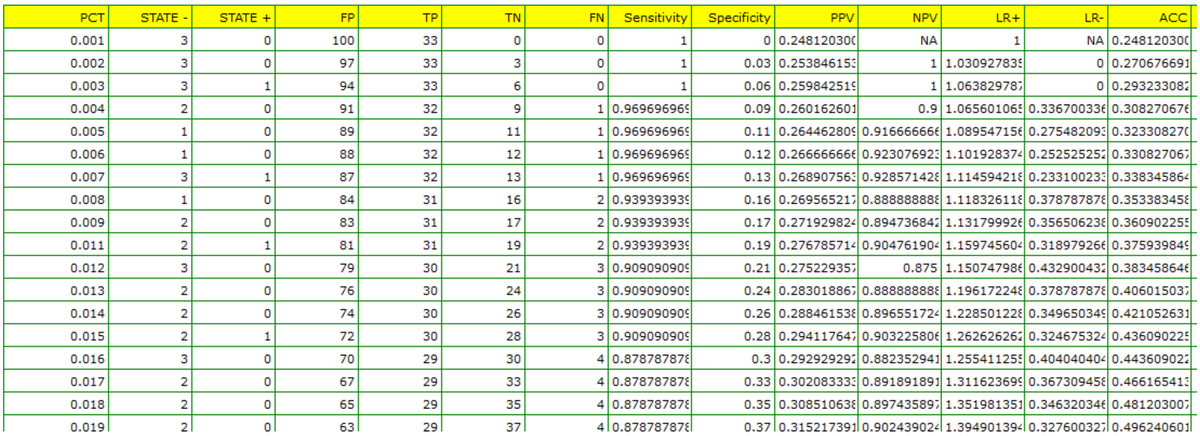
In order to check if PTC is really useful for diagnosing bacteremia we will calculate the size of the area under the ROC curve and verify the hypothesis that:

$$\begin{array}{ll}
\mathcal{H}_0: & \text{area under the constructed ROC curve} = 0.5,\\
\mathcal{H}_1: & \text{area under the constructed ROC curve} \neq 0.5.
\end{array}$$

As bacteremia is accompanied by an increased PCT level, in the test options window we will consider the indicator to be a stimulant. In the state variable we have to define which value in the bacteremia column determines its presence, then we select ”yes”. Apart from the result of the statistical test, in the report we can find an exact 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	0.248120300	NA	1	NA	0.248120300
0.002	3	0	97	33	3	0	1	0.03	0.253846153	1	1.030927835	0	0.270676691
0.003	3	1	94	33	6	0	1	0.06	0.259842519	1	1.063829787	0	0.293233082
0.004	2	0	91	32	9	1	0.969696969	0.09	0.260162601	0.9	1.065601065	0.336700336	0.308270676
0.005	1	0	89	32	11	1	0.969696969	0.11	0.264462809	0.916666666	1.089547156	0.275482093	0.323308270
0.006	1	0	88	32	12	1	0.969696969	0.12	0.266666666	0.923076923	1.101928374	0.252525252	0.330827067
0.007	3	1	87	32	13	1	0.969696969	0.13	0.268907563	0.928571428	1.114594218	0.233100233	0.338345864
0.008	1	0	84	31	16	2	0.939393939	0.16	0.269565217	0.888888888	1.118326118	0.378787878	0.353383458
0.009	2	0	83	31	17	2	0.939393939	0.17	0.271929824	0.894736842	1.131799926	0.356506238	0.360902255
0.011	2	1	81	31	19	2	0.939393939	0.19	0.276785714	0.904761904	1.159745604	0.318979266	0.375939849
0.012	3	0	79	30	21	3	0.909090909	0.21	0.275229357	0.875	1.150747986	0.432900432	0.383458646
0.013	2	0	76	30	24	3	0.909090909	0.24	0.283018867	0.888888888	1.196172248	0.378787878	0.406015037
0.014	2	0	74	30	26	3	0.909090909	0.26	0.288461538	0.896551724	1.228501228	0.349650349	0.421052631
0.015	2	1	72	30	28	3	0.909090909	0.28	0.294117647	0.903225806	1.262626262	0.324675324	0.436090225
0.016	3	0	70	29	30	4	0.878787878	0.3	0.292929292	0.882352941	1.255411255	0.404040404	0.443609022
0.017	2	0	67	29	33	4	0.878787878	0.33	0.302083333	0.891891891	1.311623699	0.367309458	0.466165413
0.018	2	0	65	29	35	4	0.878787878	0.35	0.308510638	0.897435897	1.351981351	0.346320346	0.481203007
0.019	2	0	63	29	37	4	0.878787878	0.37	0.315217391	0.902439024	1.394901394	0.327600327	0.496240601

The calculated size of the area under the ROC curve is $AUC = 0.889$. Therefore, on the basis of the adopted level $\alpha = 0.05$, based on the obtained value $p < 0.0001$ we assume that diagnosing bacteremia with the use of the PCT indicator is indeed more useful than a random distribution of patients into 2 groups: suffering from bacteremia and not suffering from it. Therefore, we return to the analysis (button Run the recent test) to define the optimal cut-off.

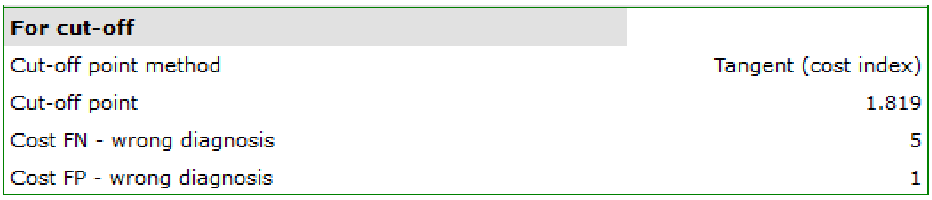
The algorithm of searching for the optimal cut-off takes into account the costs of wrong decisions and the prevalence coefficient.

(1) FN cost - wrong diagnosis is the cost of assuming that the patient does not suffer from bacteremia although in reality he or she is suffering from it (costs of a falsely negative decision)

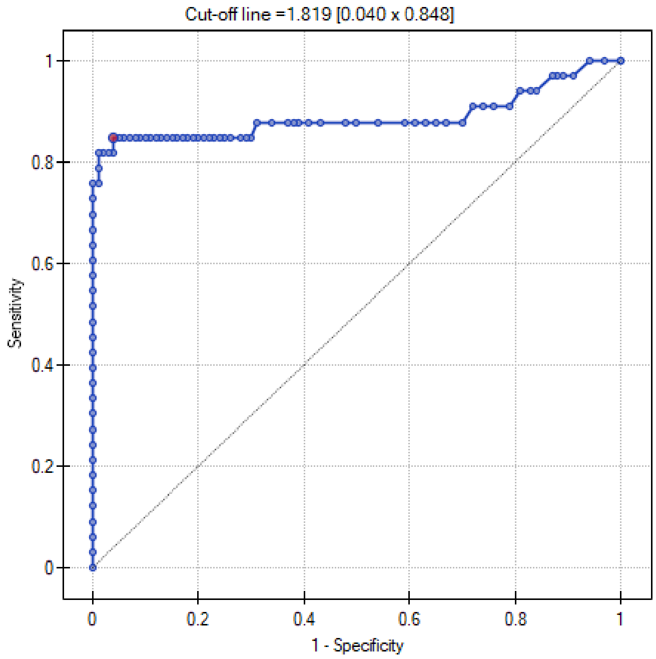
(2) FP cost - wrong diagnosis, is the cost of assuming that the patient suffers from bacteremia although in reality he or she is not suffering from it (costs of a falsely positive decision)

As the FN costs are much more serious than the FP costs, we enter a greater value in field one than in field two. We decided the value would be 5.

The PCT value is to be used in screening so we do not give the prevalence coefficient for the population (a priori prevalence coefficient) which is very low but we use the estimated coefficient from the sample. We do so in order not to move the cut-off of the PCT value too high and not to increase the number of falsely negative results.

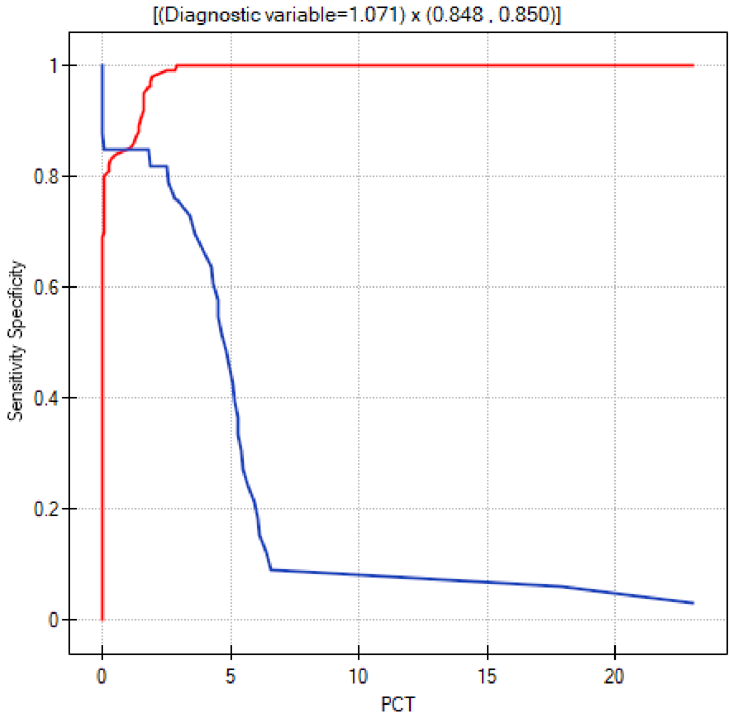


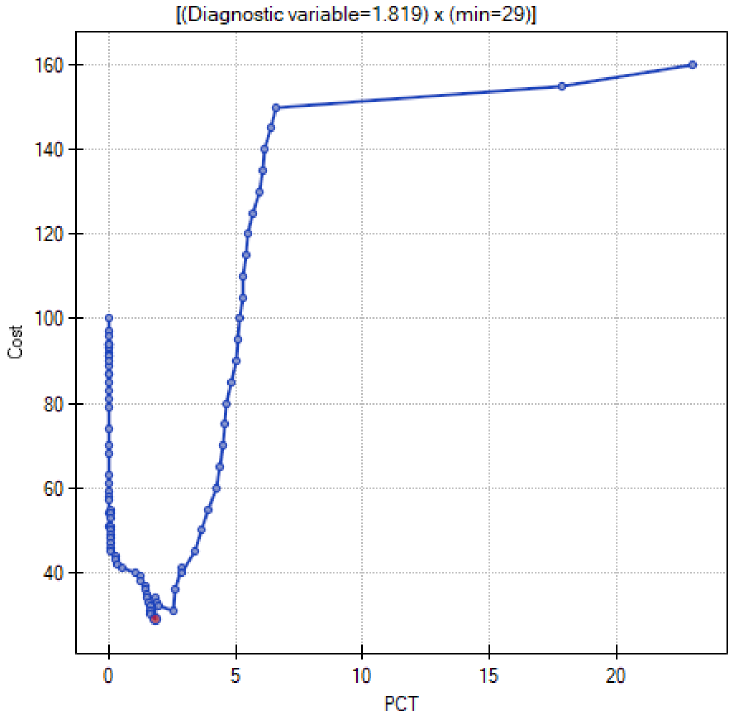
For cut-off	
Cut-off point method	Tangent (cost index)
Cut-off point	1.819
Cost FN - wrong diagnosis	5
Cost FP - wrong diagnosis	1



The optimal PCT cut-off determined in this way is 1.819. For this point sensitivity=0.85 and specificity=0.96.

Another method of selecting the cut-off is the anlysis of the costs graph and of the sensitivity intersection graph:





The analysis of the costs graph shows that the minimum of the costs of wrong decisions lies at PCT=1.819. The value of sensitivity and specificity is similar at PCT=1.071.

22.2.2 ROC curves comparison

Very often the aim of studies is the comparison of the size of the area under the ROC curve (AUC_1) with the area under another ROC curve (AUC_2). The ROC curve with a greater area usually allows a more precise classification of objects.

Methods for comparing the areas depend on the model of the study.

- Dependent model – the compared ROC curves are constructed on the basis of measurements made on the same objects.

  Hypotheses:

  $$\begin{array}{ll}\mathcal{H}_0: & AUC_1 = AUC_2,\\ \mathcal{H}_1: & AUC_1 \neq AUC_2.\end{array}$$

  The test statistics has the form presented below:

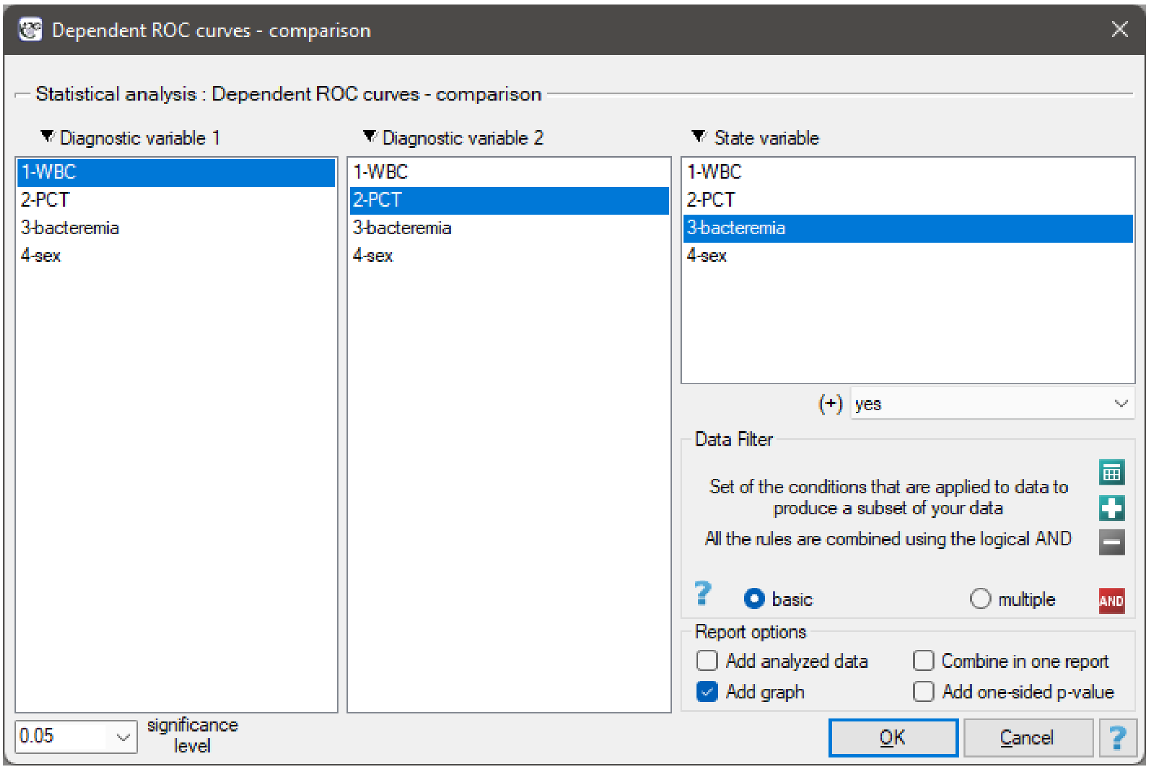
  $$Z=\frac{|AUC_1-AUC_2|}{SE_{AUC_1-AUC_2}},$$

  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[50], Hanley J.A., and Hajian-Tilaki K.O. 1997[**?**])

  Statistics Z has (for large sizes) asymptotic normal distribution.
  The p value, designated on the basis of the test statistic, is compared with the significance level α:

  $$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

  The window with settings for comparing dependent ROC curves is accessed via the menu Advanced statistics → Diagnostic tests → Dependent ROC Curves – comparison.



Dependent ROC curves - comparison

Statistical analysis : Dependent ROC curves - comparison

Diagnostic variable 1: 1-WBC (selected), 2-PCT, 3-bacteremia, 4-sex

Diagnostic variable 2: 1-WBC, 2-PCT (selected), 3-bacteremia, 4-sex

State variable: 1-WBC, 2-PCT, 3-bacteremia (selected), 4-sex

(+) yes

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

☑ Add graph □ Add one-sided p-value

0.05 significance level

- Independent model – the compared ROC curves are constructed on the basis of measurements made on different objects.

  Hypotheses:

  $$\mathcal{H}_0: \quad AUC_1 = AUC_2,$$
  $$\mathcal{H}_1: \quad AUC_1 \neq AUC_2.$$

  Test statistics (Hanley J.A. and McNeil M.D. 1983[74]) has the form:

  $$Z = \frac{|AUC_1 - AUC_2|}{\sqrt{SE^2_{AUC_1} - SE^2_{AUC_2}}},$$

  where:
  AUC_1, AUC_2 and standard errors of areas SE_{AUC_1}, SE_{AUC_2} are calculated on the basis of:

  ⋆ nonparametric method **DeLong** (DeLong E.R. et al. 1988[50], Hanley J.A., and Hajian-Tilaki K.O. 1997[?]) - **recommended**,

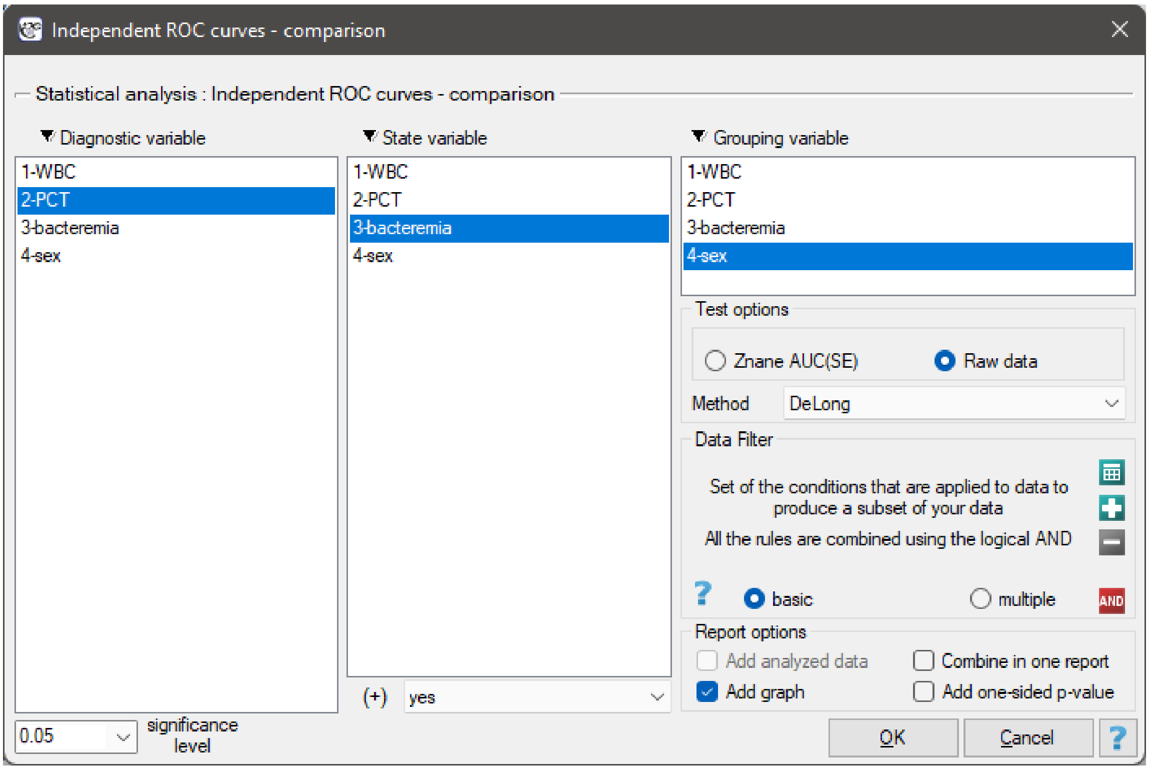
  ⋆ nonparametric **Hanley-McNeil** method (Hanley J.A. and McNeil M.D. 1982[73]),

  ⋆ method which presumes double negative exponential distribution (Hanley J.A. and McNeil M.D. 1982[73]) - computed only when groups **(+)** and **(–)** are equinumerous.

  Statistics Z has (for large sizes) asymptotic normal distribution.
  On the basis of test statistics p value is estimated and then compared with the significance level α:

  $$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no basis for rejecting } \mathcal{H}_0. \end{array}$$

The window with settings for comparing independent ROC curves is accessed via the menu Advanced statistics→Diagnostic tests→Independent ROC Curves – comparison



Independent ROC curves - comparison

Statistical analysis : Independent ROC curves - comparison

Diagnostic variable: 1-WBC, **2-PCT** (selected), 3-bacteremia, 4-sex

State variable: 1-WBC, 2-PCT, **3-bacteremia** (selected), 4-sex

Grouping variable: 1-WBC, 2-PCT, 3-bacteremia, **4-sex** (selected)

Test options: ○ Znane AUC(SE) ● Raw data

Method: DeLong

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options: □ Add analyzed data ☑ Add graph □ Combine in one report □ Add one-sided p-value

(+) yes

0.05 significance level

EXAMPLE (22.2) c.d. *(bacteriemia.pqs file)*

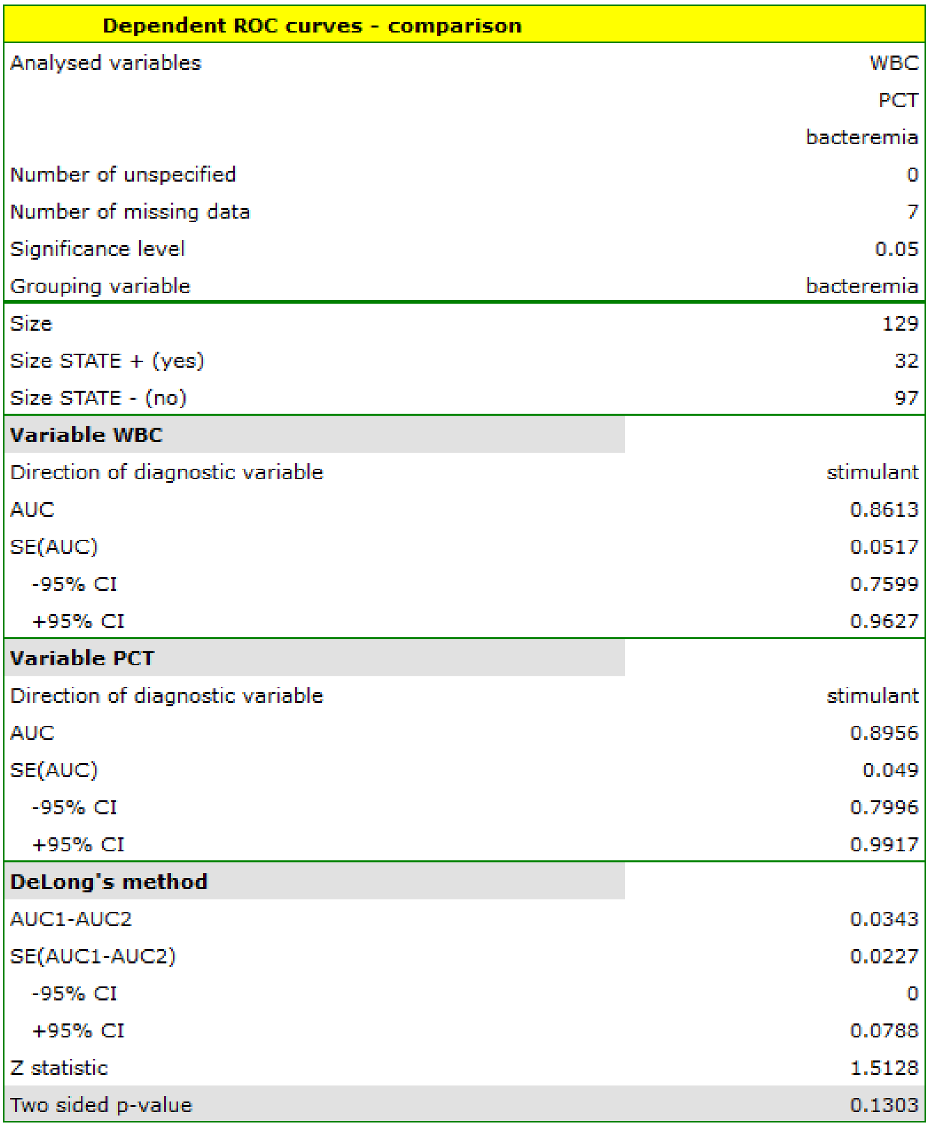
We will make 2 comparisons:

1) We will construct 2 ROC curves to compare the diagnostic value of parameters WBC and PCT;

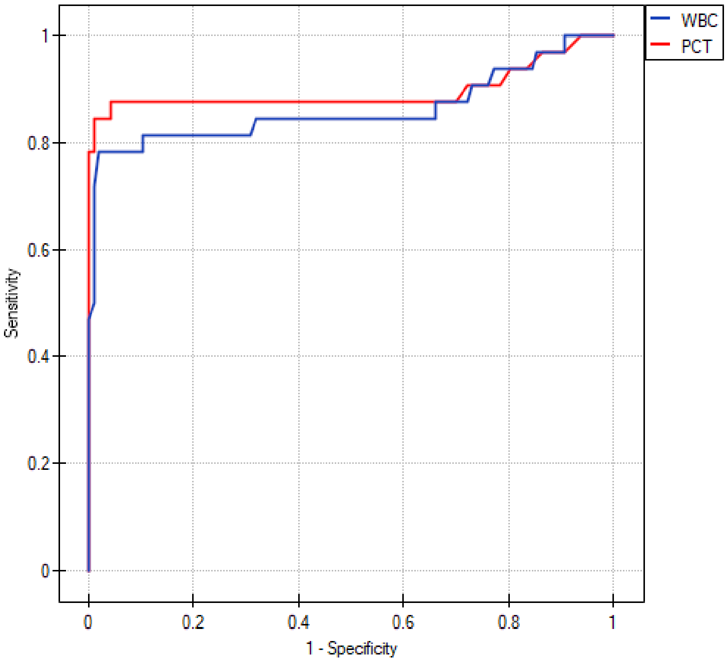
2) We will construct 2 ROC curves to compare the diagnostic value of PCT parameter for boys and girls.

ad1) Both parameters, WBC and PCT, are stimulants (in bacteremia their values are high). In the course of the comparison of the diagnostic value of those parameters we verify the following hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & \text{the area under ROC curve for WBC} = \text{the area under the ROC curve for PCT,}\\ \mathcal{H}_1: & \text{the area under ROC curve for WBC} \neq \text{the area under the ROC curve for PCT.}\end{array}$$



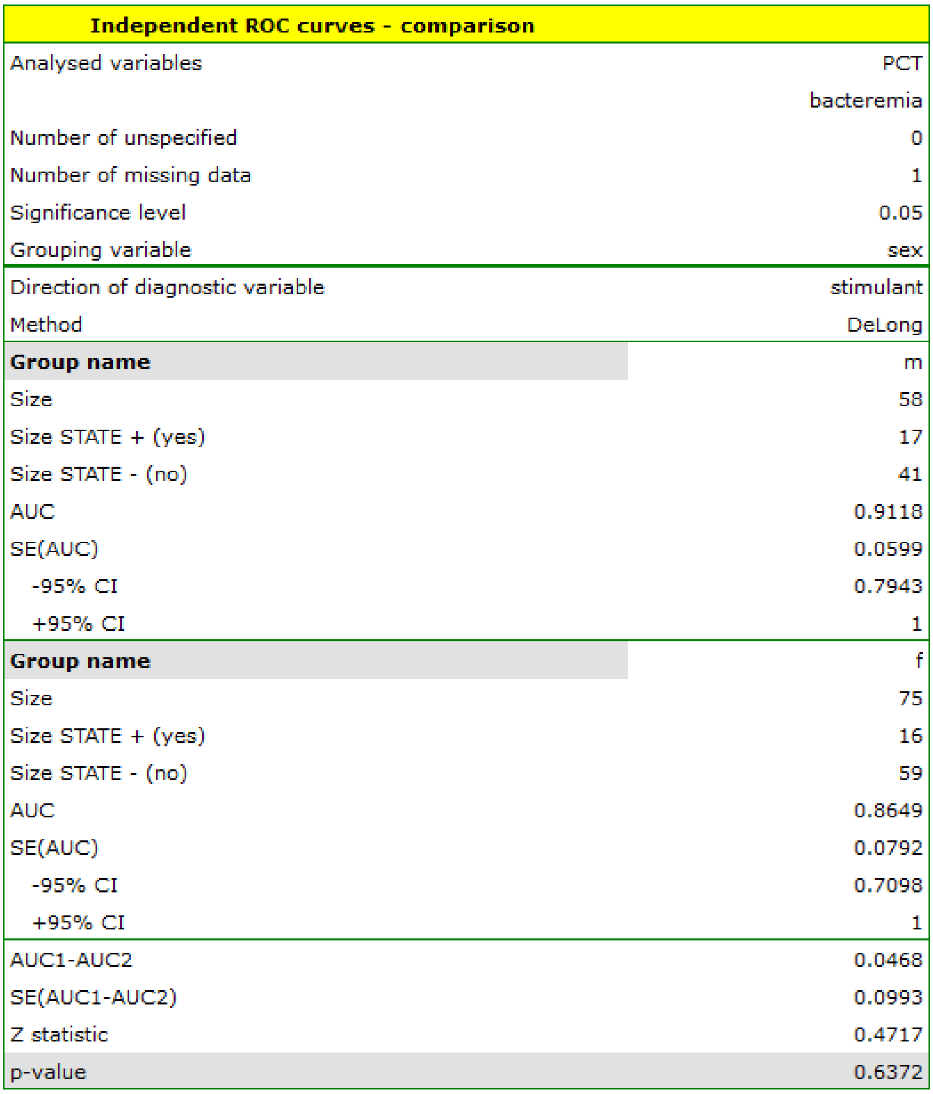
Dependent ROC curves - comparison	
Analysed variables	WBC
	PCT
	bacteremia
Number of unspecified	0
Number of missing data	7
Significance level	0.05
Grouping variable	bacteremia
Size	129
Size STATE + (yes)	32
Size STATE - (no)	97
Variable WBC	
Direction of diagnostic variable	stimulant
AUC	0.8613
SE(AUC)	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



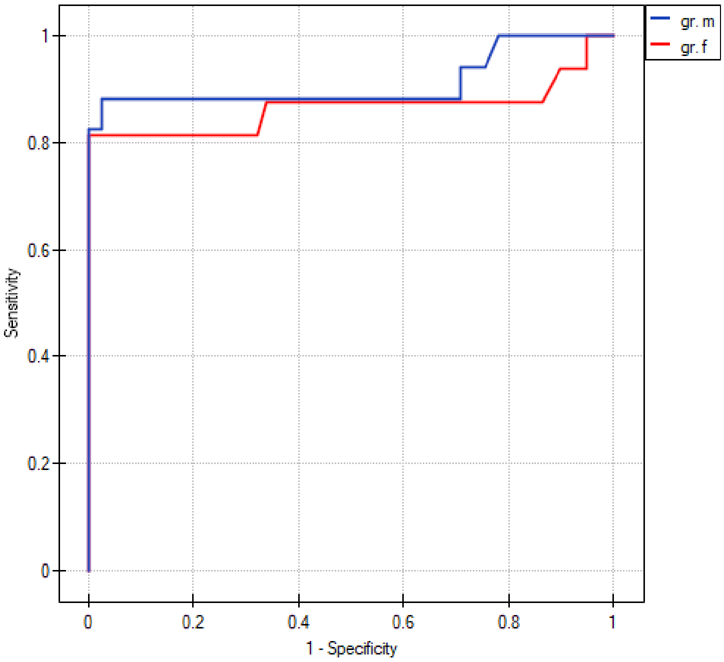
The calculated ares are $AUC_{WBC} = 086$, $AUC_{PCT} = 0.90$. On the basis of the adopted level $\alpha = 0.05$, based on the obtained value p=0.13032 we conclude that we cannot determine which of the parameters: WBC or PCT is better for diagnosing bacteremia.

ad2) PCT parameter is a stimulant (its value is high in bacteremia). In the course of the comparison of its diagnostic value for girls and boys we verify the following hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & \text{the area under ROC curve for } PCT_f = \text{the area under ROC curve for } PCT_m,\\ \mathcal{H}_1: & \text{the area under ROC curve for } PCT_f \neq \text{the area under ROC curve for } PCT_m.\end{array}$$



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



The calculated areas are $AUC_f = 0.86$, $AUC_m = 0.91$. Therefore, on the basis of the adopted level $\alpha = 0.05$, based on the obtained value p=0.6372 we conclude that we cannot select the sex for which PCT parameter is better for diagnosing bacteremia.

23 MULTIDIMENSIONAL MODELS

24 MATCHING GROUPS

Why is group matching done?
There are many answers to this question. Let us use an example of a medical situation.

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 (not dependent on confounding factors) effect of the treatment method on the outcome of the experiment. In this case, other than random group matching is not necessary.

The possibility of error arises when the difference in treatment outcome between 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 occurs when randomization is not possible for some reason, such as it is an observational study or for ethical reasons we cannot assign treatment arbitrarily. Artificial group matching may then be applicable. For example, if the people we assign to the treatment group are healthier people and the people who are in the control group are people with more severe disease, then it is not the treatment itself but the condition of the patient before treatment that may affect the outcome of the experiment. When we see such an imbalance of groups, it is good if we can decide to randomize, in this way the probem is solved, because drawing people into groups makes them similar. However, we can imagine another situation. This time the group we are interested in will not be treatment subjects 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, in order to test whether smoking does indeed increase the risk of lung cancer, it would be unethical to perform a fully randomized trial because it would mean that people randomly selected to the risk group would be forced to smoke. The solution to this situation is to establish an exposure group, i.e. to select a number of people who already smoke and then to select a control group of non-smokers. The control group should be selected because by leaving the selection to chance we may get a non-smoking group that is younger than the smokers only due to the fact that smoking is becoming less fashionable in our country, so automatically there are many young people among the non-smokers.The control should be drawn from non-smokers, but so that it is **as similar as possible to the treatment group**.In this way we are getting closer to examining the pure (independent of selected confounding factors such as age) effect of smoking/non-smoking on the outcome of the experiment, which in this case is the occurrence of lung cancer. **Such a selection can be made by the matching proposed in the program**.

One of the main advantages of investigator-controlled matching is that the control group becomes more similar to the treatment group, but this is also the biggest disadvantage of this method. It is an advantage because our study is looking more and more like a randomized study. In a randomized trial, the treatment and control groups are similar on almost all characteristics, including those we don't study - the random allocation provides us with this similarity. With investigator-controlled matching, the treatment and control groups become similar on only selected characteristics.

Ways of assessing similarity:
The first two methods mentioned are based on matching groups through Propensity Score Matching, PSM. This type of matching was proposed by Rosenbaum and Rubin [136]. In practice, it is a technique for matching a control group (untreated or minimally/standardly treated subjects) to a treatment group on the basis of a probability describing the subjects' propensity to assign treatment depending on the observed associated variables. The probability score describing propensity, called the *Propensity Score* is a balancing score, so that as a result of matching the control group to the treatment group, the distribution of 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 in terms of multiple selected characteristics.

Methods for determining similarity:

- Known probability – the Propensity Score, which is a value between 0 and 1 for each person tested, indicates the probability of being in the treatment group. This probability can be determined beforehand by various methods. For example, in a logistic regression model, through neural networks, or many other methods. If a person in the group from which we draw controls obtains a Propensity Score similar to that obtained by a person in the treatment group, then that person can enter the analysis because the two are similar in terms of the characteristics that were considered in determining the Propensity Score.

- Calculated from the logistic regression model – because logistic regression is the most commonly used matching method, PQStat provides the ability to determine a Propensity Score value based on this method automatically in the analysis window. The matching proceeds further using the Propensity Score thus obtained.

- Similarity/distance matrix – This option does not determine the value of Propensity Score, but builds a matrix indicating the distance of each person in the treatment group to the person in the control group. The user can set the boundary conditions, e.g. he can indicate that the person matched to a person from the treatment group cannot differ from him 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 method of matching the control group to the treated group is very flexible. In addition to the arbitrary choice of how the distances/dissimilarity are determined, in many metrics it allows for the indication of weights that determine how important each variable is to the researcher, i.e., the similarity of some variables may be more important to the researcher while the similarity of others is less important. However, great caution is advised when choosing a distance/ dissimilarity matrix. Many features and many sops to determine distances require prior standardization or normalization of the data, moreover, choosing the inverse of distance or similarity (rather than dissimilarity) may result in finding the most distant and dissimilar objects, whereas we normally use these methods to find similar objects. If the researcher does not have specific reasons for changing the metric, the standard recommendation is to use statistical distance, i.e. the Mahalanobia metric – It is the most universal, does not require prior standardization of data and is resistant to correlation of variables. More detailed description of distances and dissimilarity/similarity measures available in the program as well as the method of inetrpratation of the obtained results can be found in the Similarity matrix section
.

In practice, there are many methods to indicate how close the objects being compared are, in this case treated and untreated individuals. Two are proposed in the program:

- **Nearest neighbor method** – is a standard way of selecting objects not only with a similar Propensity Score, but also those whose distance/dissimilarity in the matrix is the smallest.

- **The nearest neighbor method, closer than...** – works in the same way as the nearest neighbor method, with the difference that only objects that are close enough can be matched. The limit of this closeness is determined by giving a value describing the threshold, behind which there are already objects so dissimilar to the tested objects, that we do not want to give them a chance to join the newly built control group. In the case when analysis is based on Propensity Score or matrix defined by dissimilarity, the most dissimilar objects are those distant by 1, and the most similar are those distant by 0. Choosing this method we should give a value closer to 0, when we select more restrictively, or closer to 1, when the threshold will be placed further. When we determine distances instead of dissimilarities in the matrix, then the minimum size is also 0, but the maximum size is not predetermined.

We can match without returning already drawn objects or with returning these objects again to the group from which we draw.

- **Matching without returning** – when using no-return matching, once an untreated person has been selected for matching with 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 individual is included in at most one matching set.
- **Matching with returning** – return matching allows a given untreated individual to be included more than once in a single matched set. When return matching is used, further analyses, and in particular variance estimation, must take into account the fact that the same untreated individual may be in multiple matched sets.

In the case when it is impossible to match the untreated person to the treated one, because in the group from which we choose there are more persons matching the treated one equally well, then one of these persons chosen in a random way is combined. For a renewed analysis, a fixed *seed* is set by default so that the results of a repeated draw will be the same, but when the analysis is performed again the seed is changed and the result of the draw may be different.
If it is not possible to match an untreated person to a treated one, because there are no more persons 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, then this person remains without a pair.

Most often a **1:1** match is made,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, then we can choose to match **1:k**, where k indicates the number of individuals that should be matched to each treated individual.

Matching evaluation
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, we should not assume that by applying the matching we will always obtain satisfactory results. In many situations, the group from which we draw does not have a sufficient number of such objects that are sufficiently similar to the treatment group. Therefore, the matching performed should always be evaluated. There are many methods of evaluating the matching of groups. The program uses methods based on standardized group difference and Propensity Score percentile agreement of the treatment group and the control group, more extensively described in the work of P.C Austin, among others [13][14]. This approach allows comparison of the relative balance of variables measured in different units, and the result is not affected by sample size. The estimation of concordance using statistical tests was abandoned because the matched control group is usually much smaller than the original control group, so that the obtained p-values of tests comparing the test group to the smaller control group are more likely to be left with the null hypothesis, and therefore do not show significant differences due to the reduced size.

For comparison of 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:
$\bar{x}_{treated}$, $\bar{x}_{control}$ – is the mean value of the variable in the treatment group and the mean

value of the variable in the control group,
$sd^2_{treated}$, $sd^2_{control}$ – is the variance in the treatment group and the variance in the control group.

To compare binary variables (of two categories, usually 0 and 1) we determine the standardized frequency difference:

$$d = \frac{(\hat{p}_{treated} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treated}(1-\hat{p}_{treated})+\hat{p}_{control}(1-\hat{p}_{control})}{2}}}$$

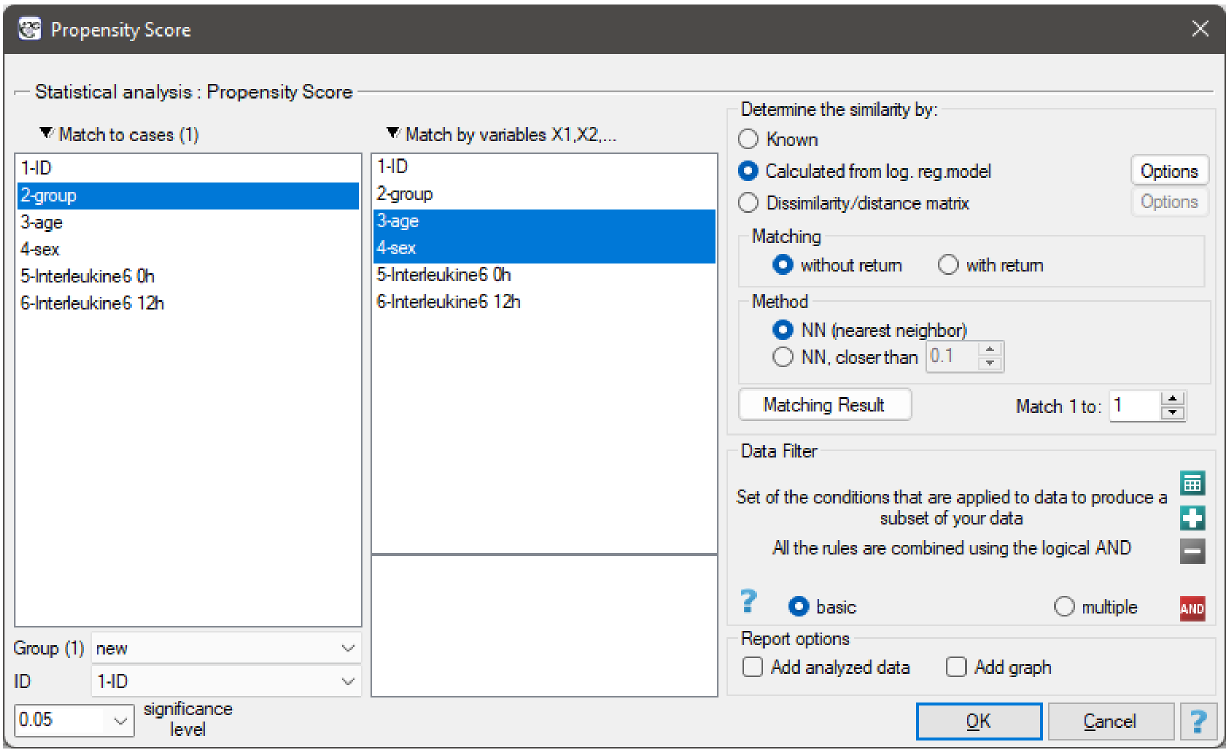
where:
$\hat{p}_{treated}$, $\hat{p}_{control}$ – is the frequency of the value described as 1 in the treatment group and the frequency of the value described as 1 in the control group.

Variables with multiple categories we should break down in logistic regression analysis into dummy variables with two categories and, by checking the fits of both groups, determine the standardized frequency difference for them.

Note
Although there is no universally agreed criterion for what threshold of standardized difference can be used to indicate significant imbalance, a standardized difference of less than 0.1 (in both mean and frequency estimation) can provide a clue. Therefore, to conclude that the groups are well matched, we should observe standardized differences close to 0, and preferably not outside the range of -0.1 to 0.1. Graphically, these results are presented in a dot plot. Negative differences indicate lower means/frequencies in the treatment group, positive in the control group.

Note
The 1:1 match obtained in the reports means the summary for the study group and the corresponding control group obtained in the first match, the 1:2 match means the summary for the study group and the corresponding control group obtained in the first + second match (i.e., not the study group and the corresponding control group obtained in the second match only), etc.

The window with the settings of group matching options is launched from the menu Advanced statistics→Multivariate models→Propensity Score



Propensity Score

Statistical analysis : Propensity Score

Match to cases (1)

- 1-ID
- 2-group
- 3-age
- 4-sex
- 5-Interleukine6 0h
- 6-Interleukine6 12h

Match by variables X1,X2,...

- 1-ID
- 2-group
- 3-age
- 4-sex
- 5-Interleukine6 0h
- 6-Interleukine6 12h

Group (1): new

ID: 1-ID

0.05 significance level

Determine the similarity by:

○ Known

● Calculated from log. reg.model

○ Dissimilarity/distance matrix

Matching: ● without return ○ with return

Method: ● NN (nearest neighbor) ○ NN, closer than 0.1

Match 1 to: 1

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

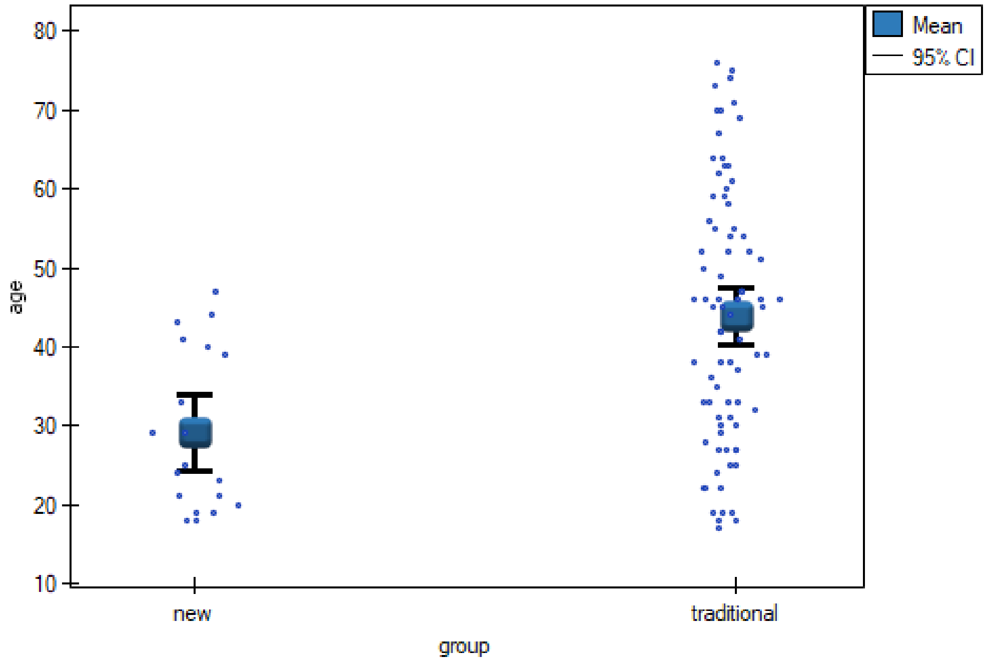
All the rules are combined using the logical AND

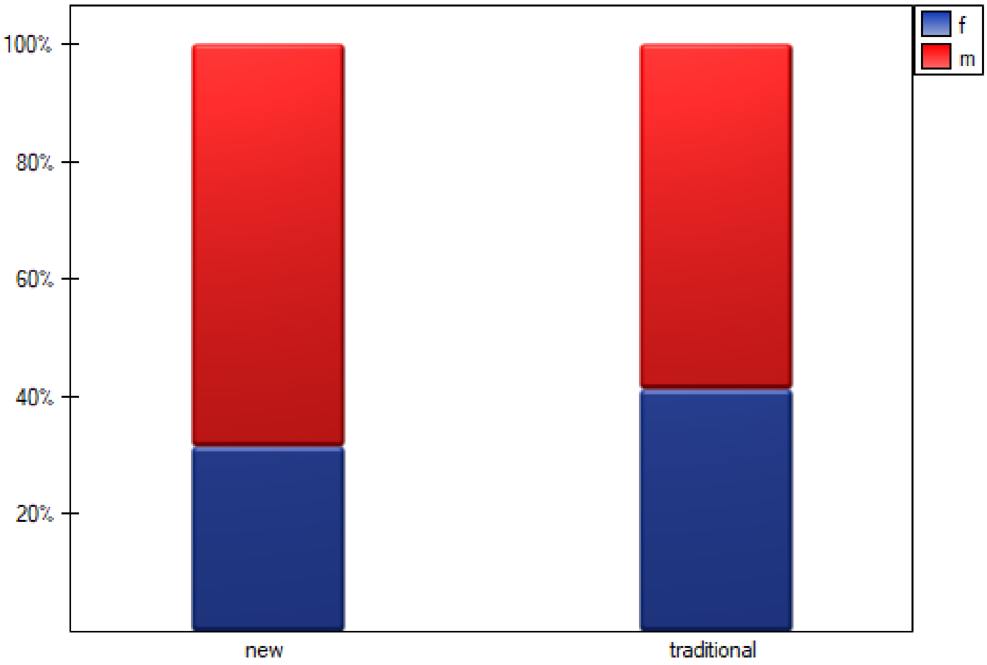
● basic ○ multiple

Report options: □ Add analyzed data □ Add graph

Example 24.1. (matching.pqs file)

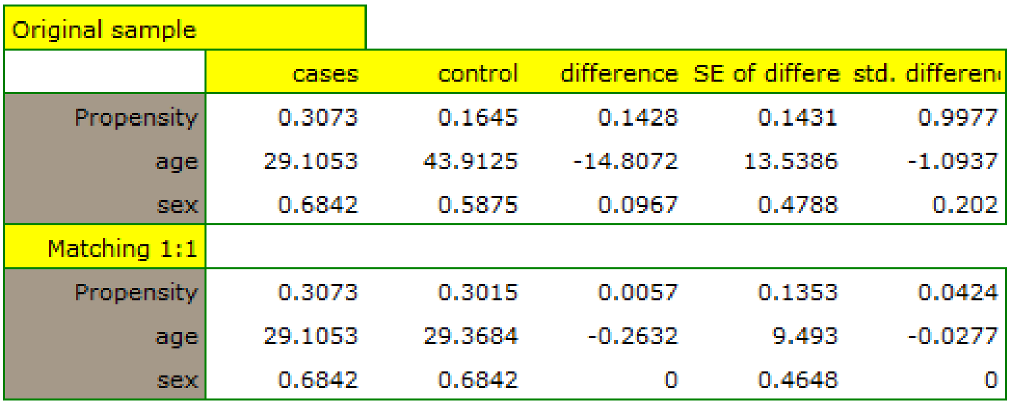
We want to compare two ways of treating patients after accidents, the traditional way and the new one. The correct effect of both treatments should be observed in the decreasing levels of selected cytokines. To compare the effectiveness of the two treatments, they should both be carried out on patients who are quite similar. Then we will be sure that any differences in the effectiveness of these treatments will be due to the treatment effect itself and not to other differences between patients assigned to different groups. The study is 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 new drug treatment group and the traditional treatment group. 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 gender structure, but not identical.





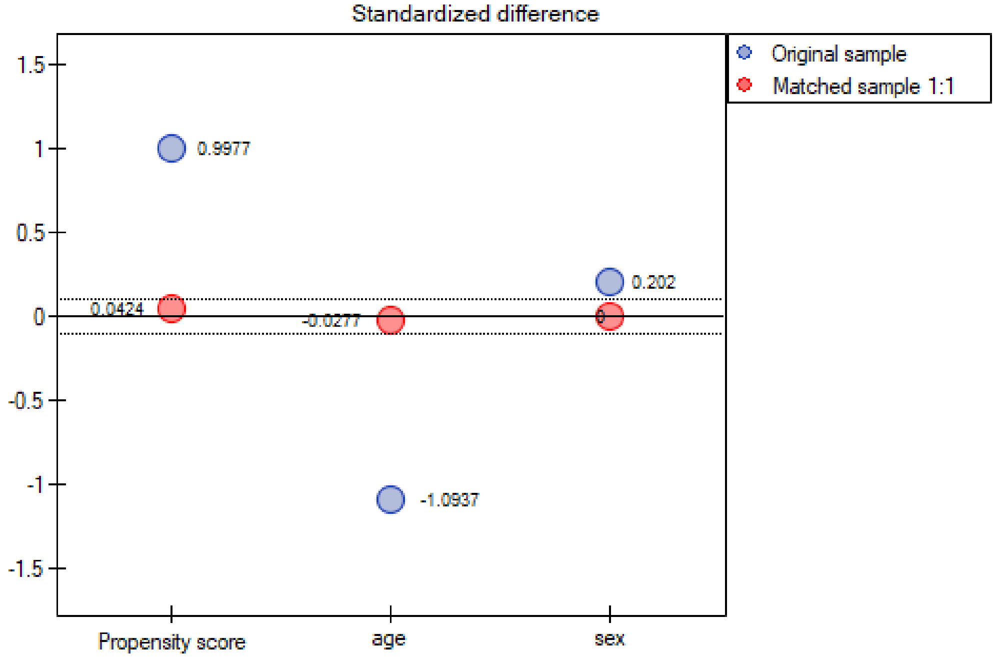
If the planned study had been carried out on such selected groups of patients, the new way would have had an easier challenge, because younger organisms might have responded better to the treatment. The conditions of the experiment would not be equal for both ways, which could falsify the results of the analyses and the conclusions drawn. Therefore, it was decided to match the group treated traditionally to be similar to the study group treated with the new way. We planned to make the matching with respect to two characteristics, i.e. age and gender. The traditional treatment group is larger (80 patients) than the new treatment group (19 patients), so there is a good chance that the groups will be similar. Random selection is performed by the logistic regression model algorithm embedded in the PSM. We remember that gender should be coded numerically, since only numerical values are involved in the logistic regression analysis. We choose nearest neighbor as the method. We want the same person to be unable to be selected duplicately, so we choose a no return randomization. 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 matching is random, so it depends on the random value of *seed* set by our computer, so the randomization performed by the reader may differ from the values presented here.
A summary of the selection can be seen in the tables and charts.



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

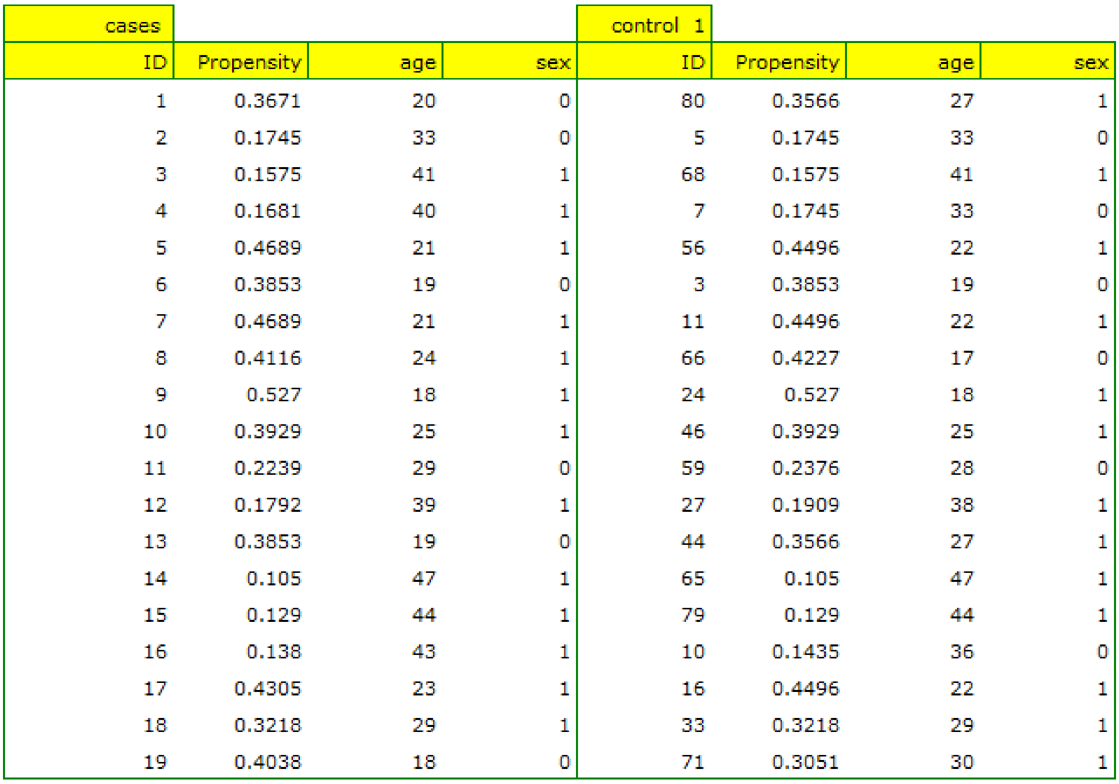
In the original sample, the mean age is more than 14 years higher in traditionally treated patients (difference between means is 14.8072), while the gender structure differs by less than 10% (0.0967). Much smaller differences are observed between patients treated with the new modality and matched patients treated traditionally. We obtain the most information about the quality of the match from the standardized differences (last column of the table and graph).



The line at 0 indicates equilibrium of the groups (difference between groups equal to 0). When the groups are in equilibrium with respect to the given characteristics, then all points on the graph are close to this line, i.e., around the interval -0.1 to 0.1. In the case of the original sample (blue color), we see a significant departure of Propensity Score. As we know, this mismatch is mainly due to age mismatch – its standardized difference is at a large distance from 0, and to a lesser extent gender mismatch.
By performing the matching we obtained groups more similar to each other (red color in the graph). The standardized difference between the groups as determined by Propensity Score is 0.0424, which is within the specified range. The age of both groups is already similar – the traditional treatment group

differs from the new treatment group by less than a year on average (the difference between the averages presented in the table is 0.2632) and the standardized difference between the averages is -0.0277. In the case of gender, the match is perfect, i.e. the percentage of females and males is the same in both groups (the standardized difference between the percentages presented in the table and the graph is now 0). We can return the data prepared in this way to the worksheet and subject it to the analyses we have planned.

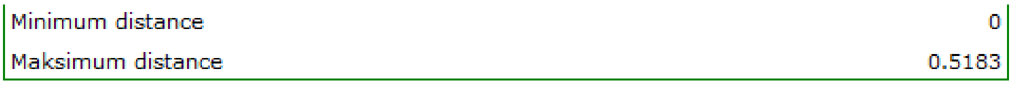
Looking at the summary we just obtained, we can see that despite the good balancing of the groups and the perfect match of many individuals, there are individuals who are not as similar as we might expect.



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

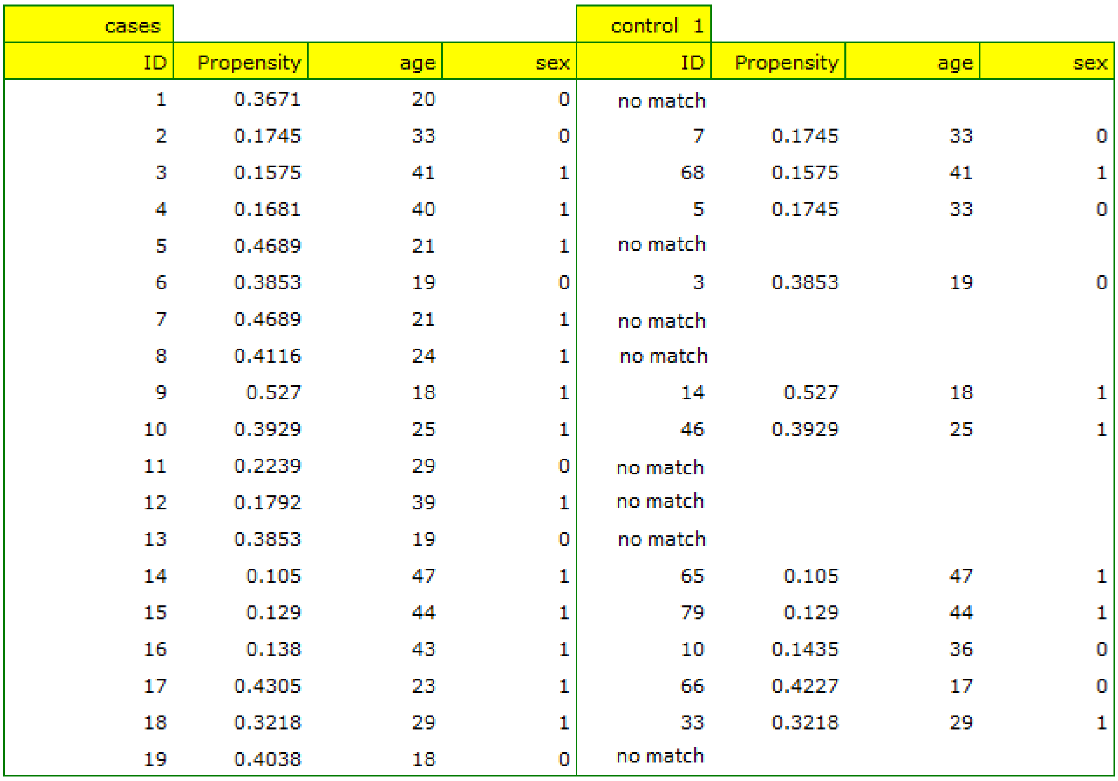
Sometimes in addition to obtaining well-balanced groups, researchers are interested in determining the exact way of selecting individuals, i.e. obtaining a greater influence on the similarity of objects as to the value of Propensity Score or on the similarity of objects as to the value of specific characteristics. Then, if the group from which we draw is sufficiently large, the analysis may yield results that are more favorable from the researcher's point of view, but if in the group from which we draw there is a lack of objects meeting our criteria, then for some people we will not be able to find a match that meets our conditions.

- Suppose that we would like to obtain such groups whose Propensity Score (i.e., propensity to take the survey) differs by no more than ...
  How to determine this value? You can take a look at the report from the earlier analysis, where the smallest and largest distance between the drawn objects is given.

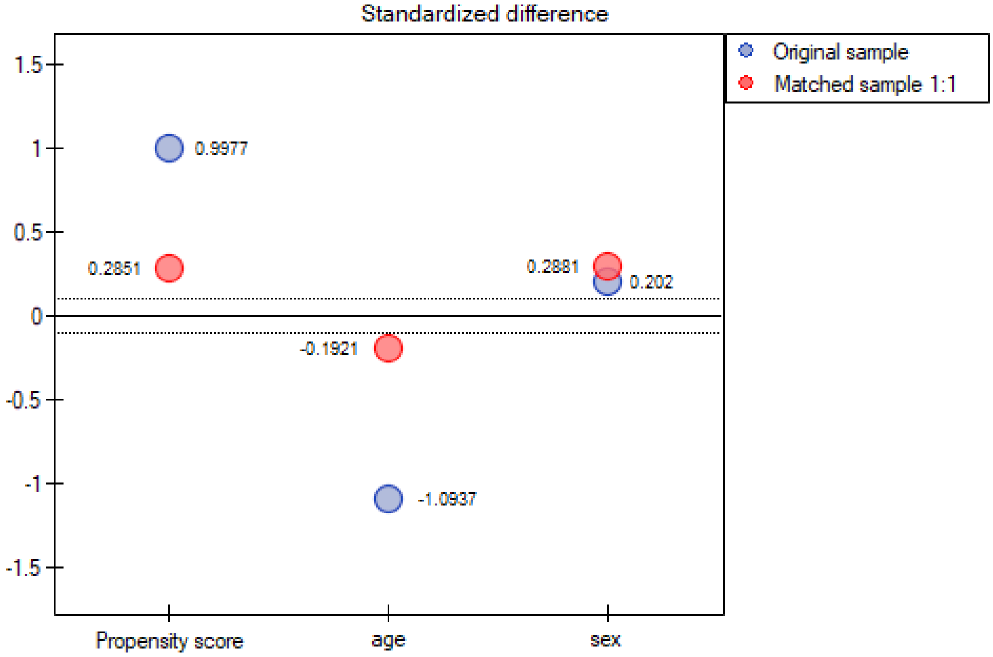
Minimum distance	0
Maksimum distance	0.5183

  In our case the objects closest to each other differ by min=0, and the furthest by max=0.5183. We will try to check what kind of selection we will obtain when we will match to people treated with the new method such people treated traditionally, whose Propensity Score will be very close to e.g. less than 0.01.

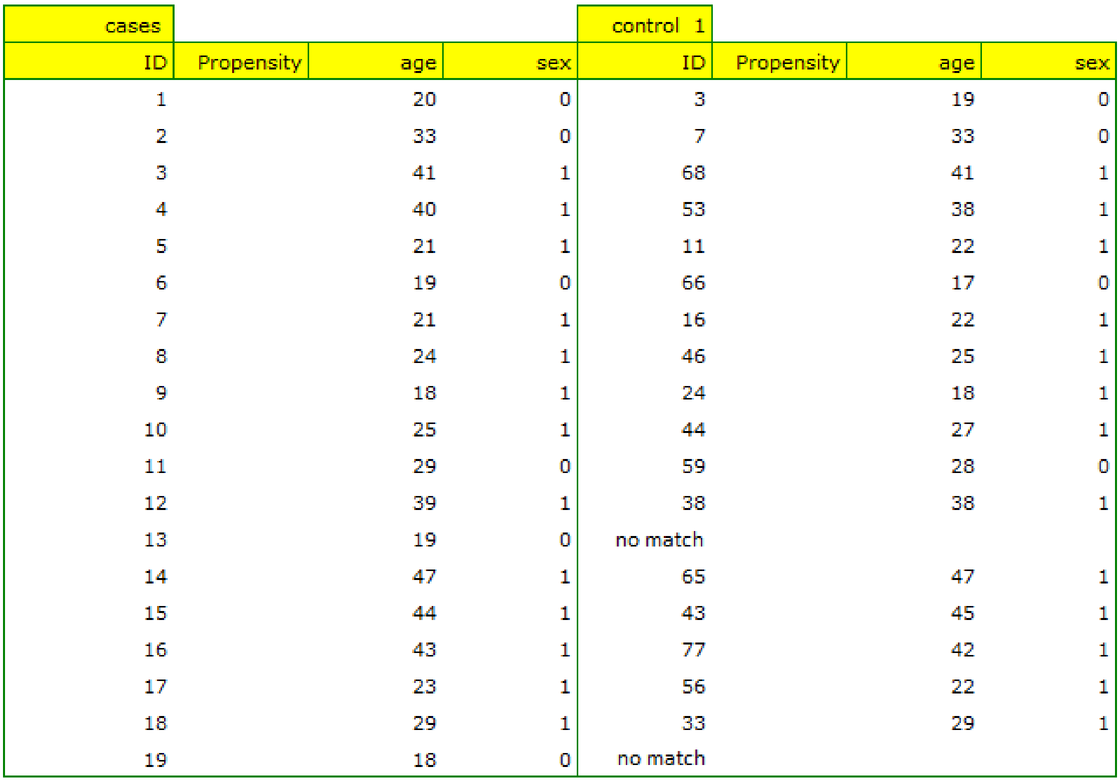


cases				control 1			
ID	Propensity	age	sex	ID	Propensity	age	sex
1	0.3671	20	0	no match			
2	0.1745	33	0	7	0.1745	33	0
3	0.1575	41	1	68	0.1575	41	1
4	0.1681	40	1	5	0.1745	33	0
5	0.4689	21	1	no match			
6	0.3853	19	0	3	0.3853	19	0
7	0.4689	21	1	no match			
8	0.4116	24	1	no match			
9	0.527	18	1	14	0.527	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 with failed to select the whole group. Comparing Propensity Score for each pair (treated with the new method and treated traditionally) we can see that the differences are really small. However, since the matched group is much smaller, to sum up the whole process we have to notice that both Propensity Score, age and sex are not close enough to the line at 0. Our will to improve the situation did not lead to the desired effect, and the obtained groups are not well balanced.

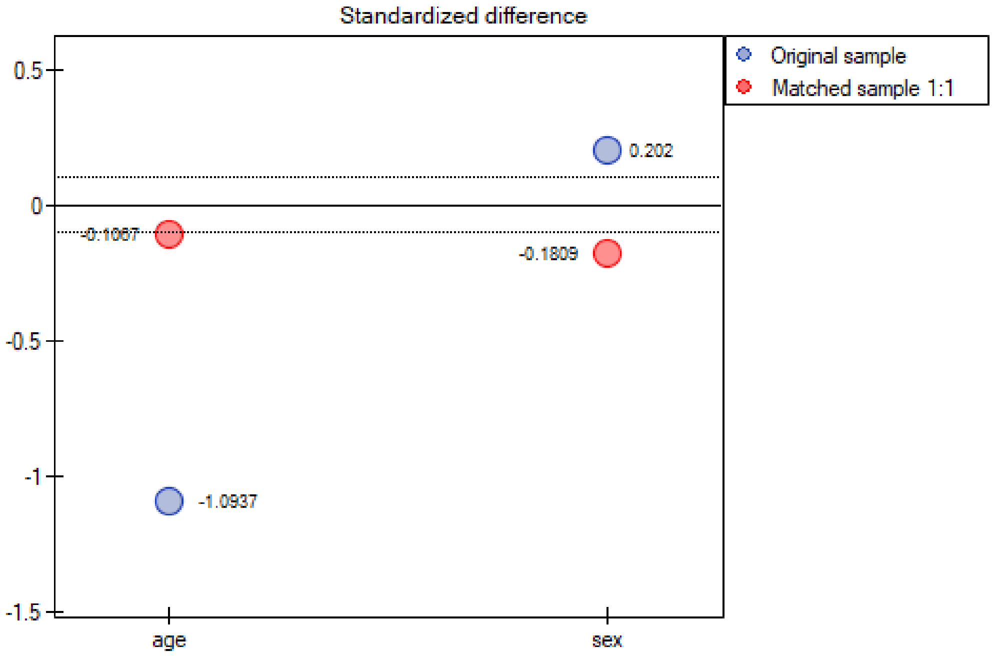


- Suppose we wanted to obtain such pairs (subjects treated with the new method and subjects treated traditionally) who are of the same sex and whose ages do not differ by more than 3 years. In the Propensity Score-based randomization, we did not have this type of ability to influence the extent of concordance of each variable. For this we will use a different method, not based on Propensity Score, but based on distance/dissimilarity matrices. After selecting the Options button, we select the proposed Mahalanobis statistical distance matrix and set the neighborhood fit to a maximum distance equal to 3 for age and equal to 0 for gender. As a result, for two people we failed to find a match, but the remaining matches meet the set criteria.

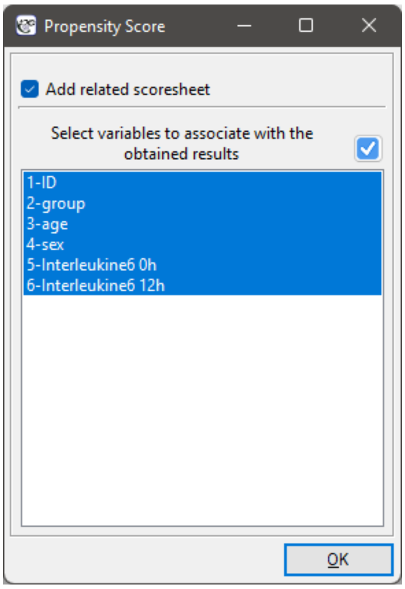


cases				control 1			
ID	Propensity	age	sex	ID	Propensity	age	sex
1		20	0	3		19	0
2		33	0	7		33	0
3		41	1	68		41	1
4		40	1	53		38	1
5		21	1	11		22	1
6		19	0	66		17	0
7		21	1	16		22	1
8		24	1	46		25	1
9		18	1	24		18	1
10		25	1	44		27	1
11		29	0	59		28	0
12		39	1	38		38	1
13		19	0	no match			
14		47	1	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			

To summarize the overall draw, we note that although it meets our assumptions, the resulting groups are not as well balanced as they were in our first draw based on Propensity Score. The points in red representing the quality of the match by age and the quality of the match by gender deviate slightly from the line of sameness set at level 0, which means that the average difference in age and sex structure is now greater than in the first matching.



It is up to the researcher to decide which way of preparing the data will be more beneficial to them. Finally, when the decision is made, the data can be returned to a new worksheet. To do this, go back to the report you selected and in the project tree under the right button select the Redo analysis menu. In the same analysis window, point to the Fit Result button and specify which other variables will be returned to the new worksheet.



☑ Add related scoresheet

Select variables to associate with the obtained results ☑

1-ID
2-group
3-age
4-sex
5-Interleukine6 0h
6-Interleukine6 12h

This will result in a new data sheet with side-by-side data for people treated with the new treatment and matched people treated traditionally.

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 - one-dimensional (univariate) models. The 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 simultaneously selected in the analysis window, it is possible to choose the model.

24.1 PREPARATION OF VARIABLES FOR ANALYSIS

24.1.1 Variables coding

When preparing data for a multidimensional analysis there is the problem of appropriate coding of nominal and ordinal variables. That is an important element of preparing data for analysis as it is a key factor in the interpretation of the coefficients of a model. The nominal or ordinal variables divide the analyzed objects into two or more categories. The dichotomous variables (in two categories, $k = 2$) must only be appropriately coded, whereas the variables with many categories ($k > 2$) ought to be divided into dummy variables with two categories and coded.

$k = 2$ If a variable is dichotomous, it is the decision of the researcher how the data representing the variable will be entered, so any numerical codes can be entered, e.g. 0 and 1. In the program one can change one's coding into effect coding by selecting that option in the window of the selected multidimensional analysis. Such coding causes a replacement of the smaller value with value -1 and of the greater value with value 1.

$k > 2$ If a variable has many categories then in the window of the selected multidimensional analysis we select the button Dummy variables and set the reference/base category for those variables which we want to break into dummy variables. The variables will be dummy coded unless the effect coding option will be selected in the window of the analysis – in such a case, they will be coded as -1, 0, and 1.

Dummy coding is employed in order to answer, with the use of multidimensional models, the question: How do the (Y) results in any analyzed category differ from the results of the reference category. The coding consists in ascribing value 0 or 1 to each category of the given variable. The category coded as 0 is, then, the **reference category**.

$k = 2$ If the coded variable is dichotomous, then by placing it in a regression model we will obtain the coefficient calculated for it, (b_i). The coefficient is the reference of the value of the dependent variable Y for category 1 to the reference category (corrected with the remaining variables in the model).

$k > 2$ If the analyzed variable has more than two categories, then k categories are represented by $k-1$ dummy variables with dummy coding. When creating variables with dummy coding one selects a category for which no dummy category is created. That category is treated as a reference category (as the value of each variable coded in the dummy coding is equal to 0. [0.2cm] When the $X_1, X_2, ..., X_{k-1}$ variables obtained in that way, with dummy coding, are placed in a regression model, then their $b_1, b_2, ..., b_{k-1}$ coefficients will be calculated.

b_1 is the reference of the Y results (for codes 1 in X_1) to the reference category (corrected with the remaining variables in the model);

b_2 is the reference of the Y results (for codes 1 in X_2) to the reference category (corrected with the remaining 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 (corrected with the remaining variables in the model);

Example
We code, in accordance with dummy coding, the sex variable with two categories (the male sex will be selected as the reference category), and the education variable with 4 categories (elementary education will be selected as the reference category).

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
...	...

	Coded education		
Education	vocational	secondary	tertiary
elementary	0	0	0
elementary	0	0	0
elementary	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 what impact the variables have on a dependent variable, e.g. Y = the amount of earnings (in thousands of PLN). As a result of such an analysis we will obtain sample coefficients for each dummy variable:
- for sex the statistically significant coefficient $b_i = -0.5$, which means that average women's wages are a half of a thousand PLN lower than men's wages, assuming that all other variables in the model remain unchanged;
- for vocational education the statistically significant coefficient $b_i = 0.6$, which means that the average wages of people with elementary education are 0.6 of a thousand PLN higher than those of people with elementary education, assuming that all other variables in the model remain unchanged;
- for secondary education the statistically significant coefficient $b_i = 1$, which means that the average wages of people with secondary education are a thousand PLN higher than those of people with elementary education, assuming that all other variables in the model remain unchanged;
- for tertiary-level education the statistically significant coefficient $b_i = 1.5$, which means that the average wages of people with tertiary-level education are 1.5 PLN higher than those of people with elementary education, assuming that all other variables in the model remain unchanged;

Effect coding is used to answer, with the use of multidimensional models, the question: How do (Y) results in each analyzed category differ from the results of the (unweighted) mean obtained from the sample. The coding consists in ascribing value -1 or 1 to each category of the given variable. The category coded as -1 is then the **base category**

$k = 2$ If the coded variable is dichotomous, then by placing it in a regression model we will obtain the coefficient calculated for it, (b_i). The coefficient is the reference of Y for category 1 to the unweighted general mean (corrected with the remaining variables in the model).

If the analyzed variable has more than two categories, then 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 made. The category is treated in the models as a base category (as in each variable made by effect coding it has values -1).

When the $X_1, X_2, ..., X_{k-1}$ variables obtained in that way, with effect coding, are placed in a regression model, then their $b_1, b_2, ..., b_{k-1}$ coefficients will be calculated.

b_1 is the reference of the Y results (for codes 1 in X_1) to the unweighted general mean (corrected by the remaining variables in the model);

b_2 is the reference of the Y results (for codes 1 in X_2) to the unweighted general mean (corrected by the remaining variables in the model);

...

b_{k-1} is the reference of the Y results (for codes 1 in X_{k-1}) to the unweighted general mean (corrected by the remaining variables in the model);

Example

With the use of effect coding we will code the sex variable with two categories (the male category will be the base category) and a variable informing about the region of residence in the analyzed country. 5 regions were selected: northern, southern, eastern, western, and central. The central region will be the base one.

	Coded
Sex	**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
...	...

Regions	Coded regions			
of residence	**western**	**eastern**	**northern**	**southern**
central	-1	-1	-1	-1
central	-1	-1	-1	-1
central	-1	-1	-1	-1
western	1	0	0	0
western	1	0	0	0
western	1	0	0	0
western	1	0	0	0
eastern	0	1	0	0
eastern	0	1	0	0
eastern	0	1	0	0
eastern	0	1	0	0
northern	0	0	1	0
northern	0	0	1	0
southern	0	0	0	1
southern	0	0	0	1
southern	0	0	0	1
...	...	...	...	...

Building on the basis of dummy variables, in a multiple regression model, we might want to check what impact the variables have on a dependent variable, e.g. Y = the amount of earnings (expressed in thousands of PLN). As a result of such an analysis we will obtain sample coefficients for each dummy variable:

- for sex the statistically significant coefficient $b_i = -0.5$, which means that the average women's wages are a half of a thousand PLN lower than the average wages in the country, assuming that the other variables in the model remain unchanged;

- for the western region the statistically significant coefficient $b_i = 0.6$, which means that the average wages of people living in the western region of the country are 0.6 thousand PLN higher than the average wages in the country, assuming that the other variables in the model remain unchanged;

- for the eastern region the statistically significant coefficient $b_i = -1$, which means that the average wages of people living in the eastern region of the country are a thousand PLN lower than the average wages in the country, assuming that the other variables in the model remain unchanged;

- for the northern region the statistically significant coefficient $b_i = 0.4$, which means that the

average wages of people living in the western region of the country are 0.4 thousand PLN higher than the average wages in the country, assuming that the other variables in the model remain unchanged;
- for the southern region the statistically significant coefficient $b_i = 0.1$, which means that the average wages of people living in the southern region of the country do not differ in a statistically significant manner from the average wages in the country, assuming that the other variables in the model remain unchanged;

24.1.2 Interctions

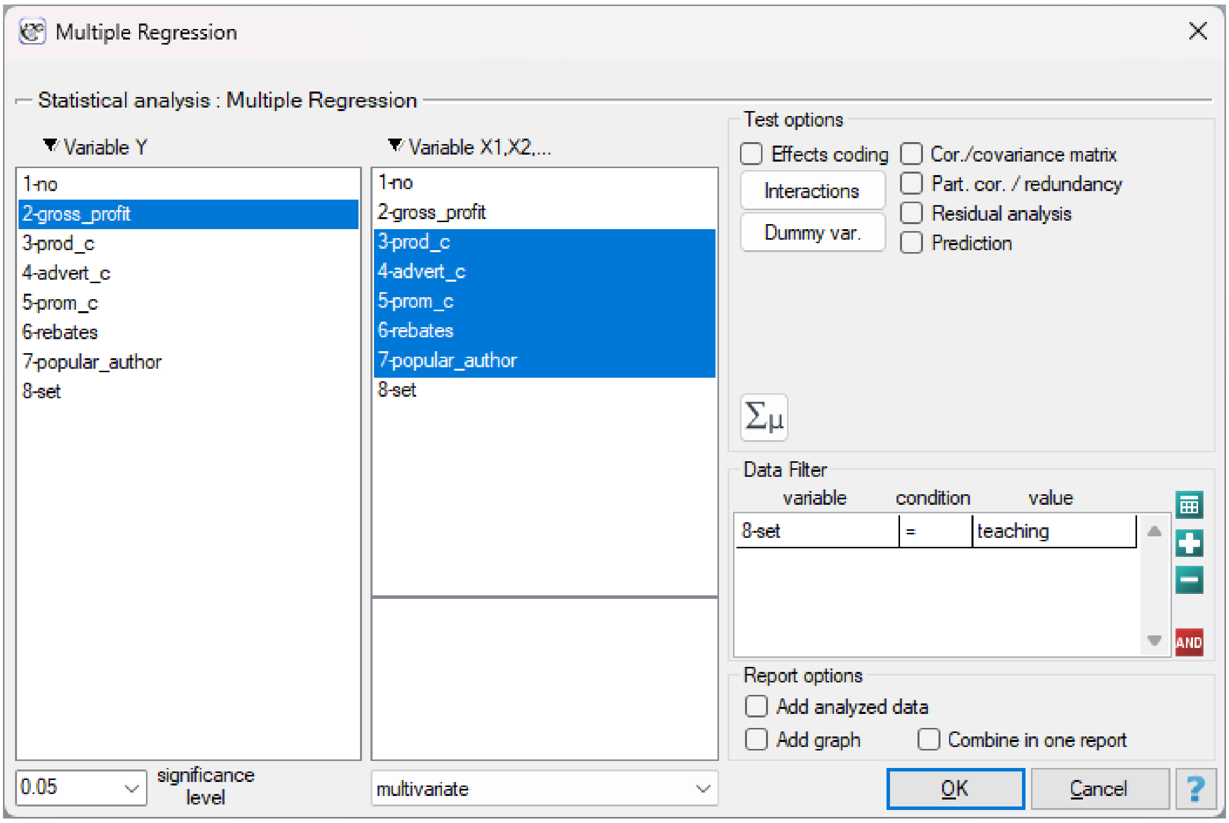
Interactions are considered in multidimensional models. Their presence means that the influence of the independent variable (X_1) on the dependent variable (Y) differs depending on the level of another independent variable (X_2) or a series of other independent variables. To discuss the interactions in multidimensional models one must determine the variables informing about possible interactions, i.e the product of appropriate variables. For that purpose we select the Interactions button in the window of the selected multidimensional analysis. In the window of interactions settings, with the CTRL button pressed, we determine the variables which are to form interactions and transfer the variables into the neighboring list with the use of an arrow. By pressing the OK button we will obtain appropriate columns in the datasheet.

In the analysis of the interaction the choice of appropriate coding of dichotomous variables allows the avoidance of the over-parametrization related to interactions. Over-parametrization causes the effects of the lower order for dichotomous variables to be redundant with respect to the confounding interactions of the higher order. As a result, the inclusion of the interactions of the higher order in the model annuls the effect of the interactions of the lower orders, not allowing an appropriate evaluation of the latter. In order to avoid the over-parametrization in a model in which there are interactions of dichotomous variables it is recommended to choose the option effect coding.

In models with interactions, remember to ”trim” them appropriately, so that when removing the main effects, we also remove the effects of higher orders that depend on them. That is: if in a model we have 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 by removing the variable X_1 from the model we must also remove the interactions in which it occurs, viz: $X_1 * X_2$, $X_1 * X_3$ and $X_1 * X_2 * X_3$.

24.2 MULTIPLE LINEAR REGRESSION

The window with settings for Multiple Regression is accessed via the menu Advanced statistics → Multidimensional Models→Multiple Regression



Multiple Regression

Statistical analysis : Multiple Regression

Variable Y

- 1-no
- 2-gross_profit (selected)
- 3-prod_c
- 4-advert_c
- 5-prom_c
- 6-rebates
- 7-popular_author
- 8-set

Variable X1,X2,...

- 1-no
- 2-gross_profit
- 3-prod_c (selected)
- 4-advert_c (selected)
- 5-prom_c (selected)
- 6-rebates (selected)
- 7-popular_author (selected)
- 8-set

Test options

- □ Effects coding
- □ Cor./covariance matrix
- □ Part. cor. / redundancy
- □ Residual analysis
- □ Prediction

Data Filter

variable	condition	value
8-set	=	teaching

Report options

- □ Add analyzed data
- □ Add graph
- □ Combine in one report

0.05 significance level

multivariate

The constructed model of linear regression allows the study of the influence of many independent variables$(X_1, X_2, \ldots, X_k)$ 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 Pearson's linear correlation coefficient. It presumes the existence of a linear relation between the studied variables. The linear model of multiple regression has the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon.$$

where:
Y - dependent variable, explained by the model,
$X_1, X_2, \ldots X_k$ - independent variables, explanatory,
$\beta_0, \beta_1, \beta_2, \ldots \beta_k$ - parameters,
ϵ - random parameter (model residual).

If the model was created on the basis of a data sample of size n the above equation can be presented in the form of a matrix:

$$Y = X\beta + \epsilon.$$

where:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

In such a case, the solution of the equation is the vector of the estimates of parameters $\beta_0, \beta_1, \ldots, \beta_k$ called **regression coefficients**:

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}.$$

Those coefficients are estimated with the help of the classical **least squares method**. On the basis of those values we can infer the magnitude of the effect of the independent variable (for which the coefficient was estimated) on the dependent variable. They inform by how many units will the dependent variable change when the independent variable is changed by 1 unit. There is a certain error of estimation for each coefficient. The magnitude of that error is estimated from the following 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 on the basis of the model).

Dummy variables and interactions in the model
A discussion of the coding of dummy variables and interactions is presented in chapter 24.1 Preparation of the variables for the analysis in multidimensional models.

Note
When constructing the model one should remember that the number of observations should meet the assumptions ($n \geq 50 + 8k$) where k is the number of explanatory variables in the model[70].

24.2.1 Model verification

- **Statistical significance of particular variables in the model**.

  On the basis of the coefficient and its error of estimation we can infer if the independent variable for which the coefficient was estimated has a significant effect on the dependent variable. For that purpose we use t-test.

  Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \beta_i = 0, \\ \mathcal{H}_1: & \beta_i \neq 0. \end{array}$$

  Let us estimate the test statistics according to the formula below:

$$t = \frac{b_i}{SE_{b_i}}$$

  The test statistics has t-Student distribution with $n - k$ degrees of freedom.
  The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \leq \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

- **The quality of the constructed model** of multiple linear regression can be evaluated with the help of several measures.

- **The standard error of estimation** – it is the measure of model adequacy:

$$SE_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-(k+1)}}.$$

The measure is based on model residuals $e_i = y_i - \widehat{y}_i$, that is on the discrepancy between the actual values of the dependent variable y_i in the sample and the values of the independent variable $\widehat{y}_i$ estimated on the basis of the constructed model. It would be best if the difference were as close to zero as possible for all studied properties of the sample. Therefore, for the model to be well-fitting, the standard error of estimation (SE_e), expressed as e_i variance, should be the smallest possible.

- **Multiple correlation coefficient** $R = \sqrt{R^2} \in < 0; 1 >$ – defines the strength of the effect of the set of variables $X_1, X_2, \ldots X_k$ on the dependent variable Y.
- **Multiple determination coefficient** R^2 – it is the measure of model adequacy.
The value of that coefficient falls within the range of $< 0; 1 >$, where 1 means excellent model adequacy, 0 – a complete lack of adequacy. The estimation is made using the following formula:

$$T_{SS} = E_{SS} + R_{SS},$$

where:
T_{SS} – total sum of squares,
E_{SS} – the sum of squares explained by the model,
R_{SS} – residual sum of squares.

The coefficient of determination is estimated from the formula:

$$R^2 = \frac{T_{SS}}{E_{SS}}.$$

It expresses the percentage of the variability of the dependent variable explained by the model.
As the value of the coefficient R^2 depends on model adequacy but is also influenced by the number of variables in the model and by the sample size, there are situations in which it can be encumbered with a certain error. That is why a corrected value of that parameter is estimated:

$$R^2_{adj} = R^2 - \frac{k(1-R^2)}{n-(k+1)}.$$

- **Information criteria** are based on the entropy of information carried by the model (model uncertainty) i.e. they estimate the information lost when a given model is used to describe the phenomenon under study. Therefore, we should choose the model with the minimum value of a given information criterion.
The AIC, $AICc$ and BIC is a kind of trade-off between goodness of fit and complexity. The second element of the sum in the information criteria formulas (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). In both cases, this element increases as the number of variables increases, and this increase is faster the smaller the number of observations.The information criterion, however, is not an absolute measure, i.e., if all the models being compared misdescribe reality in the information criterion there is no point in looking for a warning.
Akaike information criterion

$$AIC = n \cdot \ln \frac{R_{SS}}{n} + 2(k+1) + (constant)$$

where, the constant can be omitted because 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 favor models with a large number of variables.

Example of interpretation of AIC size comparison
Suppose we determined the AIC for three models AIC_1=100, AIC_2=101.4, AIC_3=110. Then the relative reliability for the model can be determined. This reliability is relative because it is determined relative to another model, usually the one with the smallest AIC value. We determine it according to the formula: $e^{(AIC_{min}-AIC_i)/2}$. Comparing model 2 to model 1, we will say that the probability that it will minimize the loss of information is about half of the probability that model 1 will do so (specifically exp((100− 101.4)/2) = 0.497). Comparing model 3 to model one, we will say that the probability that it will minimize information loss is a small fraction of the probability that model 1 will do so (specifically exp((100- 110)/2) = 0.007).

Akaike coreccted information criterion

$$AICc = AIC + \frac{2(k+3)(k+4)}{n-k}$$

Correction of Akaike's criterion relates to sample size, which makes this measure recommended also for small sample sizes. **Bayes Information Criterion (or Schwarz criterion)**

$$BIC = n \cdot \ln \frac{R_{SS}}{n} + (k+1)\ln n + (constant)$$

where, the constant can be omitted because it is the same in each of the compared models. Like Akaike's revised criterion, the BIC takes into account the sample size.

- **Error analysis for ex post forecasts:**
  MAE (mean absolute error) -– forecast accuracy specified by MAE informs how much on average the realised values of the dependent variable will deviate (in absolute value) from the forecasts.

  $$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

  MPE (mean percentage error) -– informs what average percentage of the realization of the dependent variable are forecast errors.

  $$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{y_i}$$

  MAPE (mean absolute percentage error) -– informs about the average size of forecast errors expressed as a percentage of the actual values of the dependent variable. MAPE allows you to compare the accuracy of forecasts obtained from different models.

  $$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right|$$

- **Statistical significance of all variables in the model**
  The basic tool for the evaluation of the significance of all variables in the model is the analysis of variance test (the F-test). The test simultaneously verifies 3 equivalent hypotheses:

$$\begin{array}{llll}\mathcal{H}_0: & \text{all } \beta_i = 0, & \mathcal{H}_1: & \text{exists } \beta_i \neq 0;\\ \mathcal{H}_0: & R^2 = 0, & \mathcal{H}_1: & R^2 \neq 0;\\ \mathcal{H}_0: & \text{linearity of the relation}, & \mathcal{H}_1: & \text{a lack of a linear relation}.\end{array}$$

The test statistics has the form presented below:

$$F = \frac{E_{MS}}{R_{MS}}$$

where:
$E_{MS} = \frac{E_{SS}}{df_E}$ – the mean square explained by the model,
$R_{MS} = \frac{R_{SS}}{df_R}$ – residual mean square,
$df_E = k$, $df_R = n - (k+1)$ – appropriate degrees of freedom.

That statistics is subject to F-Snedecor distribution with df_E and df_R degrees of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

24.2.2 More information about the variables in the model

- **Standardized** $b_1, b_2, \ldots, b_k$ – In contrast to raw parameters (which are expressed in different units of measure, depending on the described variable, and are not directly comparable) the standardized estimates of the parameters of the model allow the comparison of the contribution of particular variables to the explanation of the variance of the dependent variable Y.

- **Correlation matrix** – contains information about the strength of the relation between particular variables, that is the Pearson's correlation coefficient $r_p \in < -1; 1 >$. The coefficient is used for the study of the corrrelation of each pair of variables, without taking into consideration the effect of the remaining variables in the model.

- **Covariance matrix** – similarly to the correlation matrix it contains information about the linear relation among particular variables. That value is not standardized.

- **Partial correlation coefficient** – falls within the range $< -1; 1 >$ and is the measure of correlation between the specific independent variable X_i (taking into account its correlation with the remaining variables in the model) and the dependent variable Y (taking into account its correlation with the remaining variables in the model).
  The square of that coefficient is the **partial determination coefficient** – it falls within the range $< 0; 1 >$ and defines the relation of only the variance of the given independent variable X_i with that variance of the dependent variable Y which was not explained by other variables in the model.
  The closer the value of those coefficients to 0, the more useless the information carried by the studied variable, which means the variable is redundant.

- **Semipartial correlation coefficient** – falls within the range $< -1; 1 >$ and is the measure of correlation between the specific independent variable X_i (taking into account its correlation with the remaining variables in the model) and the dependent variable Y (NOT taking into account its correlation with the remaining variables in the model).
  The square of that coefficient is the **semipartial determination coefficient** – it falls within the range $< 0; 1 >$ and defines the relation of only the variance of the given independent variable X_i with the complete variance of the dependent variable Y.
  The closer the value of those coefficients to 0, the more useless the information carried by the studied variable, which means the variable is redundants.

- **R-squared ($R^2 \in < 0; 1 >$)** – it represents the percentage of variance of the given independent variable X_i, explained by the remaining independent variables. The closer to value 1 the stronger the linear relation of the studied variable with the remaining independent variables, which can mean that the variable is a redundant one.
- **Variance inflation factor ($VIF \in < 1; \infty)$)** – determines how much the variance of the estimated regression coefficient is increased due to collinearity. The closer the value is to 1, the lower the collinearity and the smaller its effect on the coefficient variance. It is assumed that strong collinearity occurs when the coefficient VIF>5 [148]. f the variance inflation factor is 5 ($\sqrt{5}$ = 2.2), this means that the standard error for the coefficient of this variable is 2.2 times larger than if this variable had zero correlation with other variables X_i.
- **Tolerance =** $1 - R^2 \in < 0; 1 >$ – it represents the percentage of variance of the given independent variable X_i, NOT explained by the remaining independent variables. The closer the value of tolerance is to 0 the stronger the linear relation of the studied variable with the remaining independent variables, which can mean that the variable is a redundant one.
- **A comparison of a full model with a model in which a given variable is removed**
  The comparison of the two model is made with by means of:
  - F test, in a situation in which one variable or more are removed from the model (see: the comparison of models),
  - t-test, when only one variable is removed from the model. It is the same test that is used for studying the significance of particular variables in the model.

  In the case of removing only one variable the results of both tests are identical.
  If the difference between the compared models is statistically significant (the value $p \leq \alpha$), the full model is significantly better than the reduced model. It means that the studied variable is not redundant, it has a significant effect on the given model and should not be removed from it.
- **Scatter plots**
  The charts allow a subjective evaluation of linearity of the relation among the variables and an identification of outliers. Additionally, scatter plots can be useful in an analysis of model residuals.

24.2.3 Analysis of model residuals

To obtain a correct regression model we should check the basic assumptions concerning model residuals.

- **Outliers**
  The study of the model residual can be a quick source of knowledge about outlier values. Such observations can disturb the equation of the regression to a large extent because they have a great effect on the values of the coefficients in the equation. If the given residual e_i deviates by more than 3 standard deviations from the mean value, such an observation can be classified as an outlier. A removal of an outlier can greatly enhance the model.
  Cook's distance - describes the magnitude of change in regression coefficients produced by omitting a case. In the program, Cook's distances for cases that exceed the 50th percentile of the F-Snedecor distribution statistic are highlighted in bold $F(0.5, k+1, n-k-1)$.
  Mahalanobis distance - is dedicated to detecting outliers - high values indicate that a case is significantly distant from the center of the independent variables. If a case with the highest Mahalanobis value is found among the cases more than 3 deviations away, it will be marked in bold as the outlier.

- **Normalność rozkładu reszt modelu**
  We check this assumption visually using a Q-Q plot of the nromal distribution. The 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 (homogeneity of variance)**
  To check if there are areas in which the variance of model residuals is increased or decreased we use the charts of:

  - the residual with respect to predicted values
  - the square of the residual with respect to predicted values
  - the residual with respect to observed values
  - the square of the residual with respect to observed values

- **Autocorrelation of model residuals**
  For the constructed model to be deemed correct the values of residuals should not be correlated with one another (for all pairs e_i, e_j). The assumption can be checked by by computing the Durbin-Watson statistic.

  $$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2},$$

  To test for positive autocorrelation on the significance level α we check the position of the statistics d with respect 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 ambiguous.

  To test for negative autocorrelation on the significance level α we check the position of the value $4 - d$ with respect to the upper ($d_{U,\alpha}$) and lower ($d_{L,\alpha}$) critical value:

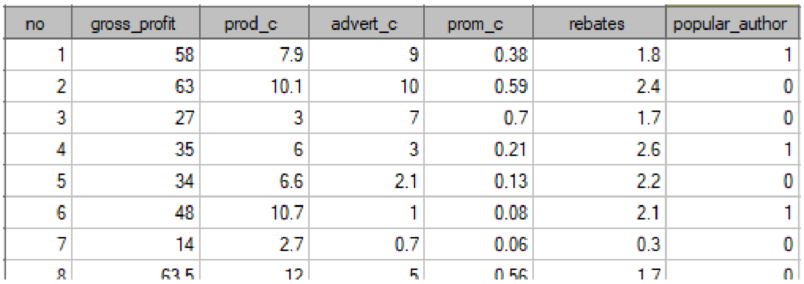
  - If $4 - d < d_{L,\alpha}$ – the errors are negatively correlated;
  - If $4 - d > d_{U,\alpha}$ – the errors are not negatively correlated;
  - If $d_{L,\alpha} < 4 - d < d_{U,\alpha}$ – the test result is ambiguous.

  The critical values of the Durbin-Watson test for the significance level $\alpha = 0.05$ are on the website www.pqstat.com – the source of the: Savin and White tables (1977)[144]

24.2.4 Example for multiple regression

Example 24.2. (publisher.pqs file)

A certain book publisher wanted to learn how was gross profit from sales influenced by such variables as: production cost, advertising costs, direct promotion cost, the sum of discounts made, and the author's popularity. For that purpose he analyzed 40 titles published during the previous year (teaching set). A part of the data is presented in the image below:

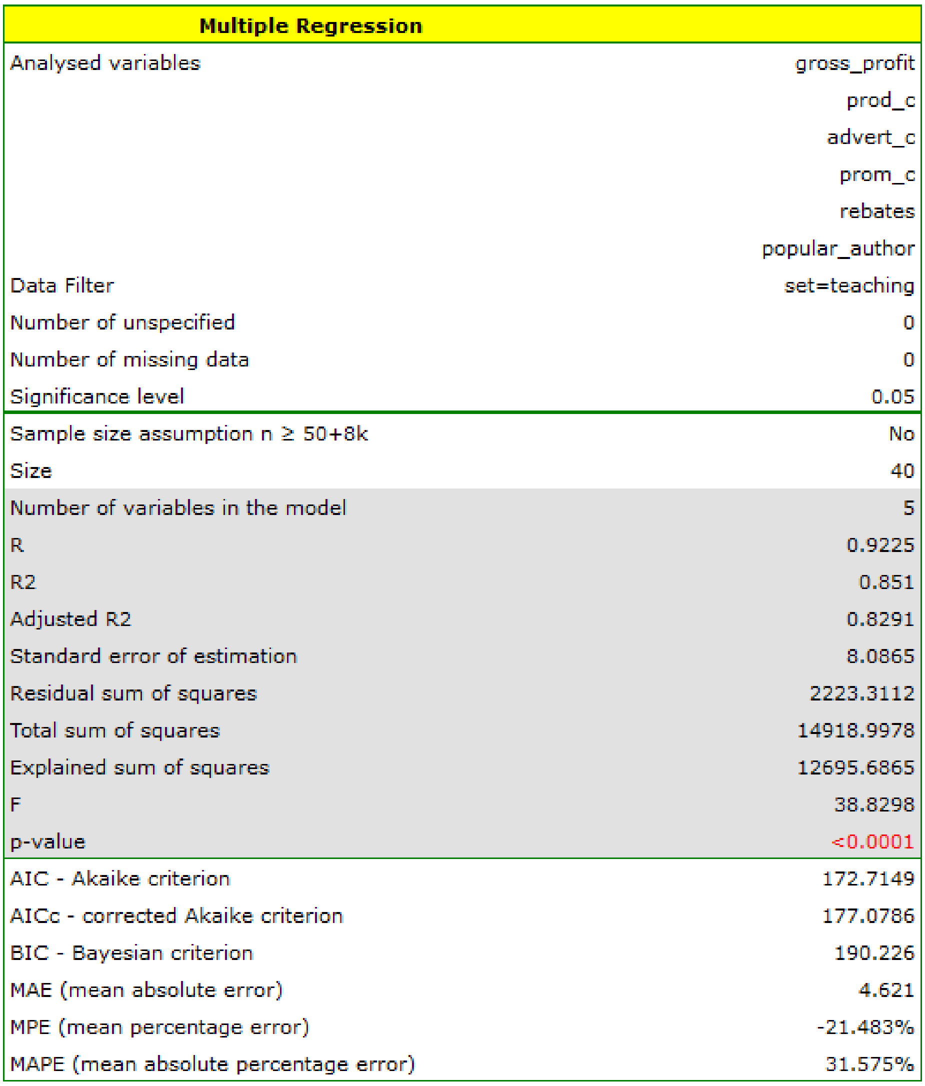


no	gross_profit	prod_c	advert_c	prom_c	rebates	popular_author
1	58	7.9	9	0.38	1.8	1
2	63	10.1	10	0.59	2.4	0
3	27	3	7	0.7	1.7	0
4	35	6	3	0.21	2.6	1
5	34	6.6	2.1	0.13	2.2	0
6	48	10.7	1	0.08	2.1	1
7	14	2.7	0.7	0.06	0.3	0
8	63.5	12	5	0.56	1.7	0

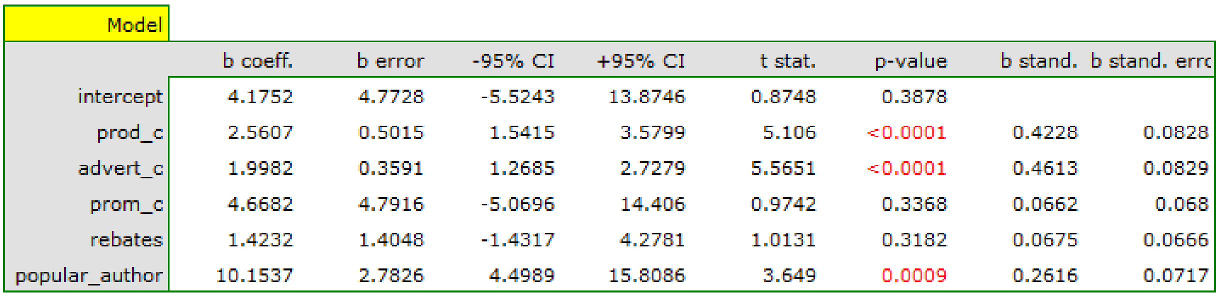
The first five variables are expressed in thousands fo dollars - so they are variables gathered on an interval scale. The last variable: the author's popularity – is a dychotomic variable, where 1 stands for a known author, and 0 stands for an unknown author.

On the basis of the knowledge gained from the analysis the publisher wants to predict the gross profit from the next published book written by a known author. The expenses the publisher will bear are: production cost ≈ 11, advertising costs ≈ 13, direct promotion costs ≈ 0.5, the sum of discounts made ≈ 0.5.

We construct the model of multiple linear regression, for teaching dataset, selecting: gross profit – as the dependent variable Y, production cost, advertising costs, direct promotion costs, the sum of discounts made, the 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 estimated, together with measures which will allow the evaluation of the quality of the model.



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. errc
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

On the basis of the estimated value of the coefficient b, the relationship between gross profit and all independent variables can be described by means of the equation:

$$profit_{gross} = 4.18{+}2.56(c_{prod}){+}2(c_{adv}){+}4.67(c_{prom}){+}1.42(discounts){+}10.15(popul_{author}){+}[8.09]$$

The obtained coefficients are interpreted in the following manner:

- If the production cost increases by 1 thousand dollars, then gross profit will increase by about 2.56 thousand dollars, assuming that the remaining variables do not change;
- If the production cost increases by 1 thousand dollars, then gross profit will increase by about 2 thousand dollars, assuming that the remaining variables do not change;
- If the production cost increases by 1 thousand dollars, then gross profit will increase by about 4.67 thousand dollars, assuming that the remaining variables do not change;
- If the sum of the discounts made increases by 1 thousand dollars, then gross profit will increase by about 1.42 thousand dollars, assuming that the remaining variables do not change;
- If the book has been written by a known author (marked as 1), then in the model the author's popularity is assumed to be the value 1 and we get the equation:

  $$profit_{gross} = 14.33 + 2.56(c_{prod}) + 2(c_{adv}) + 4.67(c_{prom}) + 1.42(discounts)$$

  If the book has been written by an unknown author (marked as 0), then in the model the author's popularity is assumed to be the value 0 and we get the equation:

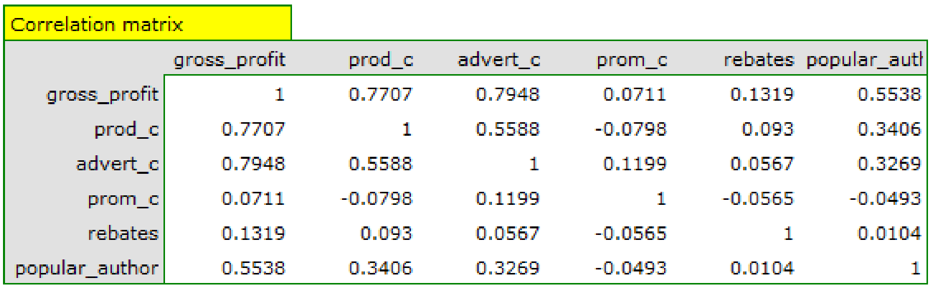
  $$profit_{gross} = 4.18 + 2.56(c_{prod}) + 2(c_{adv}) + 4.67(c_{prom}) + 1.42(discounts)$$

The result of t-test for each variable shows that only the production cost, advertising costs, and author's popularity have a significant influence on the profit gained. At the same time, that standardized coefficients b are the greatest for those variables.

Additionally, the model is very well-fitting, which is confirmed by: the small standard error of estimation $SE_e = 8.087$, the high value of the multiple determination coefficient $R^2 = 0.85$, the corrected multiple determination coefficient $R^2_{adj} = 0.83$, and the result of the F-test of variance analysis: $p < 0.0001$.

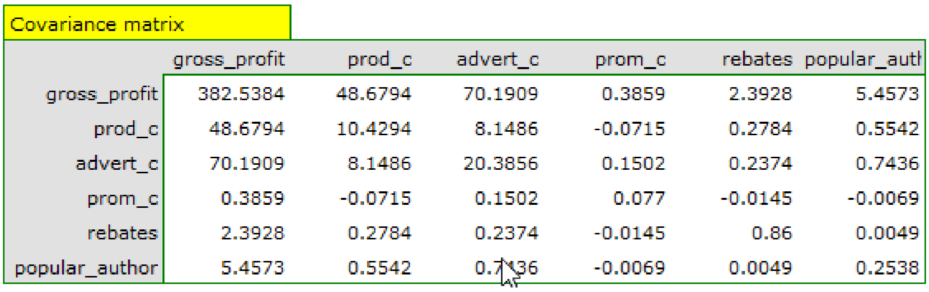
On the basis of the interpretation of the results obtained so far we can assume that a part of the variables does not have a significant effect on the profit and can be redundant.

For the model to be well formulated the interval independent variables ought to be strongly correlated with the dependent variable and be relatively weakly correlated with one another. That can be checked by computing the correlation matrix and the covariance matrix:



Correlation matrix

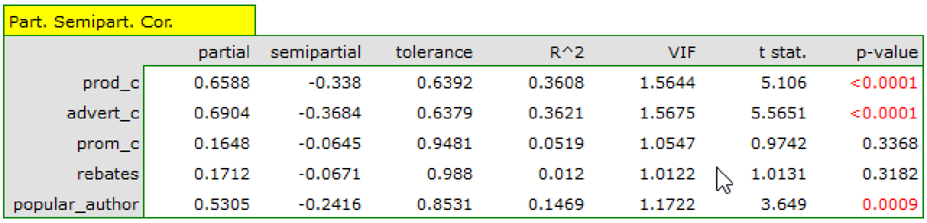
	gross_profit	prod_c	advert_c	prom_c	rebates	popular_auth
gross_profit	1	0.7707	0.7948	0.0711	0.1319	0.5538
prod_c	0.7707	1	0.5588	-0.0798	0.093	0.3406
advert_c	0.7948	0.5588	1	0.1199	0.0567	0.3269
prom_c	0.0711	-0.0798	0.1199	1	-0.0565	-0.0493
rebates	0.1319	0.093	0.0567	-0.0565	1	0.0104
popular_author	0.5538	0.3406	0.3269	-0.0493	0.0104	1



Covariance matrix

	gross_profit	prod_c	advert_c	prom_c	rebates	popular_auth
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 which allows finding those variables in the model which are redundant is given by the parial and semipartial correlation analysis as well as redundancy analysis:

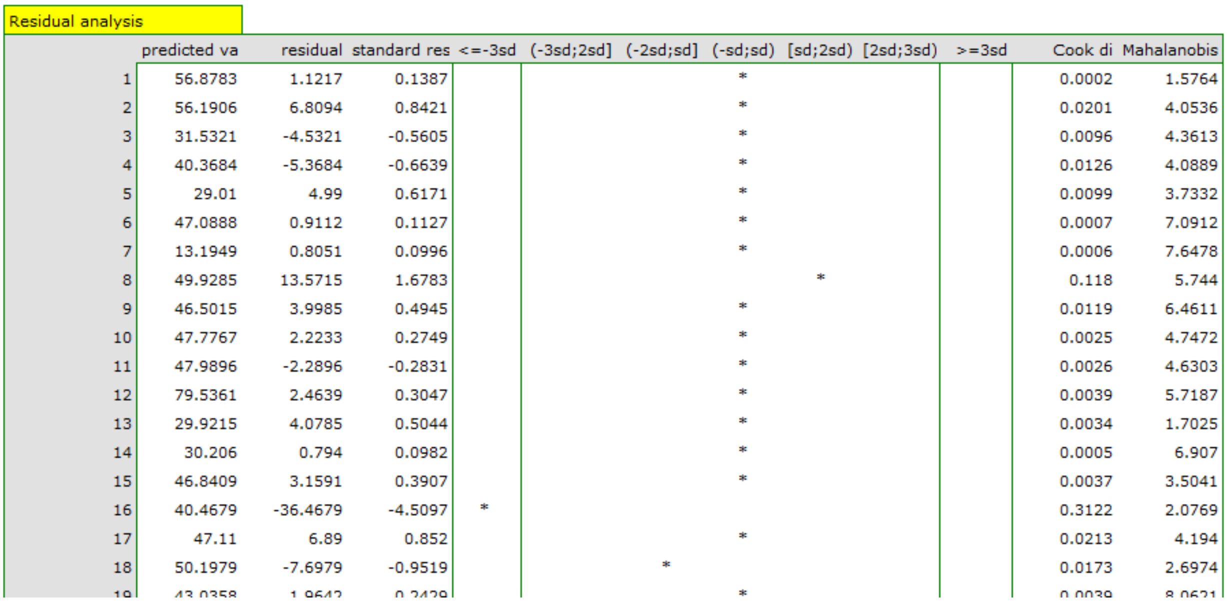


Part. Semipart. Cor.

	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 coefficients of partial and semipartial correlation indicate that the smallest contribution into the constructed model is that of direct promotion costs and the sum of discounts made. However, those variables are the least correlated with model residuals, which is indicated by the low value R^2 and the high tolerance value. All in all, from the statistical point of view, models without those variables would not be worse than the current model (see the result of t-test for model comparison). The decision about whether or not to leave that model or to construct a new one without the direct promotion costs and the sum of discounts made, belongs to the researcher. We will leave the current model.

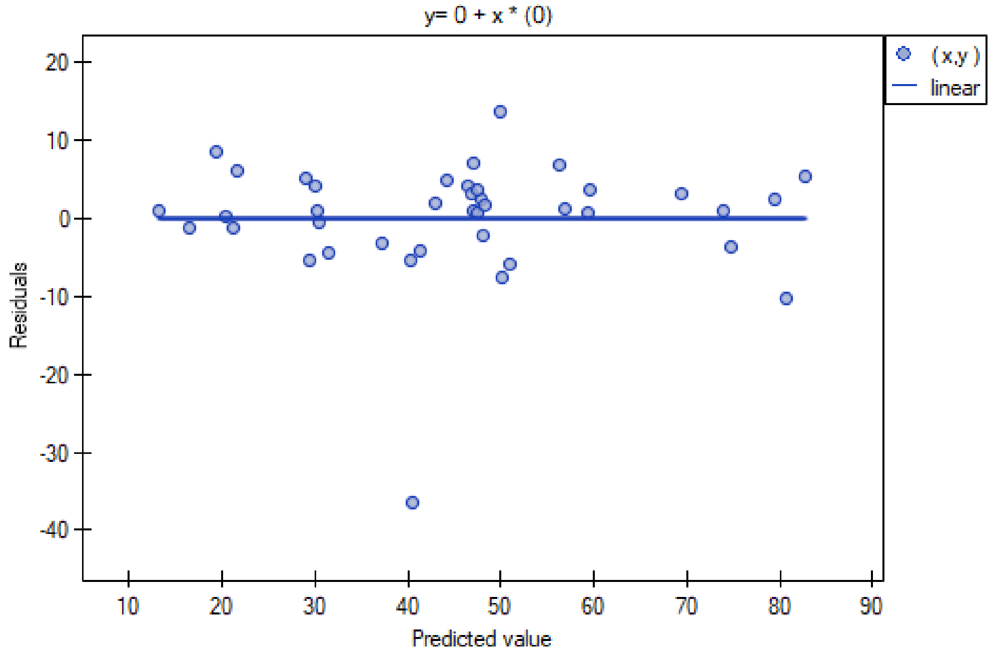
Finally, we will analyze the residuals. A part of that analysis is presented below:

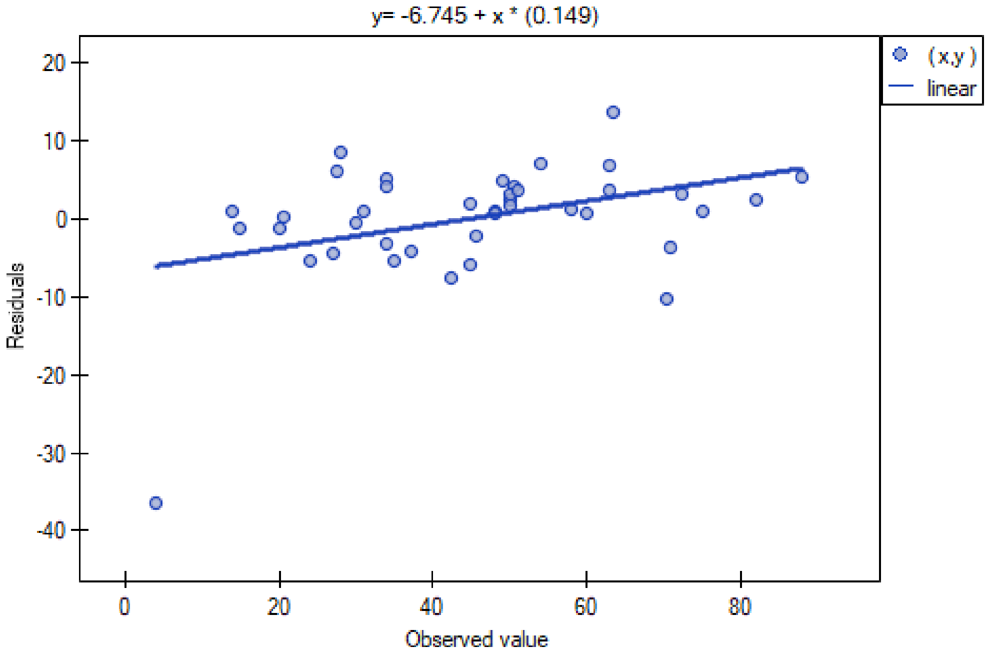


Residual analysis

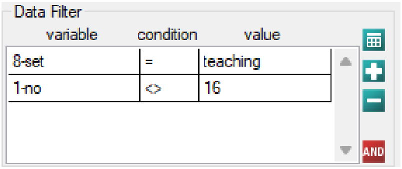
	predicted va	residual	standard res	<=-3sd	(-3sd;2sd]	(-2sd;sd]	(-sd;sd)	[sd;2sd)	[2sd;3sd)	>=3sd	Cook di	Mahalanobis
1	56.8783	1.1217	0.1387				*				0.0002	1.5764
2	56.1906	6.8094	0.8421				*				0.0201	4.0536
3	31.5321	-4.5321	-0.5605				*				0.0096	4.3613
4	40.3684	-5.3684	-0.6639				*				0.0126	4.0889
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

It is noticeable that one of the model residuals is an outlier – it deviates by more than 3 standard deviations from the mean value. It is observation number 16. The observation can be easily found by drawing a chart of residuals with respect to observed or expected values of the variable Y.





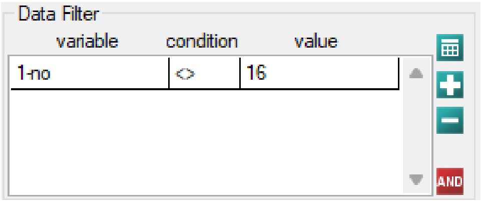
That outlier undermines the assumption concerning homoscedasticity. The assumption of homoscedasticity would be confirmed (that is, residuals variance presented on the axis Y would be similar when we move along the axis X), if we rejected that point. Additionally, the distribution of residuals deviates slightly from normal distribution (the value p of Liliefors test is $p = 0.0164$):


Data Filter

variable	condition	value
8-set	=	teaching
1-no	<>	16

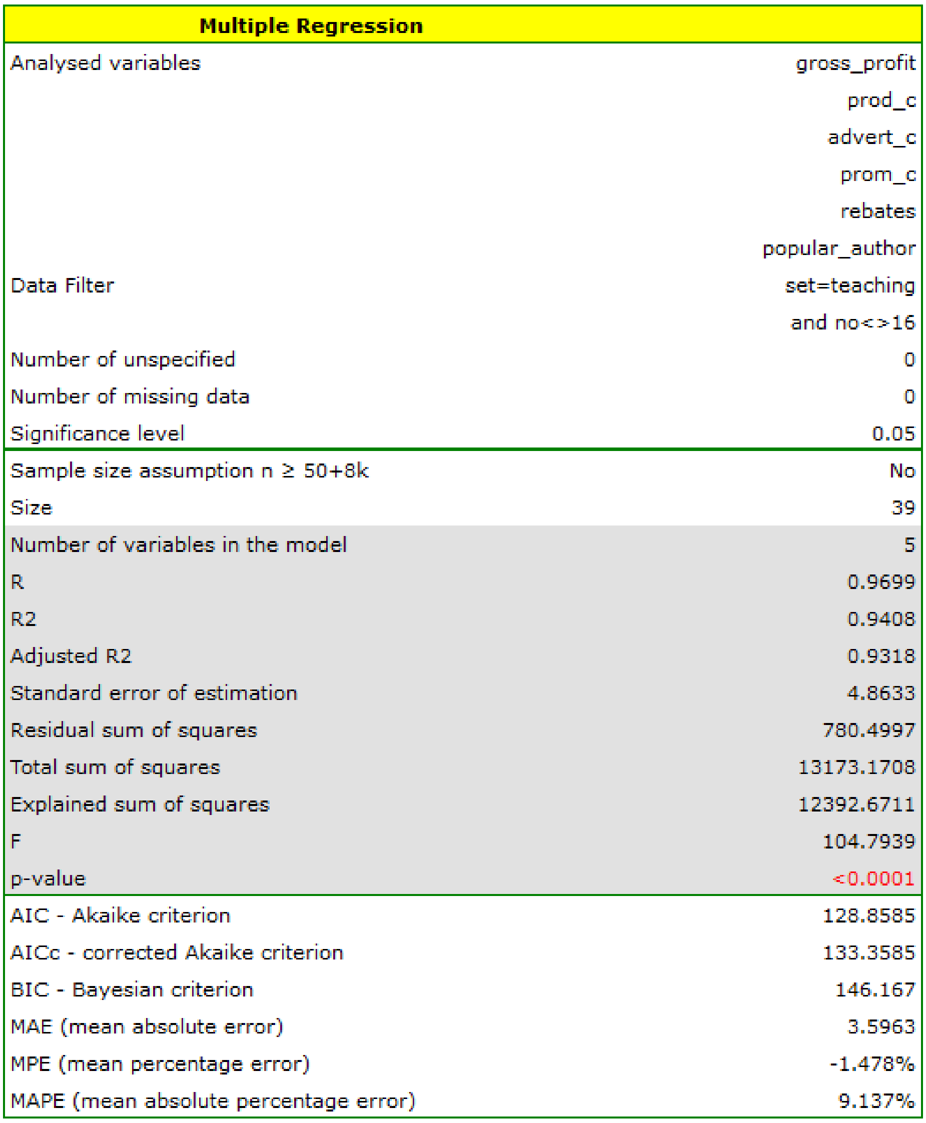
When we take a closer look of the outlier (position 16 in the data for the task) we see that the book is the only one for which the costs are higher than gross profit (gross profit=4 thousand dollars, the sum of costs = (8+6+0.33+1.6) = 15.93 thousand dollars).

The obtained model can be corrected by removing the outlier. For that purpose, another analysis has to be conducted, with a filter switched on which will exclude the outlier.

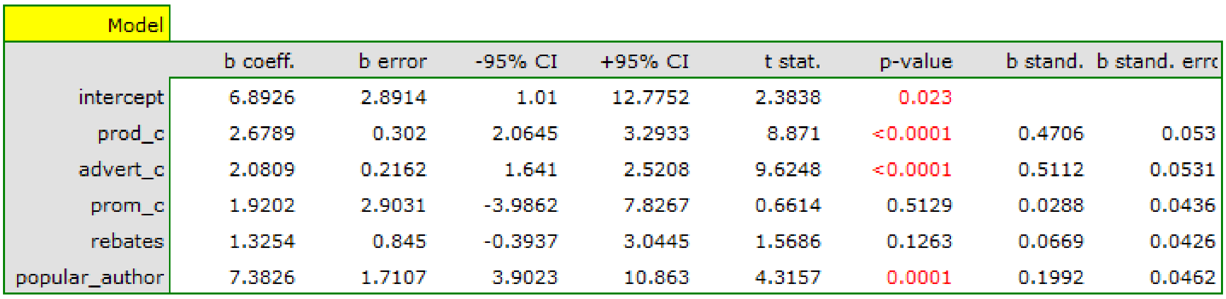

Data Filter

variable	condition	value
1-no	<>	16

As a result, we receive a model which is very similar to the previous one but is encumbered with a smaller error and is more adequate:



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%

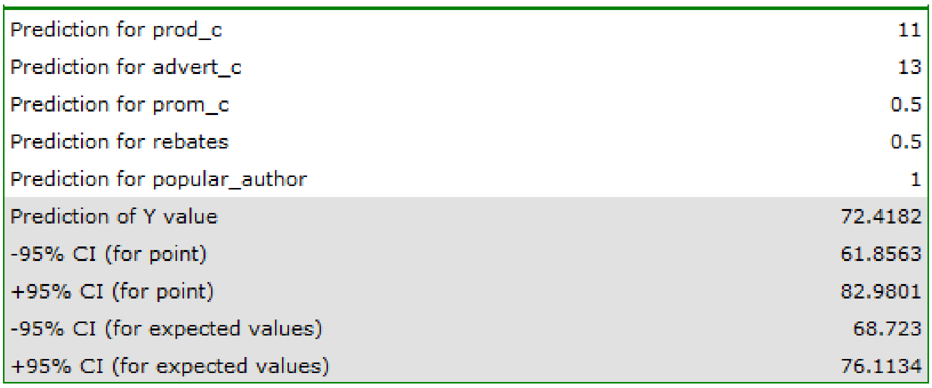


Model	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. errc
intercept	6.8926	2.8914	1.01	12.7752	2.3838	0.023		
prod_c	2.6789	0.302	2.0645	3.2933	8.871	<0.0001	0.4706	0.053
advert_c	2.0809	0.2162	1.641	2.5208	9.6248	<0.0001	0.5112	0.0531
prom_c	1.9202	2.9031	-3.9862	7.8267	0.6614	0.5129	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

$$profit_{gross} = 6.89{+}2.68(c_{prod}){+}2.08(c_{adv}){+}1.92(c_{prom}){+}1.33(discounts){+}7.38(popul_{author}){+}[4.86]$$

The final version of the model will be used for prediction. On the basis of the predicted costs amounting to:
production cost ≈ 11 thousand dollars,
advertising costs ≈ 13 thousand dollars,
direct promotion costs ≈ 0.5 thousand dollars,

the sum of discounts made ≈ 0.5 thousand dollars,
and the fact that the author is known (the 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.

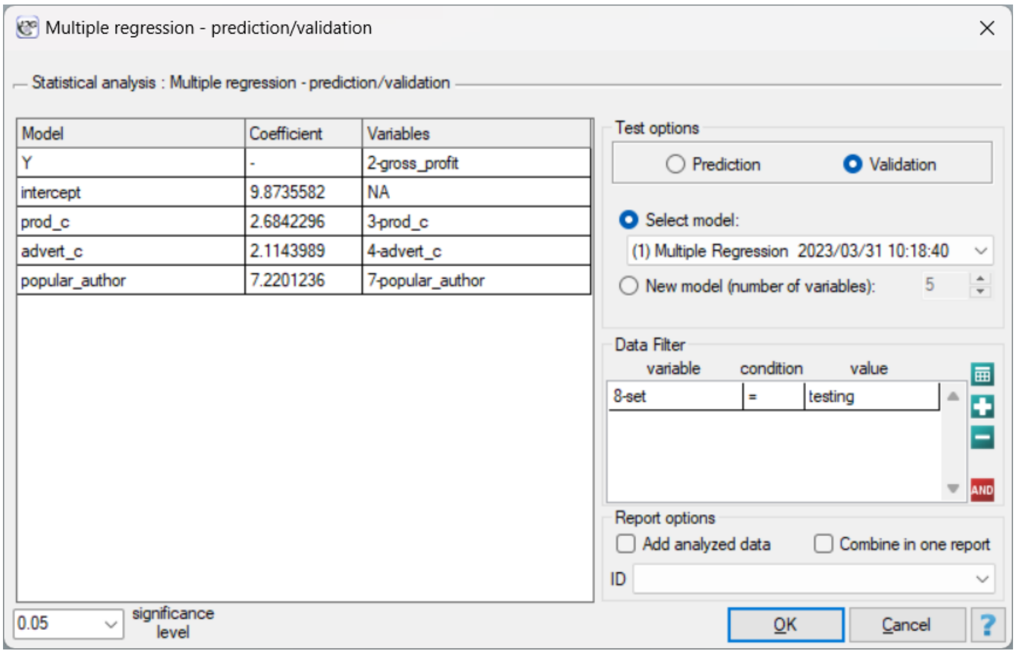
Finally, it should still be noted that this is only a preliminary model. In a proper study more data would have to be collected. The number of variables in the model is too small in relation to the number of books evaluated, i.e. n<50+8k.

24.2.5 Model-based prediction and test set validation

Validation

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 data set**), that is, it is returned in a report describing the resulting model. In order to be able to judge with greater certainty how suitable the model is for forecasting new data, an important part of the validation is to become a model to data that were not used in the model estimation. If the summary based on the treining data is satisfactory, i.e., the determined errors R^2 coefficients and information criteria are at a satisfactory level, and the summary based on the new data (the so-called **test data set**) is equally favorable, then with high probability it can be concluded that such a model is suitable for prediction. The testing 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.

The settings window with the validation can be opened in Advanced statistics→Multivariate models→Multiple regression - prediction/validation.



Multiple regression - prediction/validation

Statistical analysis : Multiple regression - prediction/validation

Model	Coefficient	Variables
Y	-	2-gross_profit
intercept	9.8735582	NA
prod_c	2.6842296	3-prod_c
advert_c	2.1143989	4-advert_c
popular_author	7.2201236	7-popular_author

Test options: ○ Prediction ● Validation

● Select model: (1) Multiple Regression 2023/03/31 10:18:40

○ New model (number of variables): 5

Data Filter

variable	condition	value
8-set	=	testing

Report options: □ Add analyzed data □ Combine in one report

ID

0.05 significance level

To perform validation, it is necessary to indicate the model on the basis of which we want to perform the validation. Validation can be done on the basis of:

- multivariate regression model built in PQStat - simply select a model from the models assigned to the 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;
- 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 enter the coefficients for each of them.

In the analysis window, indicate those new variables that should be used for validation.

Prediction

Most often, the final step in regression analysis is to use the built and previously validated model for prediction.

- Prediction for a single object can be performed along with the construction of the model, that is, in the analysis window Advanced statistics→Multivariate models→Multiple regression ,
- Prediction for a larger group of new data is done through the menu Advanced statistics→Multivariate models→Multiple regression - prediction/validation.

  To make a prediction, it is necessary to indicate the model on the basis of which we want to make the prediction. Prediction can be made on the basis of:

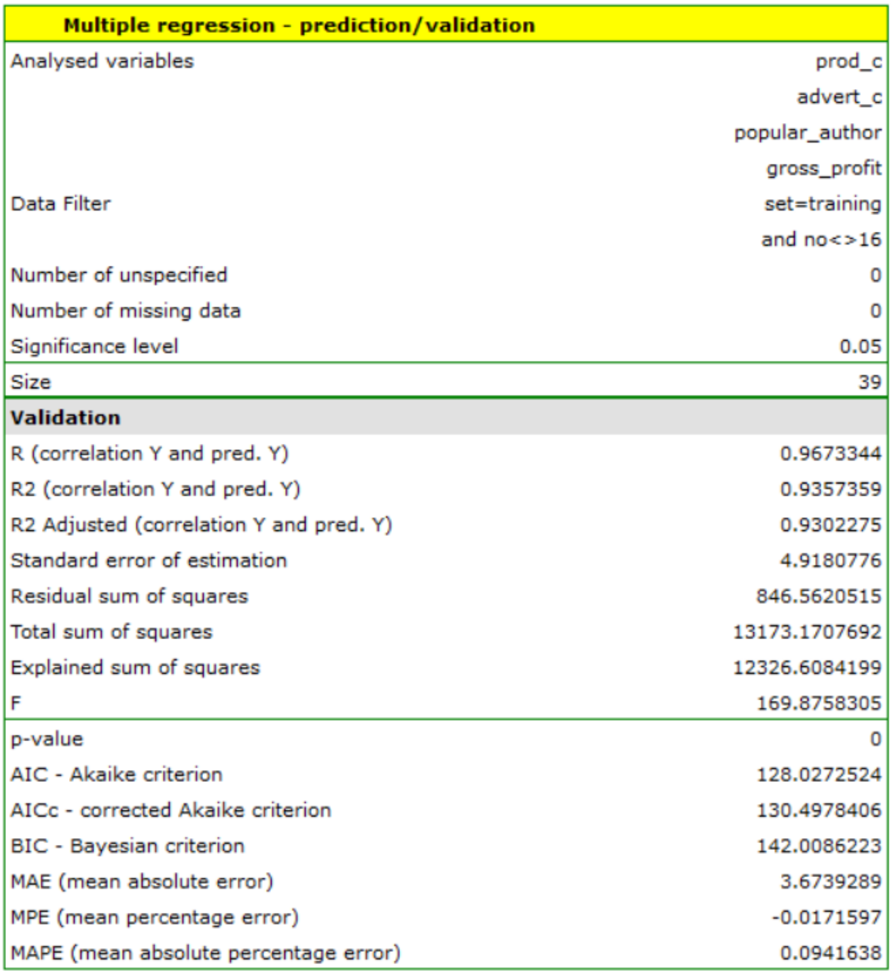
  - multivariate regression model built in PQStat -simply select a model from the models assigned to the sheet, and the number of variables and model coefficients will be set automatically; the test set should be in the same sheet as the training set;
  - model not built in PQStat but obtained from another source (e.g., described in a scientific paper we read) - in the analysis window, the number of variables and the coefficients on each of them should be entered.

  In the analysis window, indicate those new variables that should be used for prediction.

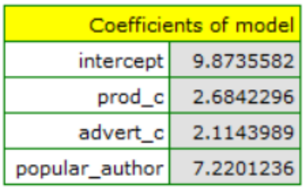
The estimated value is calculated with some error. Therefore, in addition, for the value predicted by the model, limits are set due to the error:

- confidence intervals are set for the expected value,
- For a single point, prediction intervals are determined.

Example 24.2 continued (publisher.pqs file)
To predict gross profit from book sales, the publisher built a regression model based on a training set stripped of item 16 (that is, 39 books). The model included: production costs, advertising costs and author popularity (1=popular author, 0=not). We will build the model once again based on the learning set and then, to make sure the model will work properly, we will validate it on a test data set. If the model passes this test, we will apply it to predictions for book items. To use the right collections we set a data filter each time.

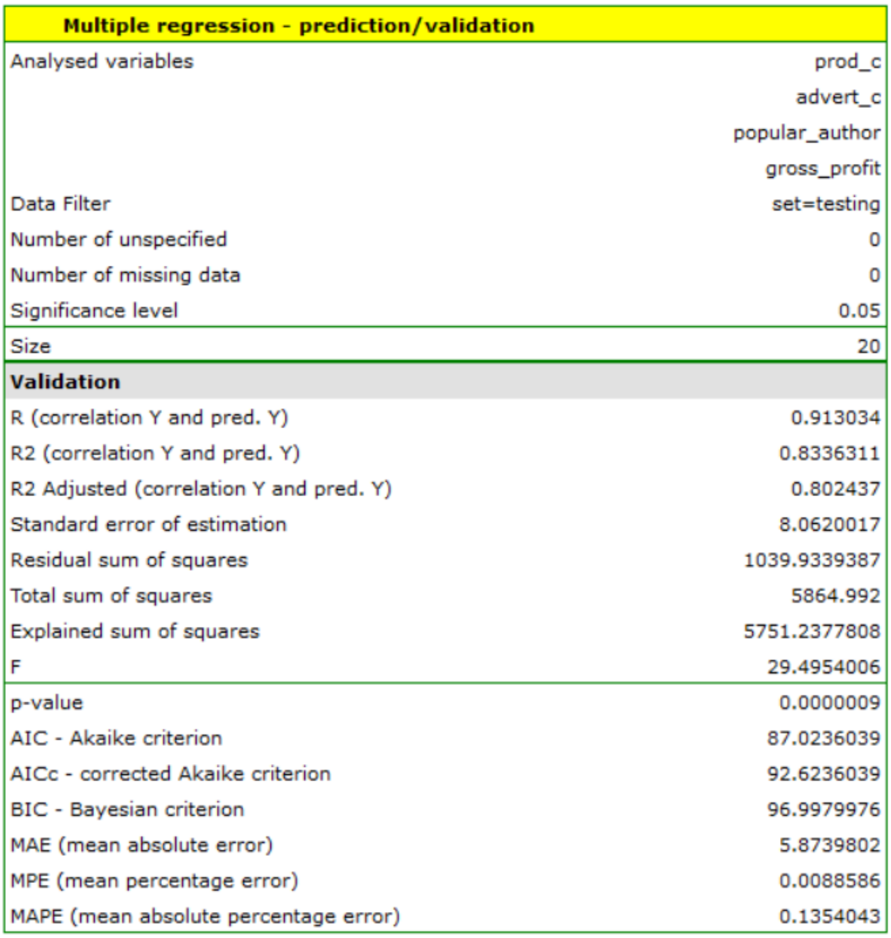


Multiple regression - prediction/validation	
Analysed variables	prod_c advert_c popular_author gross_profit
Data Filter	set=training and no<>16
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	39
Validation	
R (correlation Y and pred. Y)	0.9673344
R2 (correlation Y and pred. 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

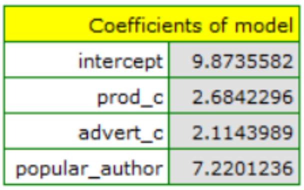


Coefficients of model	
intercept	9.8735582
prod_c	2.6842296
advert_c	2.1143989
popular_author	7.2201236

For the training set, the values describing the quality of the model's fit are very high: adjusted R^2 = 0.93 and the average forecast error (MAE) is 3.7 thousand dollars.

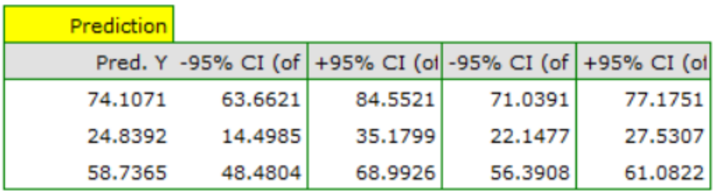


Multiple regression - prediction/validation	
Analysed variables	prod_c
	advert_c
	popular_author
	gross_profit
Data Filter	set=testing
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	20
Validation	
R (correlation Y and pred. Y)	0.913034
R2 (correlation Y and pred. Y)	0.8336311
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

For the test set, the values describing the quality of the model fit are slightly lower than for the learning set: Adjusted R^2 = 0.80 and the mean error of prediction (MAE) is 5.9 thousand dollars. Since the validation result on the test set is almost as good as on the training set, we will use the model for prediction. To do this, we will use the data of three new book items added to the end of the set. We'll select Prediction, set filter on the new dataset and use our model to predict the gross profit for these books.



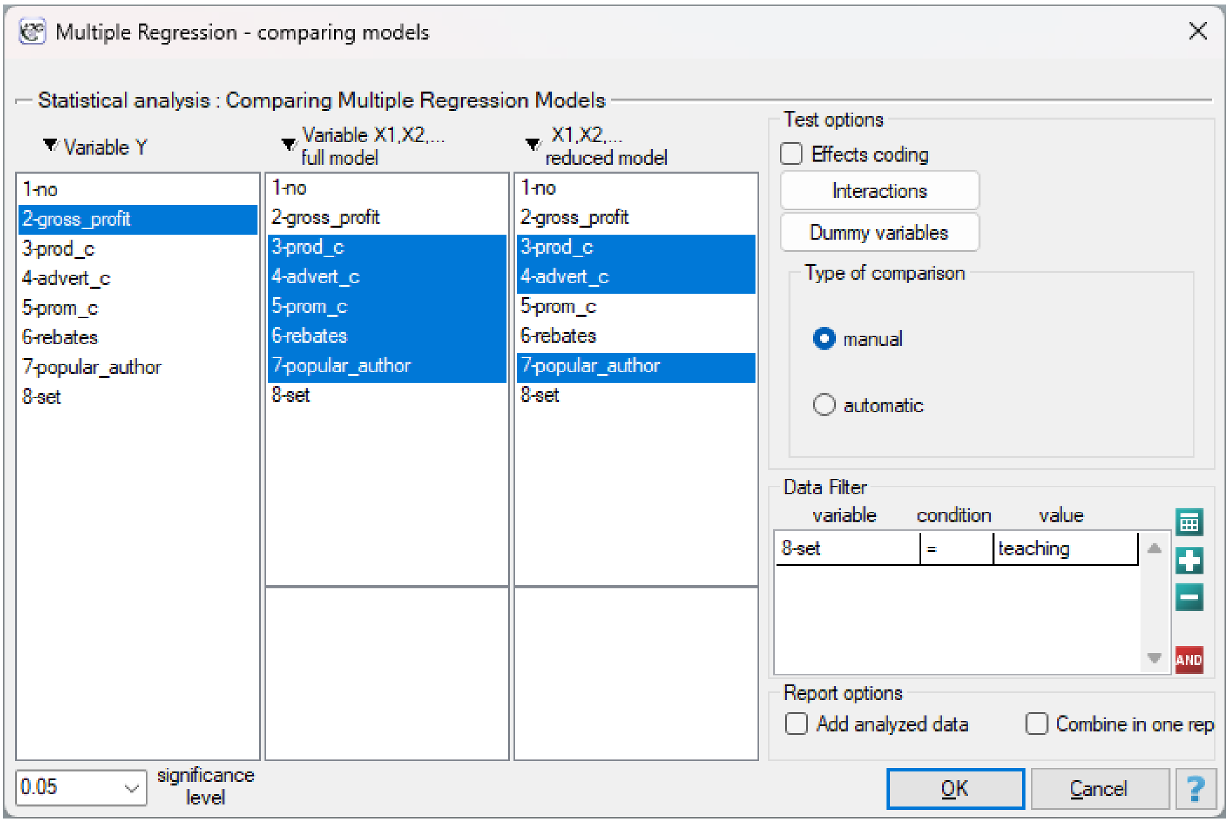
Prediction

Pred. Y	-95% CI (of	+95% CI (o	-95% CI (of	+95% CI (o
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 thousands of dollars) is projected for the first, most advertised and most expensive book published by a popular author.

24.3 COMPARISON OF MULTIPLE LINEAR REGRESSION MODELS

The window with settings for model comparison is accessed via the menu Advanced statistics → Multidimensional models→Multiple regression – model comparison



Multiple Regression - comparing models

Statistical analysis : Comparing Multiple Regression Models

Variable Y: 1-no, **2-gross_profit**, 3-prod_c, 4-advert_c, 5-prom_c, 6-rebates, 7-popular_author, 8-set

Variable X1,X2,... full model: 1-no, 2-gross_profit, **3-prod_c**, **4-advert_c**, **5-prom_c**, **6-rebates**, **7-popular_author**, 8-set

X1,X2,... reduced model: 1-no, 2-gross_profit, **3-prod_c**, **4-advert_c**, 5-prom_c, 6-rebates, **7-popular_author**, 8-set

Test options

□ Effects coding

Interactions

Dummy variables

Type of comparison

● manual

○ automatic

Data Filter

variable	condition	value
8-set	=	teaching

Report options

□ Add analyzed data □ Combine in one rep

0.05 significance level

OK Cancel ?

The multiple linear regression offers the possibility of simultaneous analysis of many independent variables. There appears, then, the problem of choosing the optimum model. Too large a model involves a plethora of information in which the important ones may get lost. Too small a model involves the risk of omitting those features which could describe the studied phenomenon in a reliable manner. Because it is not the number of variables in the model but their quality that determines the quality of the model. To make a proper selection of independent variables it is necessary to have knowledge and experience connected with the studied phenomenon. One has to remember to put into the model variables strongly correlated with the dependent variable and weakly correlated with one another.

There is no single, simple statistical rule which would decide about the number of variables necessary in the model. The measures of model adequacy most frequently used in a comparison are: R^2_{adj} – the corrected value of multiple determination coefficient (the higher the value the more adequate the model), SE_e – the standard error of estimation (the lower the value the more adequate the model) or or information criteria AIC, AICc, BIC (the lower the value, the better the model). For that purpose, the F-test based on the multiple determination coefficient R^2 can also be used. The test is used to verify the hypothesis that the adequacy of both compared models is equally good.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & R^2_{FM}=R^2_{RM},\\ \mathcal{H}_1: & R^2_{FM}\neq R^2_{RM},\end{array}$$

where:
R^2_{FM}, R^2_{RM} – multiple determination coefficients in compared models (full and reduced).

The test statistics has the form presented below:

$$F=\frac{R^2_{FM}-R^2_{RM}}{k_{FM}-k_{RM}}\cdot\frac{n-k_{FM}-1}{1-R^2_{FM}},$$

The statistics is subject to F-Snedecor distribution with $df_1=k_{FM}-k_{RM}$ and $df_2=n-k_{FM}-1$ degrees of freedom.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

If the compared models do not differ significantly, we should select the one with a smaller number of variables. Because a lack of a difference means that the variables present in the full model but absent from the reduced model do not carry significant information. However, if the difference in the quality of model adequacy is statistically significant, it means that one of them (the one with the greater number of variables, with a greater R^2 or lesser value of the information criterion) is significantly better than the other one.

In the program PQStat the comparison of models can be done manually or automatically.

- **Manual** model comparison – construction of 2 models:
  - a full model – a model with a greater number of variables,
  - a reduced model – a model with a smaller number of variables – such a model is created from the full model by removing those variables which are superfluous from the perspective of studying a given phenomenon.

  The choice of independent variables in the compared models and, subsequently, the choice of a better model on the basis of the results of the comparison, is made by the researcher.

- **Automatic** model comparison is done in several steps:

  step 1 Constructing the model with the use of all variables.

  step 2 Removing one variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.

  step 3 A comparison of the full and the reduced model.

  step 4 Removing another variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.

  step 5 A comparison of the previous and the newly reduced model.

  ...

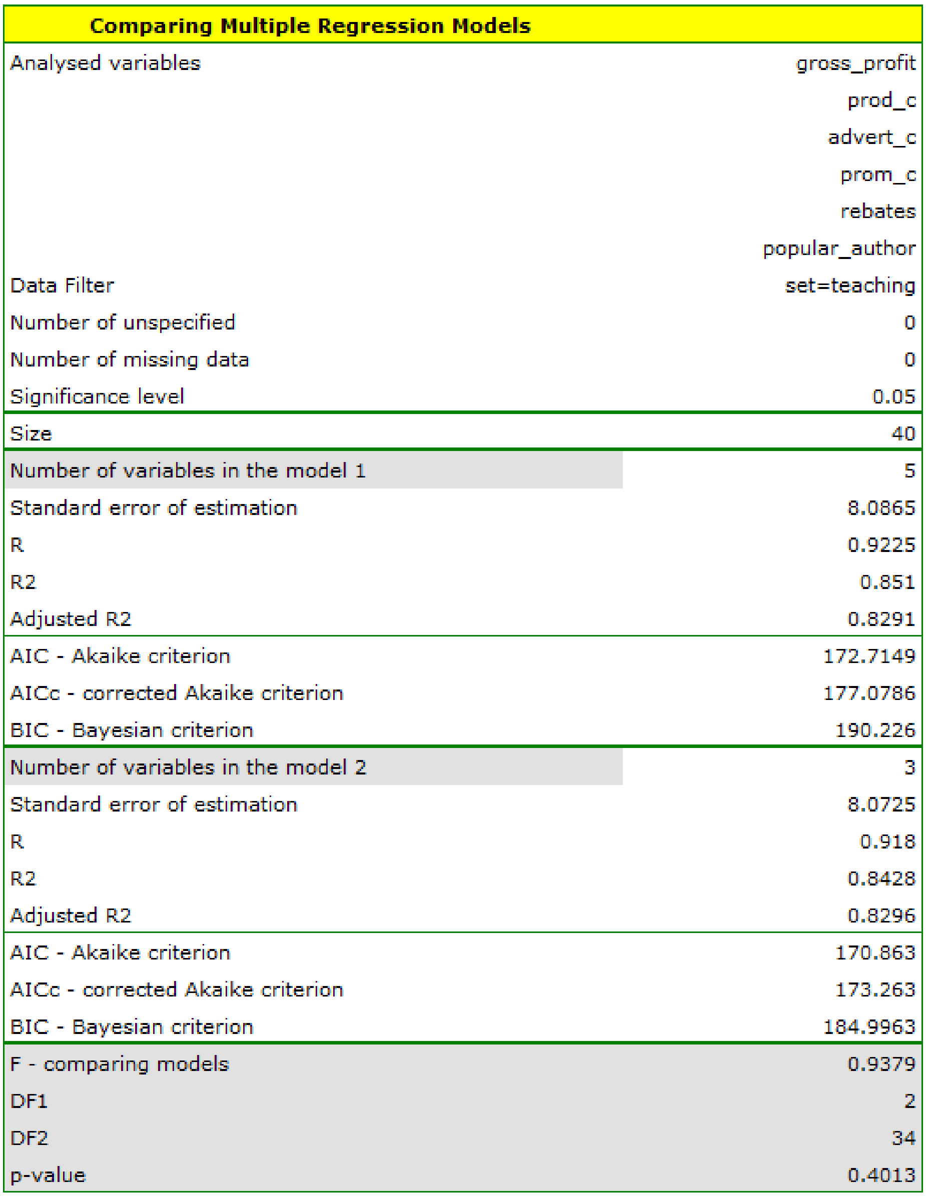
  In that way numerous, ever smaller models are created. The last model only contains 1 independent variable.

  As a result, each model is described with the help of adequacy measures (R^2_{adj}, SE_e, AIC, AICc, BIC), and the subsequent (neighboring) models are compared by means of the F-test. The model which is finally marked as statistically best is the model with the greatest R^2_{adj} and the smallest SE_e. However, as none of the statistical methods cannot give a full answer to the question which of the models is the best, it is the researcher who should choose the model on the basis of the results.

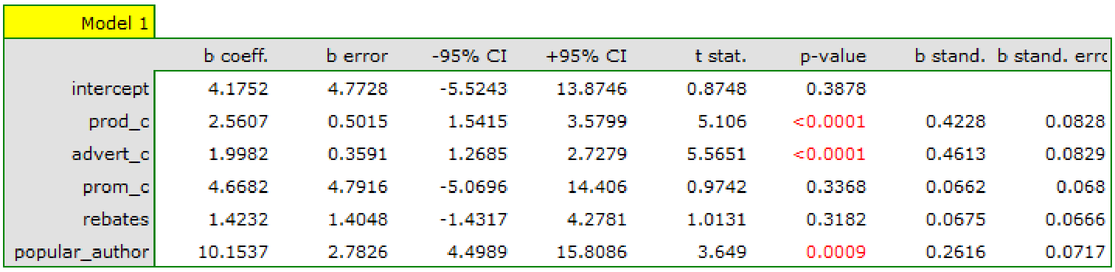
Example (24.2) continued *(publisher.pqs file)*

To predict the gross profit from book sales a publisher wants to consider such variables as: production cost, advertising costs, direct promotion cost, the sum of discounts made, and the author's popularity. However, not all of those variables need to have a significant effect on profit. Let us try to select such a model of linear regression which will contain the optimum number of variables (from the perspective of statistics). For this analysis, we will use teaching set data.

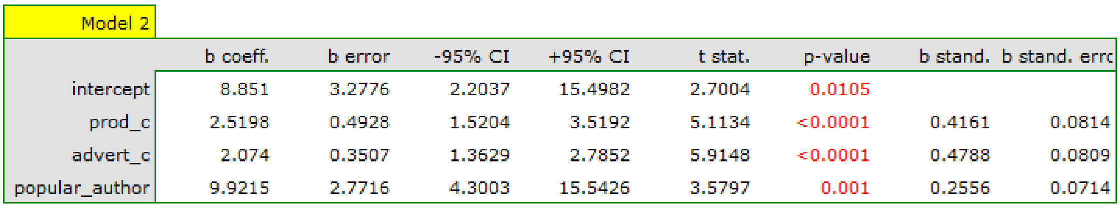
- **Manual** model comparison.
  On the basis of the earlier constructed, full model we can suspect that the variables: direct promotion costs and the sum of discounts made have a small influence on the constructed model (i.e. those variables do not help predict the greatness of the profit). We will check if, from the perspective of statistics, the full model is better than the model from which the two variables have been removed.



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



Model 1	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. errc
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. errc
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

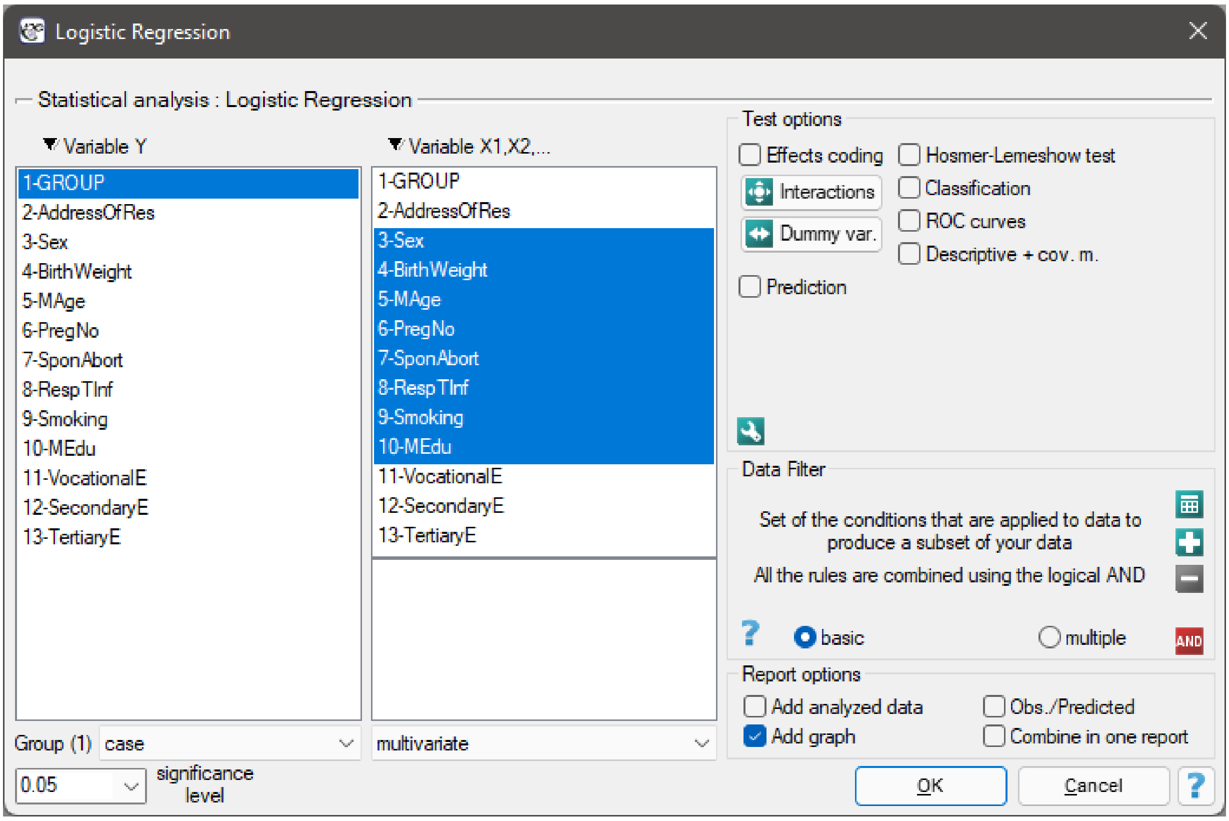
It turns out that there is no basis for thinking that the full model is better than the reduced model (the value p of F-test which is used for comparing models is $p = 0.4013$). Additionally, the reduced model is slightly more adequate than the full model (for the reduced model $R^2_{adj} = 0.8296$, for the full model $R^2_{adj} = 0.8291$ and has smaller, or more favorable, values of the information criteria AIC, AIcc, BIC.

- **Automatic** model comparison.
  In the case of automatic model comparison we receive very similar results. The best model is the one with the greatest coefficient R^2_{adj}, the smallest information criteria and the smalles standard estimation error SE_e. The best model we suggest is the model containing only 3 independent variables: the production cost, advertising costs, and the author's popularity.

On the basis of the analyses above, from the perspective of statistics, the optimum model is the model with the 3 most important independent variables: the production cost, advertising costs, and the author's popularity. However, the final decision which model to choose should be made by a person with specialist knowledge about the studied topic – in this case, the publisher. It ought to be remembered that the selected model should be constructed anew and its assumptions verified in the window Multiple regression.

24.4 LOGISTIC REGRESSION

The window with settings for Logistic Regression is accessed via the menu Advanced statistics→Multidimensional Models→Logistic Regression



Logistic Regression

Statistical analysis : Logistic Regression

Variable Y: 1-GROUP (selected), 2-AddressOfRes, 3-Sex, 4-BirthWeight, 5-MAge, 6-PregNo, 7-SponAbort, 8-RespTInf, 9-Smoking, 10-MEdu, 11-VocationalE, 12-SecondaryE, 13-TertiaryE

Variable X1,X2,...: 1-GROUP, 2-AddressOfRes, 3-Sex (selected), 4-BirthWeight (selected), 5-MAge (selected), 6-PregNo (selected), 7-SponAbort (selected), 8-RespTInf (selected), 9-Smoking (selected), 10-MEdu (selected), 11-VocationalE, 12-SecondaryE, 13-TertiaryE

Group (1) case

multivariate

0.05 significance level

Test options

□ Effects coding
□ Hosmer-Lemeshow test
□ Classification
□ ROC curves
□ Descriptive + cov. m.
□ Prediction

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data
□ Obs./Predicted
☑ Add graph
□ Combine in one report

The constructed model of logistic regression (similarly to the case of 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 assumes two values, e.g. ill/healthy, insolvent/solvent etc.

The two values are coded as (1)/(0), where:
(1) –the distinguished value –possessing the feature
(0) –not possessing the feature.

The function on which the model of logistic regression is based does not calculate the 2-level variable Y but the probability of that variable assuming 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 assuming the distinguished value (1) on condition that specific values of independent variables are achieved, the so-called probability predicted for 1.
> Z is most often expressed in the form of a linear relationship:

$$Z = \beta_0 + \sum_{i=1}^{k}\beta_i X_i,$$

$X_1, X_2, \dots X_k$ –independent variables, explanatory,
$\beta_0, \beta_1, \beta_2, \dots \beta_k$ –parameters.

Dummy variables and interactions in the model
A discussion of the coding of dummy variables and interactions is presented in chapter 24.1 Preparation of the variables for the analysis in multidimensional models.

Note
Function Z can also be described with the use of a higher order relationship, e.g. a square relationship - in such a case we introduce into the model a variable containing the square of the independent variable X_i^2.

The logit is the transformation of that model into the form:

$$\ln\left(\frac{P}{1-P}\right) = Z.$$

The matrices involved in the equation, for a sample of size n, are recorded in the following manner:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

In such a case, the solution of the equation is the vector of the estimates of parameters $\beta_0, \beta_1, \dots, \beta_k$ called **regression coefficients**:

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}.$$

The coefficients are estimated with the use of the **maximum likelihood method**, that is through the search for the maximum value of likelihood function L (in the program the Newton-Raphson iterative algorithm was used). On the basis of those values we can infer the magnitude of the effect of the independent variable (for which the coefficient was estimated) on the dependent variable.

There is a certain error of estimation for each coefficient. The magnitude of that error is estimated from the following formula:

$$SE_b = \sqrt{diag(H^{-1})_b},$$

where:

$diag(H^{-1})$ is the main diagonal of the covariance matrix.

Note
When building the model you need remember that the number of observations should be ten times greater than or equal to the number of the estimated parameters of the model ($n \geq 10(k+1)$). However, a more restrictive criterion proposed by P. Peduzzi et al. in 1996[130] is increasingly used, stating that the number of observations should be ten times or equal to the ratio of the number of independent variables (v) and the smaller of the proportion of counts (p)described from the dependent variable (i.e., proportions of sick or healthy), i.e. ($n \geq 10v/p$).

Note
When building the model you need remember that the independent variables should not be collinear. In a case of collinearity estimation can be uncertain and the obtained error values very high. The collinear variables should be removed from the model or one independent variable should be built of them, e.g. instead of the collinear variables of mother age and father age one can build the parents age variable.

Note
The criterion of convergence of the function of the Newton-Raphson iterative algorithm can be controlled with the help of two parameters: the limit of convergence iteration (it gives the maximum number of iterations in which the algorithm should reach convergence) and the convergence criterion (it gives the value below which the received improvement of estimation shall be considered to be insignificant and the algorithm will stop).

24.4.1 The Odds Ratio

Individual Odds Ratio

On the basis of many coefficients, for each independent variable in the model an easily interpreted measure is estimated, i.e. the individual Odds Ratio:

$$OR_i = e^{\beta_i}.$$

The received Odds Ratio expresses the change of the odds for the occurrence of the distinguished value (1) when the independent variable grows by 1 unit. The result is corrected with the remaining independent variables in the model so that it is assumed that they remain at a stable level while the studied variable is growing by 1 unit.

The OR value is interpreted as follows:

- $OR > 1$ means the stimulating influence of the studied independent variable on obtaining the distinguished value (1), i.e. it gives information about how much greater are the odds of the occurrence of the distinguished value (1) when the independent variable grows by 1 unit.
- $OR < 1$ means the destimulating influence of the studied independent variable on obtaining the distinguished value (1), i.e. it gives information about how much lower are the odds of the occurrence of the distinguished value (1) when the independent variable grows by 1 unit.
- $OR \approx 1$ means that the studied independent variable has no influence on obtaining the distinguished value (1).

Odds Ratio - the general formula

The PQStat program calculates the individual Odds Ratio. Its modification on the basis of a general formula makes it possible to change the interpretation of the obtained result.

The Odds Ratio for the occurrence of the distinguished state in a general case is calculated as the ratio of two odds. Therefore for the independent variable X_1 for Z expressed with a linear relationship we calculate:
the odds for the first category:

$$Odds(1) = \frac{P(1)}{1-P(1)} = e^Z(1) = e^{\beta_0+\beta_1 X_1(1)+\beta_2 X_2+\ldots+\beta_k X_k},$$

the odds for the second category:

$$Odds(2) = \frac{P(2)}{1-P(2)} = e^Z(2) = e^{\beta_0+\beta_1 X_1(2)+\beta_2 X_2+\ldots+\beta_k X_k}.$$

The Odds Ratio for variable X_1 is then expressed with the formula:

$$\begin{aligned} OR_1(2)/(1) &= \frac{Odds(2)}{Odds(1)} = \frac{e^{\beta_0+\beta_1X_1(2)+\beta_2X_2+...+\beta_kX_k}}{e^{\beta_0+\beta_1X_1(1)+\beta_2X_2+...+\beta_kX_k}} \\ &= e^{\beta_0+\beta_1X_1(2)+\beta_2X_2+...+\beta_kX_k-[\beta_0+\beta_1X_1(1)+\beta_2X_2+...+\beta_kX_k]} \\ &= e^{\beta_1X_1(2)-\beta_1X_1(1)} = e^{\beta_1[X_1(2)-X_1(1)]} = \\ &= \left(e^{\beta_1}\right)^{[X_1(2)-X_1(1)]}. \end{aligned}$$

Example

If the independent variable is age expressed in years, then the difference between neighboring age categories such as 25 and 26 years is 1 year $(X_1(2) - X_1(1) = 26 - 25 = 1)$. In such a case we will obtain the individual Odds Ratio:

$$OR = \left(e^{\beta_1}\right)^1,$$

which expresses the degree of change of the odds for the occurrence of the distinguished value if the age is changed by 1 year.

The odds ratio calculated for non-neighboring variable categories, such as 25 and 30 years, will be a five-year Odds Ratio, because the difference $X_1(2) - X_1(1) = 30 - 25 = 5$. In such a case we will obtain the five-year Odds Ratio:

$$OR = \left(e^{\beta_1}\right)^5,$$

which expresses the degree of change of the odds for the occurrence of the distinguished value if the age is changed by 5 years.

Note

If the analysis is made for a non-linear model or if interaction is taken into account, then, on the basis of a general formula, we can calculate an appropriate Odds Ratio by changing the formula which expresses Z.

24.4.2 Model verification

Statistical significance of particular variables in the model (significance of the Odds Ratio)

On the basis of the coefficient and its error of estimation we can infer if the independent variable for which the coefficient was estimated has a significant effect on the dependent variable. For that purpose we use Wald test.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \beta_i = 0, \\ \mathcal{H}_1: & \beta_i \neq 0. \end{array} \quad \text{or, equivalently:} \quad \begin{array}{cl} \mathcal{H}_0: & OR_i = 1, \\ \mathcal{H}_1: & OR_i \neq 1. \end{array}$$

The Wald test statistics is calculated according to the formula:

$$\chi^2 = \left(\frac{b_i}{SE_{b_i}}\right)^2$$

The statistic asymptotically (for large sizes) has the χ^2 distribution with 1 degree of freedom.
On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{array}{lll}\text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

The quality of the constructed model

A good model should fulfill two basic conditions: it should fit well and be possibly simple. The quality of multiple linear regression can be evaluated can be evaluated with a few general measures based on: L_{FM} –the maximum value of likelihood function of a full model (with all variables), L_0 –the maximum value of the likelihood function of a model which only contains one free word, n –the sample size.

- **Information criteria** are based on the information entropy carried by the model (model insecurity), i.e. they evaluate the lost information when a given model is used to describe the studied phenomenon. We should, then, choose the model with the minimum value of a given information criterion.
  AIC, $AICc$, and BIC is a kind of a compromise between the good fit and complexity. The second element of the sum in formulas for information criteria (the so-called penalty function) measures the simplicity of the model. That depends on the number of variables (k) in the model and the sample size (n). In both cases the element grows with the increase of the number of variables and the growth is the faster the smaller the number of observations.
  The information criterion, however, is not an absolute measure, i.e. if all the compared models do not describe reality well, there is no use looking for a warning in the information criterion.
  - Akaike information criterion

    $$AIC = -2\ln L_{FM} + 2k,$$

    It is an asymptomatic criterion, appropriate for large sample sizes.
  - Corrected Akaike information criterion

    $$AICc = AIC + \frac{2k(k+1)}{n-k-1},$$

    Because the correction of the Akaike information criterion concerns the sample size it is the recommended measure (also for smaller sizes).
  - Bayesian information criterion or Schwarz criterion

    $$BIC = -2\ln L_{FM} + k\ln(n),$$

    Just like the corrected Akaike criterion it takes into account the sample size.
- **Pseudo R^2** –the so-called McFadden R^2 is a goodness of fit measure of the model (an equivalent of the coefficient of multiple determination R^2 defined for multiple linear regression). The value of that coefficient falls within the range of $<0;1)$, where values close to 1 mean excellent goodness of fit of a model, 0 –a complete lack of fit Coefficient R^2_{Pseudo} is calculated according to the formula:

  $$R^2_{Pseudo} = 1 - \frac{\ln L_{FM}}{\ln L_0}.$$

  As coefficient R^2_{Pseudo} never assumes value 1 and is sensitive to the amount of variables in the model, its corrected value is calculated:

  $$R^2_{Nagelkerke} = \frac{1-e^{-(2/n)(\ln L_{FM}-\ln L_0)}}{1-e^{(2/n)\ln L_0}} \quad \text{lub} \quad R^2_{Cox-Snell} = 1 - e^{\frac{(-2\ln L_0)-(-2\ln L_{FM})}{n}}.$$

- **Statistical significance of all variables in the model**
  The basic tool for the evaluation of the significance of all variables in the model is **the Likelihood Ratio test**. The test verifies the hypothesis:

$$\begin{array}{cl}\mathcal{H}_0: & \text{all } \beta_i = 0,\\ \mathcal{H}_1: & \text{there is } \beta_i \neq 0.\end{array}$$

  The test statistic has the form presented below:

$$\chi^2 = -2\ln(L_0/L_{FM}) = -2\ln(L_0) - (-2\ln(L_{FM})).$$

  The statistic asymptotically (for large sizes) has the χ^2 distribution with k degrees of freedom.
  On the basis of test statistics, p value is estimated and then compared with α :

$$\begin{array}{ccl}\text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

- **Hosmer-Lemeshow test** –The test compares, for various subgroups of data, the observed rates of occurrence of the distinguished value O_g and the predicted probability E_g. If O_g and E_g are close enough then one can assume that an adequate model has been built.

  For the calculation the observations are first divided into G subgroups –usually deciles ($G = 10$).

  Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & O_g = E_g \text{ for all categories,}\\ \mathcal{H}_1: & O_g \neq E_g \text{ for at least one category.}\end{array}$$

  The test statistic has the form presented below:

$$H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - \frac{E_g}{N_g})},$$

  where:

  N_g –the number of observations in group g.

  The statistic asymptotically (for large sizes) has the χ^2 distribution with $G-2$ degrees of freedom.
  On the basis of test statistics, p value is estimated and then compared with α :

$$\begin{array}{ccl}\text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

- **AUC - the area under the ROC curve** –The ROC curve built on th ebasis of the value of the dependent variable, and the predicted probability of dependent variable P, allows to evaluate the ability of the constructed logistic regression model to classify the cases into two groups: (1) and (0). The constructed curve, especially the area under the curve, presents the classification quality of the model. When the ROC curve overlaps with the diagonal $y = x$, then the decision about classifying a case within a given class (1) or (0), made on the basis of the model, is as good as a random division of the studied cases into the groups. The classification quality of a model is good when the curve is much above the diagonal $y = x$, that is when the area under the ROC curve is much larger than the area under the $y = x$ line, i.e. it is greater than 0.5

  Hypotheses:

$$\begin{array}{cc}\mathcal{H}_0: & AUC = 0.5,\\ \mathcal{H}_1: & AUC \neq 0.5.\end{array}$$

The test statistic has the form presented below:

$$Z = \frac{AUC - 0.5}{SE_{0.5}},$$

where:

$SE_{0.5}$ –area error.

Statistics Z asymptotically (for large sizes) has the normal distribution.
On the basis of test statistics, p value is estimated and then compared with the significance level α:

$$\begin{array}{ccl}\text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

Additionally, for ROC curve the suggested value of the **cut-off point** of the predicted probability is given, together with the table of sensitivity and specificity for each possible cut-off point.

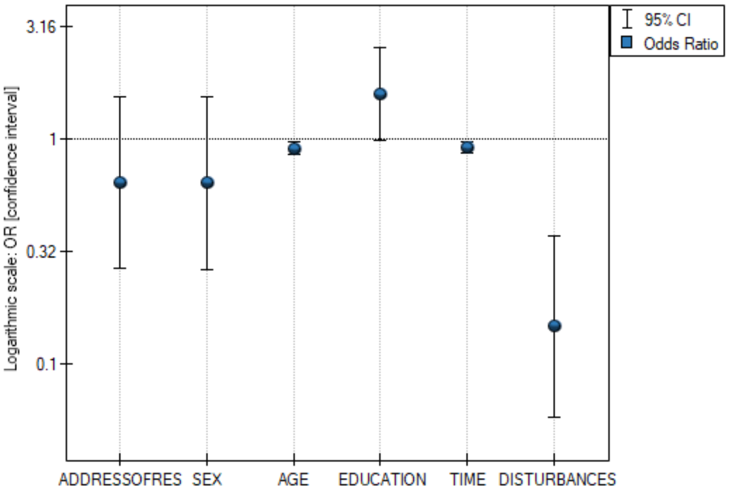
Note!
More possibilities of calculating a cut-off point are offered by module **ROC curve**. The analysis is made on the basis of observed values and predicted probability obtained in the analysis of Logistic Regression.

- **Classification**
  On the basis of the selected cut-off point of predicted probability we can change the classification quality. By default the cut-off point has the value of 0.5. The user can change the value into any value from the range of (0.1), e.g. the value suggested by the ROC curve.

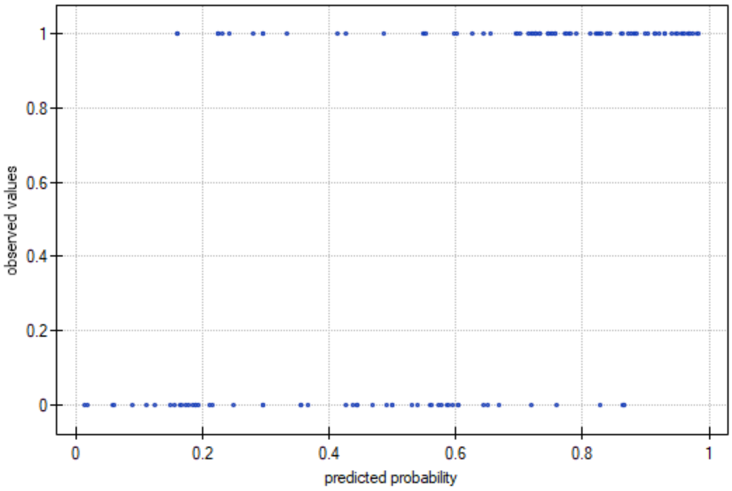
  As a result we shall obtain the classification table and the percentage of properly classified cases, the percentage of properly classified (0) –specificity, and the percentage of properly classified (1) –sensitivity.

Graphs in logistic regression

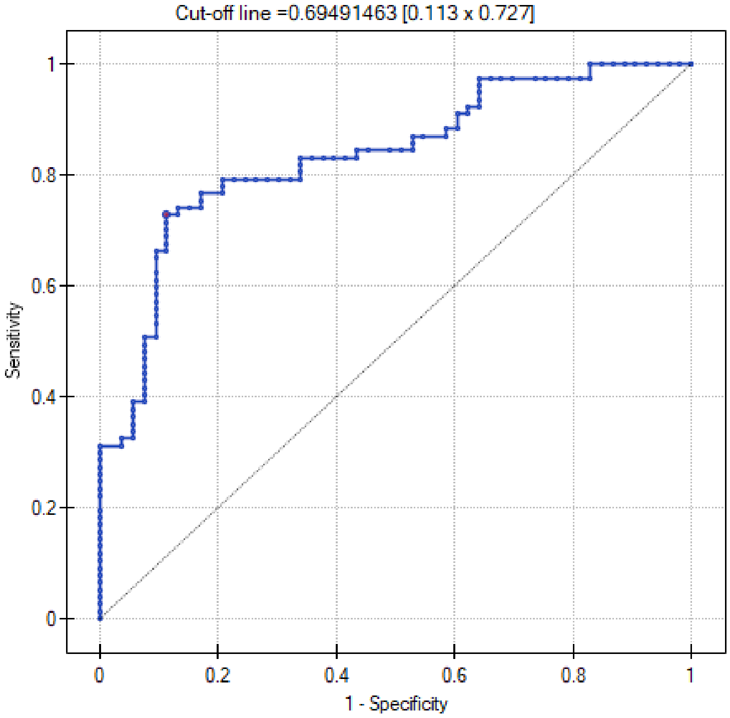
- **Odds Ratio $\pm$ confidence interval** – is a graph showing the OR along with the 95 percent confidence interval for the score of each variable returned in the constructed model. For categorical variables, the line at level 1 indicates the odds ratio value for the reference category.



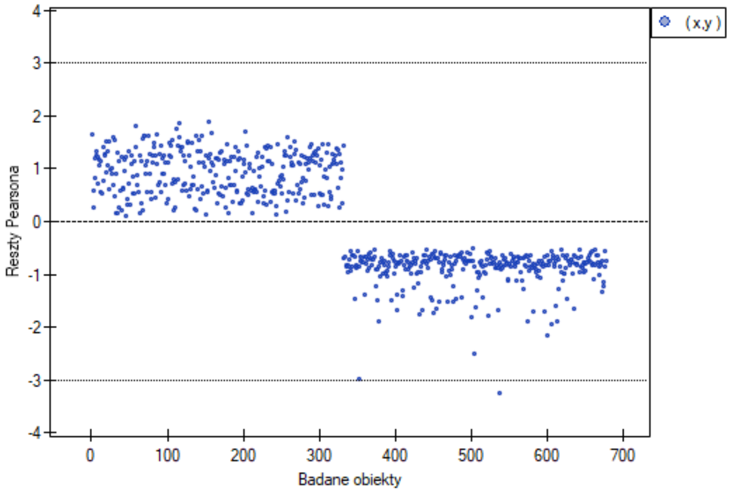
- **Observed Values / Expected Probability** – is a graph showing the results of each person's predicted probability of an event occurring (X-axis) and the true value, which is the occurrence of the event (value 1 on the Y-axis) or the absence of the event (value 0 on the Y-axis). If the model predicts very well, points will accumulate at the bottom near the left side of the graph and at the top near the right side of the graph.



- **ROC curve** – is a graph constructed based on the value of the dependent variable and the predicted probability of an event.



- **Pearson residuals plot** – is a graph that allows you to assess whether there are outliers in the data. 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 unified by determining Pearson residuals. The Pearson residual is 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 to the model, and the magnitude indicates the degree of deviation. Person's residuals less than or greater than 3 suggest that the variance of a given object is too largeu.

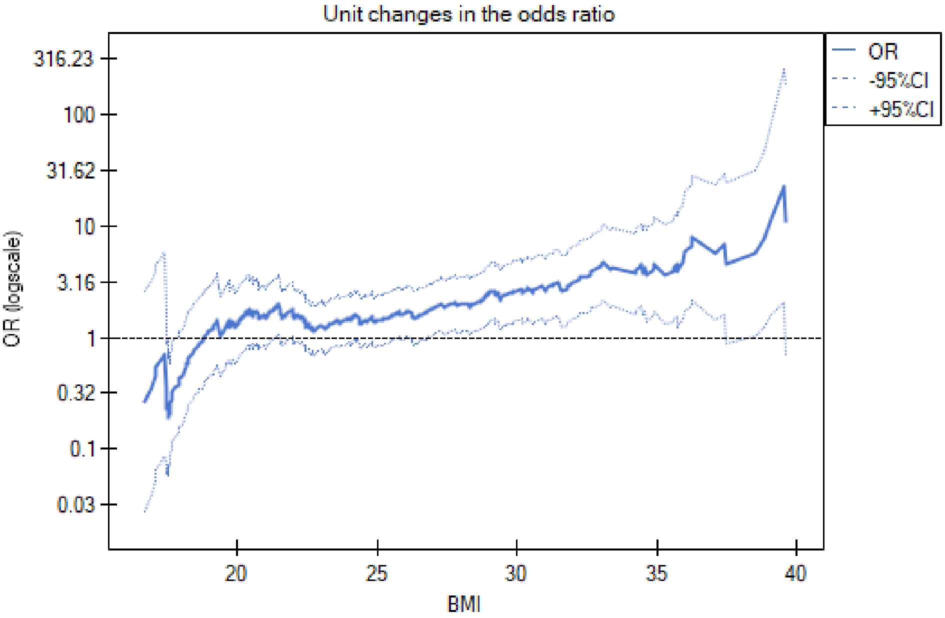


- **Unit changes in the odds ratio** – is a graph showing a series of odds ratios, along with a confidence interval, determined for each possible cut-off point of a variable placed on the X axis. It allows the user to select one good cut-off point and then build from that a new bivariate variable at which a high or low odds ratio is achieved, respectively. The chart is dedicated to the evaluation of continuous variables in univariate analysis, i.e. when only one independent variable is selected.
- **odds ratio profile** is a graph presenting series of odds ratios with confidence interval, determined for a given window size, i.e. comparing frequencies inside the window with frequencies placed outside the window. It enables the user to choose several categories into which he wants to divide the examined variable and adopt the most advantageous reference category. It works best when we are looking for a U-shaped function i.e. high risk at low and at high values of the variable under study and low risk at average values. There is no one window size that is good for every analysis, the window size must be determined individually for each variable. The size of the window indicates the number of unique values of variable X contained in the window. The wider the window, the greater the generalizability of the results and the smoother the odds ratio function. The narrower the window, the more detailed the results, resulting in a more lopsided odds ratio function. A curve is added to the graph showing the smoothed (Lowess method) odds ratio value. Setting the smoothing factor close to 0 results in a curve closely fitting to the odds ratio, whereas setting the smoothing factor closer to 1 results in more generalized odds ratio, i.e. smoother and less fitting to the odds ratio curve. The graph is dedicated to the evaluation of continuous variables in univariate analysis, i.e. when only one independent variable is selected.

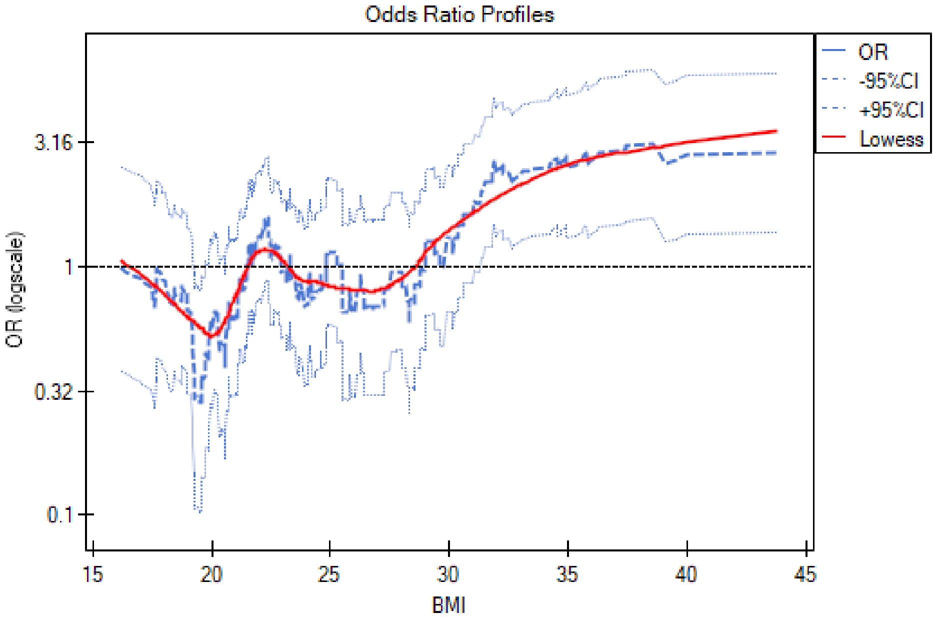
Example 24.3. (OR profiles.pqs file)

We examine the risk of disease A and disease B as a function of the patient's BMI. Since BMI is a continuous variable, its inclusion in the model results in a unit odds ratio that determines a linear trend of increasing or decreasing risk. We do not know whether a linear model will be a good model for the analysis of this risk, so before building multivariate logistic regression models, we will build some univariate models presenting this variable in graphs to be able to assess the shape of the relationship under study and, based on this, decide how we should prepare the variable for analysis. For this purpose, we will use plots of unit changes in odds ratio and odds ratio profiles, and for the profiles we will choose a window size of 100 because almost every patient has a different BMI, so about 100 patients will be in 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 get a statistically significant and positive odds ratio showing that people with a BMI above this value have a significantly higher risk of disease than people below this value.

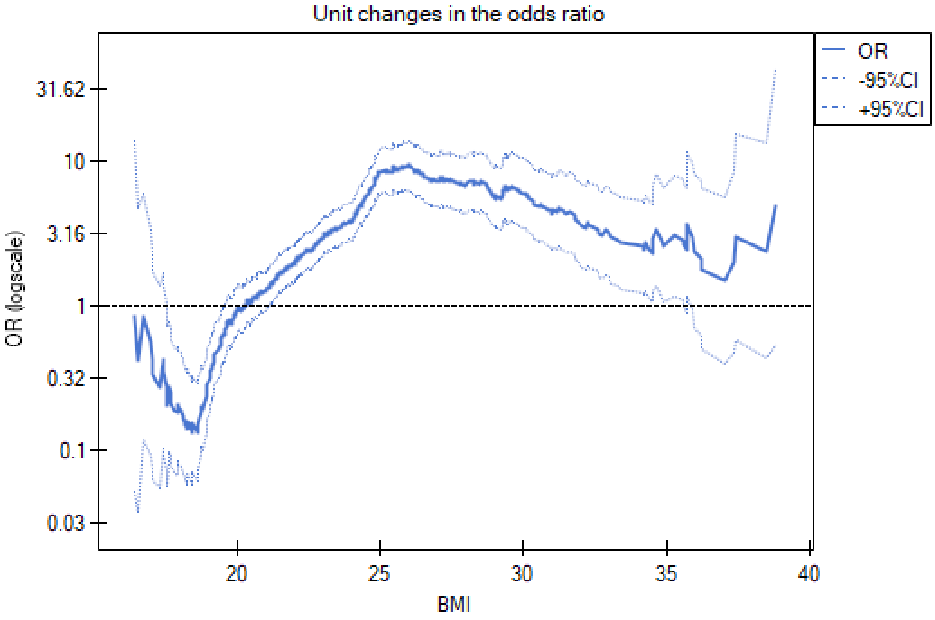


The odds ratio profiles show that the red curve is still close to 1, only the top of the curve is slightly higher, indicating that it may be difficult to divide BMI into more than 2 categories and select a good reference category, i.e., one that yields significant odds ratios.

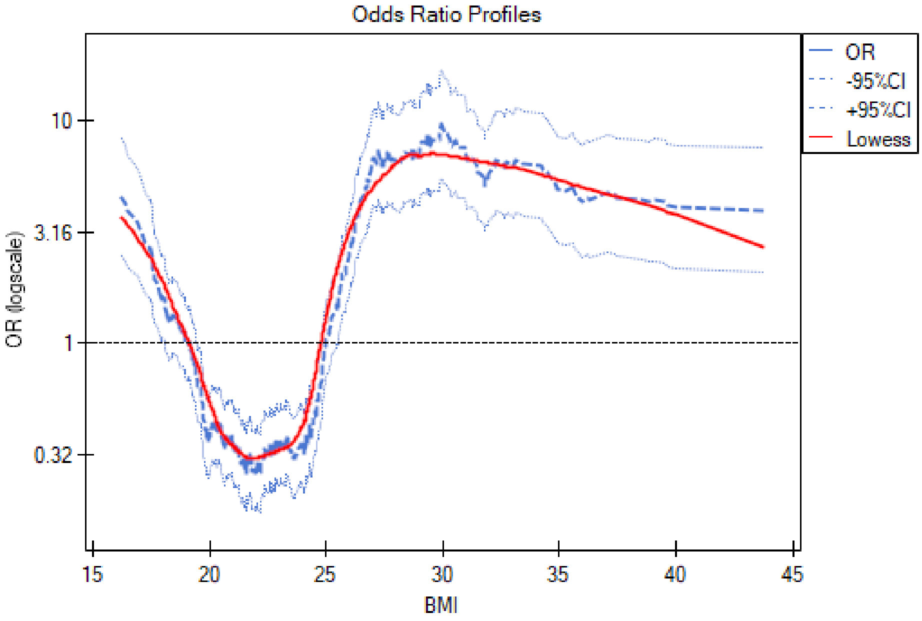


In summary, one can use a split of BMI into two values (e.g., relate those with a BMI above 30 to those with a BMI below that, in which case OR[95%CI]=[1.41, 4.90], p=0.0024) or stay with the unit odds ratio, indicating a constant increase in disease risk with an increase in BMI of one unit (OR[95%CI]=1.07[1.02, 1.13], p=0.0052).

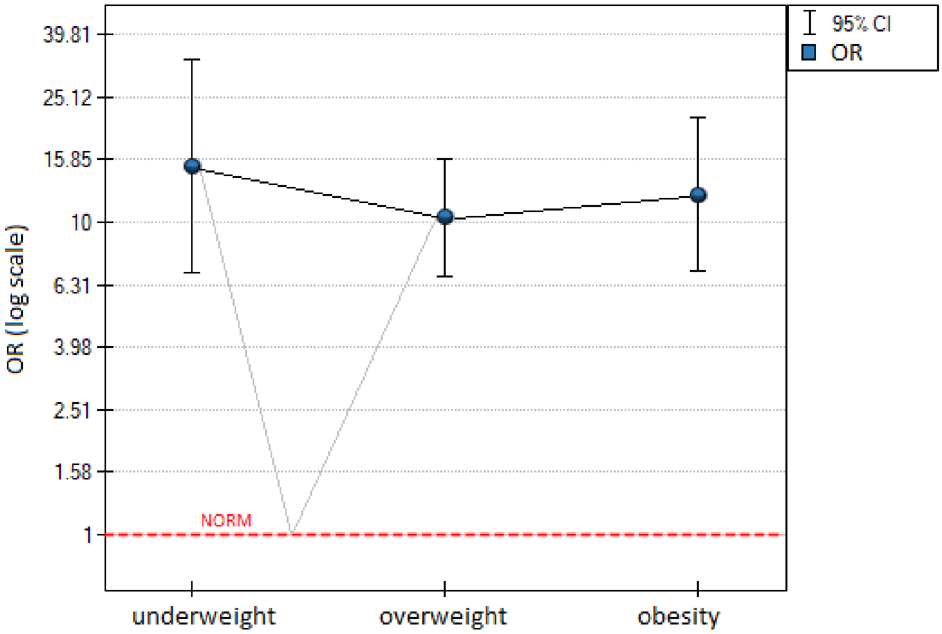
- Disease B
  Unit changes in the odds ratio show that when the BMI cut-off point is chosen somewhere between 22 and 35, we get a statistically significant and positive odds ratio showing that people with a BMI above this value have a significantly higher risk of disease than those below this value.



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 that includes a BMI somewhere between 19 and 25, as this is the category that is lowest and is far removed from the results for BMIs to the left and right of this range. We see a distinct U-like shape, meaning that disease risk is high at low BMI and at high BMI.



In summary, although the relationship for the unit odds ratio, or linear relationship, is statistically significant, it is not worth building such a model. It is much better to divide BMI into categories. The division that best shows the shape of this relationship is the one using two or three BMI categories, where the reference value will be the average BMI. Using the standard division of BMI and establishing a reference category of BMI in the normal range will result in a more than 15 times higher risk for underweight people (OR[95%CI]=15.14[6.93, 33.10]) more than 10 times for overweight people (OR[95%CI]=10.35[6.74, 15.90]) and more than twelve times for people with obesity (OR[95%CI]=12.22[6.94, 21.49]).



In the odds ratio plot, the BMI norm is indicated at level 1, as the reference category. We have drawn lines connecting the obtained ORs and also the norm, so as to show that the obtained shape of the relationship is the same as that determined previously by the odds ratio profile.

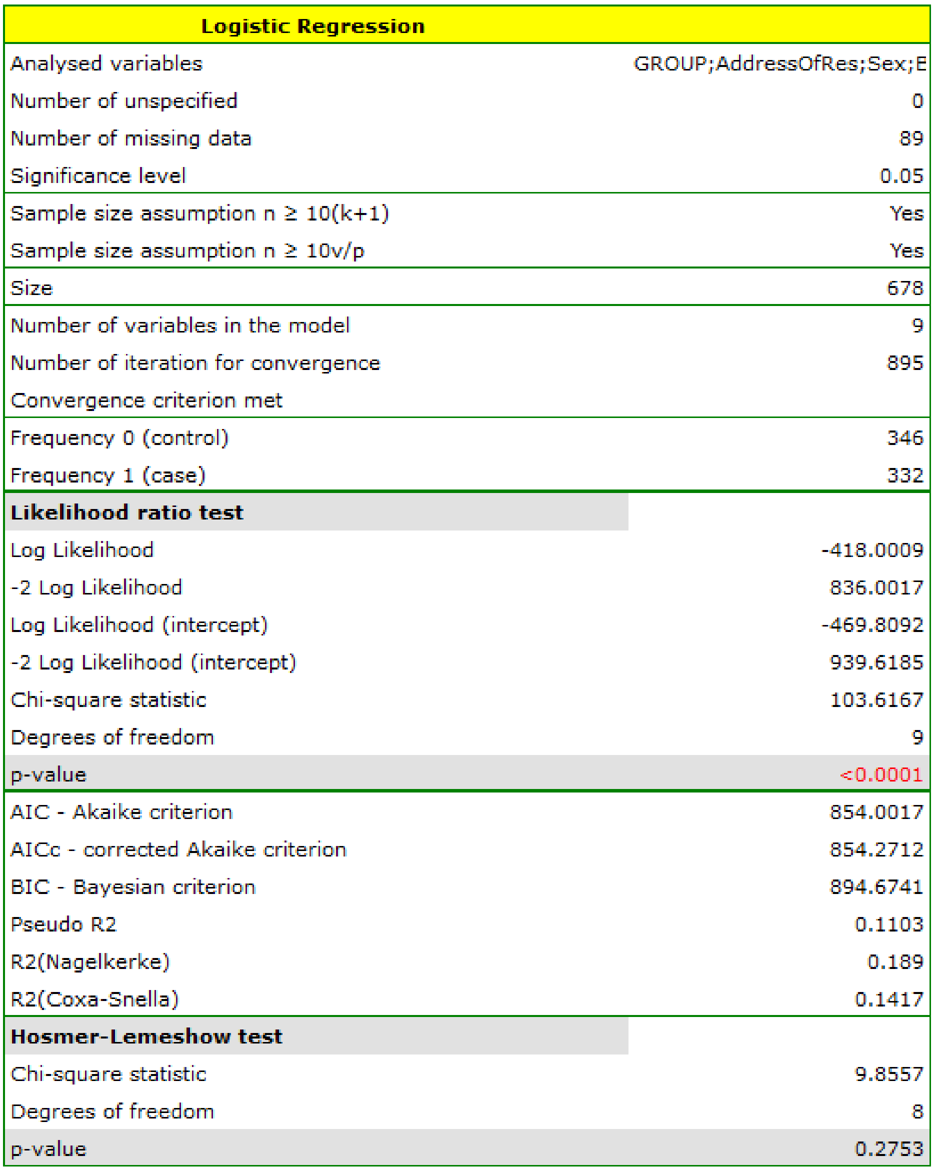
24.4.3 Examples for logistic regression

Example 24.4. (anomaly.pqs file)
Studies have been conducted for the purpose of identifying the risk factors for a certain rare congenital anomaly in children. 395 mothers of children with that anomaly and 375 of healthy children have participated in that study. The gathered data are: address of residence, child's sex, child's weight at birth, mother's age, number of pregnancy, previous spontaneous abortions, respiratory tract infections, smoking, mother's education.

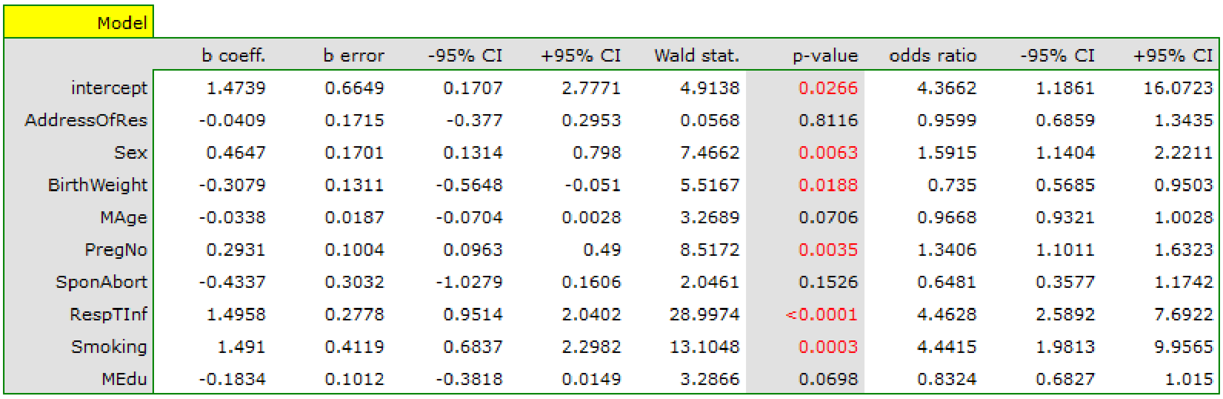
We construct a logistic regression model to check which variables may have a significant influence on the occurrence of the anomaly. The dependent variable is the column GROUP, the distinguished values in that variable as 1 are the ”cases”, that are mothers of children with anomaly. The following 9 variables are independent variables:

AddressOfRes (2=city/1=village),
Sex (1=male/0=female),
BirthWeight (in kilograms, with an accuracy of 0.5 kg),
MAge (in years),
PregNo (which pregnancy is the child 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).



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

The quality of model goodness of fit is not high ($R^2_{Pseudo} = 0.11$, $R^2_{Nagelkerke} = 0.19$ i $R^2_{Cox-Snell} = 0.14$). At the same time the model is statistically significant (value $p < 0.000001$ of the Likelihood Ratio test), which means that a part of the independent variables in the model is statistically significant. The result of the Hosmer-Lemeshow test points to a lack of significance ($p = 0.2753$). However, in the case of the Hosmer-Lemeshow test we ought to remember that a lack of significance is desired as it indicates a similarity of the observed sizes and of predicted probability.

An interpretation of particular variables in the model starts from checking their significance. In this case the variables which are significantly related to the occurrence of the anomaly are:

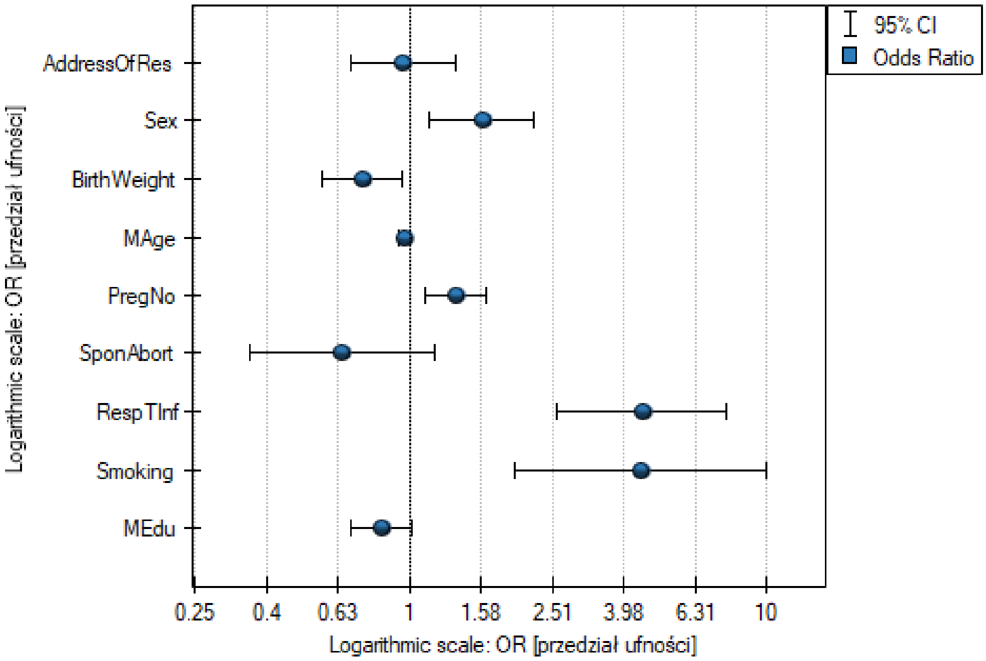
Sex: $p = 0.0063$,
BirthWeight: $p = 0.0188$,
PregNo: $p = 0.0035$,
RespTInf: $p < 0.000001$,
Smoking: $p = 0.0003$.

The studied congenital anomaly is a rare anomaly 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 occurrence of the anomaly 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 smaller the odds of the occurrence of the anomaly in a child;
- variable PregNo: $OR[95\%CI] = 1.34[1.10; 1.63]$ –the odds of the occurrence of the anomaly in a child is 1.34 times greater with each subsequent pregnancy;
- variable RespTInf: $OR[95\%CI] = 4.46[2.59; 7.69]$ –the odds of the occurrence of the anomaly in a child if the mother had a respiratory tract infection during the pregnancy is 4.46 times greater than in a mother who did not have such an infection during the pregnancy;
- variable Smoking: $OR[95\%CI] = 4.44[1.98; 9.96]$ –a mother who smokes when pregnant increases the risk of the occurrence of the anomaly in her child 4.44 times.

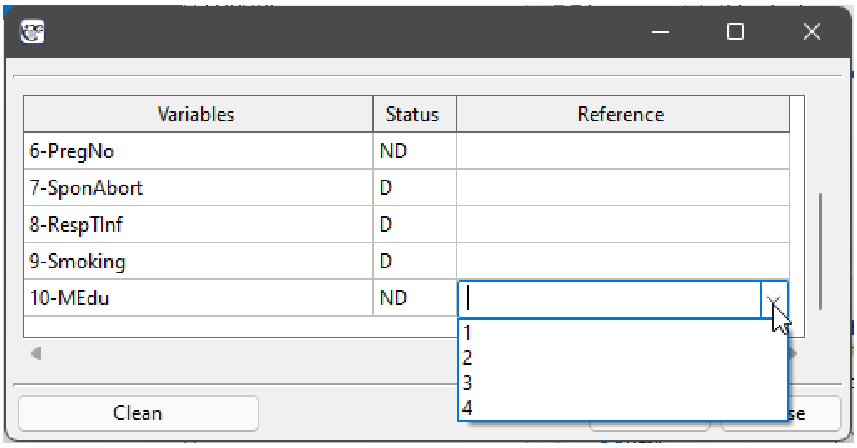
In the case of statistically insignificant variables the confidence interval for the Odds Ratio contains 1 which means that the variables neither increase nor decrease the odds of the occurrence of the studied anomaly. Therefore, we cannot interpret the obtained ratio in a manner similar to the case of statistically significant variables.

The influence of particular independent variables on the occurrence of the anomaly can also be described with the help of a chart concerning the odds ratio:



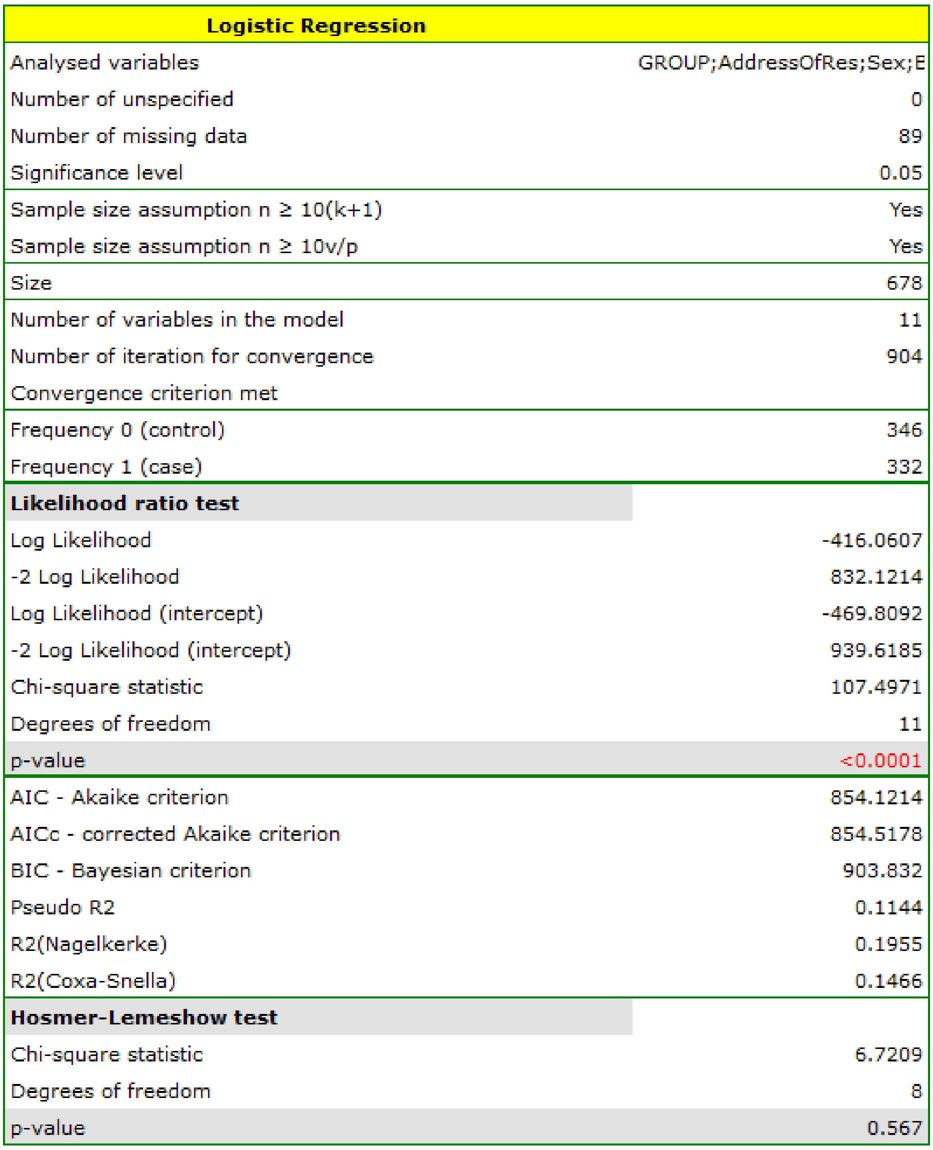
EXAMPLE 24.4 continued (anomaly.pqs file)

Let us once more construct a logistic regression model, however, this time let us divide the variable mother's education into dummy variables (with dummy coding). With this operation we lose the information about the ordering of the category of education but we gain the possibility of a more in-depth analysis of particular categories. The breakdown into dummy variables is done by selecting Dummy var. in the analysis window.:

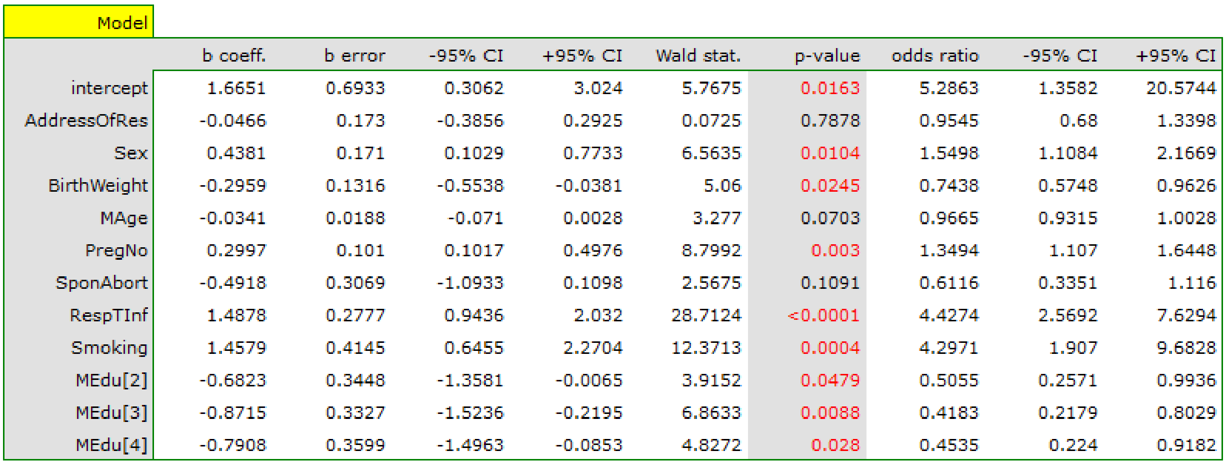


Variables	Status	Reference
6-PregNo	ND	
7-SponAbort	D	
8-RespTInf	D	
9-Smoking	D	
10-MEdu	ND	

The primary education variable is missing as it will constitute the reference category.



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

As a result the variables which describe education become statistically significant. The goodness of fit of the model does not change much but the manner of interpretation of the the odds ratio for education 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 occurrence of the studied anomaly in each education category is always compared with the odds of the occurrence of the anomaly in the case of primary education. We can see that for more educated the mother, the odds is lower. For a mother with:

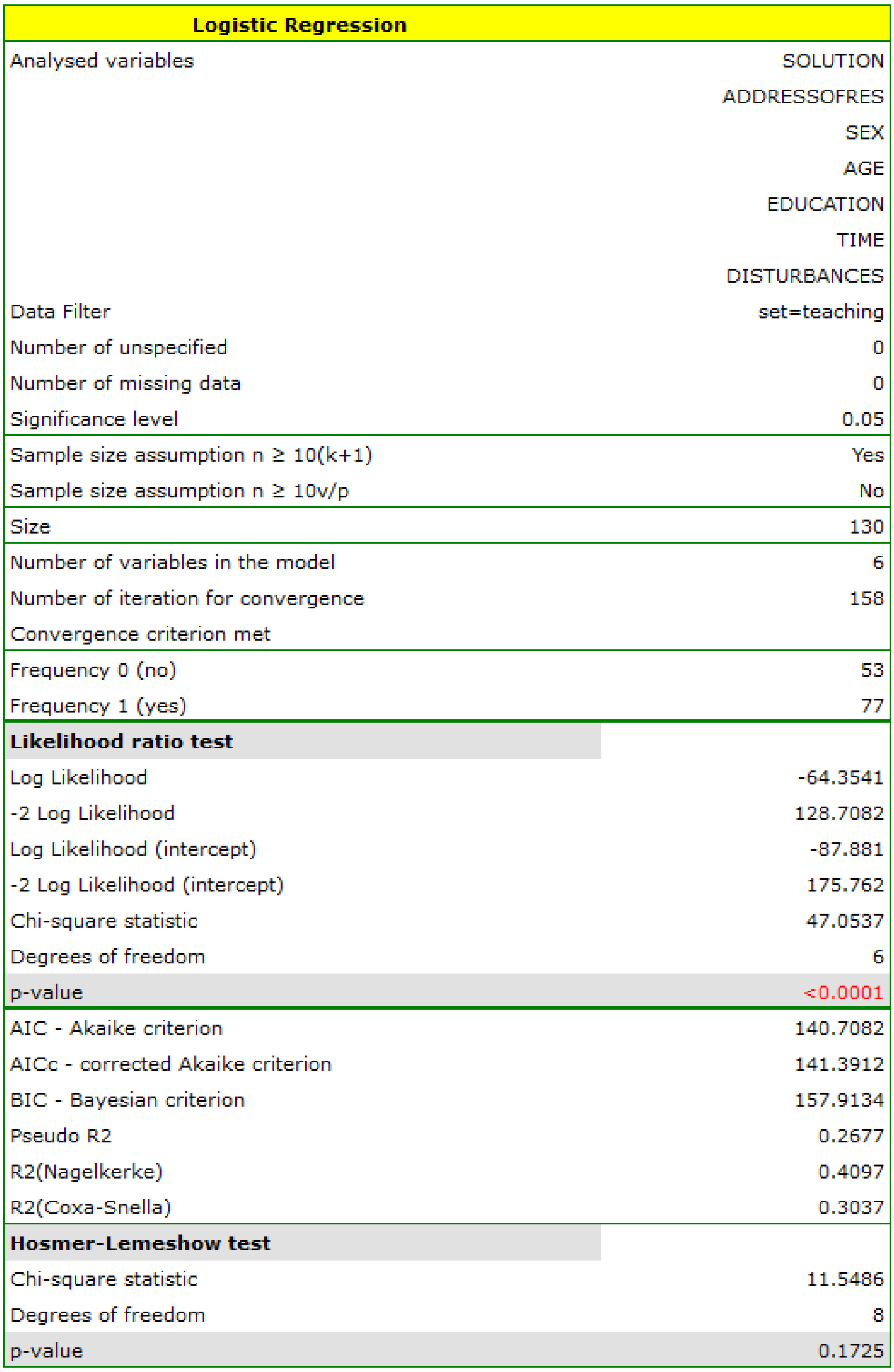
- vocational education the odds of the occurrence of the anomaly in a child is 0.51 of the odds for a mother with primary education;
- secondary education the odds of the occurrence of the anomaly in a child is 0.42 of the odds for a mother with primary education;
- tertiary education the odds of the occurrence of the anomaly in a child is 0.45 of the odds for a mother with primary education;

Example 24.5. (task.pqs file)
An experiment has been made with the purpose of studying the ability to concentrate of a group of adults in an uncomfortable situation. 130 people have taken part in the experiment. Each person was assigned a certain task the completion of which requried concentration. During the experiment some people were subject to a disturbing agent in the form of temperature increase to 32 degrees Celsius. The participants were also asked about their address of residence, sex, age, and education. The time for the completion of the task was limited to 45 minutes. In the case of participants who completed the task before the deadline, the actual time devoted to the completion of the task was recorded.

Variable SOLUTION (yes/no) contains the result of the experiment, i.e. the information about whether the task was solved correctly or not. The remaining variables which could have influenced the result of the experiment are:

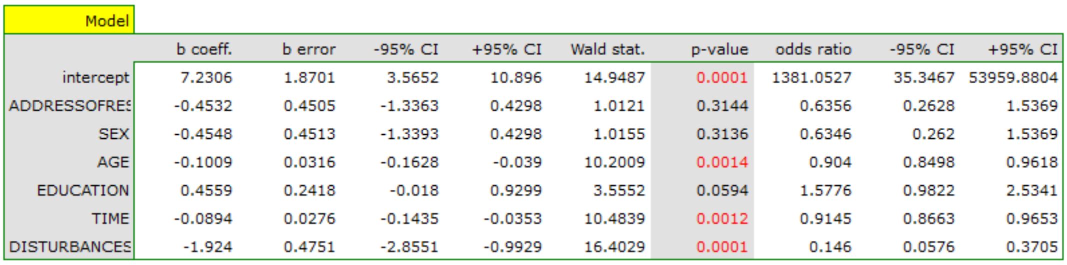
ADDRESSOFRES (1=city/0=village),
SEX (1=female/0=male),
AGE (in years),
EDUCATION (1=primary, 2=vocational, 3=secondary, 4=tertiary),
TIME needed for the completion of the task (in minutes),
DISTURBANCES (1=yes/0=no).

On the basis of all those variables a logistic regression model was built in which the distinguished state of the variable SOLUTION was set to ”yes”.



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

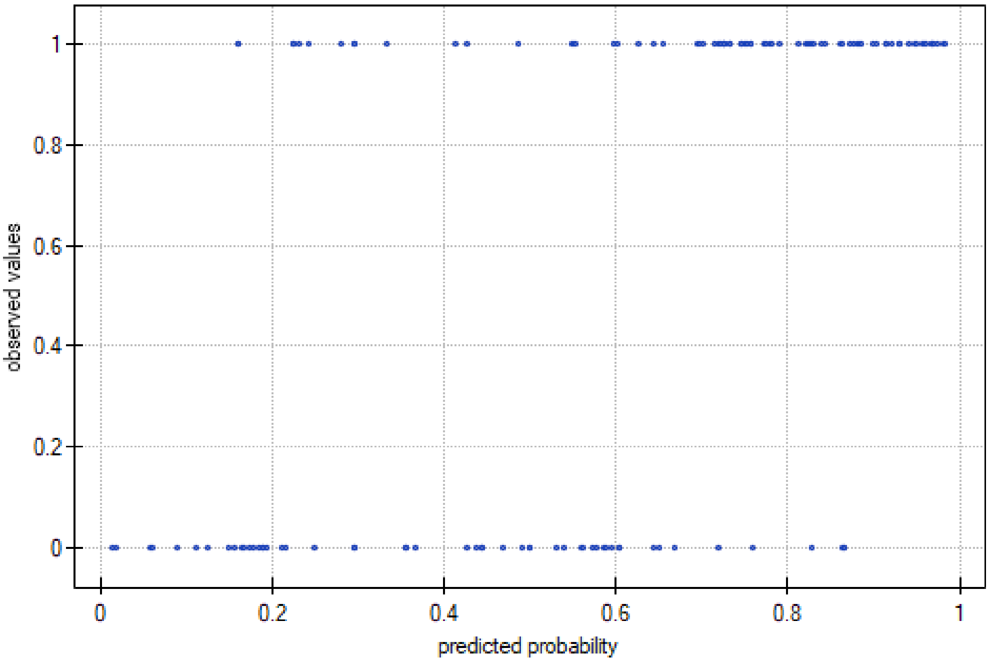
The adequacy quality is described by the coefficients: $R^2_{Pseudo} = 0.27$, $R^2_{Nagelkerke} = 0.41$ i $R^2_{Cox-Snell} = 0.30$. The sufficient adequacy is also indicated by the result of the Hosmer-Lemeshow test $(p = 0.1725)$. The whole model is statistically significant, which is indicated by the result of the Likelihood Ratio test $(p < 0.000001)$.



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

The observed values and predicted probability can be observed on the chart:



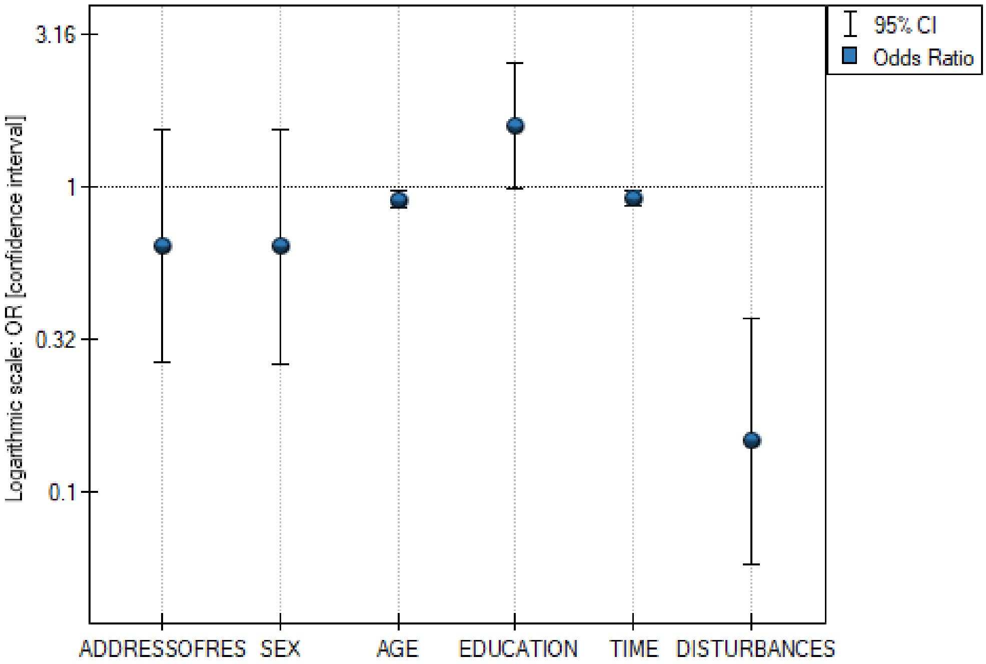
In the model the variables which have a significant influence on the result are:

AGE: $p = 0.0014$,
TIME: $p = 0.0012$,
DISTURBANCES: $p = 0.0001$.

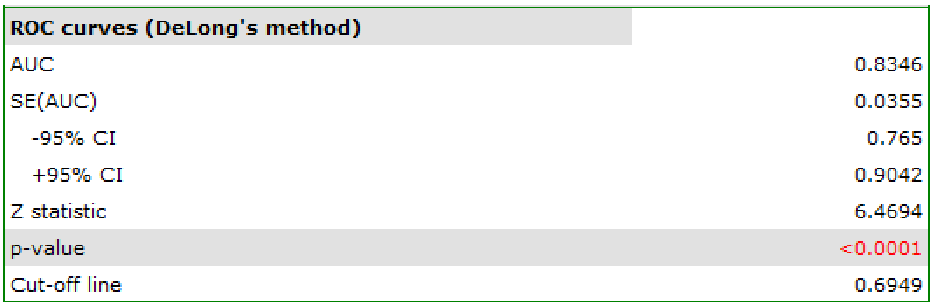
What is more, the younger the person who solves the task the shorter the time needed for the completion of the task, and if there is no disturbing agent, the probability of correct solution is greater:

AGE: $OR[95\%CI] = 0.90[0.85; 0.96]$,
TIME: $OR[95\%CI] = 0.91[0.87; 0.97]$,
DISTURBANCES: $OR[95\%CI] = 0.15[0.06; 0.37]$.

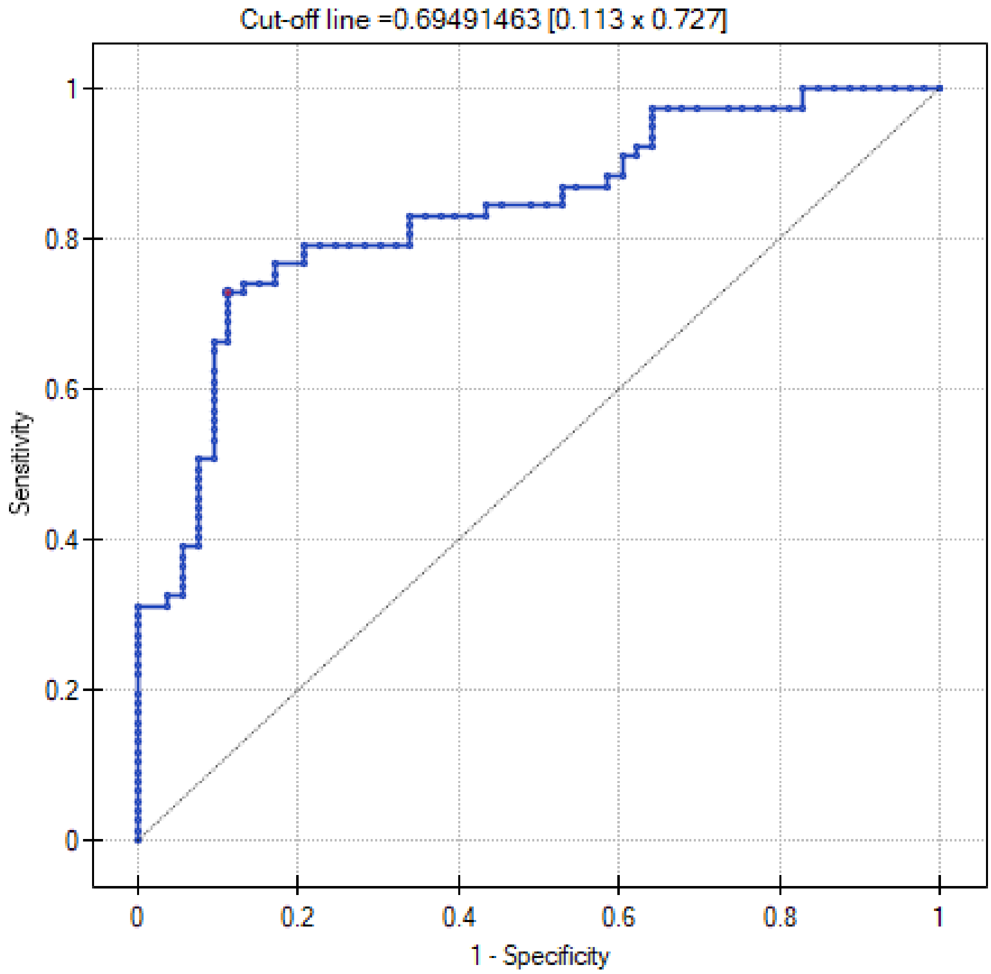
The obtained results of the Odds Ratio are presented on the chart below:



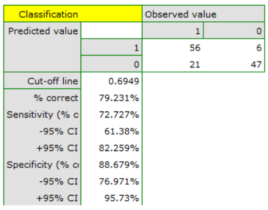
Should the model be used for prediction, one should pay attention to the quality of classification. For that purpose we calculate the ROC curves.



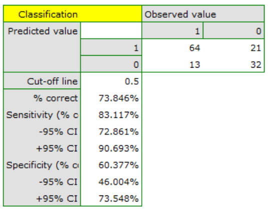
ROC curves (DeLong's method)	
AUC	0.8346
SE(AUC)	0.0355
-95% CI	0.765
+95% CI	0.9042
Z statistic	6.4694
p-value	<0.0001
Cut-off line	0.6949



The result seems satisfactory. The area under the curve is $AUC = 0.83$ and is statistically greater than 0.5 $(p < 0.0001)$, so classification is possible on the basis of the constructed model. The suggested cut-off point for the ROC curve is 0.6949 and is slightly higher than the standard level used in regression, i.e. 0.5. The classification determined from this cut-off point yields 79.23% of cases classified correctly, of which correctly classified "yes" values are 72.73% (sensitivity), "no" values are 88.68% (specificity). The classification derived from the standard value yields no less, 73.85% of cases classified correctly, but it will yield more correctly classified "yes" values are 83.12%, although less correctly classified "no" values are 60.38%.



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 (% c	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 (% c	60.377%		
-95% CI	46.004%		
+95% CI	73.548%		

We can finish the analysis of classification at this stage or, if the result is not satisfactory, we can make a more detailed analysis of the ROC curve in module ROC curve.

As we have assumed that classification on the basis of that model is satisfactory we can calculate the predicted value of a dependent variable for any conditions. Let us check what odds of solving the task has a person whose:

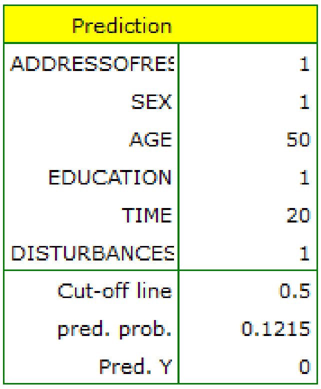
ADDRESSOFRES (1=city),
SEX (1=female),

AGE (50 years),
EDUCATION (1=primary),
TIME needed for the completion of the task (20 minutes),
DISTURBANCES (1=yes).

For that purpose, on the basis of the value of coefficient b, we calculate the predicted probability (probability of receiving the answer ”yes” on condition of defining the values of dependent variables):

$$P(Y=yes|ADDRESSOFRES,SEX,AGE,EDUCATION,TIME,DISTURBANCES)=$$
$$=\frac{e^{7.23-0.45\mathit{ADDRESSOFRES}-0.45\mathit{SEX}-0.1\mathit{AGE}+0.46\mathit{EDUCATION}-0.09\mathit{TIME}-1.92\mathit{DISTURBANCES}}}{1+e^{7.23-0.45\mathit{ADDRESSOFRES}-0.45\mathit{SEX}-0.1\mathit{AGE}+0.46\mathit{EDUCATION}-0.09\mathit{TIME}-1.92\mathit{DISTURBANCES}}}=$$
$$=\frac{e^{7.231-0.453\cdot 1-0.455\cdot 1-0.101\cdot 50+0.456\cdot 1-0.089\cdot 20-1.924\cdot 1}}{1+e^{7.231-0.453\cdot 1-0.455\cdot 1-0.101\cdot 50+0.456\cdot 1-0.089\cdot 20-1.924\cdot 1}}$$

As a result of the calculation the program will return the result:



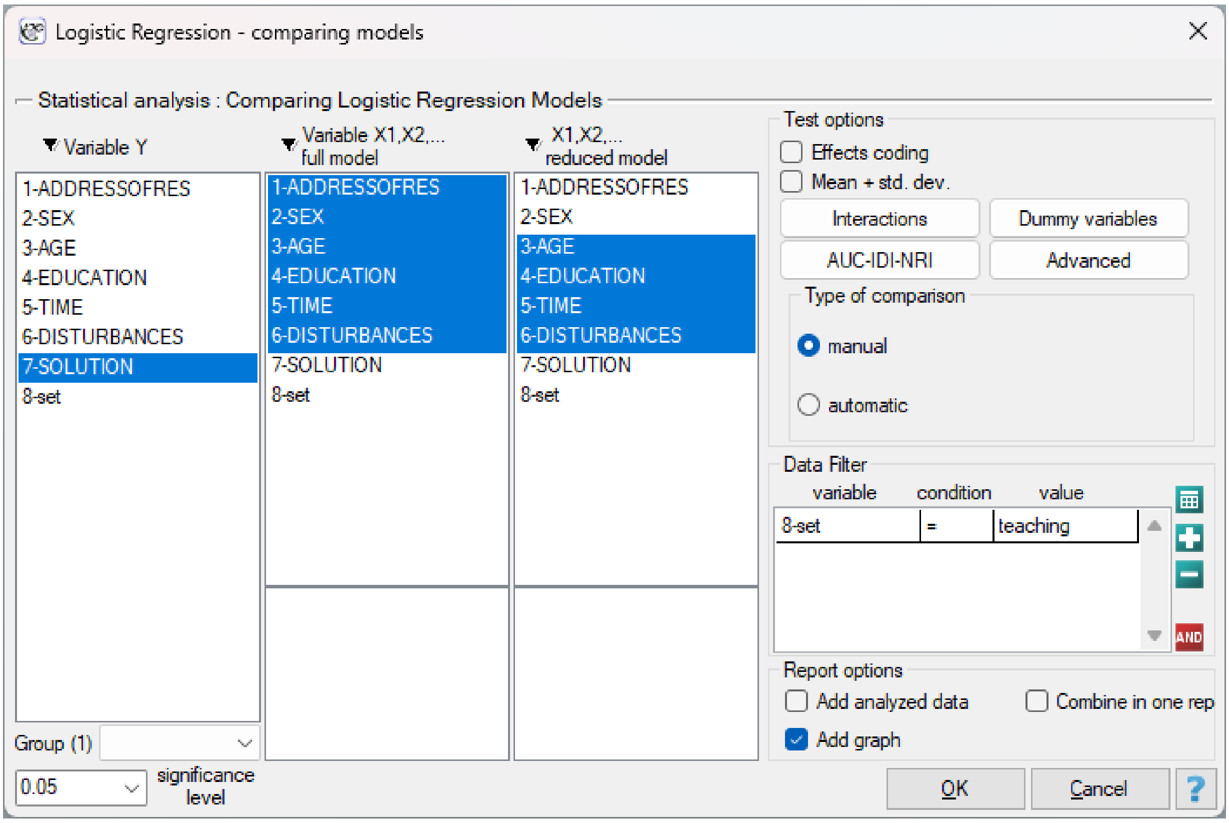
Prediction	
ADDRESSOFRES	1
SEX	1
AGE	50
EDUCATION	1
TIME	20
DISTURBANCES	1
Cut-off line	0.5
pred. prob.	0.1215
Pred. Y	0

The obtained probability of solving the task is equal to 0.1215, so, on the basis of the cut-off 0.60, the predicted result is 0 –which means the task was not solved correctly.

24.4.4 Model-based prediction and test set validation

24.5 COMPARISON OF LOGISTIC REGRESSION MODELS

The window with settings for model comparison is accessed via the menu Advanced Statistics→Multivariate models→Logistic regression – comparing models



Logistic Regression - comparing models

Statistical analysis : Comparing Logistic Regression Models

Variable Y: 1-ADDRESSOFRES, 2-SEX, 3-AGE, 4-EDUCATION, 5-TIME, 6-DISTURBANCES, **7-SOLUTION**, 8-set

Variable X1,X2,... full model: **1-ADDRESSOFRES, 2-SEX, 3-AGE, 4-EDUCATION, 5-TIME, 6-DISTURBANCES**, 7-SOLUTION, 8-set

X1,X2,... reduced model: 1-ADDRESSOFRES, 2-SEX, **3-AGE, 4-EDUCATION, 5-TIME, 6-DISTURBANCES**, 7-SOLUTION, 8-set

Group (1)

0.05 significance level

Test options
□ Effects coding
□ Mean + std. dev.

Type of comparison
● manual
○ automatic

Data Filter

variable	condition	value
8-set	=	teaching

Report options
□ Add analyzed data
□ Combine in one rep
☑ Add graph

Due to the possibility of simultaneous analysis of many independent variables in one logistic regression model, similarly to the case of multiple linear regression, there is a problem of selection of an optimum model. When choosing independent variables one has to remember to put into the model variables strongly correlated with the dependent variable and weakly correlated with one another.

When comparing models with different numbers of independent variables, we pay attention to model fit and information criteria. For each model we also calculate the maximum of the credibility function, which we then compare using the credibility quotient test.

Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & L_{FM}=L_{RM},\\ \mathcal{H}_1: & L_{FM}\neq L_{RM},\end{array}$$

where:
L_{FM}, L_{RM} – the maximum of likelihood function in compared models (full and reduced).

The test statistic has the form presented below:

$$\chi^2=-2\ln(L_{RM}/L_{FM})=-2\ln(L_{RM})-(-2\ln(L_{FM}))$$

The statistic asymptotically (for large sizes) has the χ^2 distribution with $df = k_{FM} - k_{RM}$ degrees of freedom, where k_{FM} i k_{RM} is the number of estimated parameters in compared models.

On the basis of test statistics, p value is estimated and then compared with α :

$$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

We make the decision about which model to choose on the basis of the size R^2_{Pseudo}, $R^2_{Nagelkerke}$, $R^2_{Cox-Snell}$ and the result of the Likelihood Ratio test which compares the subsequently created (neighboring) models. If the compared models do not differ significantly, we should select the one with a smaller number of variables. This is because a lack of a difference means that the variables present in the full model but absent in the reduced model do not carry significant information. However, if the difference is statistically significant, it means that one of them (the one with the greater number of variables, with a greater R^2 and a lower information criterion value of AIC, AICc or BIC) is significantly better than the other one.

Comparison the predictive value of models.
The regression models that are built allow to predict the probability of occurrence of the studied event based on the analyzed independent variables. When many variables (factors) that increase the risk of an event are already known, then an important criterion for a new candidate risk factor is the improvement in prediction performance when that factor is added to the model. To establish the point, let's use an example. Suppose we are studying risk factors for coronary heart disease. Known risk factors for this disease include age, systolic and diastolic blood pressure values, obesity, cholesterol, or smoking. However, the researchers are interested in how much the inclusion of individual factors in the regression model will significantly improve disease risk estimates. Risk factors added to a model will have predictive value if the new and larger model (which includes these factors) shows better predictive value than a model without them. The predictive value of the model is derived from the determined value of the predicted probability of an event, in this case coronary artery disease. This value is determined from the model for each individual tested. The closer the predicted probability is to 1, the more likely the disease is. Based on the predicted probability, the value of the AUC area under the ROC curve can be determined and compared between different models, as well as the NRI and IDI coefficient.

Change of the area under the ROC curve

The ROC curve in logistic regression models is constructed based on the classification of cases into a group experiencing an event or not, and the predicted probability of the dependent variable P. The larger the area under the curve, the more accurately the probability determined by the model predicts the actual occurrence of the event. If we are comparing models built on the basis of a larger or smaller number of predictors, then by comparing the size of the area under the curve we can check whether the addition of factors has significantly improved the prediction of the model.
Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & AUC_{FM} = AUC_{RM},\\ \mathcal{H}_1: & AUC_{FM} \neq AUC_{RM}.\end{array}$$

For a method of determining the test statistic based on DeLong's method, check out Comparing ROC curves.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

Net reclassification improvement

This measure is denoted by the acronym NRI The NRI focuses on the reclassification table describing the upward or downward shift in probability values when a new factor is added to the model. It is determined based on two separate factors, i.e., a factor determined separately for subjects experiencing the event (1) and separately for those not experiencing the event (0). The NRI can be determined with a given division of the predicted probability into categories (categorical NRI) or without the need to determine the categories (continuous NRI).

- NRI **categorical** requires an arbitrary determination of the division of probability values predicted from the model. There can be a maximum of 9 split points given, and thus a maximum of 10 predicted categories. However, one or two split points are most commonly used. At the same time, it should be noted that the values of the categorical NRI can be compared with each other only if they were based on the same split points. To illustrate the situation, let's establish two example probability split points: 0.1 and 0.3. If a test person in the ”old” (smaller) model received a probability below 0.1, and in the ”new model” (increased by a new potential risk factor) probability located between 0.1 and 0.3, it means that the person was reclassified upwards (table, situation 1). If the probability values from both models are in the same range, then the person is not reclassified (table, situation 2), whereas if the probability from the ”new” model is lower than that from the ”old” model, then the person is reclassified downwards (table, situation 3).

regression models	situation 1		situation 2		situation 3	
predicted probability	”old”	”new”	”old”	”new”	”old”	”new”
[0.3 do 1]					$\oplus$	
[0.1; 0.3)		$\oplus$	$\oplus$	$\oplus$		$\oplus$
[0; 0.1)	$\oplus$					

- NRI **continuous** does not require an arbitrary designation of categories, since any, even the smallest, change in probability up or down from the probability designated in the ”old model” is treated as a transition to the next category. Thus, there are infinitely many categories, just as there are many possible changes.

Note

Use of continuous NRI does not require arbitrary definition of probability split points, but even small changes in risk (not reflected in clinical observations) can increase or decrease this ratio. The categorical NRI factor allows only changes that are important to the investigator to reflect changes that involve exceeding preset event risk values (predicted probability values).

To determine NRI we define:

$$\hat{p}_{up,events} = \frac{\#events_{up}}{\#events}$$

$$\hat{p}_{down,events} = \frac{\#events_{down}}{\#events}$$

$$\hat{p}_{up,nonevents} = \frac{\#nonevents_{up}}{\#nonevents}$$

$$\hat{p}_{down,nonevents} = \frac{\#nonevents_{down}}{\#nonevents}$$

where:

$\#events_{up}$ – the number of subjects in the group experiencing the event for whom there was an upward change of at least one category in the predicted probability,

$\#events_{down}$ – the number of subjects in the group experiencing the event for whom there was at least one downward change in predicted probability,

$\#events$ – number of objects in the group experiencing the event,

$\#nonevents_{up}$ – The number of subjects in the group not experiencing the event for whom there was an upward change of at least one category in the predicted probability,

$\#nonevents_{down}$ – The number of subjects in the group not experiencing the event for whom there was at least one downward change in predicted probability,

$\#nonevents$ – number of objects in the group not experiencing the event.

The overall NRI and coefficients expressing the percentage change in classification are determined from the formula:

$$NRI = (\hat{p}_{up,events} - \hat{p}_{down,events}) - (\hat{p}_{up,nonevents} - \hat{p}_{down,nonevents})$$

$$NRI_{events} = \hat{p}_{up,events} - \hat{p}_{down,events}, \quad NRI_{nonevents} = \hat{p}_{down,nonevents} - \hat{p}_{up,nonevents},$$

The NRI_{events} coefficient can be interpreted as the **net percentage** of correctly reclassified individuals with an event, and $NRI_{nonevents}$ as the **net percentage** of correctly reclassified individuals without an event. The overall coefficient NRI is expressed as the sum of the coefficients NRI_{events} and $NRI_{nonevents}$ making it a coefficient implicitly weighted by event frequency and cannot be interpreted as a percentage.

The NRI_{events} coefficientsbelong to the range from-1 to 1 (from -100% to 100%), and the overall coefficients of NRI belong to the range from -2 to 2. Positive values of the coefficients indicate favorable reclassification, and negative values indicate unfavorable reclassification due to the addition of a new variable to the model.

Z test for significance of NRI coefficient

Using this test, we examine whether the change in classification expressed by the NRI coefficient was significant.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & NRI = 0, \\ \mathcal{H}_1: & NRI \neq 0. \end{array}$$

The test statistic has the form:

$$Z = \frac{NRI}{SE(NRI)}$$

where:

$$\begin{aligned} SE(NRI) = \; & [\left(\frac{\#events_{up}+\#events_{down}}{\#events^2} - \frac{(\#events_{up}+\#events_{down})^2}{\#events^3}\right) + \\ & + \left(\frac{\#nonevents_{down}+\#nonevents_{up}}{\#nonevents^2} - \frac{(\#nonevents_{down}+\#nonevents_{up})^2}{\#nonevents^3}\right)]^{1/2} \end{aligned}$$

The Z statistic asymptotically (for large sample sizes) has the normal distribution.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \leq \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Integrated Discrimination Improvement

This measure is denoted by the abbreviation IDI. The IDI coefficients indicate the difference between the value of the average change in the predicted probability between the group of objects experiencing the event and the group of objects that did not experience the event.

$$IDI = \overline{p(diff)}_{events} - \overline{p(diff)}_{nonevents}$$

where:
$\overline{p(diff)}_{events}$ – The mean of the difference in predicted probability values between the regression models (”old” and ”new”) for objects that experienced the event,
$\overline{p(diff)}_{nonevents}$ – The mean of the difference in predicted probability values between the regression models (”old” and ”new”) for objects that did not experience the event.

Z test for significance of IDI coefficient

Using this test, we examine whether the difference between the value of the mean change in predicted probability between the group of subjects experiencing the event and the subjects not experiencing the event, as expressed by the IDI coefficient, was significant.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & IDI=0,\\ \mathcal{H}_1: & IDI\neq 0.\end{array}$$

The test statistic is of 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, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl}\text{if } p\le\alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

In the program PQStat the comparison of models can be done manually or automatically.

- **Manual** model comparison – construction of 2 models:
  - a full model – a model with a greater number of variables,
  - a reduced model – a model with a smaller number of variables – such a model is created from the full model by removing those variables which are superfluous from the perspective of studying a given phenomenon.

  The choice of independent variables in the compared models and, subsequently, the choice of a better model on the basis of the results of the comparison, is made by the researcher.

- **Automatic** model comparison is done in several steps:

  step 1 Constructing the model with the use of all variables.

  step 2 Removing one variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.

  step 3 A comparison of the full and the reduced model.

  step 4 Removing another variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.

  step 5 A comparison of the previous and the newly reduced model.

  ...

In that way numerous, ever smaller models are created. The last model only contains 1 independent variable.

Example 24.5 c.d. (task.pqs file)
In the experiment made with the purpose to study, for 130 people of the teaching set, the concentration abilities a logistic regression model was constructed on the basis of the following variables:

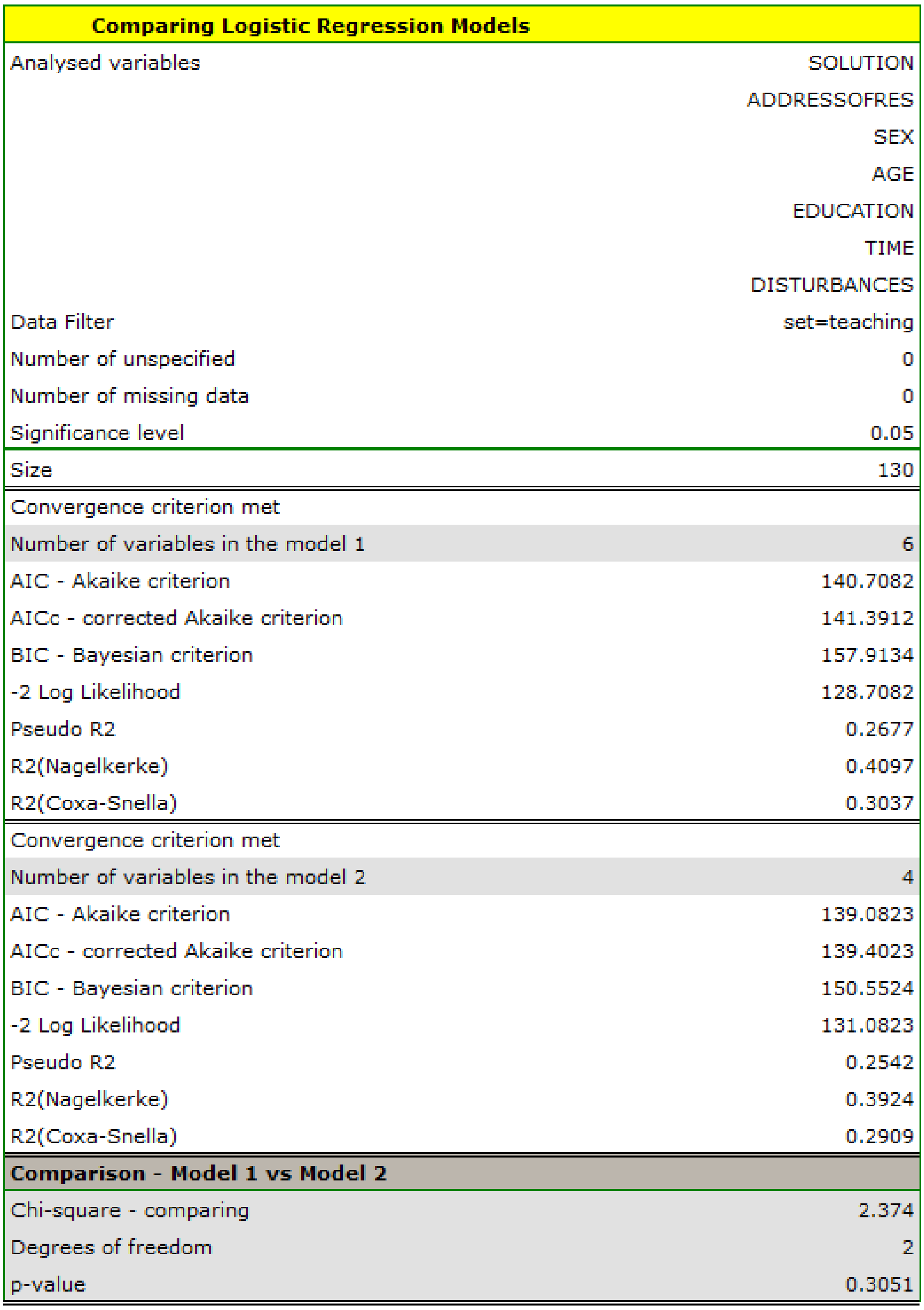
dependent variable: SOLUTION (yes/no) - information about whether the task was correctly solved or not;

independent variables:

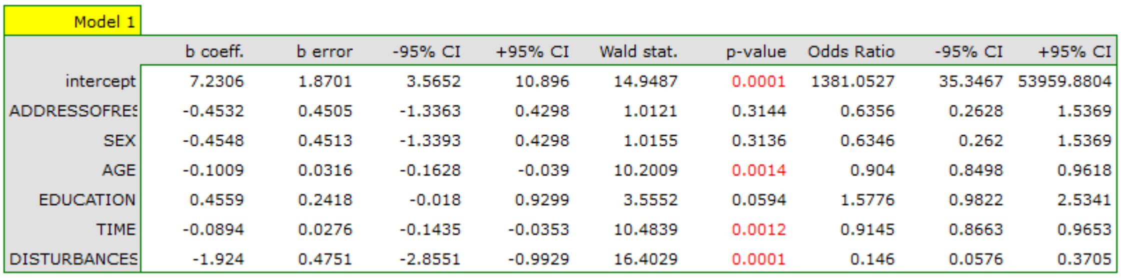
ADDRESSOFRES (1=city/0=village),
SEX (1=female/0=male),
AGE (in years),
EDUCATION (1=primary, 2=vocational, 3=secondary, 4=tertiary),
TIME needed for the completion of the task (in minutes),
DISTURBANCES (1=yes/0=no).

Let us check if all independent variables are indispensible in the model.

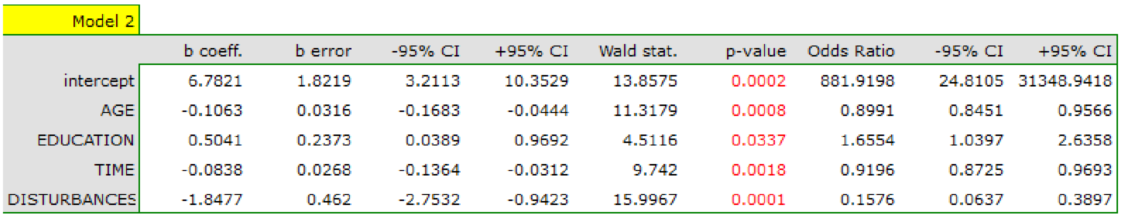
- **Manual** model comparison.
  On the basis of the previously constructed full model we can suspect that the variables: ADDRESSOFRES and SEX have little influence on the constructed model (i.e. we cannot successfully make classifications on the basis of those variables). Let us check if, from the statistical point of view, the full model is better than the model from which the two variables have been removed.



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

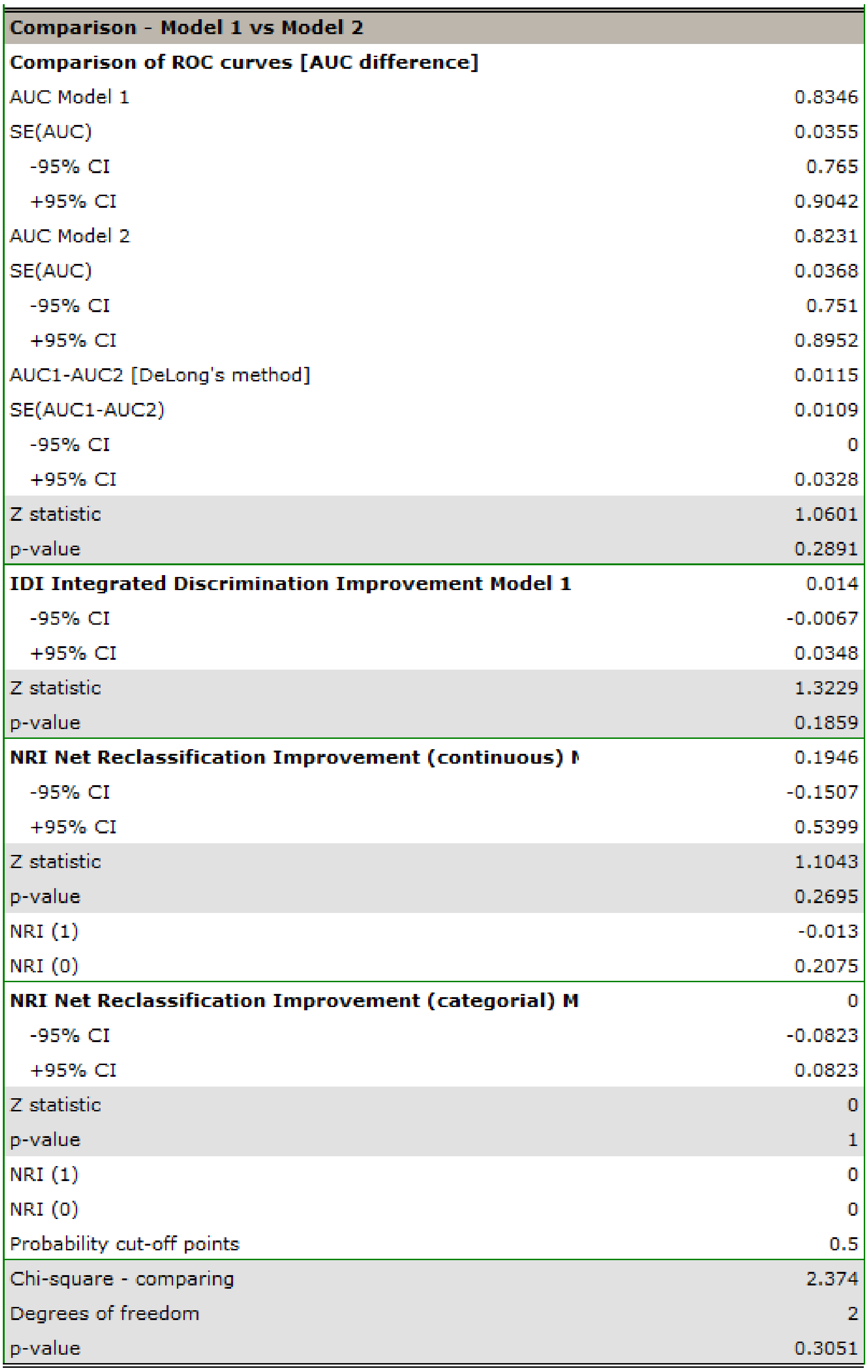


Model 2

	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

The results of the Likelihood Ratio test ($p = 0.3051$) indicates that there is no basis for believing that a full model is better than a reduced one. Therefore, with a slight worsening of model adequacy, the address of residence and the sex can be omitted.

We can compare the two models in terms of classification ability by comparing the ROC curves for these models, NRI and IDI value. To do so, 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 i.e. the p-values for the comparison of ROC curves and for the evaluation of NRI and IDI indices are statistically insignificant. We therefore decide to omit gender and place of residence from the final model.



Comparison - Model 1 vs Model 2	
Comparison of ROC curves [AUC difference]	
AUC Model 1	0.8346
SE(AUC)	0.0355
-95% CI	0.765
+95% CI	0.9042
AUC Model 2	0.8231
SE(AUC)	0.0368
-95% CI	0.751
+95% CI	0.8952
AUC1-AUC2 [DeLong's method]	0.0115
SE(AUC1-AUC2)	0.0109
-95% CI	0
+95% CI	0.0328
Z statistic	1.0601
p-value	0.2891
IDI Integrated Discrimination Improvement Model 1	0.014
-95% CI	-0.0067
+95% CI	0.0348
Z statistic	1.3229
p-value	0.1859
NRI Net Reclassification Improvement (continuous) M	0.1946
-95% CI	-0.1507
+95% CI	0.5399
Z statistic	1.1043
p-value	0.2695
NRI (1)	-0.013
NRI (0)	0.2075
NRI Net Reclassification Improvement (categorial) M	0
-95% CI	-0.0823
+95% CI	0.0823
Z statistic	0
p-value	1
NRI (1)	0
NRI (0)	0
Probability cut-off points	0.5
Chi-square - comparing	2.374
Degrees of freedom	2
p-value	0.3051

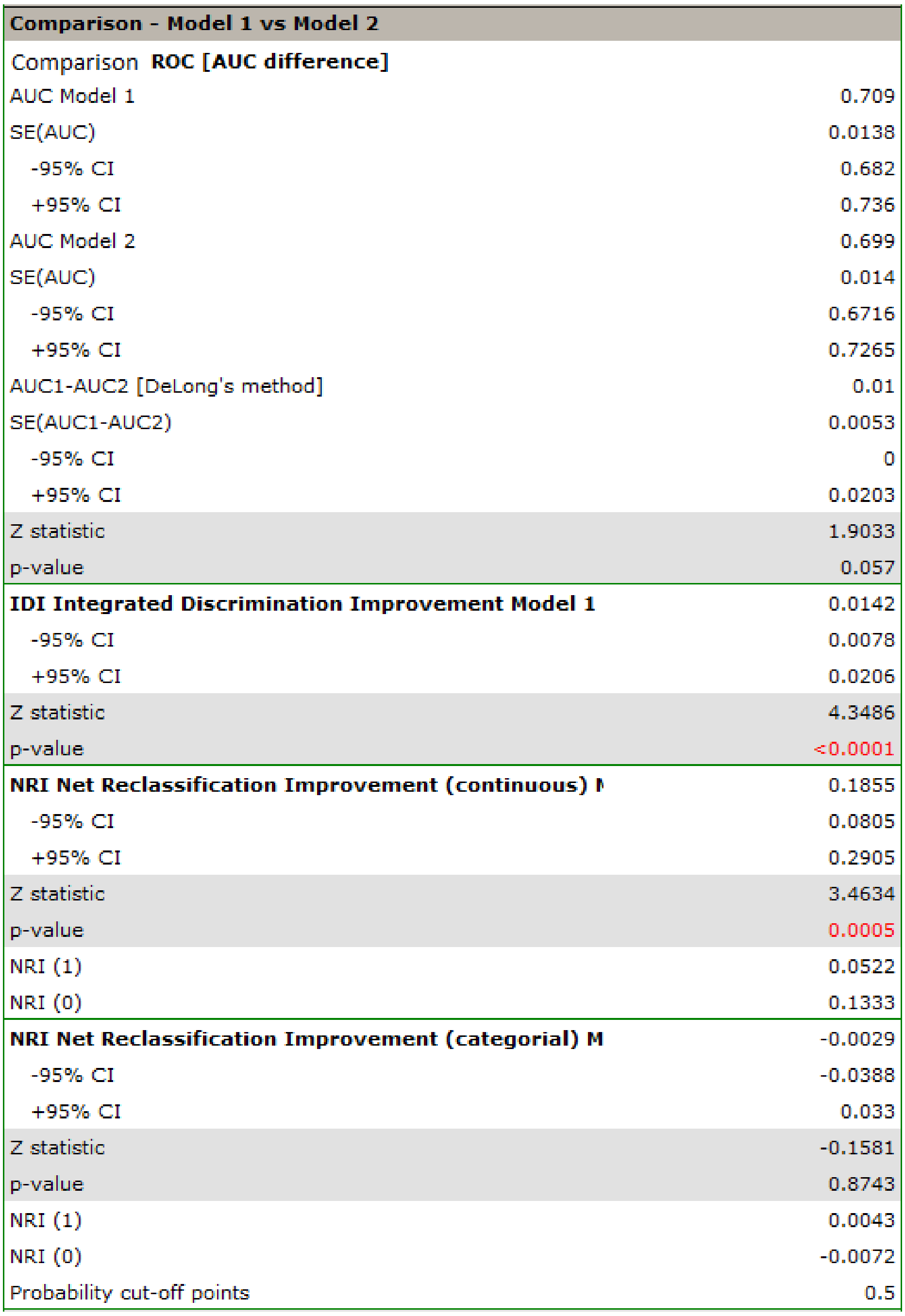
- **Automatic** model comparison.
  For automatic model comparison, we obtained very similar results. The best model is the model built on the independent variables: AGE, EDUCATION, TIME OF SOLUTION, and DISTURBANCES.

Based on the above analyses, from a statistical point of view, the optimal model is one containing the 4 most important independent variables: AGE, EDUCATION, RESOLUTION TIME, and DISTURBANCES. Its detailed analysis can be done in the Logistic Regression module. However, the final decision which model to choose is up to the experimenter.

Example 24.6. **Example**(Heart.pqs file)

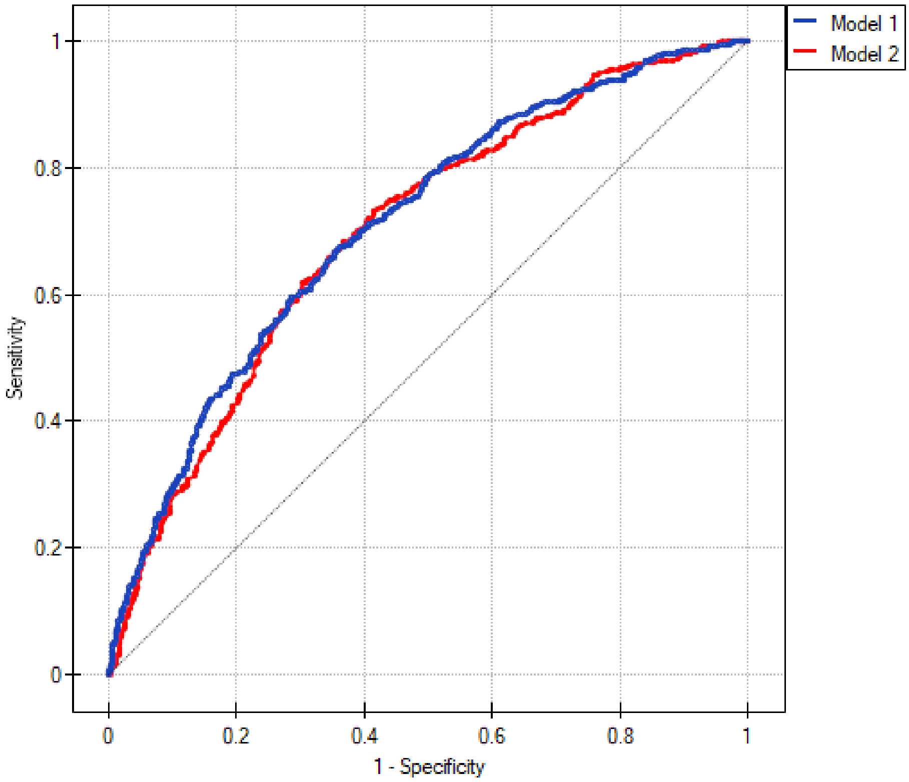
Risk factors for certain heart disease such as age, bmi, smoking, LDL fraction cholesterol, HDL fraction cholesterol, and hypertension were examined. From the researcher's point of view, it was interesting to determine how much information about smoking could improve the prediction of the occurrence of the disease under study.

We compare a logistic regression model describing the risk of heart disease based on all study variables with a model without smoking information. In the analysis window, we select the options related to the prediction evaluation, namely the ROC curve and the NRI coefficients. In addition, we indicate to include all proposed graphs in the report.

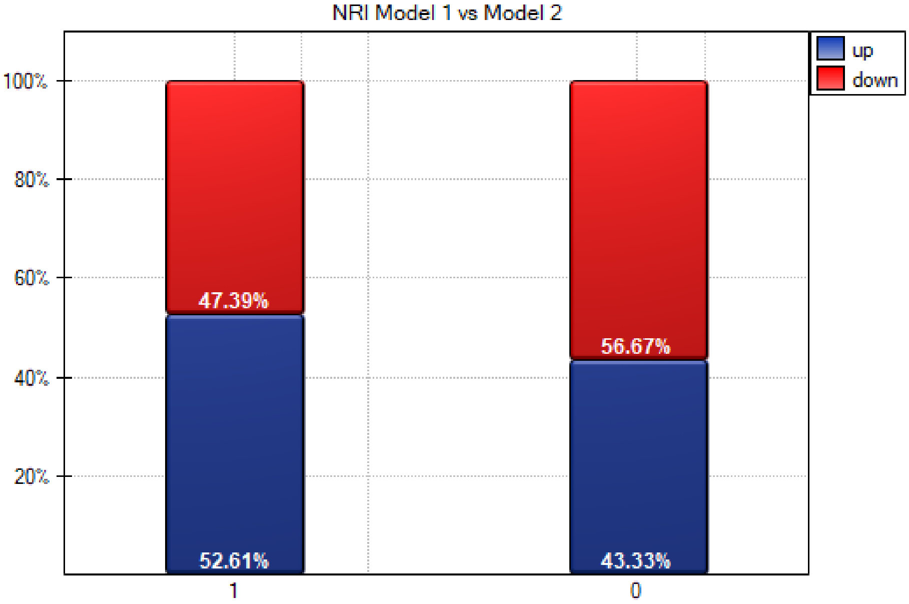


Comparison - Model 1 vs Model 2	
Comparison **ROC [AUC difference]**	
AUC Model 1	0.709
SE(AUC)	0.0138
-95% CI	0.682
+95% CI	0.736
AUC Model 2	0.699
SE(AUC)	0.014
-95% CI	0.6716
+95% CI	0.7265
AUC1-AUC2 [DeLong's method]	0.01
SE(AUC1-AUC2)	0.0053
-95% CI	0
+95% CI	0.0203
Z statistic	1.9033
p-value	0.057
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

Analysis of the report indicates important differences in prediction as a result of adding smoking information to the model, although they are not significant in describing the ROC curve (p=0.057).



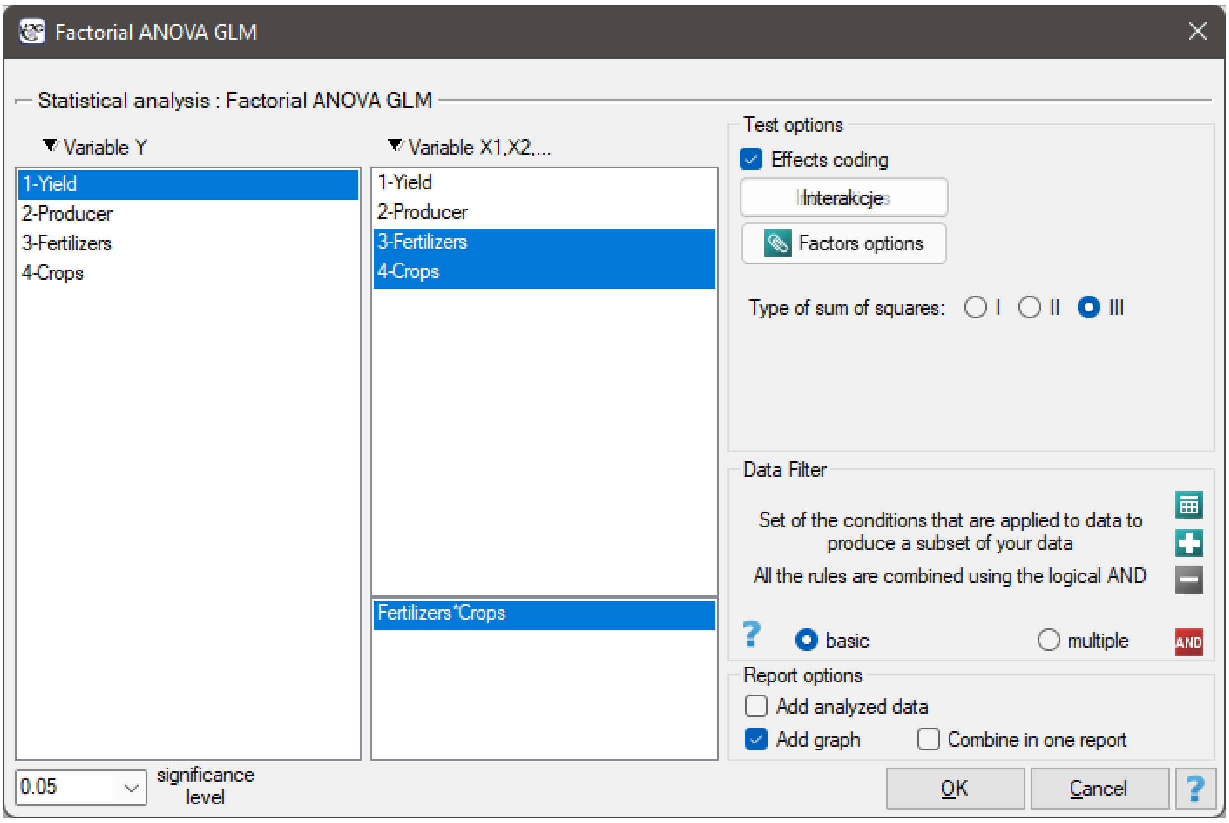
The continuous IDI and NRI coefficient values indicate a statistically significant and favorable change (the values of these coefficients are positive with p<0.05). The prognosis for those with heart disease improved by more than 5% and those without heart disease by more than 13% (NRI(sick)=0.0522, NRI(healthy)=0.1333)) as a result of including information about smoking.



We also see the conclusions drawn from the NRI in the graph. We see an increase in the model-predicted probability of disease in diseased individuals (more individuals were reclassified upward than downward 52.61It is also possible to determine a categorical NRI, but to do so, one would first need to determine the model-determined probability cut-off points accepted in the heart disease literature.

24.6 FACTORIAL ANOVA - GLM

The settings window with the GLM factorial ANOVA can be opened in Advanced statistics menu→Multivariate models→Factorial GLM ANOVA.



Factorial ANOVA GLM

Statistical analysis : Factorial ANOVA GLM

Variable Y: 1-Yield

Variable X1,X2,...: 3-Fertilizers, 4-Crops

Fertilizers*Crops

Test options

☑ Effects coding

Type of sum of squares: ○ I ○ II ● III

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data

☑ Add graph □ Combine in one report

0.05 significance level

Factor analysis of variance GLM is an extension of univariate analysis of variance (ANOVA) for independent groups and linear multiple regression. The acronym GLM (general linear model) reads as General Linear Model. GLM analysis typically involves the use of linear regression models in the calculation of various composite ANOVA comparisons.

Example

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 is about the *income* of a certain group of people. About the surveyed people we have some additional information like gender and education..

income and...	ANOVA	Multiple regression
gender	**t-test for independent groups** ⋆ comparing income for women with income for men	**gender (X) dependence of income (Y)** ⋆ gender as a column with two categories
education	**One-way ANOVA** ⋆ income comparison for individuals with different education (primary, vocational, secondary, higher)	**education (X) dependence of income (Y)** ⋆ education broken down into dummy variables
gender, education	**Multifactorial ANOVA** ⋆ comparing income for women with income for men with education as a confounding variable ⋆ income comparison for people with different education (primary, vocational, secondary, higher) with gender as a confounding variable	**gender (X_1) and education (X_2) dependence of income (Y)** ⋆ gender as a column with two categories ⋆ education broken down into dummy variables both variables are mutually confusing
gender, education, gender*education	**Multifactorial ANOVA** ⋆ comparing women's income with men's income taking into account other variables in the model ⋆ income comparison for people with different education (primary, vocational, secondary, higher) taking into account other variables in the model ⋆ income comparison for people with different education and gender (in interaction) taking into account other variables in the model	**gender (X_1), education (X_2), interaction of gender and education ($X_1 * X_2$) dependence of income (Y)** ⋆ gender as a column with two categories ⋆ education broken down into dummy variables ⋆ interaction is the product of gender and education variables in the model are mutually confusing

GLM analysis can be used in any of the above cases; however, because multivariate regression analysis as well as one-way ANOVA have been discussed in separate chapters, in this section we will present the use of GLM in multifactorial ANOVA.

Factorial ANOVA is a kind of analysis of variance, in which we can use both one and many factors to separate compared groups. Variables that are interactions of the indicated factors may also be involved in the analysis. When ANOVA includes more than one factor, the factors are entangled with each other.

Influence of confounding factors

Although all factors involved in the analysis are confounded with each other, their influence on the significance of individual factors can be controlled. There are three ways by which the influence of the entangling variables can be taken into account when examining the significance of individual factors. They depend on how the sum of squares is determined:

- **Type I sums of squares**

  Type I sums of squares depend on the order in which the factors are placed in the model. This type of sum of squares means that the significance of the factor that we interpret is adjusted by those variables whose order in the model was earlier, the other variables in the model only indirectly affect the result of the analysis. For example: if we place 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 into account the whole model (through sums of squares for error) but only factors are used explicitly as confounding variables: A, B, $A * B$, C.

  The sums of squares for the factor $A * C$ are then calculated as follows:

  $$SS(A * C) = SS(A, B, A * B, C, A * C, B * C, A * B * C, D) - SS(A, B, A * B, C)$$

  Using the sum of squares type I

  > **Indications:** When the study is fully balanced, with equal or proportional counts of each category, including when there are interactions.

Contraindications: When the study is unbalanced (different counts of each category) and/or there are interactions.

- **Type II sums of squares**
  This type of sum of squares means that the significance of the factor we interpret is corrected for those variables whose order is the same or lower, the other variables in the model only indirectly affect the result of the analysis. For example: if we include factors in the model: A, B, $A*B$, C, $A*C$, $B*C$, $A*B*C$, D, then the significance for the factor $A*C$ takes into account the whole model (via sums of squares for the error) but the first order variables are used explicitly as confounding variables: A, B, C, D and all other second order variables: $A*B$, $B*C$.
  The sums of squares for the factor $A*C$ are then calculated as follows:

$$SS(A*C) = SS(A, B, A*B, C, A*C, B*C, A*B*C, D) - SS(A, B, A*B, C, B*C, D)$$

Using the sum of squares type II

Indications: When the study is fully balanced, with equal or proportional counts of each category, including when there are interactions.
Contraindications: When the study is unbalanced (different counts of each category) and/or there are interactions.

- **Type III sums of squares**
  We recommend using this type of coding when effect coding is selected.
  This type of sum of squares causes the significance of the factor we interpret to be adjusted for all other variables in the model. For example: if we include factors in the model: A, B, $A*B$, C, $A*C$, $B*C$, $A*B*C$, D, then the significance for the variable $A*C$ takes into account the entire model (via sums of squares for error) and all factors except the one under study are used explicitly as the confounding variables: A, B, $A*B$, C, $B*C$, $A*B*C$, D.
  The sums of squares for the factor $A*C$ are then calculated as follows:

$$SS(A*C) = SS(A, B, A*B, C, A*C, B*C, A*B*C, D) - SS(A, B, A*B, C, B*C, A*B*C, D)$$

Using the sum of squares type III

Indications: When the study is balanced or unbalanced, including when there are interactions.
Contraindications: When the test contains subclasses with missing observations.

In PQStat, Type III sums of squares are selected by default because of their universality. Also selected by default is the effect coding option described in the 24.1 section. Note that the choice of the appropriate coding affects both the interpretation of model coordinates and the significance of individual factors in a factorial ANOVA - especially with unbalanced systems.

Basic assumptions:

- measurement on an interval scale,
- the samples come from a population with a normal distribution (normality of the variables or residuals of the model),
- an independent model,
- equality of variances of an analysed variable in all populations.

A factorial ANOVA requires that the factors be divided into categories (i.e., independent populations), e.g., factor A: gender is divided into male and female, factor B: education into primary, vocational, secondary and higher education. The interaction of factor $A*B$ is also divided into categories, in this case we will get eight categories:
1) female with primary education,
2) female with vocational education,
3) female with secondary education,
4) female with higher education,
5) male with primary education,
6) male with vocational education,
7) male with secondary education,
8) male with higher education,

ANOVA type analysis and regression models are treated equivalently, and in the general case their hypotheses converge. We will present hypotheses for the main effects A and B and the interaction effect $A*B$ in both of these approaches. In interpreting these hypotheses, it is important to remember that the hypotheses for the factors in question are corrected for those of the other factors that a given analysis includes

ANOVA approach

Hypotheses for the factor A:

$$\begin{array}{cl}\mathcal{H}_0: & \mu_1=\mu_2=...=\mu_a,\\ \mathcal{H}_1: & \text{not all } \mu_i \text{are equal to each other } (i=1,2,...,a),\end{array}$$

where:
μ_1,μ_2,...,μ_a – averages of the factor A for its individual categories.

Hypotheses for the factor B:

$$\begin{array}{cl}\mathcal{H}_0: & \mu_1=\mu_2=...=\mu_b,\\ \mathcal{H}_1: & \text{not all } \mu_i \text{are equal to each other } (i=1,2,...,a),\end{array}$$

where:
μ_1,μ_2,...,μ_b – śaverages of the factor B for its individual categories.

Hypotheses for the interaction of factors $A*B$:

$$\begin{array}{cl}\mathcal{H}_0: & \mu_1=\mu_2=...=\mu_{ab},\\ \mathcal{H}_1: & \text{not all } \mu_i \text{are equal to each other } (i=1,2,...,a),\end{array}$$

where:
μ_1,μ_2,...,μ_{ab} – the average interaction of the factors $A*B$ for their respective categories.

Regression approach
The model approach assumes that the regression model works

$$Y=\mu+\alpha_iA_i+\beta_jB_j+(\alpha\beta)_kB_k+\epsilon.$$

where:
Y – dependent variable, explained by the model,
μ – Overall mean of the Y variable (if effect coding was used)
A_i,B_j,AB_k – factors - independent, explanatory variables,
$\alpha_i,\beta_i,\alpha\beta_i,$ – parameters,
ϵ – random component (model residual).

Hypotheses for the factor A:

$$\begin{array}{ll}\mathcal{H}_0: & \alpha_1=\alpha_2=...=\alpha_a=0,\\ \mathcal{H}_1: & \text{not all } \alpha_i=0,\end{array}$$

Hypotheses for the factor B:

$$\begin{array}{ll}\mathcal{H}_0: & \beta_1=\beta_2=...=\beta_b=0,\\ \mathcal{H}_1: & \text{not all } \beta_j=0\text{ ,}\end{array}$$

Hypotheses for the interaction of factors $A*B$:

$$\begin{array}{ll}\mathcal{H}_0: & (\alpha\beta)_1=(\alpha\beta)_2=...=(\alpha\beta)_{ab}=0,\\ \mathcal{H}_1: & \text{not all } (\alpha\beta)_k=0,\end{array}$$

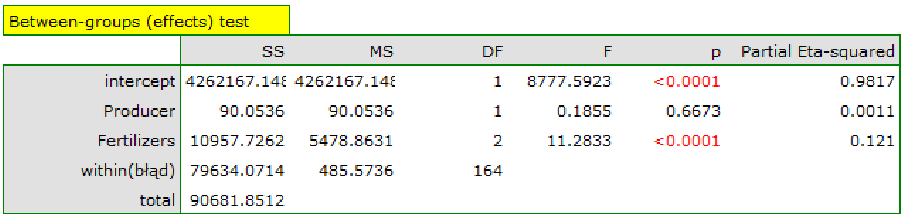
Coding
The analysis results obtained (in particular the regression model built) and the interpretation of the hypotheses also depend on the coding method. The PQStat program offers **zdummy coding** i **effect coding**. For a detailed description of the coding, check the section Preparing variables for analysis in multivariate models. By default, the program selects effect coding. Unchecking this option is equivalent to selecting dummy coding.
Note
When using type III sum of squares, when there are interactions, it is advisable to use effect coding.

Example 24.7. (yield.pqs file)
In order to increase the yield of crops, fertilizers are being developed according to newer and newer technologies. Based on an experiment, the researchers want to find out which of the three blends of new fertilizers is the most effective. The crops were grown by two different farms and involved sowing wheat, rye, oats and barley. Yield was reported in % (compared to the yield obtained without fertilization).

First, we want to check whether:
1) H0: The average yields obtained when fertilizing with compound X are the same as those obtained when fertilizing with compound Y and the same as those obtained when fertilizing with compound Z (regardless of the crop farm).
In addition, although it is of urban interest in this case, we will check whether:
2) H0: The average yields obtained in farm 1 are the same as in farm 2 (regardless of the blend of fertilizer used).
Equivalently, these hypotheses can be written using a regression approach:
1) H0: The coefficients indicating the change in yield with a change in fertilizer application are zero (regardless of the farm growing the crop).
2) H0: The coefficient indicating the change in the yield obtained when changing crop farms is zero (regardless of the fertilizer blend used).

In a second application of GLM, we will check whether:
3) H0: The average yields for each cereal are the same when using different fertilizer applications.

Hypotheses 1) i 2)
ANOVA approach
We will conduct the analysis using the third type of sum of squares and effect coding.

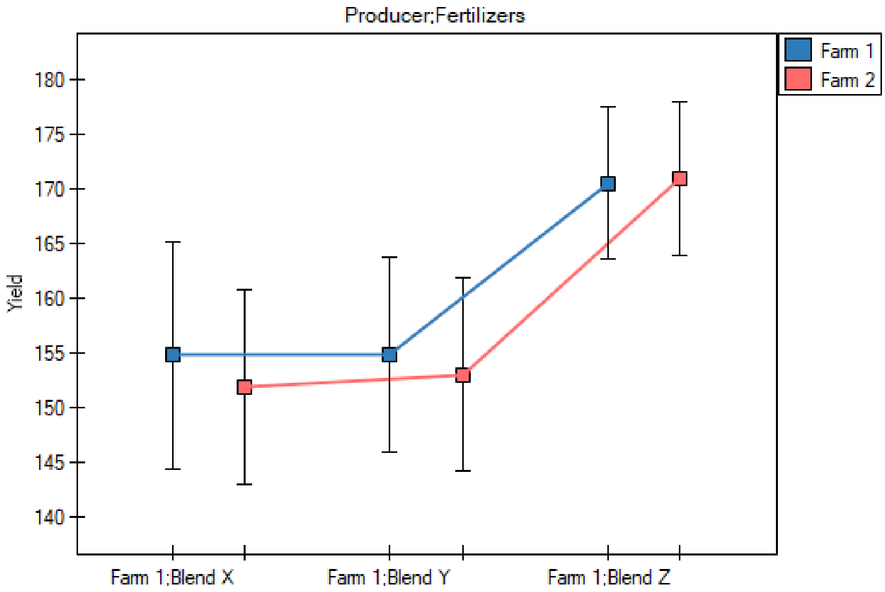


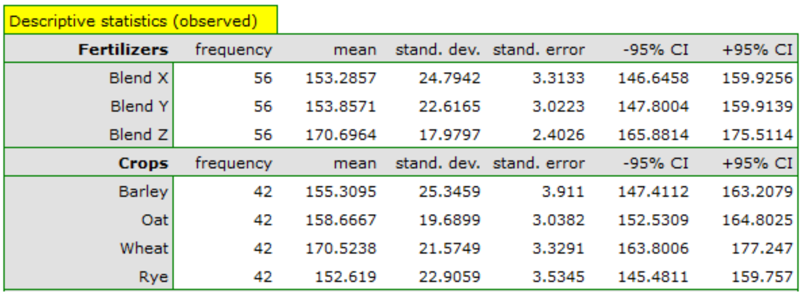
Between-groups (effects) test

	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					

We observe statistically significant differences between the yield obtained with different fertilizer blends (p<0.0001). The fertilizer blend used explains the variation in yield obtained in about 12% as evidenced by the value of partial Eta-square. In contrast, the yields obtained did not depend on the farm where the crop was grown (p=0.6673, partial Eta-square = 0,1%).

After selecting the observed or expected averages in the Factors Options, indow, we can graphically represent these differences in graphs showing the average yields when each fertilizer blend is applied. Exact values of the averages can be read from the table of descriptive statistics.

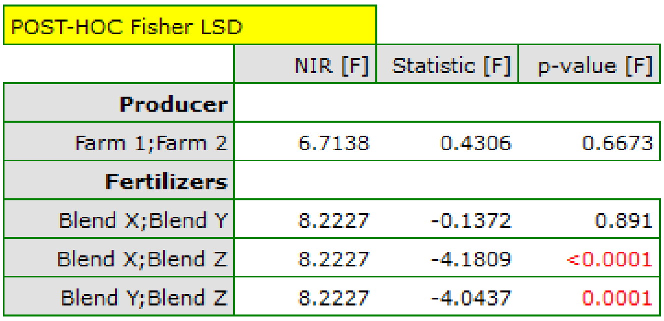




Descriptive statistics (observed)

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

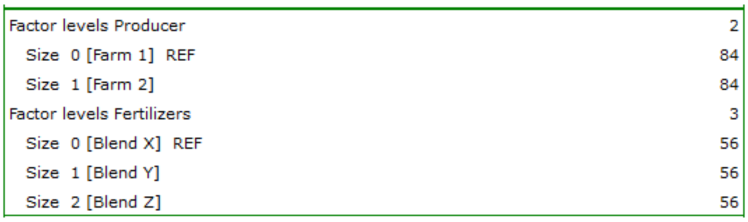
We can check where the differences are located by using post-hoc tests. Fisher's NIR post-hoc test indicates that the most favorable results are obtained with the use of blend Z - the yield obtained is on average 170.7 % of the yield that would be obtained without the use of fertilization. The remaining blends did not differ statistically significantly in the yield obtained. Since in the model, the farm where the crops were grown was analysed simultaneously, we can say that the advantage of the Z blend is independent of the farm where the sowing was done.



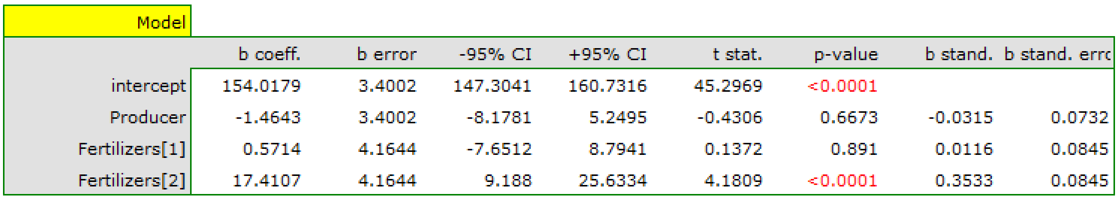
POST-HOC Fisher LSD	NIR [F]	Statistic [F]	p-value [F]
Producer			
Farm 1;Farm 2	6.7138	0.4306	0.6673
Fertilizers			
Blend X;Blend Y	8.2227	-0.1372	0.891
Blend X;Blend Z	8.2227	-4.1809	<0.0001
Blend Y;Blend Z	8.2227	-4.0437	0.0001

Podejście regresyjne

An analogous interpretation will be obtained using the regression model, although here the interpretation 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 unselecting the effect coding option. The analysis automatically took alphabetically the first level as the reference level. For fertilizers, this level was compound X, for farms it was farm 1.



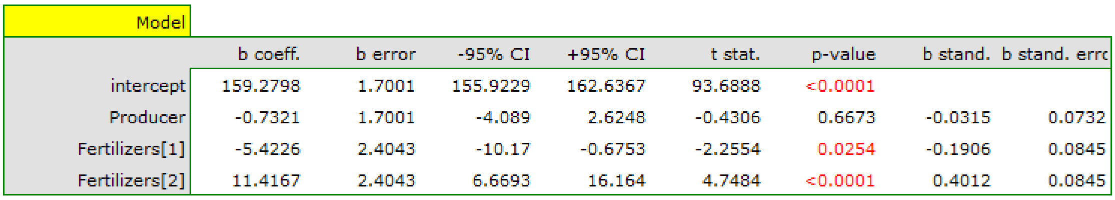
Factor levels Producer	2
Size 0 [Farm 1] REF	84
Size 1 [Farm 2]	84
Factor levels Fertilizers	3
Size 0 [Blend X] REF	56
Size 1 [Blend Y]	56
Size 2 [Blend Z]	56



Model	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. errc
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, except that we compare only to the reference category. So if we compare all the fertilizer blends to blend X we can see that only using blend Z produced significantly higher results (p< 0.0001). These results are higher by 17.4107 (recall that the means were respectively (153.285714 - for blend X, 170.696429 - for blend Z). When comparing farms, the matter is simple, because we have only two farms to compare, and the result is the result of comparing farm 2 with farm 1, which was the reference category. This time the obtained difference was small (-1.4643) and not statistically significant (0.6673).

Using effect coding, we also choose a reference category, but the magnitude of the coefficients and their significance is not related to the chosen reference category but to the overall average yield obtained, recorded in the model as an intercept (159.2798).

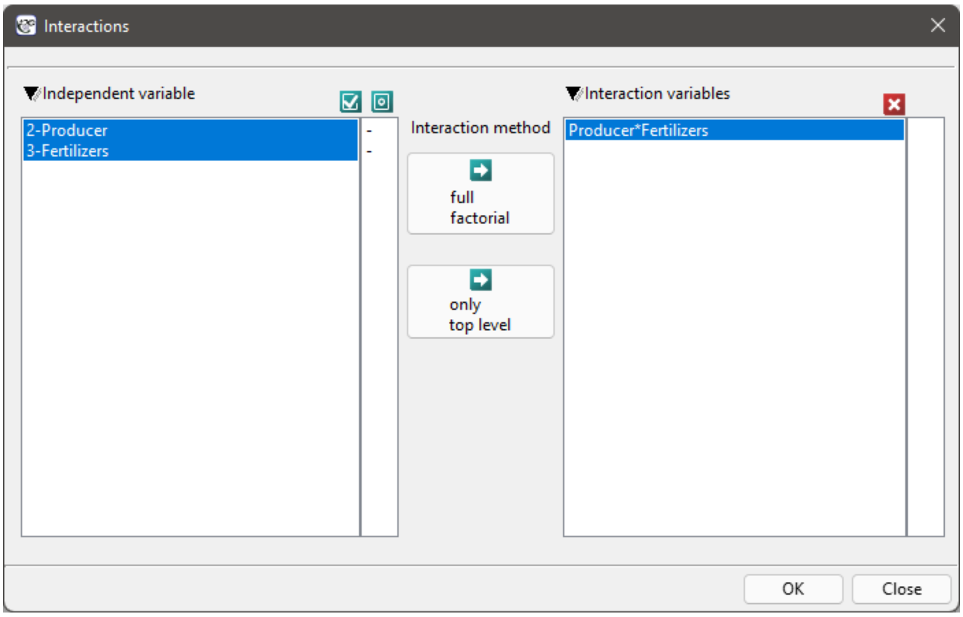


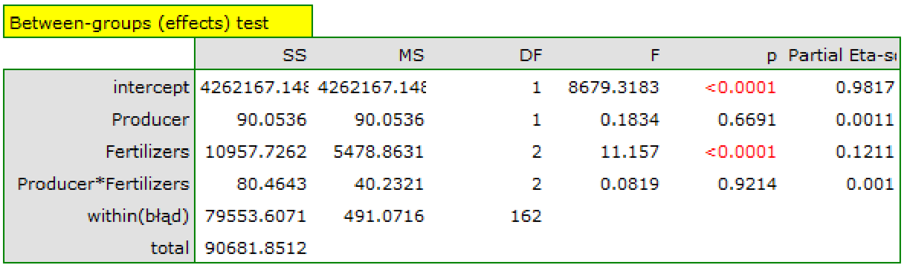
Model	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. errc
intercept	159.2798	1.7001	155.9229	162.6367	93.6888	<0.0001		
Producer	-0.7321	1.7001	-4.089	2.6248	-0.4306	0.6673	-0.0315	0.0732
Fertilizers[1]	-5.4226	2.4043	-10.17	-0.6753	-2.2554	0.0254	-0.1906	0.0845
Fertilizers[2]	11.4167	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 by 5.4226 lower than the overall average, and with blend Z by 11.4167 higher. Both differences are statistically significant.

An indisputable advantage of building a regression model is the possibility of using its formula in predicting the yields obtained. The built models are presented as follows:
For dummy coding:
$yield = 154.0179 + 0.5714 \cdot BlendY + 17.4107 \cdot BlendZ - 1.4643 \cdot Farm2$
Dla kodowania efektów:
$yield = 159.2798 - 5.4226 \cdot BlendY + 11.4167 \cdot BlendZ - 0.7321 \cdot Farm2$
To be able to use the selected model in forecasting, you need to go to the menu multiple regression - prediction and make a prediction based on the new data. The preparation of the data depends on how it is encoded.

Based on all the results obtained, we do not suspect that yield is dependent on the interaction between the types of fertilizers used and the crop farm. Most often, the presence of an interaction can be seen in the graph as clearly intersecting lines. Here, the two lines were nearly parallel and close enough to each other that the difference between farms was not statistically significant. Although intersecting lines usually indicate the presence of interaction, it should be remembered that when the lines are close to each other their accidental crossing is very likely and as a result 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 located on the right side of the window and then repeat the analysis.



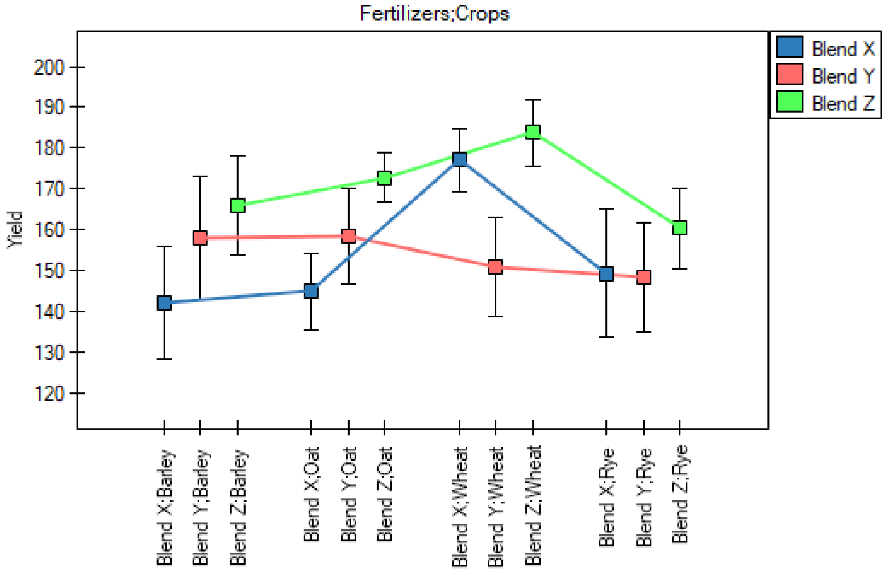


Between-groups (effects) test

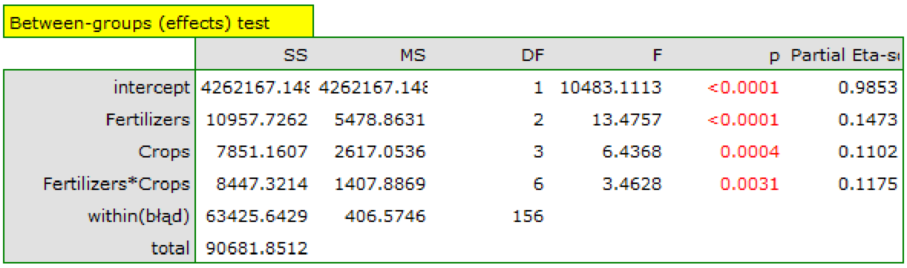
	SS	MS	DF	F	p	Partial Eta-s
intercept	4262167.148	4262167.148	1	8679.3183	<0.0001	0.9817
Producer	90.0536	90.0536	1	0.1834	0.6691	0.0011
Fertilizers	10957.7262	5478.8631	2	11.157	<0.0001	0.1211
Producer*Fertilizers	80.4643	40.2321	2	0.0819	0.9214	0.001
within(błąd)	79553.6071	491.0716	162			
total	90681.8512					

The result confirmed our assumption of no significant interaction (p=0.9214). Thus, a simpler model, i.e. without interaction, is recommended in this case.

Hypothesis 3)
The situation is different when we examine the yield obtained depending on the dose of fertilizer applied and the type of cereal grown.



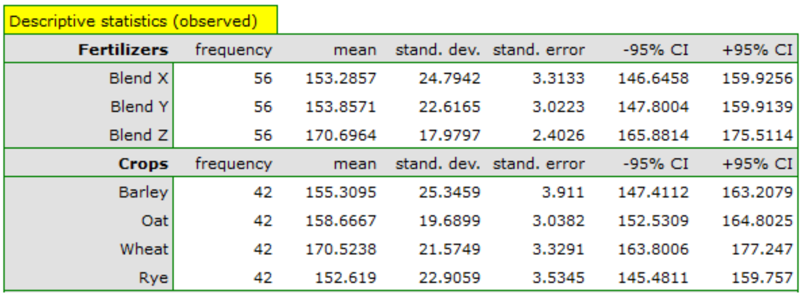
We will perform an analysis that considers interactions in addition to main effects.



Between-groups (effects) test

	SS	MS	DF	F	p	Partial Eta-s
intercept	4262167.148	4262167.148	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					

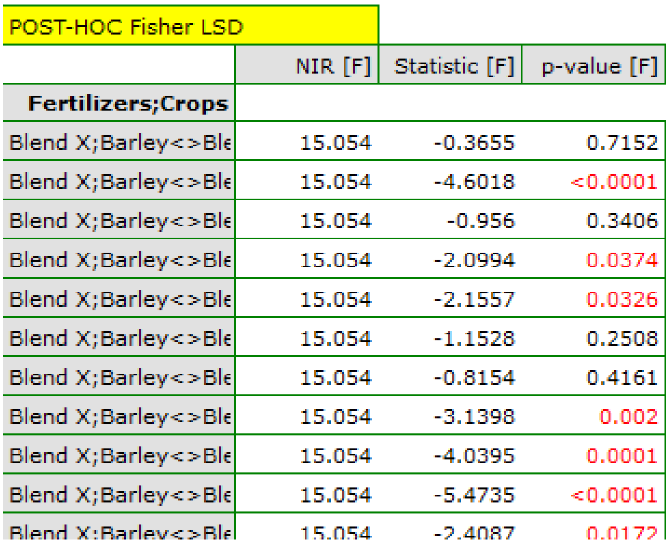
Since the interactions in the built model are statistically significant (p=0.0031), it is the model with interactions that we should use and the description of the obtained results should focus on this interaction.:
H0: The average yields obtained when fertilizing wheat with blend X are the same as when fertilizing wheat with blend Y and the same as when fertilizing wheat with blend Z and the same as when fertilizing rye with blend X and the same as when fertilizing rye with blend Y and the same as when fertilizing rye with blend Z and the same as fertilizing oats with blend X and the same as fertilizing oats with blend Y and the same as fertilizing oats with blend Z and the same as fertilizing barley with blend X and the same as fertilizing barley with blend Y and the same as fertilizing barley with blend Z.
In the regression approach we will say that:
H0: he coefficients that determine the change in yield obtained with a change in fertilizer applied and a change in crop type are zero.
On the basis of the graph (and the averages in the table) we can see that by far the best yields are obtained with the Z blend, irrespective of the type of cereal grown.



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

In contrast, blend X and blend Y yield worse than blend Z and, in addition, there is an interaction effect between them. It manifests itself in the fact that a wheat crop yields abnormally high when blend X is applied compared to the wheat yield obtained when blend Y is applied, while a barley and oat crop yields better when blend Y is applied. We can check the differences obtained more precisely by performing post-hoc tests. An excerpt from this report is given below:



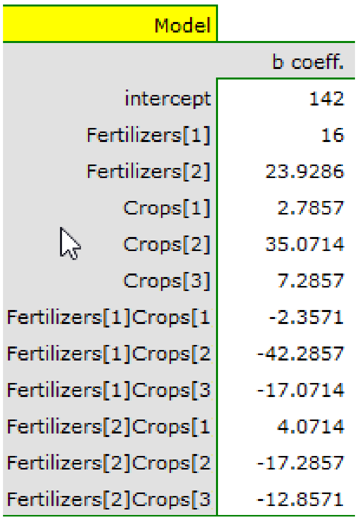
POST-HOC Fisher LSD

Fertilizers;Crops	NIR [F]	Statistic [F]	p-value [F]
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 yield advantage obtained when using blend Z for any crop and blend Y for wheat crop.

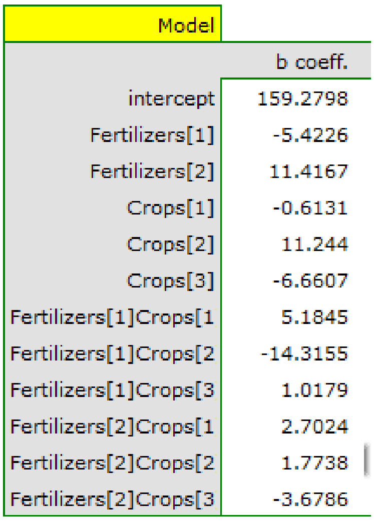
We can use the coefficients of the he regression model for prediction via the multiple regression - prediction menu remembering to code the new data appropriately depending on the model selected..

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

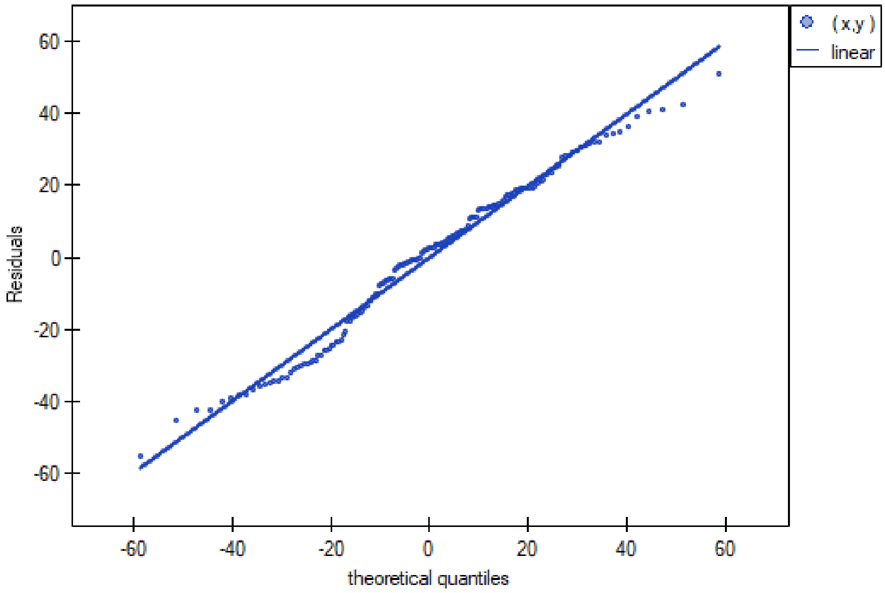
effect coding

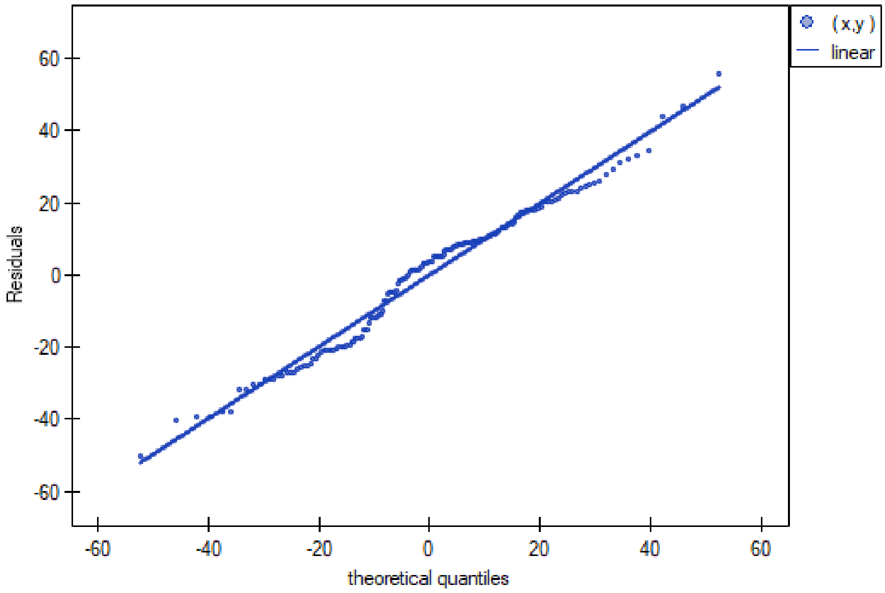


Model	b coeff.
intercept	159.2798
Fertilizers[1]	-5.4226
Fertilizers[2]	11.4167
Crops[1]	-0.6131
Crops[2]	11.244
Crops[3]	-6.6607
Fertilizers[1]Crops[1	5.1845
Fertilizers[1]Crops[2	-14.3155
Fertilizers[1]Crops[3	1.0179
Fertilizers[2]Crops[1	2.7024
Fertilizers[2]Crops[2	1.7738
Fertilizers[2]Crops[3	-3.6786

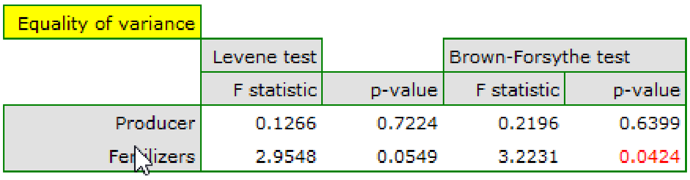
VERIFICATION OF ASSUMPTIONS

Checking the main assumptions will involve comparing the variances and visually determining the normality of the model residuals.

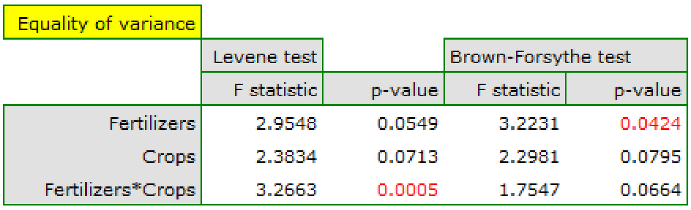




The normality plot of the residuals (Q-Q plot) for the first and for the second analysis shows the residuals of the model well distributed around a straight line, indicating a good fit of the residuals to the normal distribution. The comparison of variances is performed by the Levenea or Brown-Forsythe 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



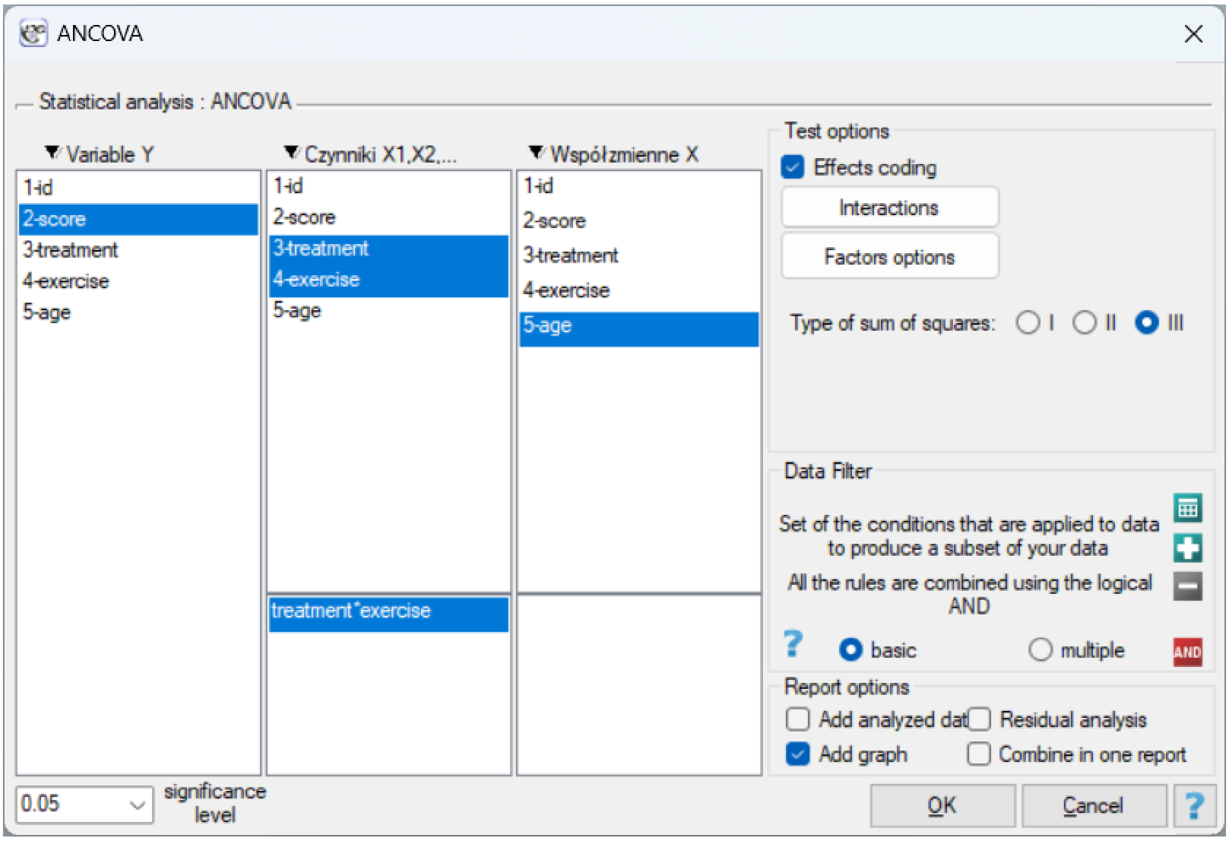
Equality of variance	Levene test		Brown-Forsythe test	
	F statistic	p-value	F statistic	p-value
Fertilizers	2.9548	0.0549	3.2231	0.0424
Crops	2.3834	0.0713	2.2981	0.0795
Fertilizers*Crops	3.2663	0.0005	1.7547	0.0664

24.7 ANCOVA

Analysis of covariance (ANCOVA) is a method of testing the hypothesis that the means of two or more populations are equal, in correction for other continuous variables. These adjustments result in effects more readily seen by researchers than those obtained through ANOVA, i.e., narrower confidence intervals and greater statistical power.
Suppose an experiment is conducted to evaluate the effects of two treatments. The groups randomly assigned to treatment differ slightly in mean age, which also affects the treatment effect. Differences between groups in achievement will be quite ambiguous to interpret, since the groups differ in both age and treatment conditions. Analysis of covariance will provide ”adjusted averages”, which estimate what the mean scores would be if the groups were exactly the same in terms of age. At the same time, the within-group variability of the results due to the variable (age) will be removed from the error variability to increase the precision of the test of the differences between the adjusted averages.

The label ”analysis of covariance” is now seen as anachronistic by some research methodologists and statisticians, since this analysis is not a separate analysis but a variant of the general linear model (GLM). However, the term is still useful because it immediately conveys to most researchers the notion that a categorical variable (e.g., treatment conditions) and a continuous variable (e.g., age) are involved in a single analysis that determines treatment outcome.

The settings window with the ANCOVA can be opened in Advanced statistics→Multivariate models→ANCOVA



ANCOVA

Statistical analysis : ANCOVA

Variable Y: 1-id, 2-score (selected), 3-treatment, 4-exercise, 5-age

Czynniki X1,X2,...: 1-id, 2-score, 3-treatment (selected), 4-exercise (selected), 5-age

treatment*exercise

Współzmienne X: 1-id, 2-score, 3-treatment, 4-exercise, 5-age (selected)

Test options

☑ Effects coding

Type of sum of squares: ○ I ○ II ● III

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed dat □ Residual analysis

☑ Add graph □ Combine in one report

0.05 significance level

Note!!!
How to take into account the study factors and confounding variables is described in the section on multivariate ANOVA (Influence of confounding factors). The recommended way is to choose Sum of Squares type III and effects coding.

Basic application conditions:

- measurement on an interval scale,
- the samples come from a population with a normal distribution (normality of the variables or residuals of the model),
- an independent model,
- equality of variances of an analysed variable in all populations,
- Equality of the slopes of the regression lines (regression coefficients between each confounding variable and the dependent variable) for each possible factor level.
  Note!
  Equality of the slopes of the regression lines is tested using the F test comparing the model containing the analyzed factors with the same model, but augmented by interactions with the confounding factors. A statistically significant result means that the assumption of equal slopes is violated, because the interaction becomes significant, so the different slopes of the simple.

ANCOVA hypotheses for a single factor A:

$$\begin{array}{cl}\mathcal{H}_0: & \mu_1=\mu_2=...=\mu_a,\\ \mathcal{H}_1: & \text{not all } \mu_i \text{ are equal } (i=1,2,...,a),\end{array}$$

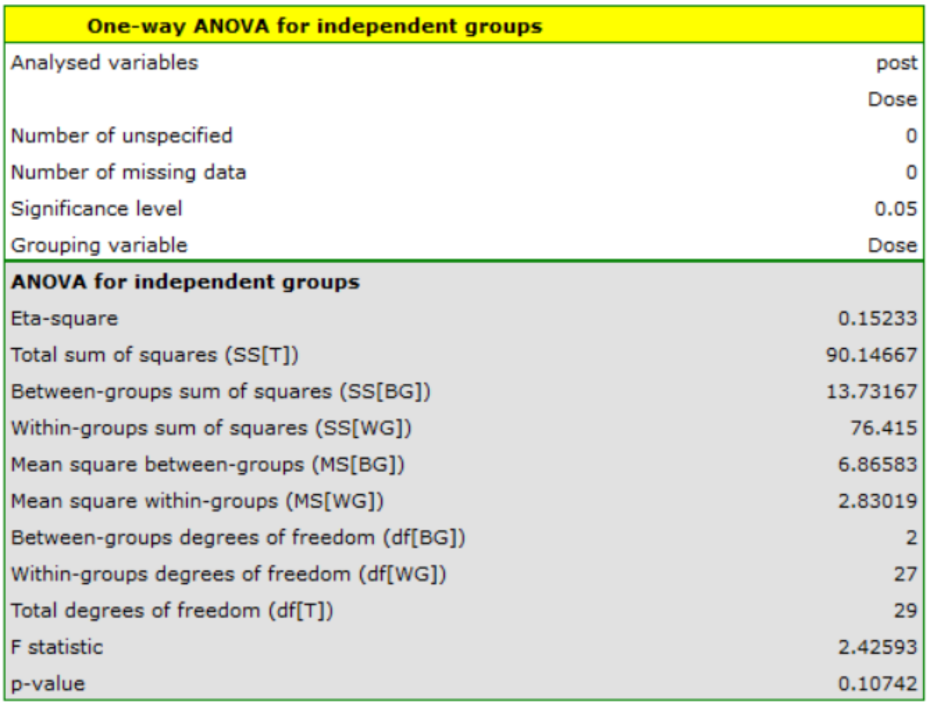
where:
μ_1,μ_2,...,μ_a - expected averages of the factor A for each of its categories.

ANCOVA hypotheses for factor interactions $A*B$:

$$\begin{array}{cl}\mathcal{H}_0: & \mu_1=\mu_2=...=\mu_{ab},\\ \mathcal{H}_1: & \text{not all } \mu_k \text{ are equal } (k=1,2,...,ab),\end{array}$$

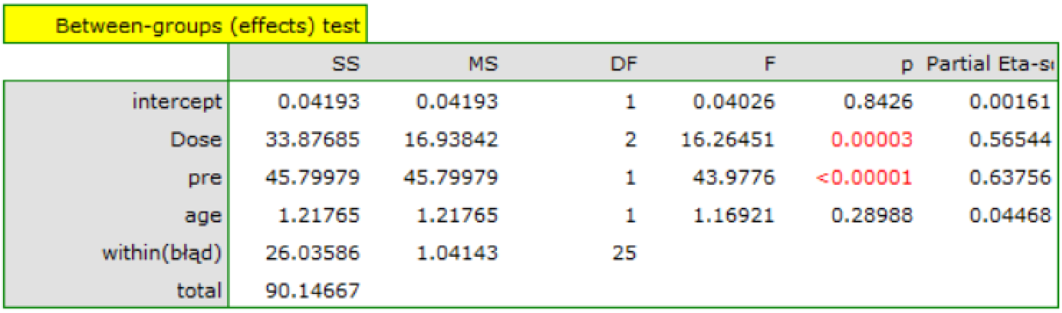
where:
μ_1,μ_2,...,μ_{ab} - expected average interactions of $A*B$ factors for their respective categories.

Example 24.8. (drug cholesterol.pqs file)
Imagine that a researcher was conducting a study on a new cholesterol-lowering drug. The study was designed so that the dose of the drug occurred at three levels:
high, low and placebo. The researcher tested (using ANOVA independent) whether cholesterol after treatment differed according to the dose of the drug.



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

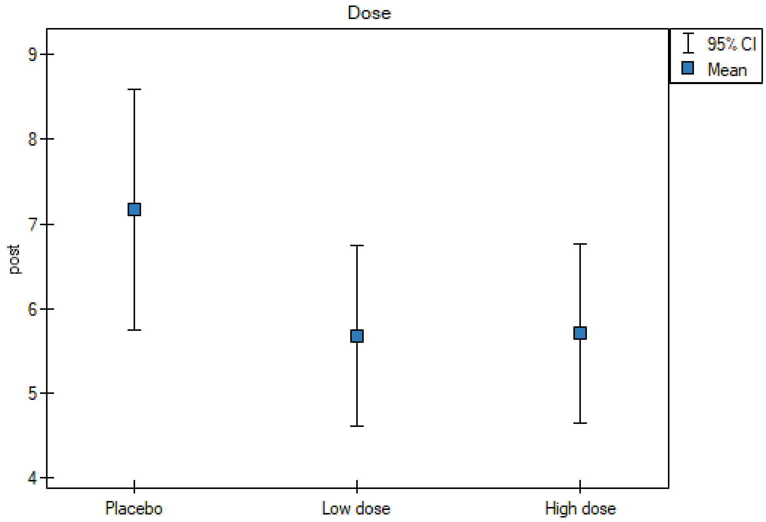
Unfortunately, the researcher did not get confirmation of the differences between the results.
Let's imagine that the researcher, realized that whether a drug would change cholesterol levels might be related to the patient's baseline cholesterol level and age. Therefore, he decided to perform a univariate ANCOVA (the factor is the dose of the drug) taking into account pre-treatment cholesterol levels and age as co-variables.
This time, the ANOCVA result indicated that there were significant differences between cholesterol levels after different doses of the drug (p=0.00003):

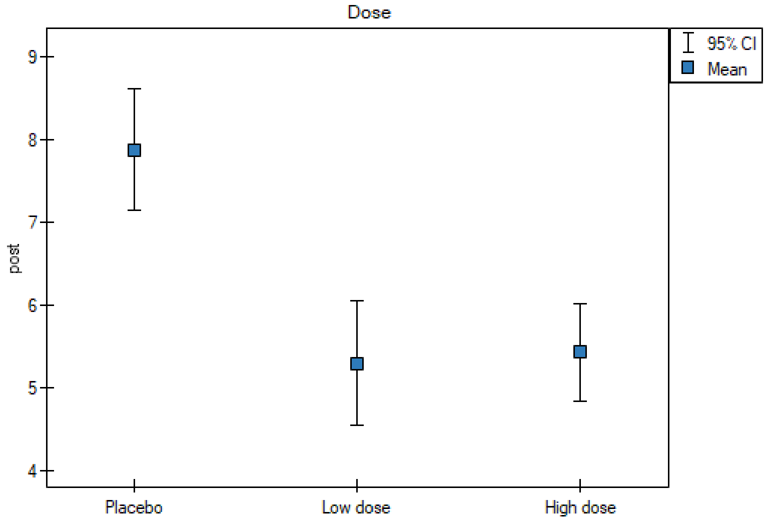


Between-groups (effects) test

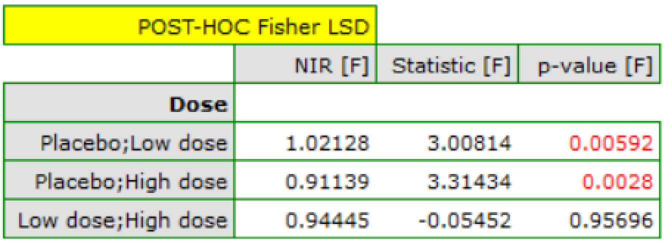
	SS	MS	DF	F	p	Partial Eta-s
intercept	0.04193	0.04193	1	0.04026	0.8426	0.00161
Dose	33.87685	16.93842	2	16.26451	0.00003	0.56544
pre	45.79979	45.79979	1	43.9776	<0.00001	0.63756
age	1.21765	1.21765	1	1.16921	0.28988	0.04468
within(błąd)	26.03586	1.04143	25			
total	90.14667					

Including pre-test cholesterol levels reduced the obtained errors for the averages and narrowed the confidence intervals. To display the observed or expected averages, I choose the appropriate settings via Factor Options, to which I select the error graph. The first graph shows the observed averages with confidence interval, i.e., not including the effect of age and pre-treatment cholesterol levels; the second graph is the expected averages based on the built model with confidence intervals, i.e., after accounting for the effect of these two co-variables:

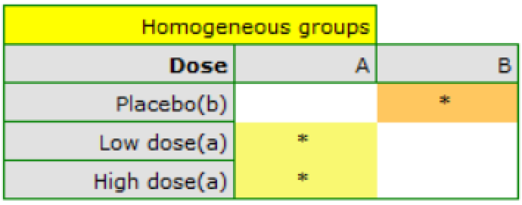




As a result, by taking into account the cholesterol level before treatment, the researcher was able to demonstrate the effectiveness of the new treatment. Cholesterol levels before treatment and age explain to some extent the changes in cholesterol levels after treatment, but we can attribute the rest of the changes in 57% to the drug dose used (partial Eta-square =0.565437). Post-hoc tests (selected by Factor options) suggested the formation of two homogeneous groups, the placebo group and the drug patient group, indicating that raising the dose to a high one does not make a difference, since the cholesterol levels obtained will be similar.

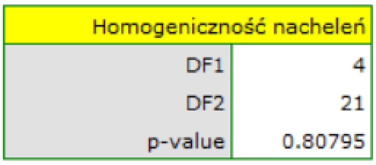


POST-HOC Fisher LSD			
	NIR [F]	Statistic [F]	p-value [F]
Dose			
Placebo;Low dose	1.02128	3.00814	0.00592
Placebo;High dose	0.91139	3.31434	0.0028
Low dose;High dose	0.94445	-0.05452	0.95696

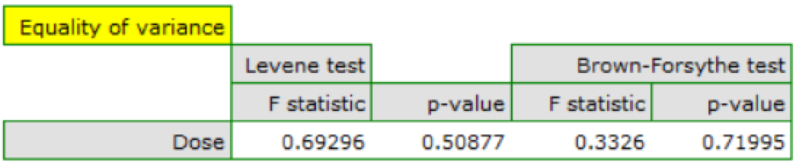


Homogeneous groups		
Dose	A	B
Placebo(b)		*
Low dose(a)	*	
High dose(a)	*	

ANCOVA assumptions remained to be tested. Homogeneity of variance and constancy of slopes of simple regressions were confirmed using tests.

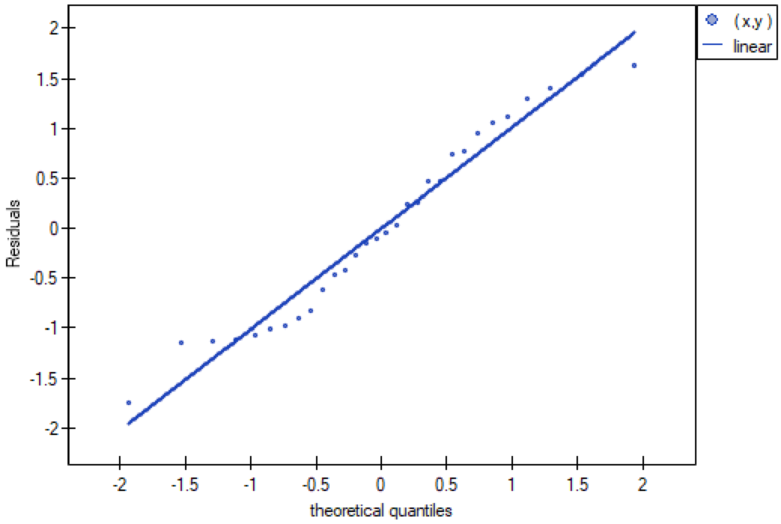


Homogeniczność nacheleń	
DF1	4
DF2	21
p-value	0.80795



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

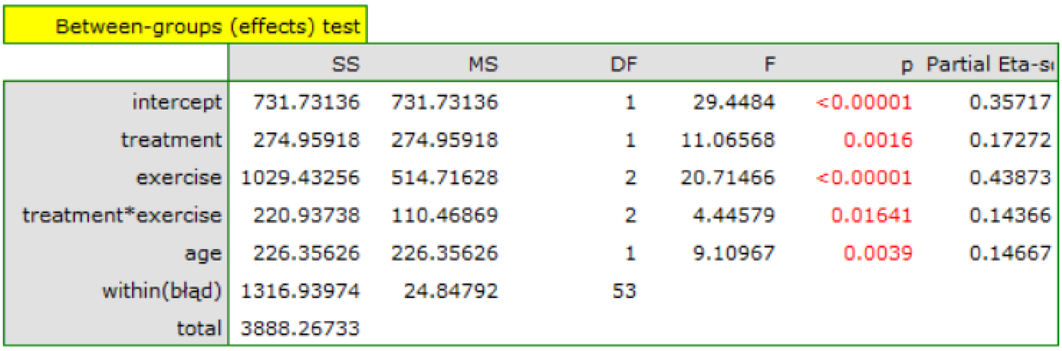
The normality of the rhesus distribution was assessed visually by plotting Q-Q plots:



Example 24.9. (stress.pqs file)
The example comes from the Datarium R-Cran package.
Researchers want to evaluate the effect of a new treatment and exercise on stress reduction after accounting for differences in age. The value of the stress measure is the interval outcome variable Y. Because the variables ”treatment” and ”exercise” have 2 and 3 categories, respectively, we will conduct a two-way ANCOVA to determine whether the interaction between exercise and treatment, while accounting for the subjects’ age, is related to stress.

Factors	
treatment X_1	exercise X_2
yes	low
	moderate
	high
no	low
	moderate
	high

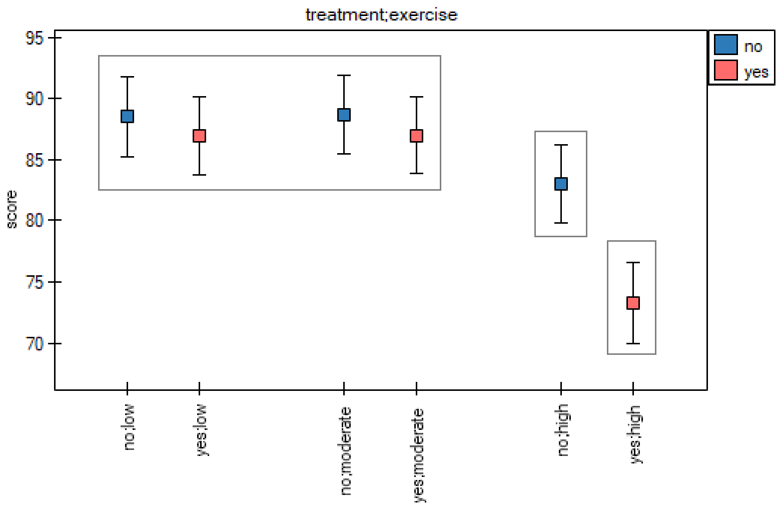
In the analysis window, I set ”stress” as the dependent variable, ”treatment” and ”exercise” as factors, and add the interaction of these two variables, the continuous co-variable is ”age.”

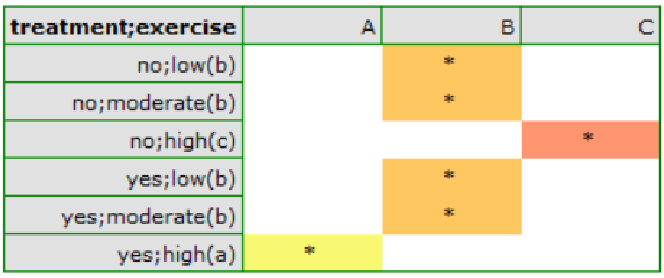


Between-groups (effects) test

	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					

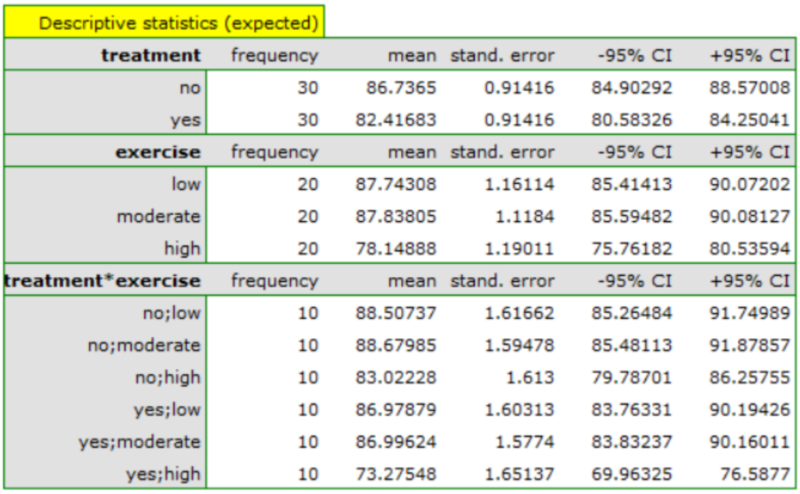
The result shows that the effect of treatment on stress varies with exercise intensity - indicated by a significant interaction of the two variables (p=0.016409). We plot a graph showing the expected mean stress levels for each of the six subgroups into which the interaction divided our data, and determine post-hoc tests.





treatment;exercise	A	B	C
no;low(b)		*	
no;moderate(b)		*	
no;high(c)			*
yes;low(b)		*	
yes;moderate(b)		*	
yes;high(a)	*		

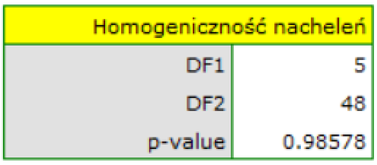
According to the results of the post-hoc test, we can speak of three different homogeneous groups: (B) the high-stress group is the group that exercises little or on average (whether or not they are treated sauropods), (C) the lower-stress group is the group that exercises a lot and is not treated, (A) the lowest-stress group is the group that exercises a lot and is treated. The values of the individual averages with confidence intervals are shown in the table



Descriptive statistics (expected)

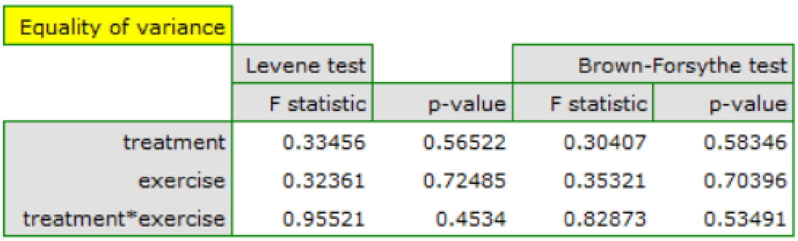
treatment	frequency	mean	stand. error	-95% CI	+95% CI
no	30	86.7365	0.91416	84.90292	88.57008
yes	30	82.41683	0.91416	80.58326	84.25041
exercise	frequency	mean	stand. error	-95% CI	+95% CI
low	20	87.74308	1.16114	85.41413	90.07202
moderate	20	87.83805	1.1184	85.59482	90.08127
high	20	78.14888	1.19011	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

Assumptions regarding equality of variances, slopes of regression lines and normality of model residuals are met.



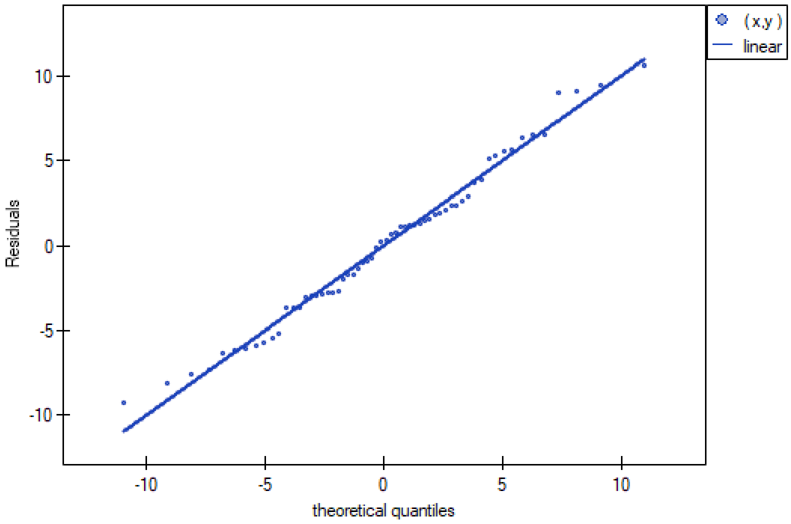
Homogeniczność nacheleń

DF1	5
DF2	48
p-value	0.98578



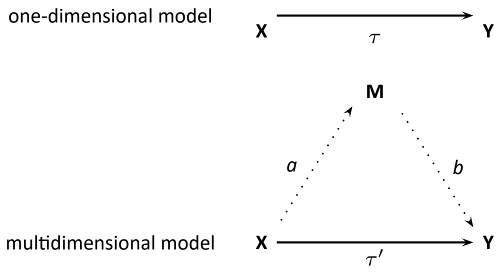
Equality of variance

	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



25 Mediation effect

Baron and Kenny (1986)[16] defined a mediator (M) as a variable that significantly explains the relationship between the independent variable (X) and the outcome variable (Y). In mediation, the relationship between the independent variable and the dependent variable is assumed to be an indirect effect that exists due to the influence of a third variable (mediator).



We determine the magnitude of change by the difference in the coefficients describing the relationship between variable X and variable Y in the univariate model:
$Y = \tau \cdot X + c$
and in the multivariate 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 that determines the relationship between the variable X and Y, the influence of the independent variable τ is reduced to τ'.

Tests to evaluate the mediation effect
The Sobel (1982)[153] test, the Aroian (1947)[7] test popularized by Baron and Kenny [16], and the Goodman (1960)[68] test are tests that determine whether the reduction in the effect of the independent variable on the outcome variable, when a mediator is included in the model, is a significant reduction and therefore whether the mediation effect is statistically significant.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \tau = \tau' \\ \mathcal{H}_1: & \tau \neq \tau', \end{array}$$

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:

$$Z = \frac{a \cdot b}{\sqrt{b^2 \cdot SE_a^2 + a^2 \cdot SE_b^2 - SE_a^2 \cdot SE_b^2}}$$

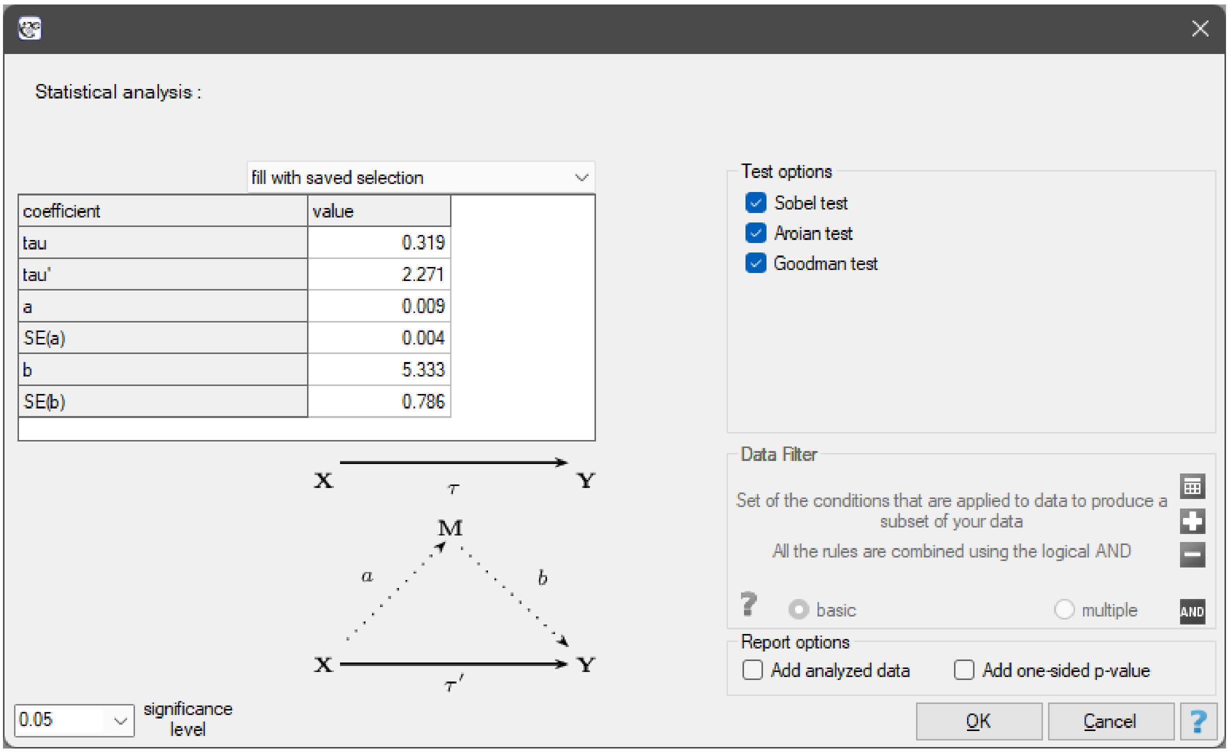
These statistics have an asymptotically (for large sizes) normal distribution.

The p value, designated on the basis of the test statistic, is compared with the

$$\begin{array}{ll} \text{if } p \le \alpha & \Longrightarrow \quad \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow \quad \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note
The Sobel test, as well as the Aroian and Goodman test, are very conservative tests and are intended only for large samples (greater than 100 items).

The mediation effect analysis window is invoked by Advanced Statistics→Multidivariate models→Mediation effect.



Statistical analysis :

fill with saved selection

coefficient	value
tau	0.319
tau'	2.271
a	0.009
SE(a)	0.004
b	5.333
SE(b)	0.786

Test options
☑ Sobel test
☑ Aroian test
☑ Goodman test

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
□ Add analyzed data □ Add one-sided p-value

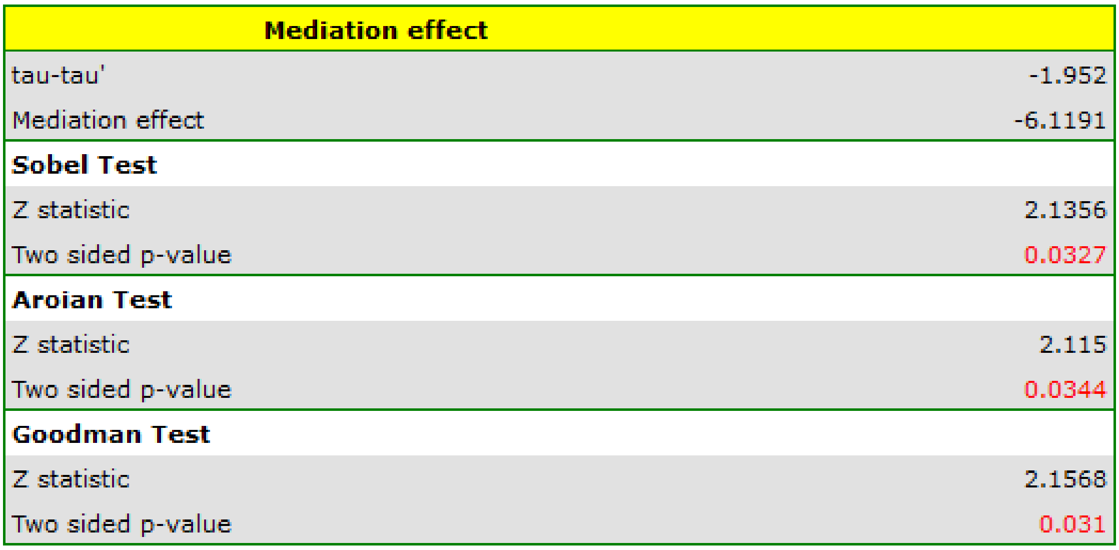
0.05 significance level

Example 25.1. Based on the paper by Mimar Sinan Fine (2017) [120].
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 purpose of this study is to investigate the relationship between age and systolic blood pressure and to present the effect of frequency of alcohol consumption on this relationship.

- A one-dimensional model was built that did not account for the potential mediator:
  $Y = 0.319 \cdot X + c$.
  The effect size of variable X (age) on variable Y (systolic blood pressure) was **tau**=0.319.
- A one-dimensional model was constructed that did not include a potential mediator::
  $Y = 2.271 \cdot X + 5.333 \cdot M + c$.
  The effect size of variable X (age) on variable Y (systolic blood pressure) was **tau'**=2.271. We also know from this model that **b**=5.333, and error SE_b=0.786

The difference between the coefficients is **tau-tau'**= a*b=0.048. The effect of mediation 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 by decreasing the coefficient by about 15%.

- A one-dimensional model was built to examine the effect of variable X on the mediator: $M = 0.009 \cdot X + c$.
  We know from this model that the coefficient **a**=0.009, and the error SE_a=0.004. We enter all this information in the analysis window obtaining 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

Based on the coefficients a and b and their standard errors, the result of Sobel (p=0.0327), Aroian (p=0.0344) and Goodman (p=0.0310) tests are determined. The obtained p-values indicate a statistically significant mediator. Thus, we confirmed that frequency of alcohol consumption affects the association of age with diastolic blood pressure so noticeably that it is worth explaining why this effect occurs.

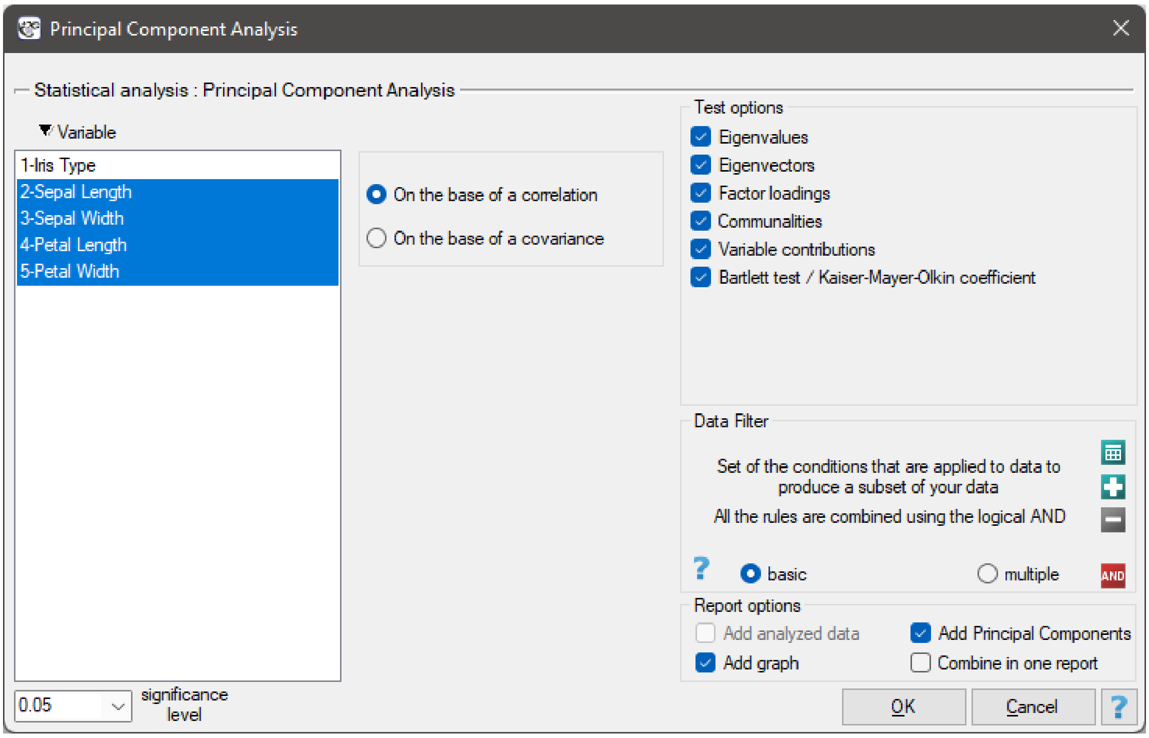
26 DIMENSION REDUCTION AND GROUPING

As the number of variables subjected to a statistical analysis grows, their precision grows, but so does the level of complexity and difficulty in interpreting the obtained results. Too many variables increase the risk of their mutual correlation. The information carried by some variables can, then, be redundant, i.e. a part of the variables may not bring in new information for analysis but repeat the information already given by other variables. The need for dimension reduction (a reduction of the number of variables) has inspired a whole group of analyses devoted to that issue, such as: factor analysis, principal component analysis, cluster analysis or discriminant analysis. Those methods allow the detection of relationships among the variables. On the basis of those relationships one can distinguish, for further analysis, groups of similar variables and select only one representative (one variable) of each group, or a new variable the values of which are calculated on the basis of the remaining variables in the group. As a result, one can be certain that the information carried by each group is included in the analysis. In this manner we can reduce a set of variables p to a set of variables k where $k < p$.

Similarly to grouping variables, we may be interested in grouping objects. Having at our disposal information about certain characteristics of objects, we are often able to distinguish groups of objects similar in terms of those characteristics, e.g. based on information about the amounts spent by customers of a certain store chain on particular items, we can divide customers in such a way as to distinguish customer segments with similar shopping preferences. As a result, your offer/advertising can be prepared not for the general public, but separately for each segment, so as to more precisely meet the needs of a potential customer.

26.1 PRINCIPAL COMPONENT ANALYSIS

The window with settings for Principal component analysis is accessed via the menu Advenced statistics → Multivariate Models → Principal Component Analysis.



Principal Component Analysis

Statistical analysis : Principal Component Analysis

▼ Variable

1-Iris Type
2-Sepal Length
3-Sepal Width
4-Petal Length
5-Petal Width

● On the base of a correlation
○ On the base of a covariance

Test options
☑ Eigenvalues
☑ Eigenvectors
☑ Factor loadings
☑ Communalities
☑ Variable contributions
☑ Bartlett test / Kaiser-Mayer-Olkin coefficient

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
□ Add analyzed data ☑ Add Principal Components
☑ Add graph □ Combine in one report

0.05 significance level

Principal component analysis involves defining completely new variables (**principal components**) which are a linear combination of the observed (original) variables. An exact analysis of the principal components makes it possible to point to those original variables which have a big influence on the appearance of particular principal components, that is those variables which constitute a homogeneous group. A principal component is then a representative of that group. Subsequent components are mutually orthogonal (uncorrelated) and their number (k) is lower than or equal to the number of original variables (p).

Particular principal components are a linear combination of original variables:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{in}X_p$$

where:

$X_1, X_2, ..., X_p$ – original variables,
$a_{i1}, a_{i2}, ..., a_{ip}$ – coefficients of the ith principal component

Each principal component explains a certain part of the variability of the original variables. They are, then, naturally based on such measures of variability as covariance (if the original variables are of similar size and are expressed in similar units) or correlation (if the assumptions necessary in order to use covariance are not fulfilled).

Mathematical calculations which allow the distinction of principal components include defining the eigenvalues and the corresponding eigenvectors from the following matrix equation:

$$(M - \lambda I)a = 0$$

where:

λ – eigenvalues,
$a_i = (a_{i1}, a_{i2}, ..., a_{ip})$ – eigenvector corresponding to the ith eigenvalue,
M – the variance matrix or covariance matrix of original variables $X_1, X_2, ..., X_p$,
I – identity matrix (1 on the main diagonal, 0 outside of it).

26.1.1 Interpretation of coefficients related to the analysis

Every principal component is described by:

Eigenvalue

An eigenvalue informs about which part of the total variability is explained by a given principal component. The first principal component explains the greatest part of variance, the second principal component explains the greatest part of that variance which has not been explained by the previous component, and the subsequent component explains the greatest part of that variance which has not been explained by the previous components. As a result, each subsequent principal component explains a smaller and smaller part of the variance, which means that the subsequent values are smaller and smaller.

Total variance is a sum of the eigenvalues, which allows the calculation of the variability percentage defined by each component.

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p} \cdot 100\%$$

Consequently, one can also calculate the cumulative variability and the cumulative variability percentage for the subsequent components.

Eigenvector

An eigenvector reflects the influence of particular original variables on a given principal component. It contains the $a_{i1}, a_{i2}, ..., a_{ip}$ coefficients of a linear combination which defines a component. The sign of those coefficients points to the direction of the influence and is accidental which does not change the value of the carried information.

Factor loadings

Factor loadings, just as the coefficients included in the eigenvector, reflect the influence of particular variables on a given principal component. Those values illustrate the part of the variance of a given component is constituted by the original variables. When an analysis is based on the correlation matrix, we interpret those values as correlation coefficients between original variables and a given principal value.

Variable contributions

They are based on the determination coefficients between 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 particular original variables.

Communalities

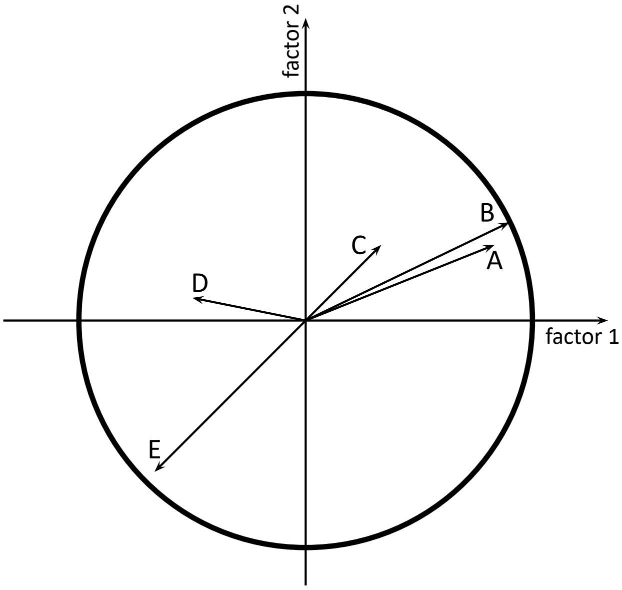
They are based on the determination coefficients between original variables and a given principal component. They show what percentage of a given original variable can be explained by the variability of a few initial principal components. For example: the result concerning the second variable contained in the column concerning the fourth principal component tells us what percent of the variability of the second variable can be explained by the variability of four initial principal components.

26.1.2 Graphical interpretation

A lot of information carried by the coefficients returned in the tables can be presented on one chart. The ability to read charts allows a quick interpretation of many aspects of the conducted analysis. The charts gather in one place the information concerning the mutual relationships among the components, the original variables, and the cases. They give a general picture of the principal components analysis which makes them a very good summary of it.

Factor loadings graph

The graph shows vectors connected with the beginning of the coordinate system, which represent original variables. The vectors are placed on a plane defined by the two selected principal components.



The coordinates of the terminal points of the vector are the corresponding factor loadings of the variables.

Vector length represents the information content of an original variable carried by the principal components which define the coordinate system. The longer the vector the greater the contribution of the original variable to the components. In the case of an analysis based on a correlation matrix the loadings are correlations between original variables and principal components. In such a case points fall into the unit circle. It happens because the correlation coefficient cannot exceed one. As a result, the closer a given original variable lies to the rim of the circle the better the representation of such a variable by the presented principal components.

The sign of the coordinates of the terminal point of the vector i.e. the sign of the loading factor, points to the positive or negative correlation of an original variable and the principal components forming the coordination system. If we consider both axes (2 components) together then original variables can fall into one of four categories, depending on the combination of signs ($+/-$) and their loading factors.

The angle between vectors indicates the correlation of original values:

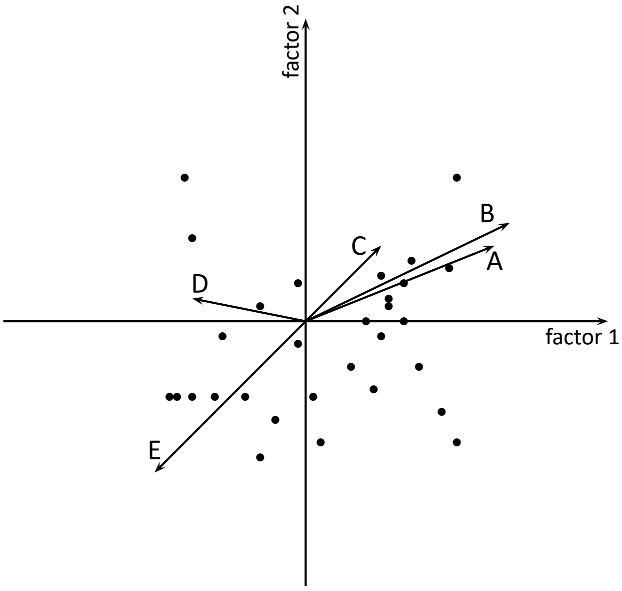
$0 < \alpha < 90^0$ – the smaller the angle between the vectors representing original variables, the stronger the positive correlation among these variables.

$\alpha = 90^0$ – the vectors are perpendicular, which means that the original variables are not correlated.

$90^0 < \alpha < 180^0$ – the greater the angle between the vectors representing the original variables, the stronger the negative correlation among these variables.

Biplot

The graph presents 2 series of data placed in a coordinate system defined by 2 principal components. The first series on the graph are data from the first graph (i.e. the vectors of original variables) and the second series are points presenting particular cases.



Point coordinates should be interpreted as standardized values, i.e. positive coordinates pointing to a value higher than the mean value of the principal component, negative ones to a lower value, and the higher the absolute value the further the points are from the mean. If there are untypical observations on the graph, i.e. outliers, they can disturb the analysis and should be removed, and the analysis should be made again.

The distances between the points show the similarity of cases: the closer (in the meaning of Euclidean distance) they are to one another, the more similar information is carried by the compared cases.

Orthographic projection of points on vectors are interpreted in the same manner as point coordinates, i.e. projections onto axes, but the interpretation concerns original variables and not principal components. The values placed at the end of a vector are greater than the mean value of the original variable, and the values placed on the extension of the vector but in the opposite direction are values smaller than the mean.

26.1.3 The criteria of dimension reduction

There is not one universal criterion for the selection of the number of principal components. For that reason it is recommended to make the selection with the help of several methods.

The percentage of explained variance

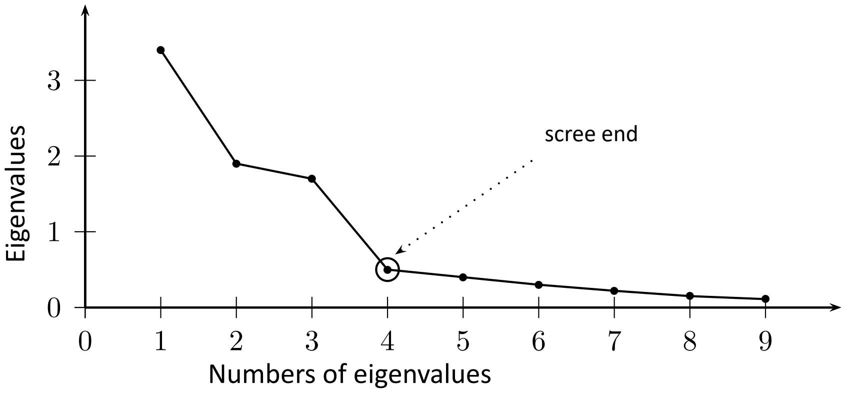
The number of principal components to be assumed by the researcher depends on the extent to which they represent original variables, i.e. on the variance of original variables they explain. All principal components explain 100% of the variance of original variables. If the sum of the variances for a few initial components constitutes a large part of the total variance of original variables, then principal components can satisfactorily replace original variables. It is assumed that the variance should be reflected in principal components to the extent of over 80 percent.

Kaiser criterion

According to the Kaiser criterion the principal components we want to leave for interpretation should have at least the same variance as any standardized original variable. As the variance of every standardized original variable equals 1, according to Kaiser criterion the important principal components are those the eigenvalue of which exceeds or is near value 1.

Scree plot

The graph presents the pace of the decrease of eigenvalues, i.e. the percentage of explained variance.



The moment on the chart in which the process stabilizes and the decreasing line changes into a horizontal one is the so-called end of the scree (the end of sprinkling of the information about the original values carried by principal components). The components on the right from the point which ends the scree represent a very small variance and are, for the most part, random noise.

26.1.4 Defining principal components

When we have decided how many principal components we need we can start generating them. In the case of principal components created on the basis of a correlation matrix they are computed as a linear combination of standardized original values. If, however, principal components have been created on the basis of a covariance matrix, they are computed as a linear combination of eigenvalues which have been centralized with respect to the mean of the original values.

The obtained principal components constitute new variables with certain advantages. First of all, the variables are not collinear. Usually there are fewer of them than original variables, sometimes much fewer, and they carry the same or a slightly smaller amount of information than the original values. Thus, the variables can easily be used in most multidimensional analyses.

26.1.5 The advisability of using the Principal Component Analysis

If the variables are not correlated (the Pearson's correlation coefficient is near 0), then there is no use to conduct a principal component analysis, as in such a situation every variable is already a separate component.

Bartlett's test

The test is used to verify the hypothesis that the correlation coefficients between variables are zero (i.e. the correlation matrix is an identity matrix).

Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & M = I,\\ \mathcal{H}_1: & M \neq I.\end{array}$$

where:
M – the variance matrix or covariance matrix of original variables $X_1, X_2, ..., X_p$,
I – the identity matrix (1 on the main axis, 0 outside of it).

The test statistic has the form presented below:

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right)\sum_{i=1}^{k} \ln \lambda_i,$$

where:
p – the number of original variables,
n – size (the number of cases),
λ_i – ith eigenvalue.

That statistic has, asymptotically (for large expected frequencies), the distribution χ^2 with $p(p-1)/2$ degrees of freedom.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

The Kaiser-Meyer-Olkin coefficient

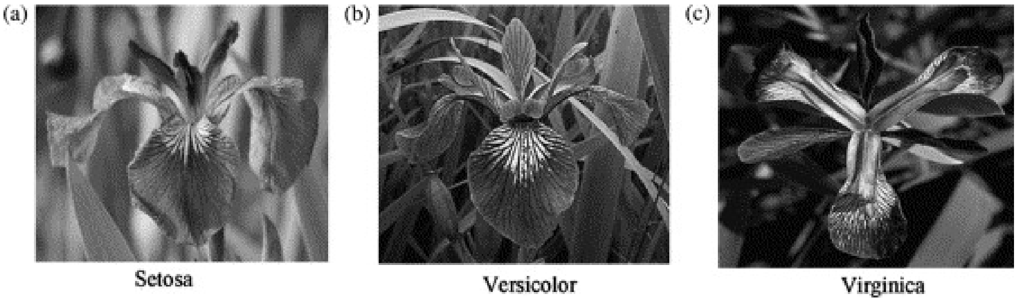
The coefficient is used to check the degree of correlation of original variables, i.e. the strength of the evidence testifying to the relevance of conducting a principal component analysis.

$$KMO = \frac{\sum_{i\neq j}^{p}\sum_{j\neq i}^{p} r_{ij}^2}{\sum_{i\neq j}^{p}\sum_{j\neq i}^{p} r_{ij}^2 + \sum_{i\neq j}^{p}\sum_{j\neq i}^{p} \hat{r}_{ij}^2},$$

r_{ij} – the correlation coefficient between the ith and the jth variable,
$\hat{r}_{ij}$ – the partial correlation coefficient between the ith and the jth variable.

The value of the Kaiser coefficient belongs to the range $< 0, 1 >$ where low values testify to the lack of a need to conduct a principal component analysis, and high values are a reason for conducting such an analysis.

Example 26.1. (file: iris.pqs) That classical set of data was first published in Ronald Aylmer Fisher's 1936[58] work in which discriminant analysis was presented. The file contains the measurements (in centimeters) of the length and width of the petals and sepals for 3 species of irises. The studied species are setosa, versicolor, and virginica. It is interesting how the species can be distinguished on the basis of the obtained measurements.

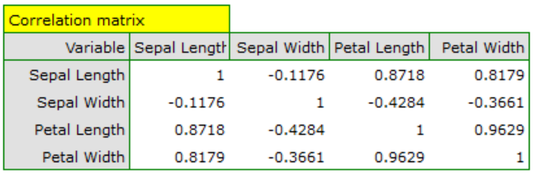


Setosa Versicolor Virginica

The photos are from scientific paper: Lee, et al. (2006r), "Application of a noisy data classification technique to determine the occurrence of flashover in compartment fires"

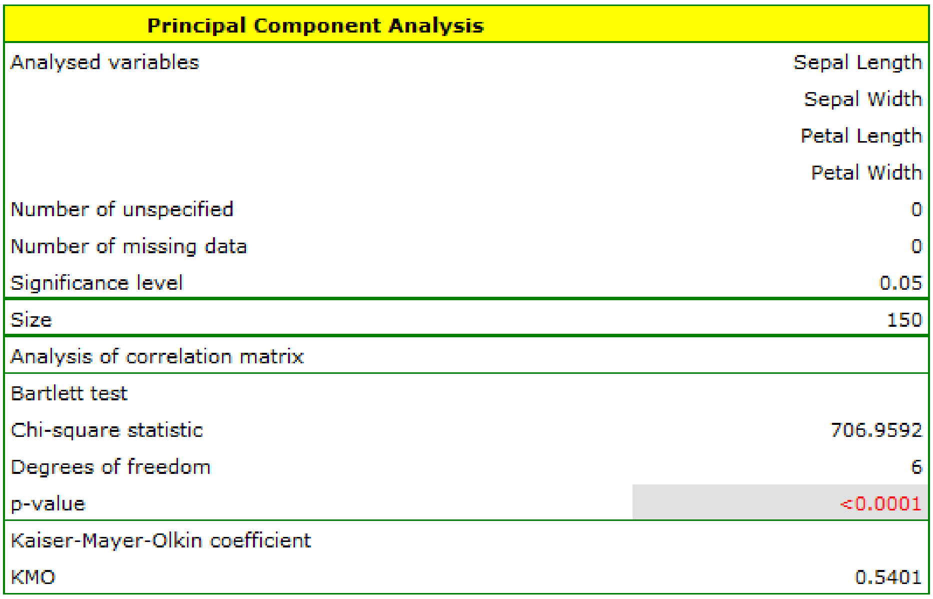
Principal component analysis will allow us to point to those measurements (the length and the width of the petals and sepals) which give the researcher the most information about the observed flowers.

The first stage of work, done even before defining and analyzing principal components, is checking the advisability of conducting the analysis. We start, then, from defining a correlation matrix of the variables and analyzing the obtained correlations with the use of Bartlett's test and the KMO coefficient.



Correlation matrix

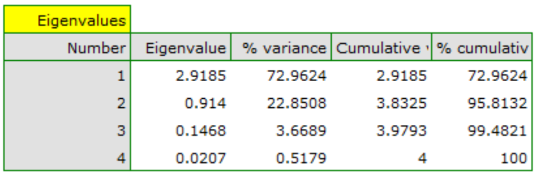
Variable	Sepal Lengt	Sepal Width	Petal Length	Petal Width
Sepal Length	1	-0.1176	0.8718	0.8179
Sepal Width	-0.1176	1	-0.4284	-0.3661
Petal Length	0.8718	-0.4284	1	0.9629
Petal Width	0.8179	-0.3661	0.9629	1



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

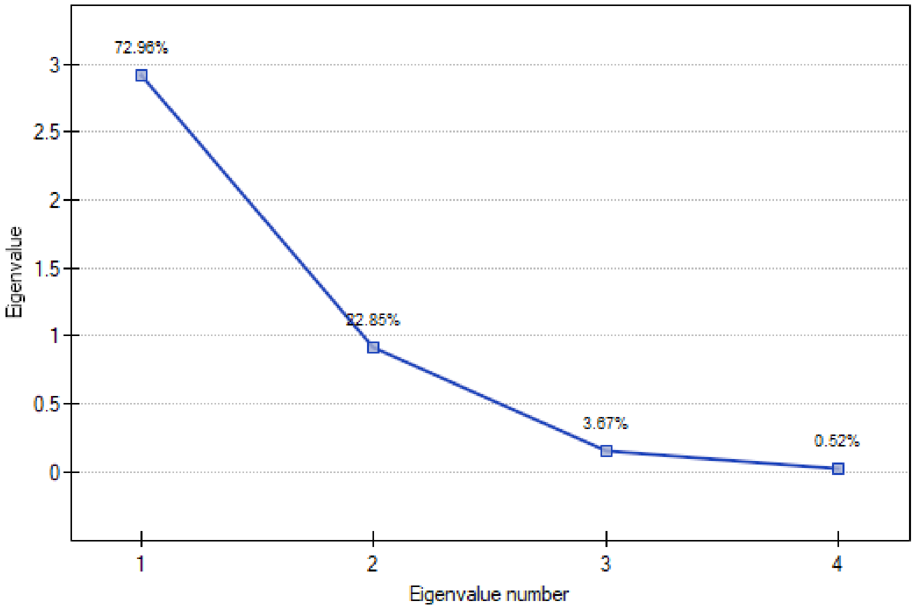
The value p of Bartlett's statistics points to the truth of the hypothesis that there is a significant difference between the obtained correlation matrix and the identity matrix, i.e. that the data are strongly correlated. The obtained KMO coefficient is average and equals 0.54. We consider the indications for conducting a principal component analysis to be sufficient.

The first result of that analysis which merits our special attention are eigenvalues:



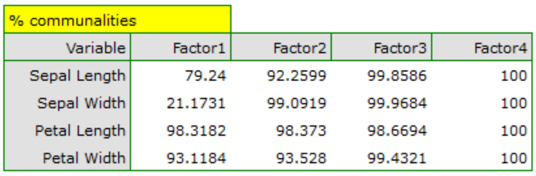
Eigenvalues				
Number	Eigenvalue	% variance	Cumulative '	% cumulativ
1	2.9185	72.9624	2.9185	72.9624
2	0.914	22.8508	3.8325	95.8132
3	0.1468	3.6689	3.9793	99.4821
4	0.0207	0.5179	4	100

The obtained eigenvalues show that one or even two principal components will describe our data well. The eigenvalue of the first component is 2.92 and the percent of the explained variance is 72.96. The second component explains much less variance, i.e. 22.85%, and its eigenvalue is 9.91. According to Kaiser criterion, one principal component is enough for an interpretation, as only for the first principal component the eigenvalue is greater than 1. However, looking at the graph of the scree we can conclude that the decreasing line changes into a horizontal one only at the third principal component.



From that we may infer that the first two principal components carry important information. Together they explain a great part, as much as 95.81%, of the variance (see the cumulative % column).

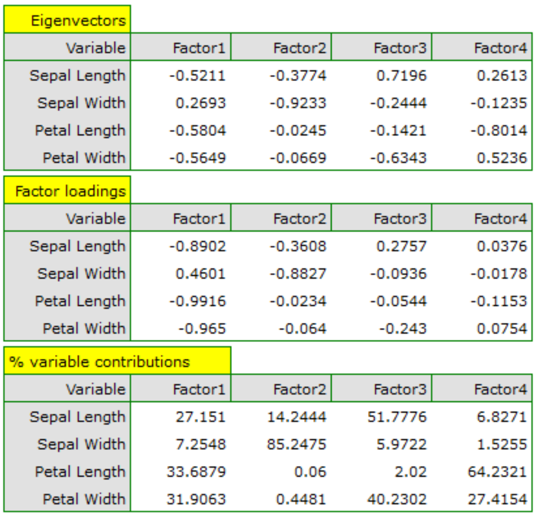
The communalities for the first principal component are high for all original variables except the variable of the width of the sepal, for which they equal 21.17%. That means that if we only interpret the first principal component, only a small part of the variable of the width of the sepal would be reflected.



% 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 principal components the communalities are at a similar, very high level and they exceed 90% for each of the analyzed variables, which means that with the use of those components the variance of each variability is represented in over 90%.
In the light of all that knowledge it has been decided to separate and interpret 2 components.

In order to take a closer look at the relationship of principal components and original variables, that is the length and the width of the petals and sepals, we interpret: eigenvectors, factor loadings, and contributions of original variables.



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

% 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

Particular original variables have differing effects on the first principal component. Let us put them in order according to that influence:

1. The length of a petal is negatively correlated with the first component, i.e. the longer the petal, the lower the values of that component. The eigenvector of the length of the petal is the greatest in that component and equals -0.58. Its factor loading informs that the correlation between the first principal component and the length of the petal is very high and equals -0.99 which constitutes 33.69% of the first component;

2. The width of the petal has an only slightly smaller influence on the first component and is also negatively correlated with it;

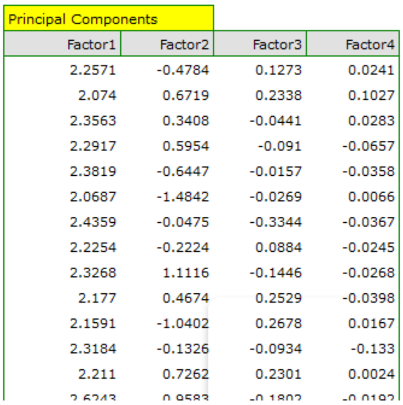
3. We interpret the length of the sepal similarly to the two previous variables but its influence on the first component is smaller;

4. The correlation of the width of the sepal and the first component is the weakest, and the sign of that correlation is positive.

The second component represents chiefly the original variable ”sepal width”; the remaining original variables are reflected in it to a slight degree. The eigenvector, factor loading, and the contribution of the variable ”sepal width” is the highest in the second component.

Each principal component defines a homogeneous group of original values. We will call the first component ”petal size” as its most important variables are those which carry the information about the petal, although it has to be noted that the length of the sepal also has a significant influence on the value of that component. When interpreting we remember that the greater the values of that component, the smaller the petals.
We will call the second component ”sepal width” as only the width of the sepal is reflected to a greater degree here. The greater the values of that component, the narrower the sepal.

Finally, we will generate the components by choosing, in the analysis window, the option: Add Principal Components. A part of the obtained result is presented below:



Principal Components

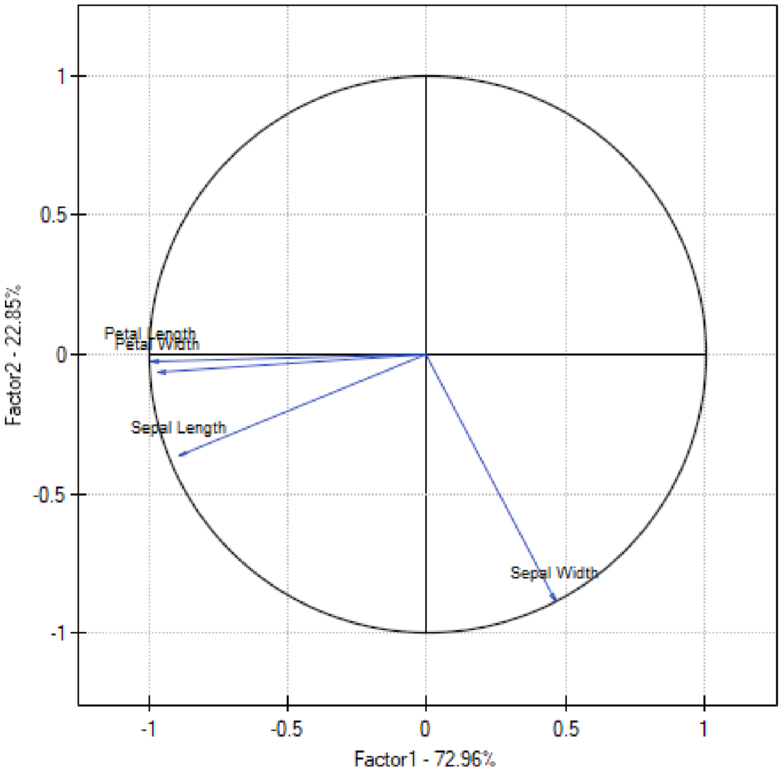
	Factor1	Factor2	Factor3	Factor4
	2.2571	-0.4784	0.1273	0.0241
	2.074	0.6719	0.2338	0.1027
	2.3563	0.3408	-0.0441	0.0283
	2.2917	0.5954	-0.091	-0.0657
	2.3819	-0.6447	-0.0157	-0.0358
	2.0687	-1.4842	-0.0269	0.0066
	2.4359	-0.0475	-0.3344	-0.0367
	2.2254	-0.2224	0.0884	-0.0245
	2.3268	1.1116	-0.1446	-0.0268
	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

In order to be able to use the two initial components instead of the previous four original values, we copy and paste them into the data sheet. Now, the researcher can conduct the further statistics on two new, uncorrelated variables.

Analysis of the graphs of the two initial components

The analysis of the graphs not only leads the researcher to the same conclusions as the analysis of the tables but will also give him or her the opportunity to evaluate the results more closely.

Factor loadings graph



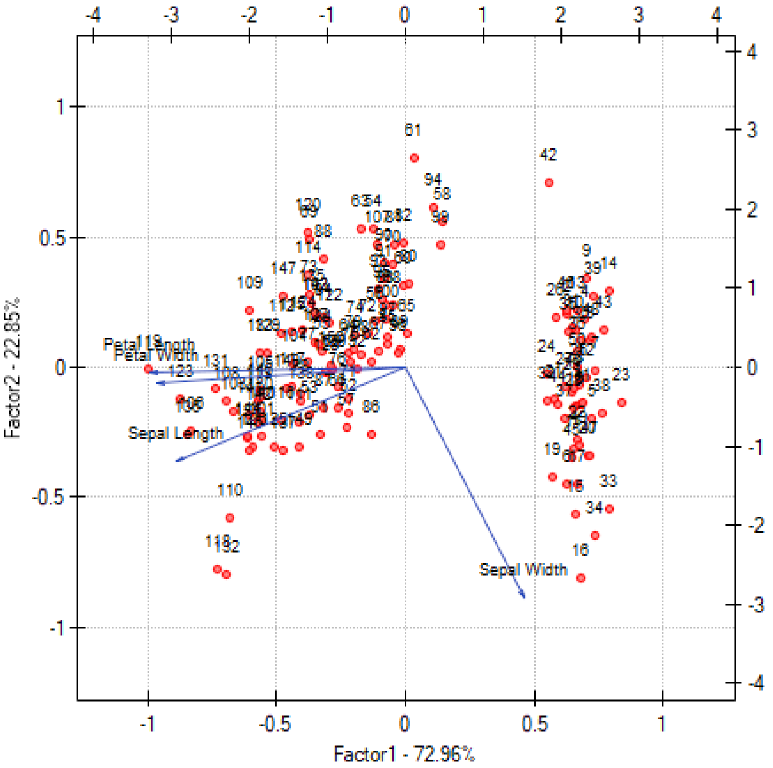
The graph shows the two first principal components which represent 72.96% of the variance and 22.85% of the variance, together amounting to 95.81% of the variance of original values

The vectors representing original values almost reach the rim of the unit circle (a circle with the radius of 1), which means they are all well represented by the two initial principal components which form the coordinate system.

The angle between the vectors illustrating the length of the petal, the width of the petal, and the length of the sepal is small, which means those variables are strongly correlated. The correlation of those variables with the components which form the system is negative, the vectors are in the third quadrant of the coordinate system. The observed values of the coordinates of the vector are higher for the first component than for the second one. Such a placement of vectors indicates that they comprise a uniform group which is represented mainly by the first component.

The vector of the width of the sepal points to an entirely different direction. It is only slightly correlated with the remaining original values, which is shown by the inclination angle with respect to the remaining original values – it is nearly a right angle. The correlation of that vector with the first component is positive and not very high (the low value of the first coordinate of the terminal point of the vector), and it is negative and high (the high value of the second coordinate of the terminal point of the vector) in the case of the second component. From that we may infer that the width of the sepal is the only original variable which is well represented by the second component.

Biplot



The biplot presents two series of data spread over the first two components. One series are the vectors of original values which have been presented on the previous graph and the other series are the points which carry the information about particular flowers. The values of the second series are read on the upper axis X and the right axis Y. The manner of interpretation of vectors, that is the first series, has been discussed with the previous graph. In order to understand the interpretation of points let us focus on flowers number 33, 34, and 109.

Flowers number 33 and 34 are similar – the distance between points 33 and 34 is small. For both points the value of the first component is much greater than the average and the value of the second component is much smaller than the average. The average value, i.e. the arithmetic mean of both components, is 0, i.e. it is the middle of the coordination system. Remembering that the first component is mainly the size of the petals and the second one is mainly the width of the sepal we can say that flowers number 33 and 34 have small petals and a large width of the sepal. Flower number 109 is represented by a point which is at a large distance from the other two points. It is a flower with a negative first component and a positive, although not high second component. That means the flower has relatively large petals while the width of the sepal is a bit smaller than average.

Similar information can be gathered by projecting the points onto the lines which extend the vectors of original values. For example, flower 33 has a large width of the sepal (high and positive values on the projection onto the original value ”sepal width”) but small values of the remaining original values (negative values on the projection onto the extension of the vectors illustrating the remaining original values).

26.2 CLUSTER ANALYSIS

Cluster analysis is a series of methods for dividing objects or features (variables) into similar groups. In general, these methods are divided into two classes: hierarchical methods and non-hierarchical methods such as the k-means method. In their algorithms, both methods use a similarity matrix to create clusters based on it.

Object grouping and variable grouping are done in cluster analysis in exactly the same way. In this chapter, clustering methods will be explained using object clustering as an example.

Note!
In order to ensure balanced influence of all variables on similarity matrix elements, data should be standardized by choosing appropriate option in the analysis window. Lack of standardization gives more influence on obtained result to variables expressed with higher numbers.

26.2.1 Hierarchical methods

Hierarchical cluster analysis methods involve building a hierarchy of clusters, starting from the smallest (consisting of single objects) and ending with the largest (consisting of the maximum number of objects). Clusters are created on the basis of object similarity matrix.

AGGLOMERATION PROCEDURE

1. By following the indicated **linkage method**, the algorithm finds a pair of similar objects in the **similarity** matrix and combines 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.

Object similarity

In the process of working with cluster analysis, similarity or distance measures play an essential role. The mutual similarity of objects is placed in the similarity matrix. A large variety of methods for determining the distance/similarity between objects allows to choose such measures that best reflect the actual relation. Distance and similarity measures are described in more detail in the section similarity matrix.

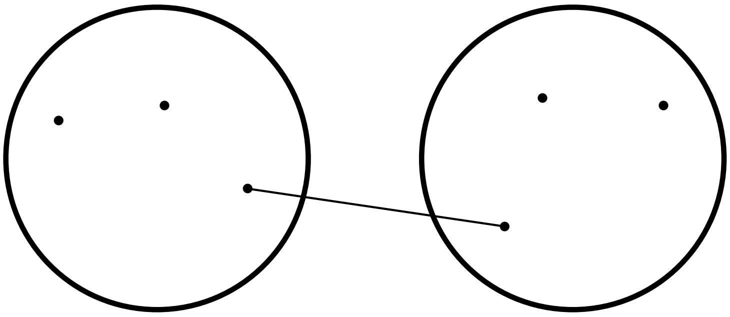
Cluster analysis is based on finding clusters inside a similarity matrix. Such a matrix is created in the course of performing cluster analysis. For the cluster analysis to be successful, it is important to remember that higher values in the similarity matrix should indicate greater variation of objects, and lower values should indicate their similarity.

Note!
To increase the influence of the selected variables on the elements of the similarity matrix, indicate the appropriate weights when defining the distance while remembering to standardize the data.

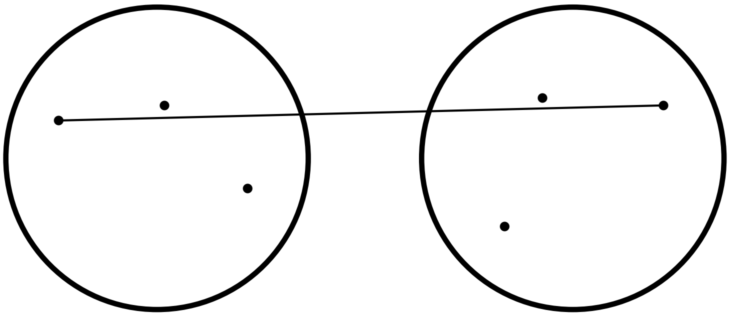
> For example, for people wanting to take care of a dog, grouping dogs according to size, coat, tail length, character, breed, etc. will make the choice easier. However, treating all characteristics identically may put completely dissimilar dogs into one group. For most of us, on the other hand, size and character are more important than tail length, so the similarity measures should be set so that size and character are most important in creating clusters.

Object and cluster linkage methods

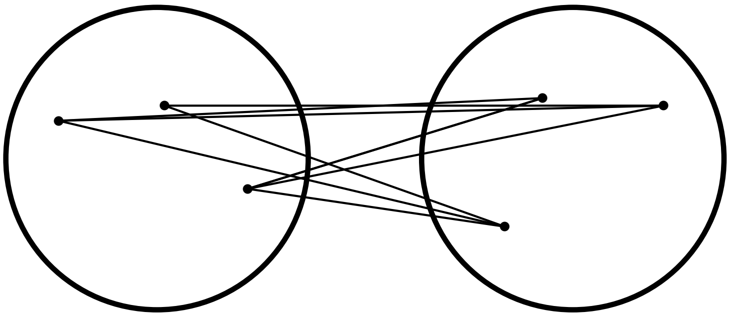
- Single linkage method - the distance between clusters is determined by the distance of those objects of each cluster that are closest to each other.



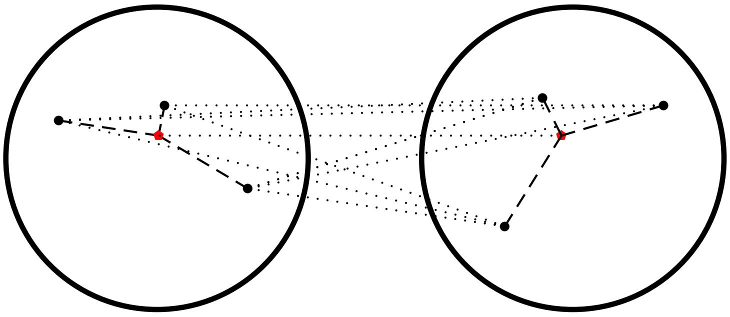
- Complete linkage method - the distance between clusters is determined by the distance of those objects of each cluster that are farthest apart.



- Unweighted pair-group method using arithmetic averages - the distance between clusters is determined by the average distance between all pairs of objects located within two different clusters.

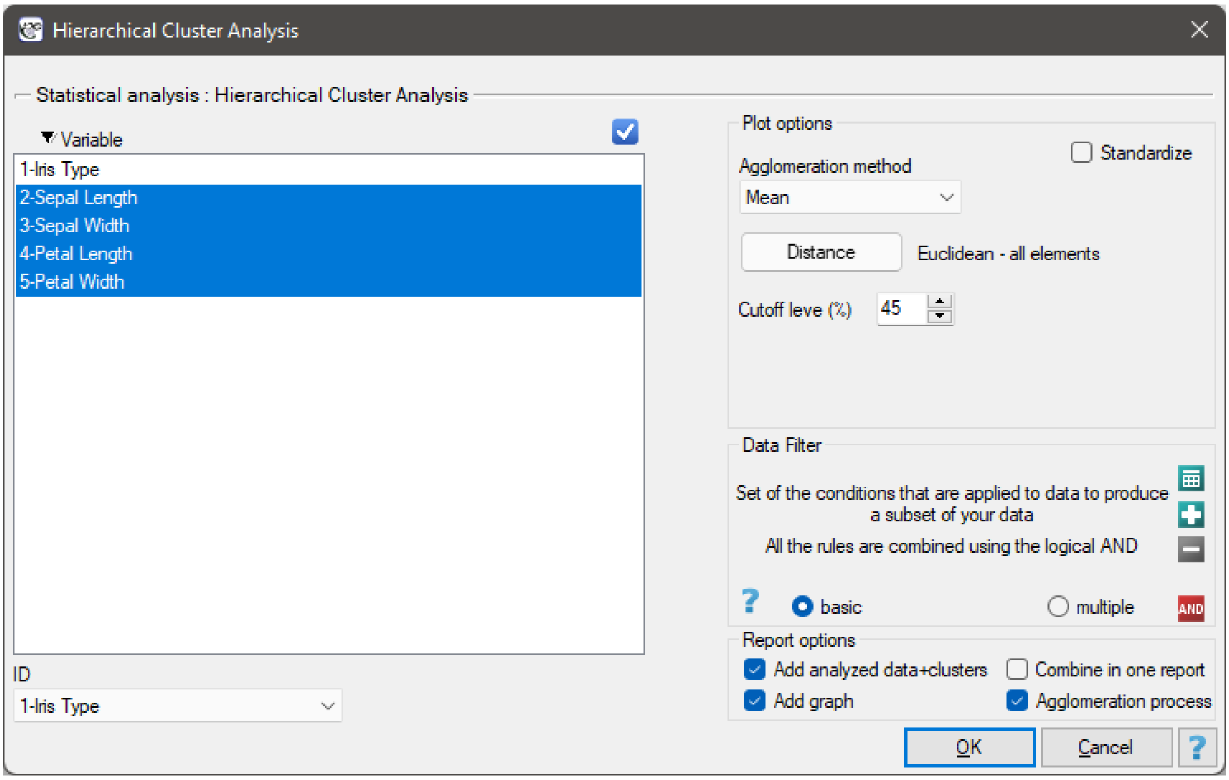


- Weighted pair-group metod using arithmetic averages - similarly to the unweighted pair-group method using arithmetic averages method it involves calculating the average distance, but this average is weighted by the number of elements in each cluster. As a result, we should choose this method when we expect to get clusters with similar sizes.
- Ward's method - is based on the variance analysis concept - it calculates the difference between the sums of squares of deviations of distances of individual objects from the center of gravity of clusters, to which these objects belong. This method is most often chosen due to its quite universal character.



The result of a cluster analysis conducted using the hierarchical method is represented using a dendogram. A **Dendogram** is a form of a tree indicating the relations between particular objects obtained from the similarity matrix analysis. The cutoff level of the dendogram determines the number of clusters into which we want to divide the collected objects. The choice of the cutoff is determined by specifying the length of the bond at which the cutoff will occur as a percentage, where 100% is the length of the last and also the longest bond in the dendogram.

Settings window of the hierarchical cluster analysis is opened via menu Advanced Statistics→Reduction and grouping→Hierarchical Cluster Analysis.



Hierarchical Cluster Analysis

Statistical analysis : Hierarchical Cluster Analysis

Variable ☑
- 1-Iris Type
- 2-Sepal Length
- 3-Sepal Width
- 4-Petal Length
- 5-Petal Width

ID: 1-Iris Type

Plot options

Agglomeration method: Mean

□ Standardize

Distance: Euclidean - all elements

Cutoff leve (%): 45

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

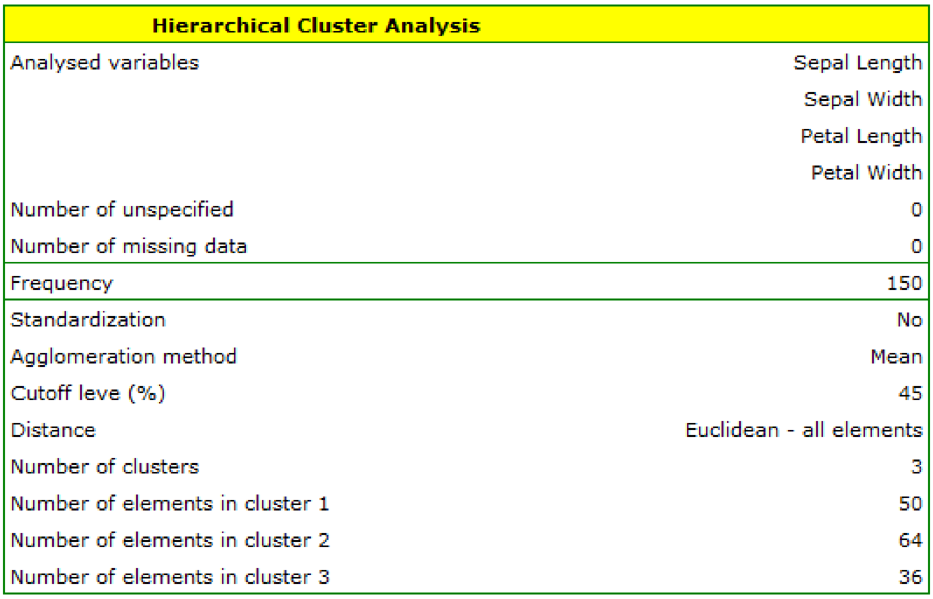
Report options

☑ Add analyzed data+clusters □ Combine in one report

☑ Add graph ☑ Agglomeration process

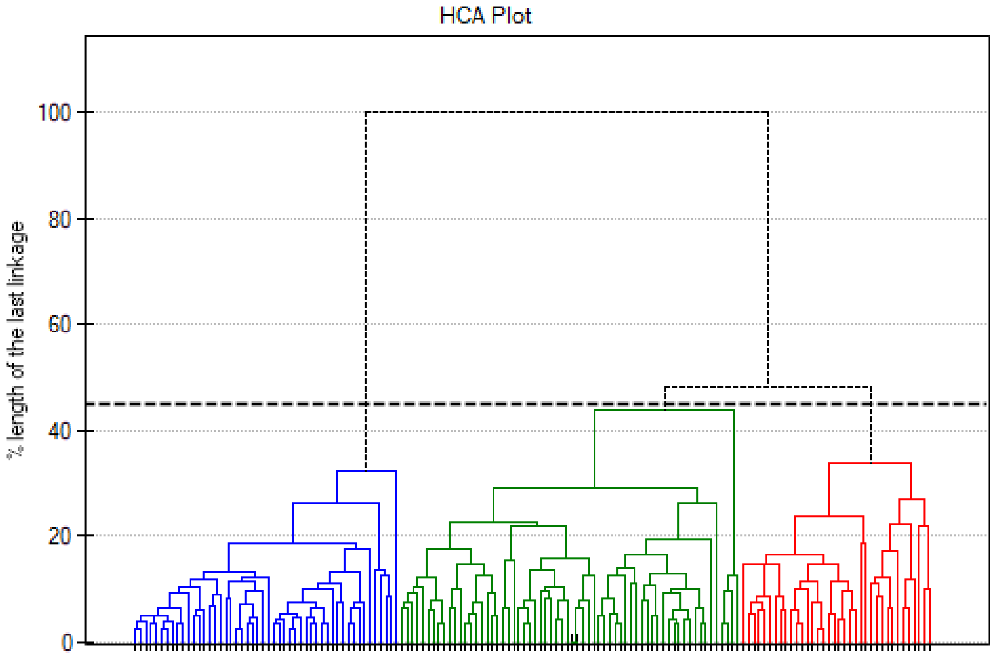
Example (26.1) c.d. *(iris.pqs file)*
The analysis will be performed on the classic data set of dividing iris flowers into 3 varieties based on the width and length of the petals and sepal sepals (R.A. Fisher 1936[58]). Because this data set contains information about the actual variety of each flower, after performing a cluster analysis it is possible to determine the accuracy of the division made.

We assign flowers to particular groups on the basis of columns from 2 to 5. We choose the way of calculating distances e.g. Euclidean distance and the linkage method. Specifying the cutoff level of clusters will allow us to cut off the dendogram in such a way that clusters will be formed - in the case of this analysis we want to get 3 clusters and to achieve this we change the cutoff level to 45. We will also attach data+clusters to the report..

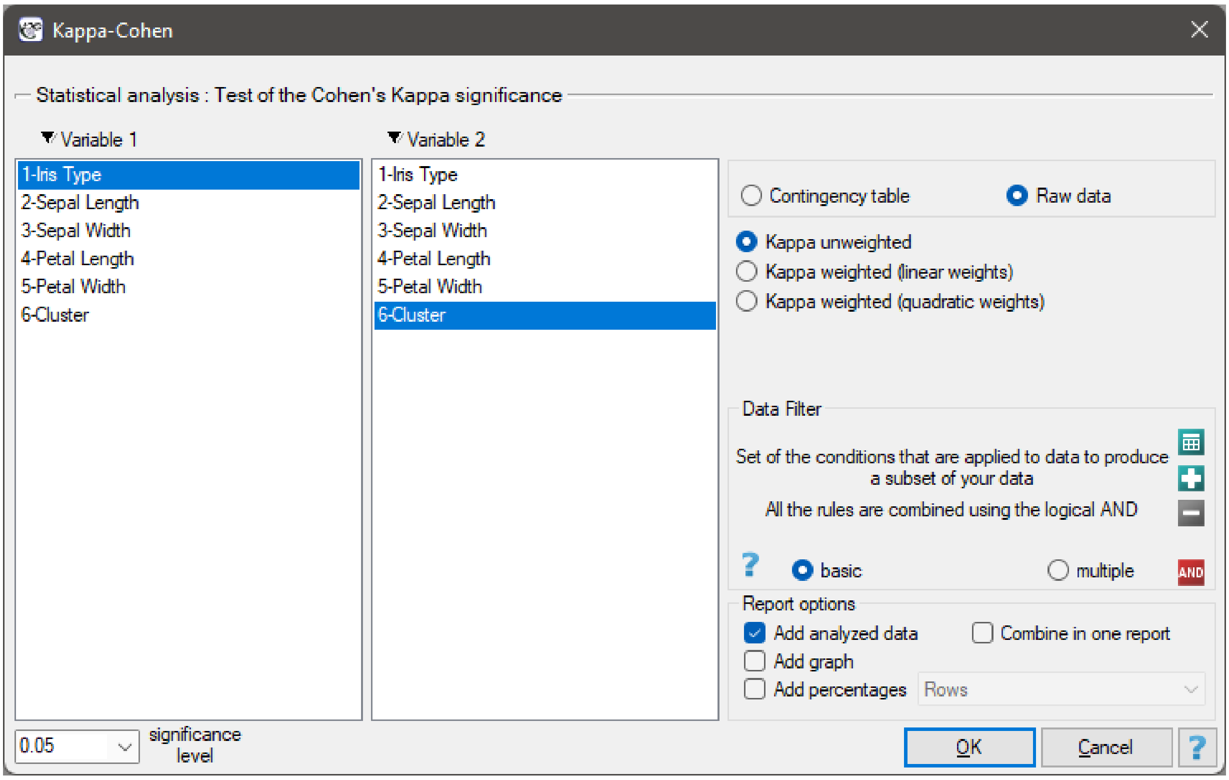


Hierarchical Cluster Analysis	
Analysed variables	Sepal Length Sepal Width Petal Length Petal Width
Number of unspecified	0
Number of missing data	0
Frequency	150
Standardization	No
Agglomeration method	Mean
Cutoff leve (%)	45
Distance	Euclidean - all elements
Number of clusters	3
Number of elements in cluster 1	50
Number of elements in cluster 2	64
Number of elements in cluster 3	36

In the dendogram, the order of the bonds and their lengths are shown.



To examine whether the extracted clusters represent the 3 actual varieties of iris flowers, we can copy the column containing the information about cluster belonging from the report and paste it into the datasheet. Like the clusters, the varieties are also described numerically by Codes/Labels/Format, so we can easily perform a concordance analysis. We will check the concordance of our results with the actual belonging of a given flower to the corresponding species using the Cohen’s Kappa method .



Kappa-Cohen

Statistical analysis : Test of the Cohen's Kappa significance

Variable 1
- 1-Iris Type (selected)
- 2-Sepal Length
- 3-Sepal Width
- 4-Petal Length
- 5-Petal Width
- 6-Cluster

Variable 2
- 1-Iris Type
- 2-Sepal Length
- 3-Sepal Width
- 4-Petal Length
- 5-Petal Width
- 6-Cluster (selected)

○ Contingency table ● Raw data

● Kappa unweighted
○ Kappa weighted (linear weights)
○ Kappa weighted (quadratic weights)

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

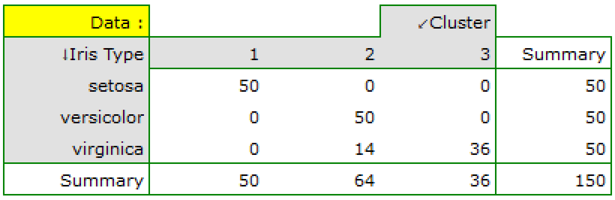
● basic ○ multiple

Report options

☑ Add analyzed data ☐ Combine in one report
☐ Add graph
☐ Add percentages Rows

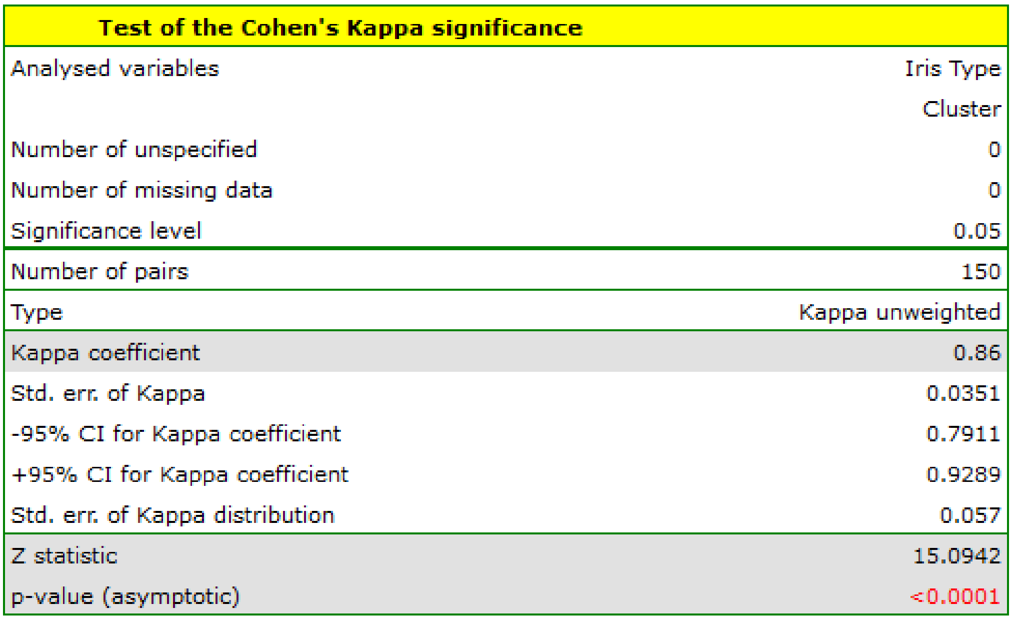
0.05 significance level

For this example, the observed concordance is shown in 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

We conclude from it that the virginica variety can be confused with the versicolor variety, hence we observe 14 misclassifications. However, the Kappa concordance coefficient is statistically significant at 0.86, indicating that the clusters obtained are highly consistent with the actual flower variety.



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

26.2.2 K-means method

K-means method is based on an algorithm initially proposed by Stuart Lloyd and published in 1982 [104]. In this method, objects are divided into a predetermined number of k clusters. The initial clusters are adjusted during the agglomeration procedure by moving objects between them so that the variation of objects within the cluster is as small as possible and the cluster distances are as large as possible. The algorithm works on the basis of the matrix of Euclidean distances between objects, and the parameters necessary in the procedure of agglomeration of the k-means method are: starting centers and stopping criterion. The starting centers are the objects from which the algorithm will start building clusters, and the stopping criterion is the definition of how to stop the algorithm.

AGGLOMERATION PROCEDURE

1. Selection of starting centers
2. Based on the similarity matrix, the algorithm assigns each object to the nearest center
3. For the clusters obtained, the adjusted centers are determined.
4. Steps 2-3 are repeated until the stop criterion is met.

Starting centers
The choice of starting centers has a major impact on the convergence of the k-means algorithm for obtaining appropriate clusters. The starting centers can be selected in two ways:

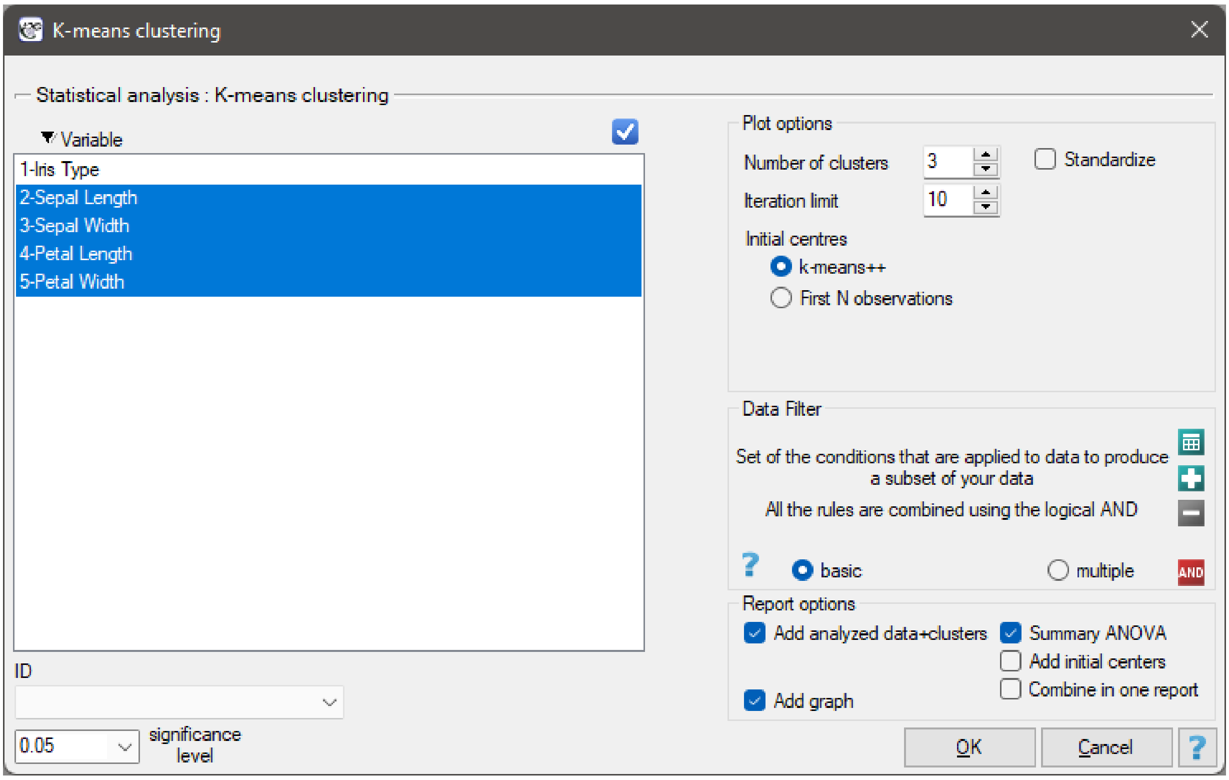
k-means++ - is the optimal selection of starting points by using the k-means++ algorithm proposed in 2007 by David Arthur and Sergei Vassilvitskii [1]. It ensures that the optimal solution of the k-means algorithm is obtained with as few iterations as possible. The algorithm uses an element of randomness in its operation, so the results obtained may vary slightly with successive runs of the analysis. If the data do not form into natural clusters, or if the data cannot be effectively divided into disconnected clusters, using k-means++ will result in completely different results in subsequent runs of the k-means analysis. High reproducibility of the results, on the other hand, demonstrates the possibility of good division into separable clusters.

n firsts - allows the user to indicate points that are start centers by placing these objects in the first positions in the data table.

Stop criterion is the moment when the belonging of the points to the classes does not change or the number of iterations of steps 2 and 3 of the algorithm reaches a user-specified number of iterations.

Because of the way the k-means cluster analysis algorithm works, a natural consequence of it is to compare the resulting clusters using a one-way analysis of variance (**ANOVA**) for independent groups.

Settings window of the k-means cluster analysis is opened via menu Advanced Statistics→Reduction and grouping→K-means cluster analysis.



K-means clustering

Statistical analysis : K-means clustering

▼ Variable ☑

1-Iris Type
2-Sepal Length
3-Sepal Width
4-Petal Length
5-Petal Width

ID

0.05 significance level

Plot options

Number of clusters 3

Iteration limit 10

□ Standardize

Initial centres

● k-means++

○ First N observations

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

☑ Add analyzed data+clusters

☑ Summary ANOVA

□ Add initial centers

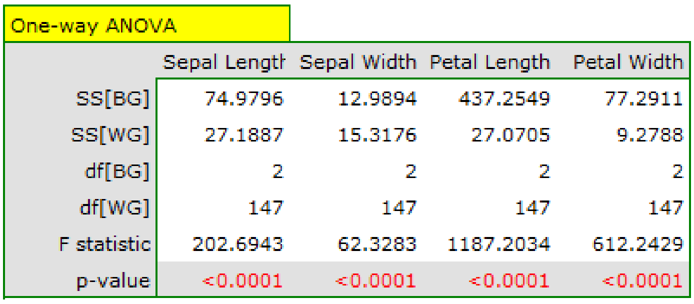
□ Combine in one report

☑ Add graph

Example (26.1) c.d. *(iris.pqs file)*
The analysis will be performed on the classic data set of dividing iris flowers into 3 varieties based on the width and length of the petals and sepal sepals (R.A. Fisher 1936[58]).

We assign flowers to particular groups on the basis of columns from 2 to 5. We also indicate that we want to divide the flowers into 3 groups. As a result of the analysis, 3 clusters were formed, which differed statistically significantly in each of the examined dimensions (ANOVA results), i.e. petal width, petal length, sepal width as well as sepal length.

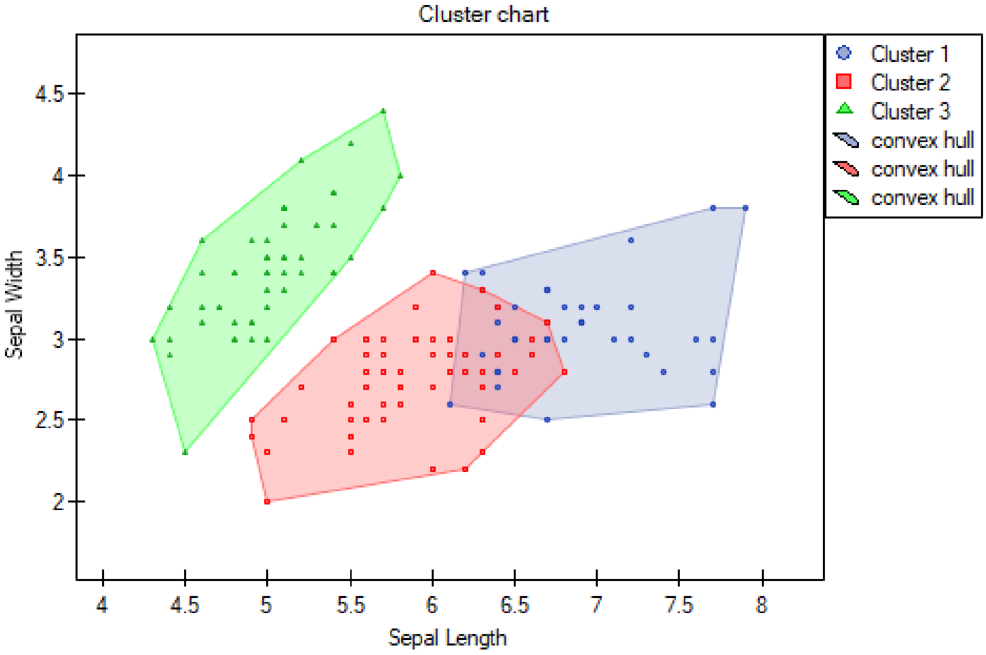
K-means clustering	
Analysed variables	Sepal Length Sepal Width Petal Length Petal Width
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Frequency	150
standardization	No
Iteration limit	10
Initial centers based on	k-means++
Number of clusters	3
Number of elements in cluster 1	39
Number of elements in cluster 2	61
Number of elements in cluster 3	50

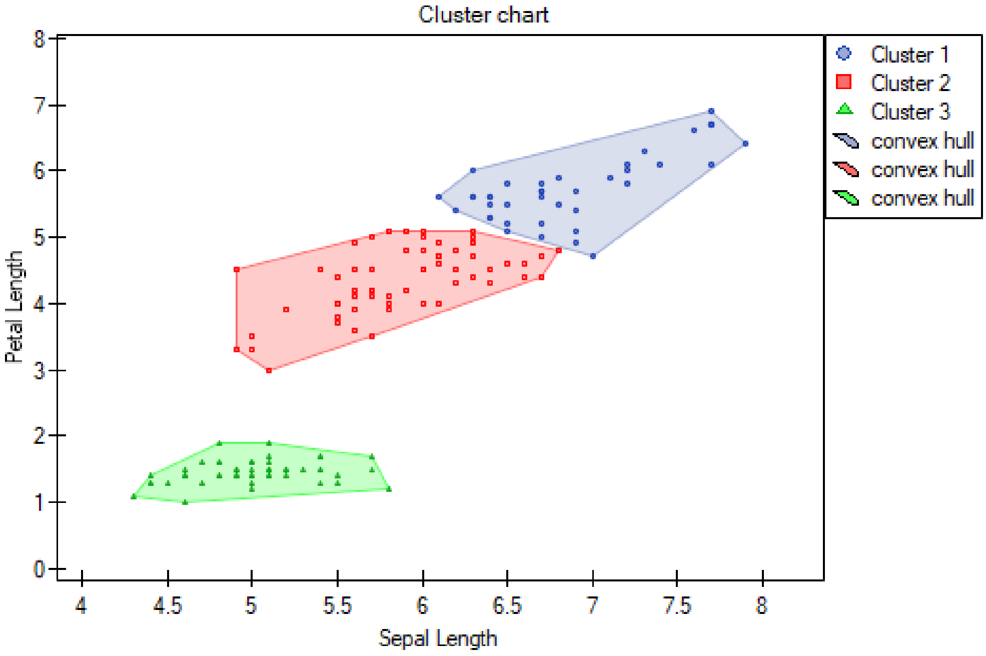


One-way ANOVA

	Sepal Lengtł	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

The difference can be observed in the graphs where we show the belonging of each point to a cluster in the two dimensions selected:





By repeating the analysis we may get slightly different results, but the variation obtained will not be large. It proves that data subjected to analysis form natural clusters and conducting a cluster analysis is justified in this case.

Note no.1!

After running the analysis a few times, we can select the result we are most interested in, and then set those data that are the starting centers at the beginning of the worksheet - then the analysis performed based on the starting centers selected as N first observations will consistently produce that result we selected.

Note no.2!

To find out if the clusters represent the 3 actual varieties of iris flowers, we can copy the information about belonging to a cluster from the report and paste it into the datasheet. We can check the consistency of our results with the actual affiliation of a given flower to the corresponding variety in the same way as for the hierarchical cluster analysis.

27 SURVIVAL ANALYSIS

Survival analysis is often used in medicine. In other fields of study it is also called reliability analysis, duration analysis, or event history analysis. Its main goal is to evaluate the remaining time of the survival of, for example, patients after an operation. Its main purpose is to evaluate the survival time of e.g. patients after surgery - the tools used here are life tables and Kaplan-Meier curves. Another interesting aspect is the comparison of survival times e.g. survival times after different treatments - for this purpose methods of comparing 2 or more survival curves are used. A number of methods (regression models) have also been developed to study the influence of various variables on survival time.

To help understand the issue, basic definitions will be given using an example describing the life expectancy of heart transplant patients:

Event – is the change interesting to the researcher, e.g. death;

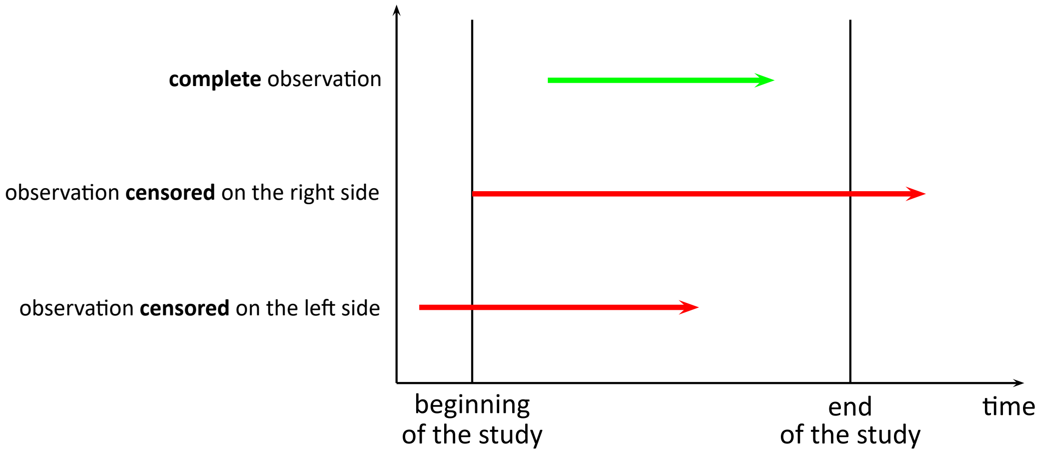
Survival time – is the period of time between the initial state and the occurrence of a given event, e.g. the length of a patient's life after a heart transplantation.

Note!
In the analysis one column with the calculated time ought to be marked. When we have at our disposal two points in time: the initial and the final ones, before the analysis we calculate the time between the two points, using the datasheet formulas.

Censored observations – are the observations for which we only have incomplete pieces of information about the survival time.

Censored and complete observations – an example concerning the survival time after a heart transplantation:

- **a complete observation** – we know the date of the transplantation and the date of the patient's death so we can establish the exact survival time after the transplantation.
- **observation censored on the right side** – the date of the patient's death is not known (the patient is alive when the study finishes) so the exact survival time cannot be established.
- **observation censored on the left side** – the date of the heart transplantation is not known but we know it was before this study started, and we cannot establish the exact survival time.

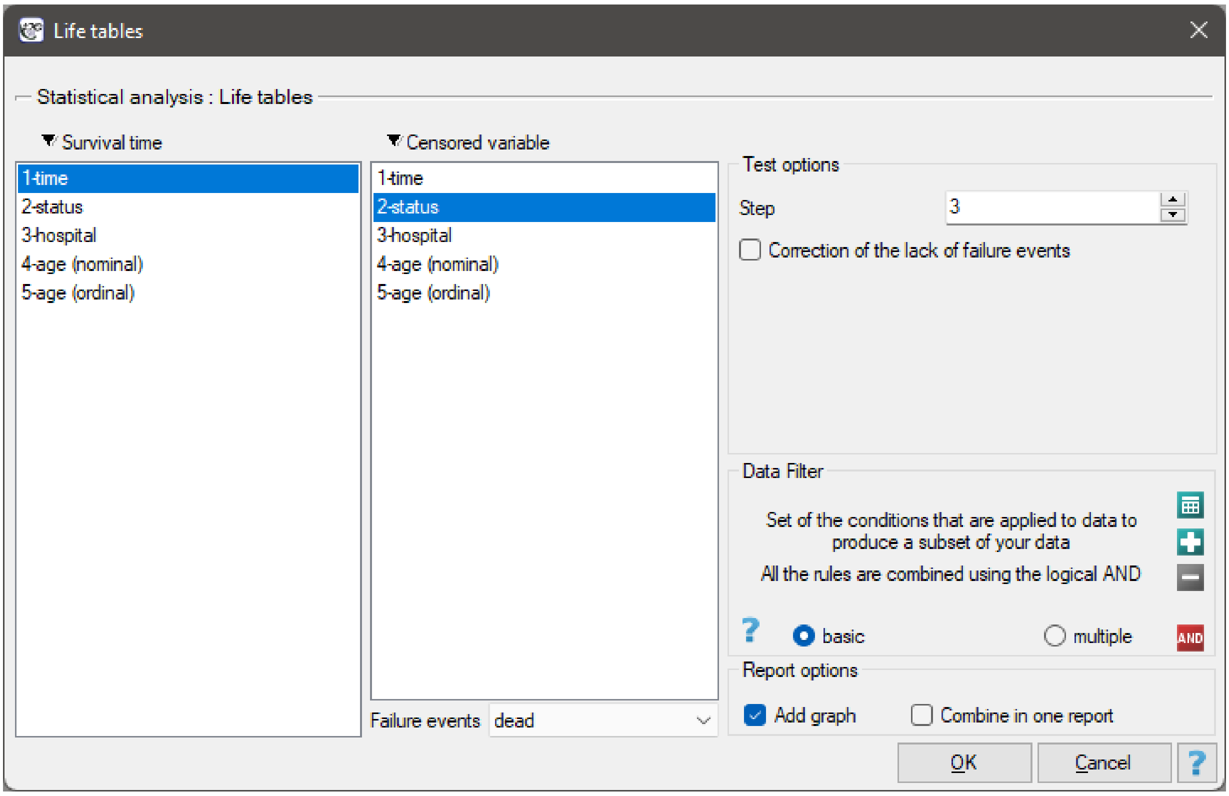

complete observation
observation **censored** on the right side
observation **censored** on the left side
beginning of the study
end of the study
time

Note

The end of the study means the end of the observation of the patient. It is not always the same moment for all patients. It can be the moment of losing touch with the patient (so we do not now the patient's survival time). Analogously, the beginning of the study does not have to be the same point in time for all patients.

27.1 LIFE TABLES

The window with settings for life tables is accessed via the menu Advanced statistics→Survival analysis→Life tables



Life tables

Statistical analysis : Life tables

Survival time: 1-time

Censored variable: 2-status

Failure events: dead

Test options

Step: 3

☐ Correction of the lack of failure events

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

☑ Add graph ☐ Combine in one report

Life tables are created for time ranges with equal spans, provided by the researcher. The ranges can be defined by giving the step. For each range PQStat calculates:

- **the number of entered cases** – the number of people who survived until the time defined by the range;
- **the number of censored cases** – the number of people in a given range qualified as censored cases;
- **the number of cases at risk** – the number of people in a given range minus a half of the censored cases in the given range;
- **the number of complete cases** – the number of people who experienced the event (i.e. died) in a given range;
- **proportions of complete cases** – the proportion of the number of complete cases (deaths) in a given range to the number of the cases at risk in that range;

- **proportions of the survival cases** – calculated as 1 minus the proportion of complete cases in a given range;
- **cumulative survival proportion (survival function)** – the probability of surviving over a given period of time. Because to survive another period of time, one must have survived all the previous ones, the probability is calculated as the product of all the previous proportions of the survival cases.

  $\pm$ standard error of the survival function;
- **probability density** – the calculated probability of experiencing the event (death) in a given range, calculated in a period of time;

  $\pm$ standard error of the probability density;
- **hazard rate** – probability (calculated per a unit of time) that a patient who has survived until the beginning of a given range will experience the event (die) in that range;

  $\pm$ standard error of the hazard rate

Note

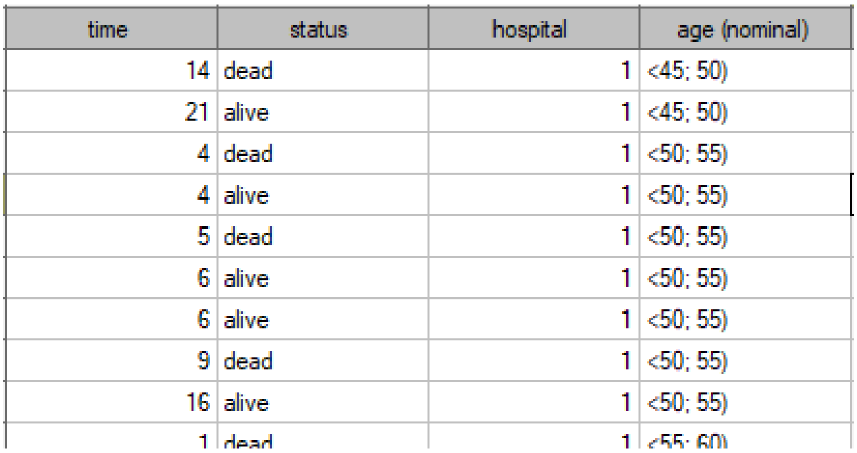
In the case of a lack of complete observations in any range of survival time range there is the possibility of using correction. The zero number of complete cases is then replaced with value 0.5.

Graphic interpretation

We can illustrate the information obtained thanks to the life tables with the use of several charts:

- a cumulative survival proportion graph,
- a probability density graph,
- a hazard rate graph.

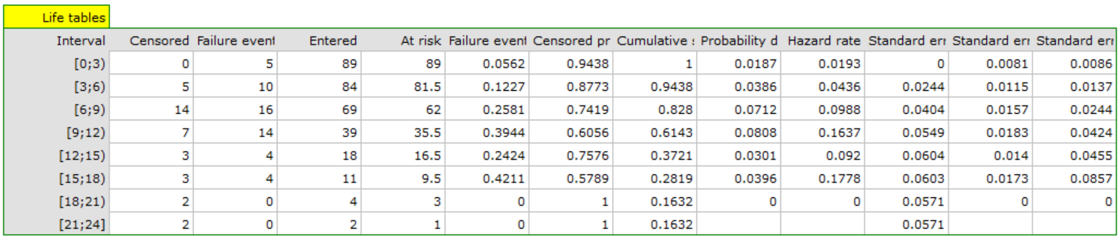
Example 27.1. (transplant.pqs file)
Patients' survival rate after the transplantation of a liver was studied. 89 patients were observed over 21 years. The age of a patient at the time of the transplantation was in the range of $\langle 45\text{years}; 60\text{years})$. A fragment of the collected data is presented in the table below:



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 the analysis are those as to which we have complete information about the length of life after the transplantation, i.e. described as "death" (it concerns 53 people which constitutes

59.55% of sample). The censored data are those about which we do not have that information because at the time when the study was finished the patients were alive (36 people, i.e. 40.45% of them). We build the life tables of those patients by creating time periods of 3 years:



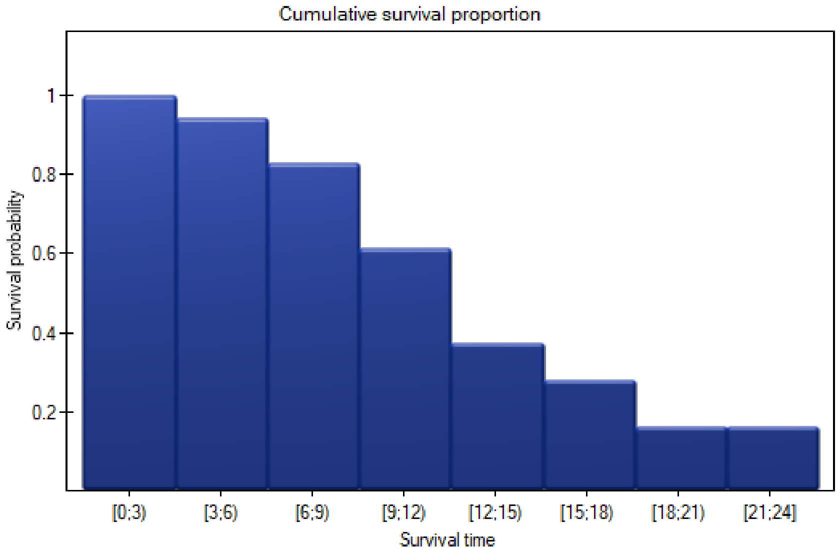
Life tables

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		

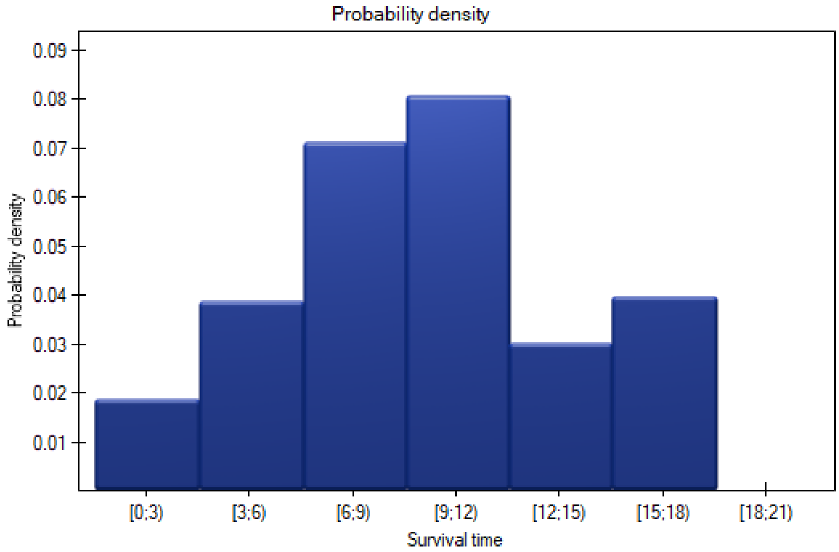
For each 3-year period of time we can interpret the results obtained in the table, for example, for people living for at least 9 years after the transplantation who are included in the range [9;12):

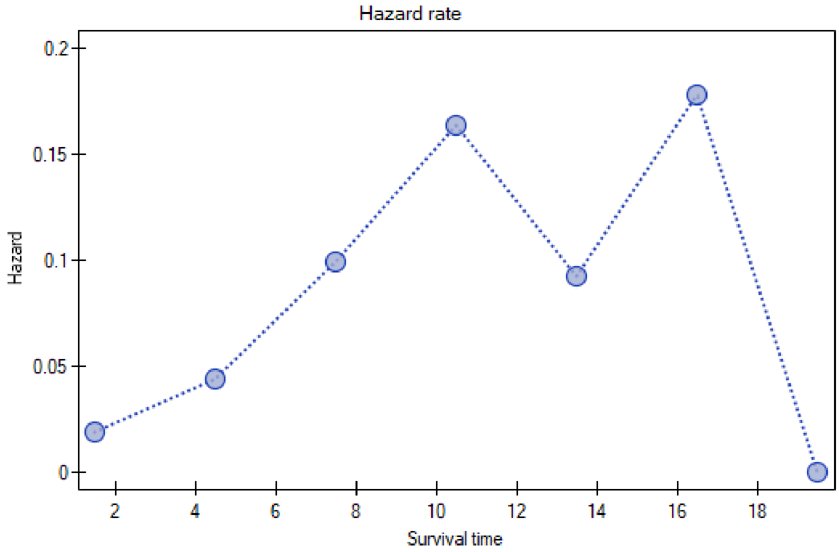
- the number of people who survived 9 years after the transplantation is 39,
- there are 7 people about whom we know they had lived at least 9-12 years at the moment the information about them was gathered but we do not know if they lived longer as they were left out of the study after that time,
- the number of people at the risk of death in that age range is 36,
- there are 14 people about whom we know they died 9 to 12 years after the transplantation,
- 39.4% of the endangered patients died 9 to 12 years after the transplantation,
- 60.6% of the endangered patients lived 9 to 12 years after the transplantation,
- the percent of survivors 9 years after the transplantation is 61.4% $\pm$ 5%,
- 0,08 $\pm$ 0.02 is the death probability for each year from the 9-12 range.

The results will be presented on a few graphs:



The probability of survival decreases with the time passed since the transplantation. We do not, however, observe a sudden plunge of the survival function, i.e. a period of time in which the probability of death would rise dramatically.

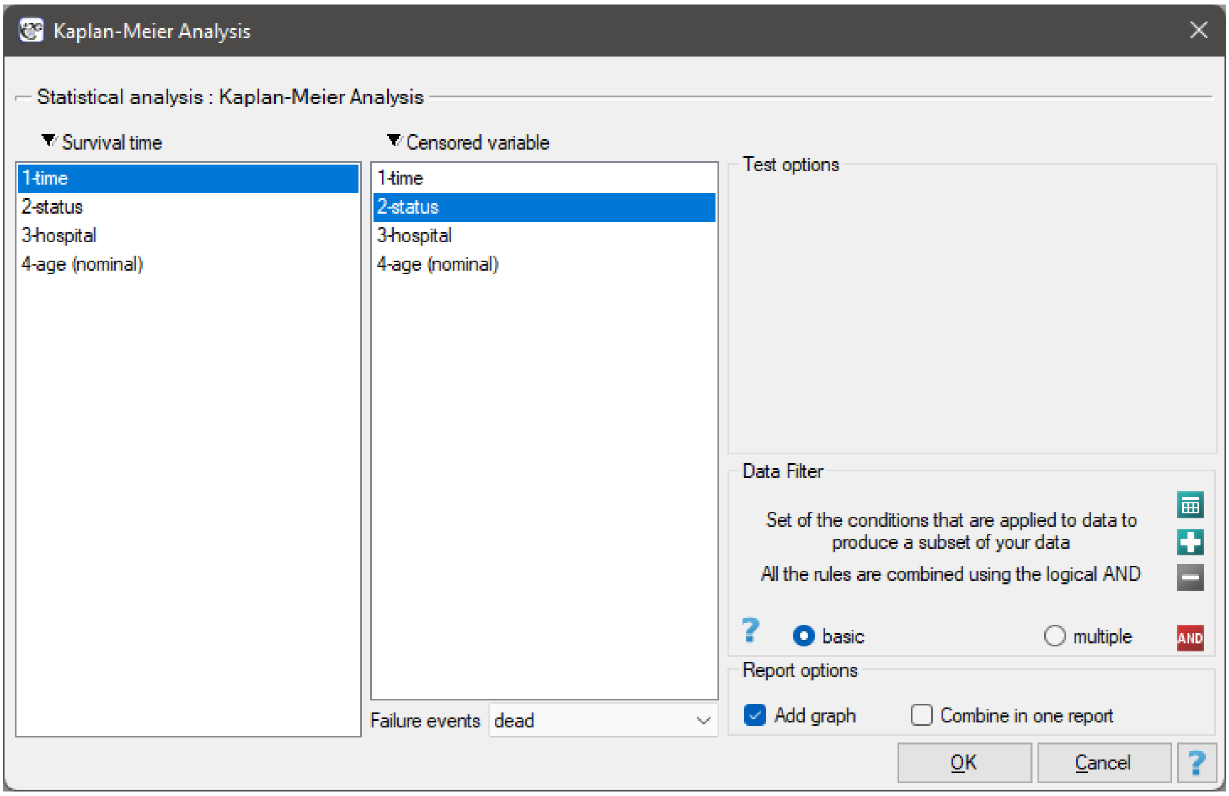




27.2 KAPLAN-MEIER CURVES

Kaplan-Meier curves allow the evaluation of the survival time without the need to arbitrarily group the observations like in the case of life tables. The estimator was introduced by Kaplan and Meier (1958)[88].

The window with settings for Kaplan-Meier curve is accessed via the menu Advanced statistics→Survival analysis→Kaplan-Meier Analysis



Kaplan-Meier Analysis

Statistical analysis : Kaplan-Meier Analysis

Survival time: 1-time (selected), 2-status, 3-hospital, 4-age (nominal)

Censored variable: 1-time, 2-status (selected), 3-hospital, 4-age (nominal)

Failure events: dead

Test options

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

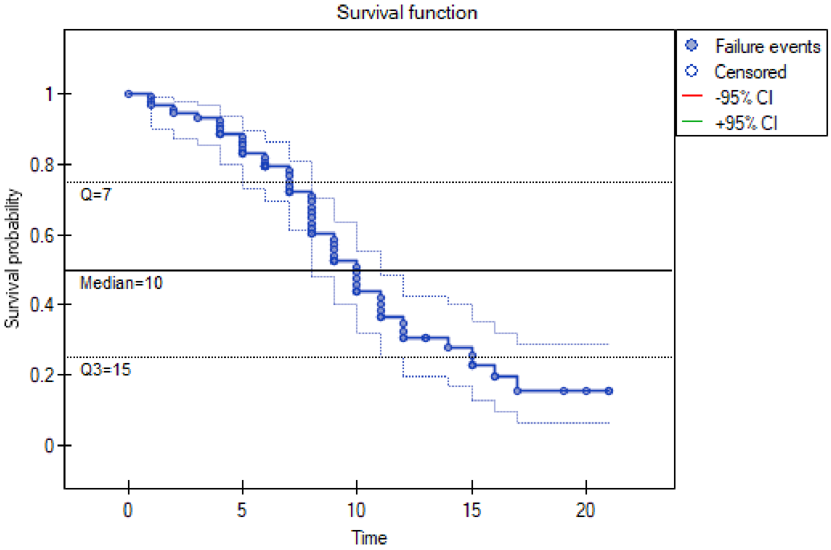
Report options

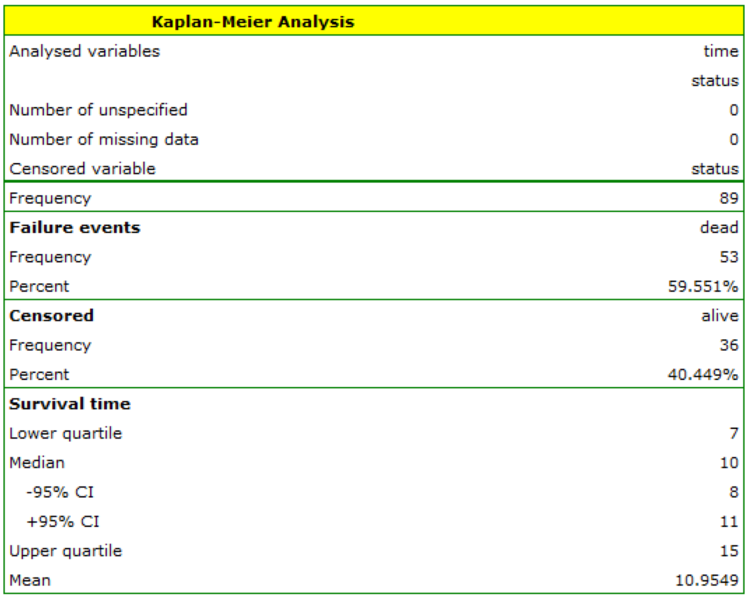
☑ Add graph □ Combine in one report

As with survival tables we calculate the survival function, i.e. the probability of survival until a certain time. The graph of the Kaplan-Meier survival function is created by a step function. Based on the standard error (Greenwood formula) and the logarithmic transformation (log-log), confidence intervals around this curve are constructed. The point of time at which the value of the function is 0.5 is the **survival time median**. The median indicates 50% risk of mortality, it means we can expect that half of the patients will die within a specific time. Both the median and other percentiles are determined as the shortest survival time for which the survival function is smaller or equal to a given percentile. For the median, a confidence interval is determined based on the ”test-based” method by Brookmeyer and Crowley (1982)[28]. **The survival time mean** is determined as the field under the survival curve.

The data concerning the survival time are usually very heavily skewed so in the survival analysis the median is a better measure of the central tendency than the mean.

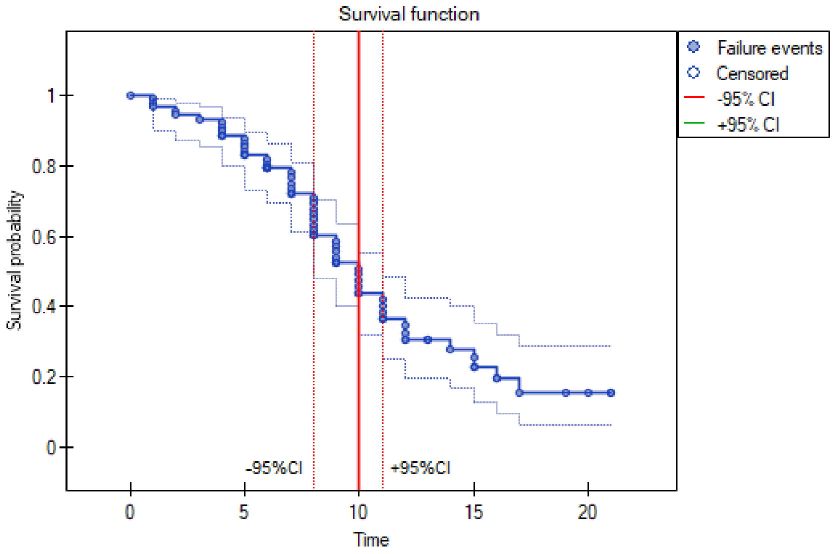
Przykład (27.1) c.d. *(transplant.pqs file)*
We present the survival time after a liver transplantation, with the use of the Kaplan-Meier curve





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 suddenly plunge right after the transplantation. Therefore, we conclude that the initial period after the transplantation does not carry a particular risk of death. The value of the median shows that within 10 years after the transplant, we expect that half of the patients will die. The value is marked on the graph by drawing a line in point 0.5 which signifies the median. In a similar manner we mark the quartiles in the graph.

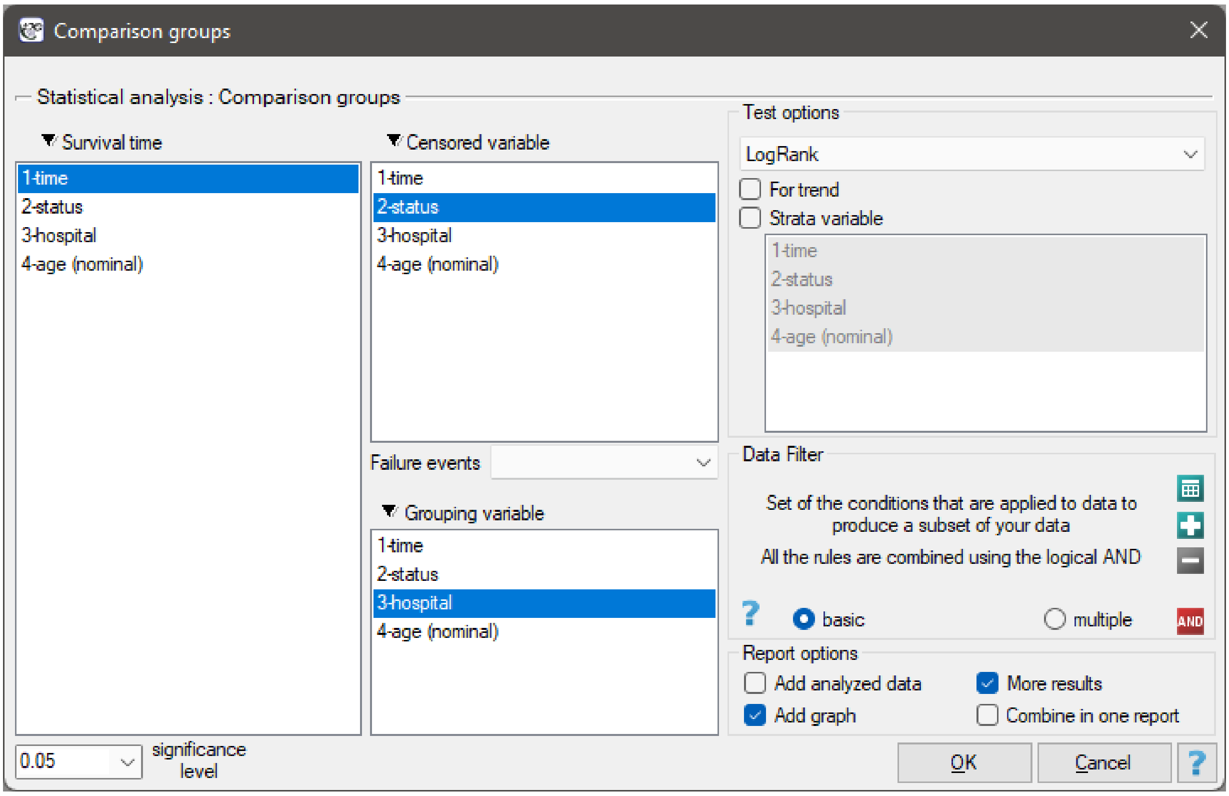


We can visualize the confidence interval for the median on a graph by drawing vertical lines based on the confidence interval around the curve and lines at the 0.5 level.

27.3 COMPARISON OF SURVIVAL CURVES

The survival functions can be built separately for different subgroups, e.g. separately for women and men, and then compared. Such a comparison may concern two curves or more.

The window with settings for the comparison of survival curves is accessed via the menu Advanced statistics→Survival analysis→Comparison groups


Comparison groups

Statistical analysis : Comparison groups

Survival time: 1-time (selected), 2-status, 3-hospital, 4-age (nominal)

Censored variable: 1-time, 2-status (selected), 3-hospital, 4-age (nominal)

Failure events

Grouping variable: 1-time, 2-status, 3-hospital (selected), 4-age (nominal)

0.05 significance level

Test options: LogRank

□ For trend

□ Strata variable: 1-time, 2-status, 3-hospital, 4-age (nominal)

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data ☑ More results

☑ Add graph □ Combine in one report

Comparisons of k survival curves $S_1, S_2, ..., S_k$, at particular points of the survival time t, in the program can be made with the use of three tests:

Log-rank test the most popular test drawing on the Mantel-Heanszel procedure for many 2 x 2 tables (Mantel-Heanszel 1959[109], Mantel 1966[111], Cox 1972[47]),
Gehan's generalization of Wilcoxon's test deriving from Wilcoxon's test (Breslow 1970, Gehan 1965[66][67]),
Tarone-Ware test deriving from Wilcoxon's test (Tarone and Ware 1977[156]).

The three tests are based on the same test statistic, they only differ in **weights** w_j the particular points of the timeline on which the test statistic is based.

Log-rank test: $w_j = 1$ – all the points of the timeline have the same weight which gives the later values of the timeline a greater influence on the result;
Gehan's generalization of Wilcoxon's test: $w_j = n_j$ – time moments are weighted with the number of observations in each of them, so greater weights are ascribed to the initial values of the time line;
Tarone-Ware test: $w_j = \sqrt{n_j}$ – time moments are weighted with the root of the number of observations in each of them, so the test is situated between the two tests described earlier.

An important condition for using the tests above is the proportionality of hazard. Hazard, defined as the slope of the survival curve, is the measure of how quickly a failure event takes place. Breaking the principle of hazard proportionality does not completely disqualify the tests above but it carries some risks. First of all, the placement of the point of the intersection of the curves with respect to the timeline has a decisive influence on decreasing the power of particular tests.

27.3.1 Differences among the survival curves

Hypotheses:

$$\begin{array}{ll} \mathcal{H}_0: & S_1(t) = S_2(t) = ... = S_k(t), \quad \text{for all } t, \\ \mathcal{H}_1: & \text{not all } S_i(t) \text{ are equal.} \end{array}$$

In calculations was used chi-square statistics form:

$$\chi^2 = U'V^{-1}U$$

where:

$U_i = \sum_{j=1}^m w_j(d_{ij} - e_{ij})$

V - covariance matrix of dimensions $(k-1) \times (k-1)$

where:
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 – number of moments in time with failure event (death),

$d_j = \sum_{i=1}^k d_{ij}$ – observed number of failure events (deaths) in the j-th moment of time,

d_{ij} – observed number of failure events (deaths) in the w i-th group w in the j-th moment of time,

$e_{ij} = \frac{n_{ij}d_j}{n_j}$ – expected number of failure events (deaths) in the w i-th group w in the j-th moment of time,

$n_j = \sum_{i=1}^{k} n_{ij}$ – the number of cases at risk in the j-th moment of time.

The statistic asymptotically (for large sizes) has the χ^2 distribution with $df = k - 1$ degrees of freedom.

The p value, estimated nn the basis of test statistics, is compared with the significance level α :

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Hazard ratio

In the log-rank test the observed values of failure events (deaths) $O_i = \sum_{j=1}^{m} d_{ij}$ and the appropriate expected values $E_i = \sum_{j=1}^{m} e_{ij}$ are given.

The measure for describing the size of the difference between a pair of survival curves is the hazard ratio (HR).

$$HR = \frac{O_1/E_1}{O_2/E_2}$$

If the hazard ratio is greater than 1, e.g. $HR = 2$, then the degree of the risk of a failure event in the first group is twice as big as in the second group. The reverse situation takes place when HR is smaller than one. When HR is equal to 1 both groups are equally at risk.

Note

The confidence interval for HR is calculated on the basis of the standard deviation of the HR logarithm (Armitage and Berry 1994[11]).

27.3.2 Survival curves trend

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0 : & \text{In the studied population there is no trend in the placement of the } S_1, S_2, ..., S_k \text{ curves,} \\ \mathcal{H}_1 : & \text{In the studied population there is a trend in in the placement of the } S_1, S_2, ..., S_k \text{ curves.} \end{array}$$

In the calculation the chi-square statistic was used, in the following form:

$$\chi^2 = \frac{(c'U)^2}{c'Vc}$$

where:

$c = (c_1, c_2, ..., c_k)$ – vector of the weights for the compared groups, informing about their natural order (usually the subsequent natural numbers).

The statistic asymptotically (for large sizes) has the χ^2 distribution with 1 degree of freedom.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

In order to conduct a trend analysis in the survival curves the grouping variable must be a numerical variable in which the values of the numbers inform about the natural order of the groups. The numbers in the analysis are treated as the $c_1, c_2, ..., c_k$ weights.

27.3.3 Survival curves for the stratas

Often, when we want to compare the survival times of two or more groups, we should remember about other factors which may have an impact on the result of the comparison. An adjustment (correction) of the analysis by such factors can be useful. For example, when studying rest homes and comparing the length of the stay of people below and above 80 years of age, there was a significant difference in the results. We know, however, that sex has a strong influence on the length of stay and the age of the inhabitants of rest homes. That is why, when attempting to evaluate the impact of age, it would be a good idea to stratify the analysis with respect to sex.

Hypotheses for the differences in survival curves:

$$\begin{array}{ll} \mathcal{H}_0: & S_1^*(t) = S_2^*(t) = ... = S_k^*(t), \quad \text{for all } t, \\ \mathcal{H}_1: & \text{not all } S_i^*(t) \text{ are equal.} \end{array}$$

Hypotheses for the analysis of trends in survival curves:

$\mathcal{H}_0:$ In the studied population there is no trend in the placement of the $S_1^*, S_2^*, ..., S_k^*,$ curves,
$\mathcal{H}_1:$ In the studied population there is a trend in in the placement of the $S_1^*, S_2^*, ..., S_k^*$ curves.

where $S_1^*(t), S_2^*(t), ..., S_k^*(t)$ -are the survival curves after the correction by the variable determining the strata.

The calculations for test statistics are based on formulas described for the tests, not taking into account the strata, with the difference that matrix U and V is replaced with the sum of matrices $\sum_{l=1}^{L} U$ and $\sum_{l=1}^{L} V$. The summation is made according to the strata created by the variables with respect to which we adjust the analysis l=1,2,...,L

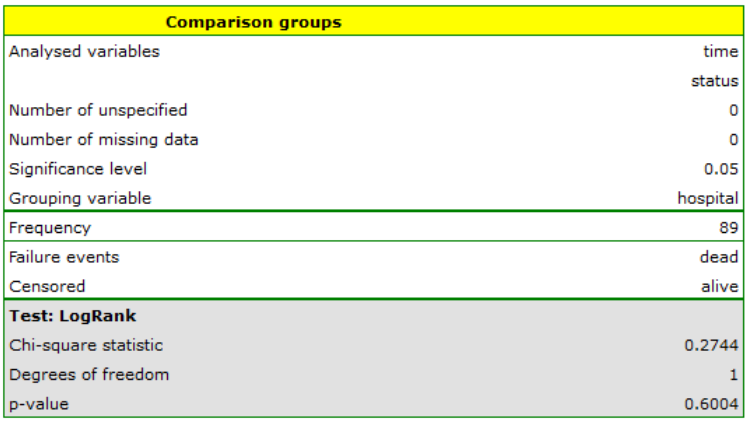
On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

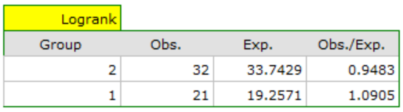
Example (27.1) continued *(transplant.pqs file)*
The differences for two survival curves

Liver transplantations were made in two hospitals. We will check if the patients' survival time after transplantations depended on the hospital in which the transplantations were made. The comparisons of the survival curves for those hospitals will be made on the basis of all tests proposed in the program for such a comparison.
Hypotheses:

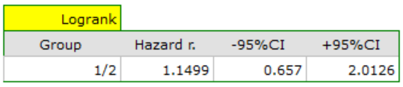
$\mathcal{H}_0:$ the survival curve of the patients of hospital no. 1 $=$ the survival curve of the patients of hospital no. 2,
$\mathcal{H}_1:$ the survival curve of the patients of hospital no. 1 $\neq$ the survival curve of the patients of hospital no. 2.



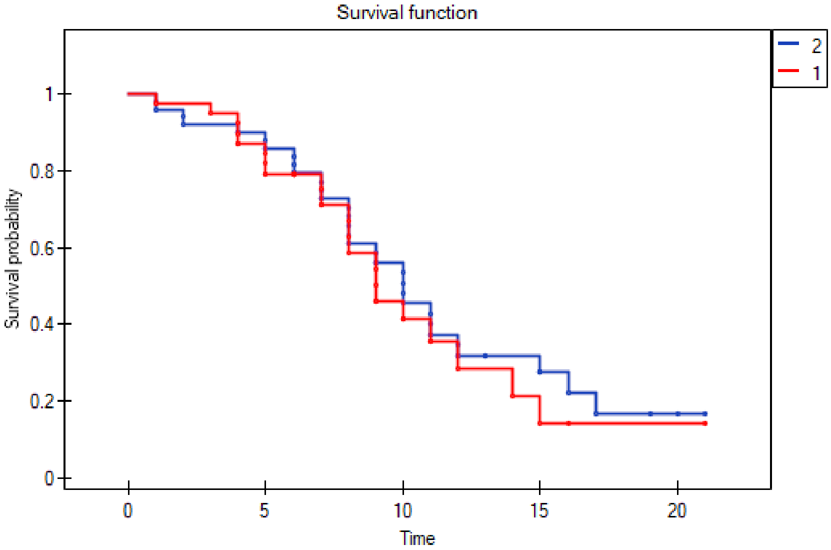
Comparison groups	
Analysed variables	time status
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	hospital
Frequency	89
Failure events	dead
Censored	alive
Test: LogRank	
Chi-square statistic	0.2744
Degrees of freedom	1
p-value	0.6004



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



On the basis of the significance level $\alpha = 0.05$, based on the obtained value p=0.6004 for the log-rank test (p=0.6959 for Gehan's and 0.6465 for Tarone-Ware) we conclude that there is no basis for rejecting the hypothesis $\mathcal{H}_0$. The length of life calculated for the patients of both hospitals is similar.

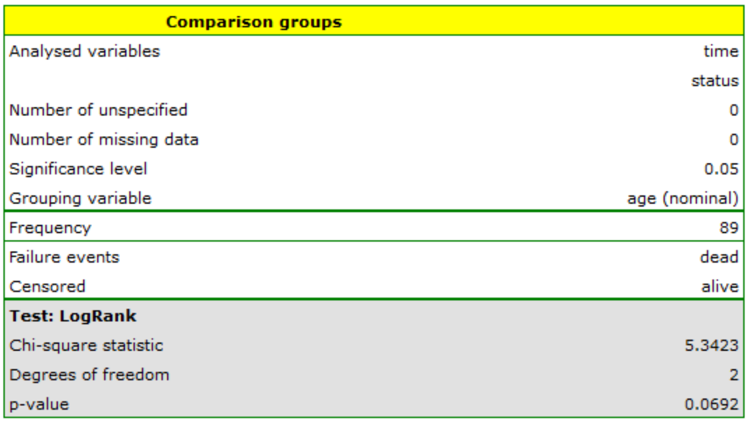
The same conclusion will be reached when comparing the risk of death for those hospitals by determining the risk ratio. The obtained estimated value is $HR = 1.1499$ and 95% of the confidence interval for that value contains 1: $\langle$0.6570, 2.0126$\rangle$.

Differences for many survival curves

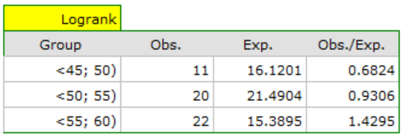
Liver transplantations were made for people at different ages. 3 age groups were distinguished: $\langle 45$ years; 50 years), $\langle 50$ years; 55 years), $\langle 55$ years; 60 years). We will check if the patients' survival time after transplantations depended on their age at the time of the transplantation.
Hypotheses:

$\mathcal{H}_0$: survival rates of patients aged $\langle 45$ years; 50 years), $\langle 50$ years; 55 years), $\langle 55$ years; 60 years) are similar,

$\mathcal{H}_1$: at least one survival curve out of the 3 curves above differs from the other curves.

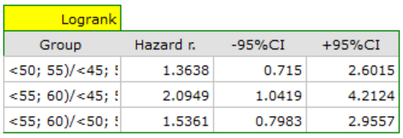


Comparison groups	
Analysed variables	time status
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	age (nominal)
Frequency	89
Failure events	dead
Censored	alive
Test: LogRank	
Chi-square statistic	5.3423
Degrees of freedom	2
p-value	0.0692



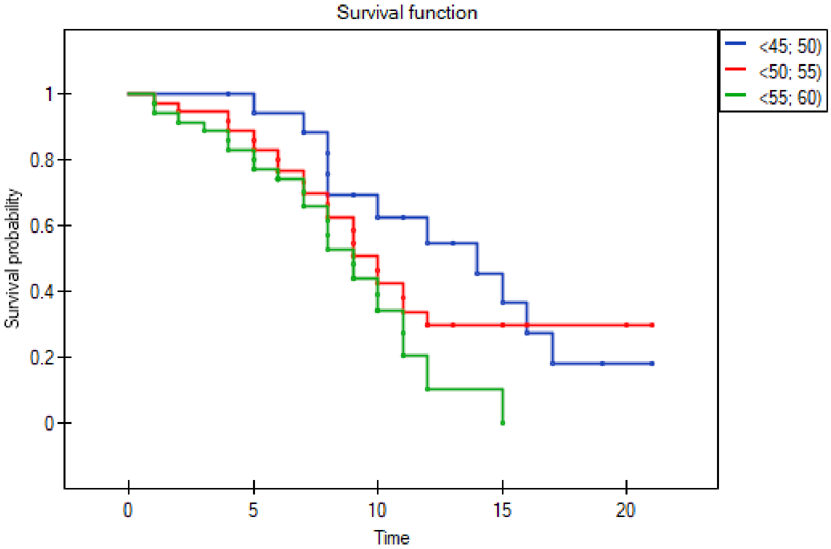
Logrank

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; 5	1.3638	0.715	2.6015
<55; 60)/<45; 5	2.0949	1.0419	4.2124
<55; 60)/<50; 5	1.5361	0.7983	2.9557



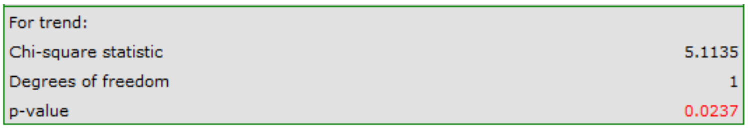
On the basis of the significance level $\alpha = 0.05$, based on the obtained value p=0.0692 in the log-rank test (p=0.0928 for Gehan's and p=0.0779 for Tarone-Ware) we conclude that there is no basis for the rejection of the hypothesis $\mathcal{H}_0$. The length of life calculated for the patients in the three compared age groups is similar. However, it is noticeable that the values are quite near to the standard significance level 0.05.

When examining the hazard values (the ratio of the observed values and the expected failure events) we notice that they are a little higher with each age group (0.68, 0.93, 1.43). Although no statistically significant differences among them are seen it is possible that a growth trend of the hazard value (trend in the position of the survival rates) will be found.

Trend for many survival curves

If we introduce into the test the information about the ordering of the compared categories (we will use the age variable in which the age ranges will be numbered, respectively, 1, 2, and 3), we will be able to check if there is a trend in the compared curves. We will study the following hypotheses:

$\mathcal{H}_0$: a lack of a trend in the survival time curves of the patients after a transplantation
(a trend dependent on the age of the patients at the time of a transplantation),
$\mathcal{H}_1$: the older the patients at the time of a transplantation, the greater/smaller
the probability of their survival over a given period of time.

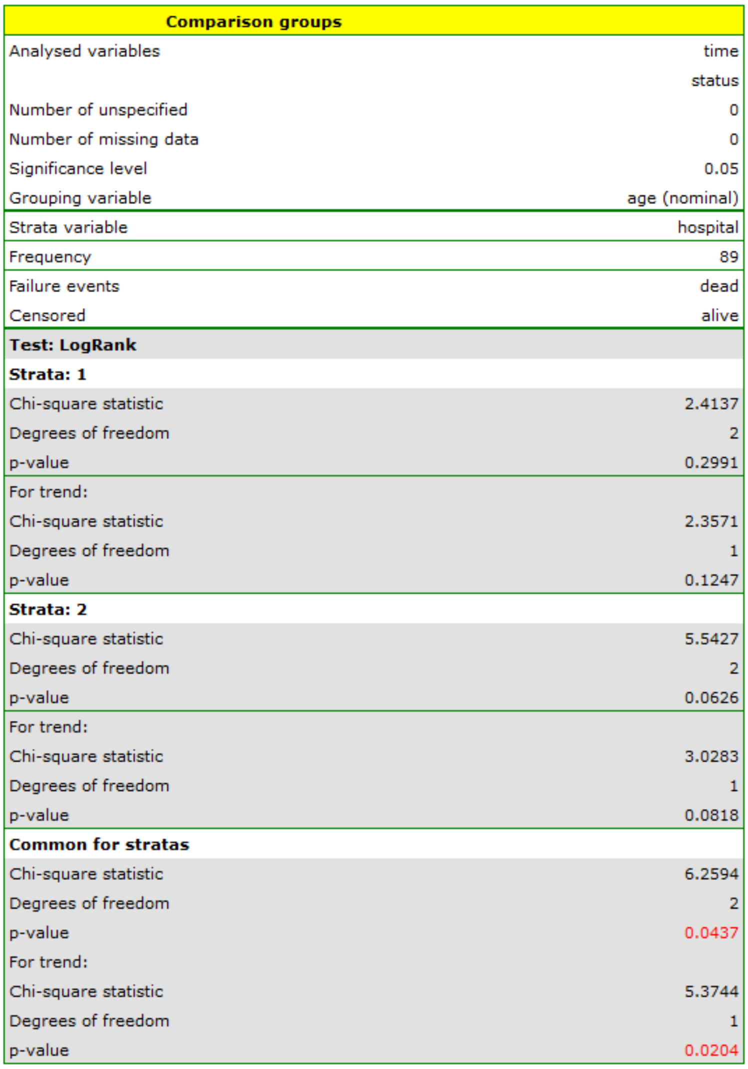


For trend:	
Chi-square statistic	5.1135
Degrees of freedom	1
p-value	0.0237

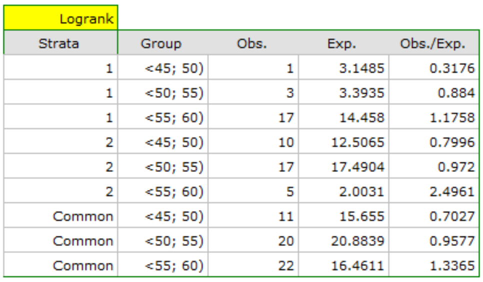
On the basis of the significance level $\alpha = 0.05$, based on the obtained value p=0.0237 in the log-rank test (p=0.0317 for Gehan's and p=0.0241 for Tarone-Ware) we conclude that the survival curves are positioned in a certain trend. On the Kaplan-Meier graph the curve for people aged $\langle$55 years; 60 years) is the lowest. Above that curve there is the curve for patients aged $\langle$50 years; 55 years). The highest curve is the one for patients aged $\langle$45 years; 50 years). Thus, the older the patient at the time of a transplantation, the lower the probability of survival over a certain period of time.

Survival curves for stratas

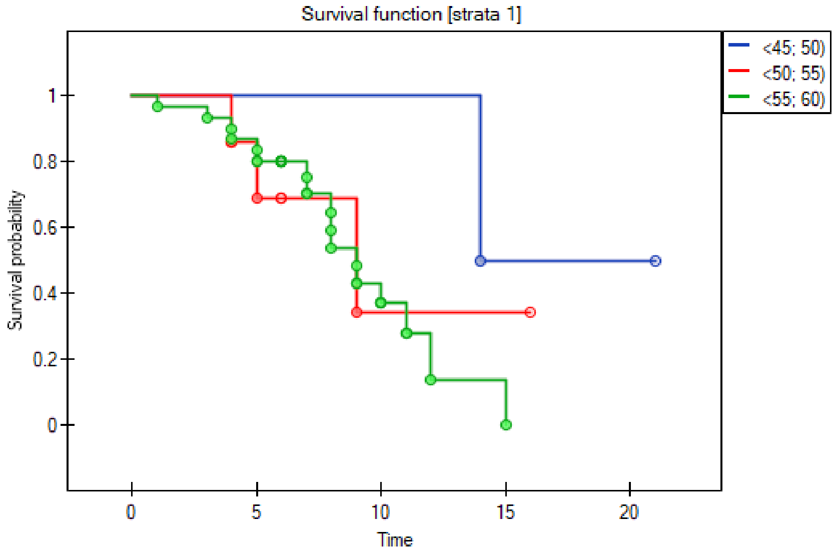
Let us now check if the trend observed before is independent of the hospital in which the transplantation took place. For that purpose we will choose a hospital as the stratum variable.

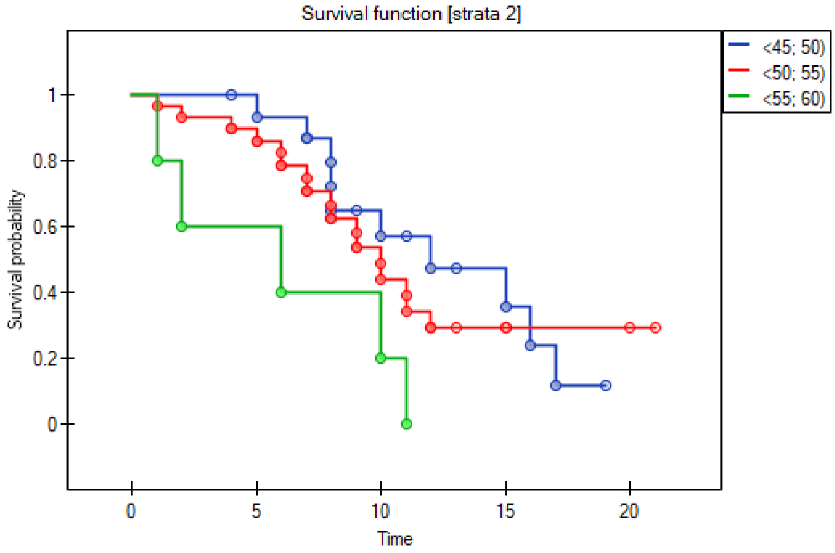


Comparison groups	
Analysed variables	time status
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	age (nominal)
Strata variable	hospital
Frequency	89
Failure events	dead
Censored	alive
Test: LogRank	
Strata: 1	
Chi-square statistic	2.4137
Degrees of freedom	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



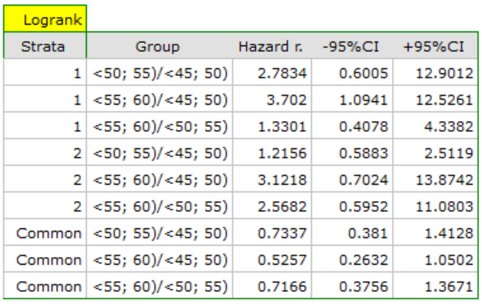
Logrank				
Strata	Group	Obs.	Exp.	Obs./Exp.
1	<45; 50)	1	3.1485	0.3176
1	<50; 55)	3	3.3935	0.884
1	<55; 60)	17	14.458	1.1758
2	<45; 50)	10	12.5065	0.7996
2	<50; 55)	17	17.4904	0.972
2	<55; 60)	5	2.0031	2.4961
Common	<45; 50)	11	15.655	0.7027
Common	<50; 55)	20	20.8839	0.9577
Common	<55; 60)	22	16.4611	1.3365





The report contains, firstly, an analysis of the strata: both the test results and the hazard ratio. In the first stratum the growing trend of hazard is visible but not significant. In the second stratum a trend with the same direction (a result bordering on statistical significance) is observed. A cumulation of those trends in a common analysis of strata allowed the obtainment of the significance of the trend of the survival curves. Thus, the older the patient at the time of a transplantation, the lower the probability of survival over a certain period of time, independently from the hospital in which the transplantation took place.

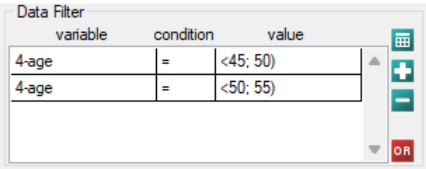
A comparative analysis of the survival curves, corrected by strata, yields a result significant for the log-rank and Tarone-Ware tests and not significant for Gehan's test, which might mean that the differences among the curves are not so visible in the initial survival periods as in the later ones. By looking at the hazard ratio of the curves compared in pairs



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 significant differences. For the comparison of the curve of the youngest group with the curve of the oldest group the hazard ratio is the smallest, 0.53, the 95% confidence interval for that ratio, $\langle 0.26\ ;\ 1.05\rangle$, does contain value 1 but is on the verge of that value, which can suggest that there are significant differences between the respective curves. In order to confirm that supposition an inquisitive researcher can, with the use of the data filter in the analysis window, compare the curves in pairs.

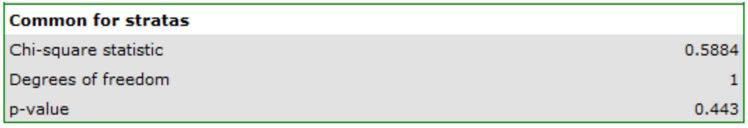


Data Filter

variable	condition	value
4-age	=	<45; 50)
4-age	=	<50; 55)

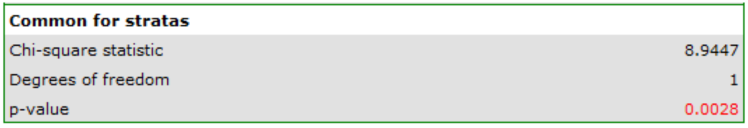
However, it ought to be remembered that one of the corrections for multiple comparisons should be used and the significance level should be modified. In this case, for Bonferroni's correction, with three comparisons, the significance level will be 0.017. For simplicity, we will only avail ourselves of the log-rank test.

$\langle$45 lat; 50 lat) vs $\langle$50 lat; 55 lat)



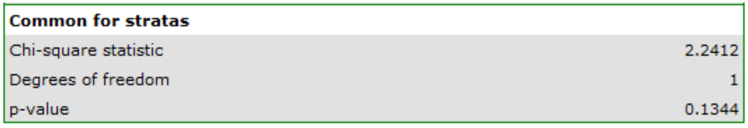
Common for stratas	
Chi-square statistic	0.5884
Degrees of freedom	1
p-value	0.443

$\langle$45 lat; 50 lat) vs $\langle$55 lat; 60 lat)



Common for stratas	
Chi-square statistic	8.9447
Degrees of freedom	1
p-value	0.0028

$\langle$50 lat; 55 lat) vs $\langle$55 lat; 60 lat)

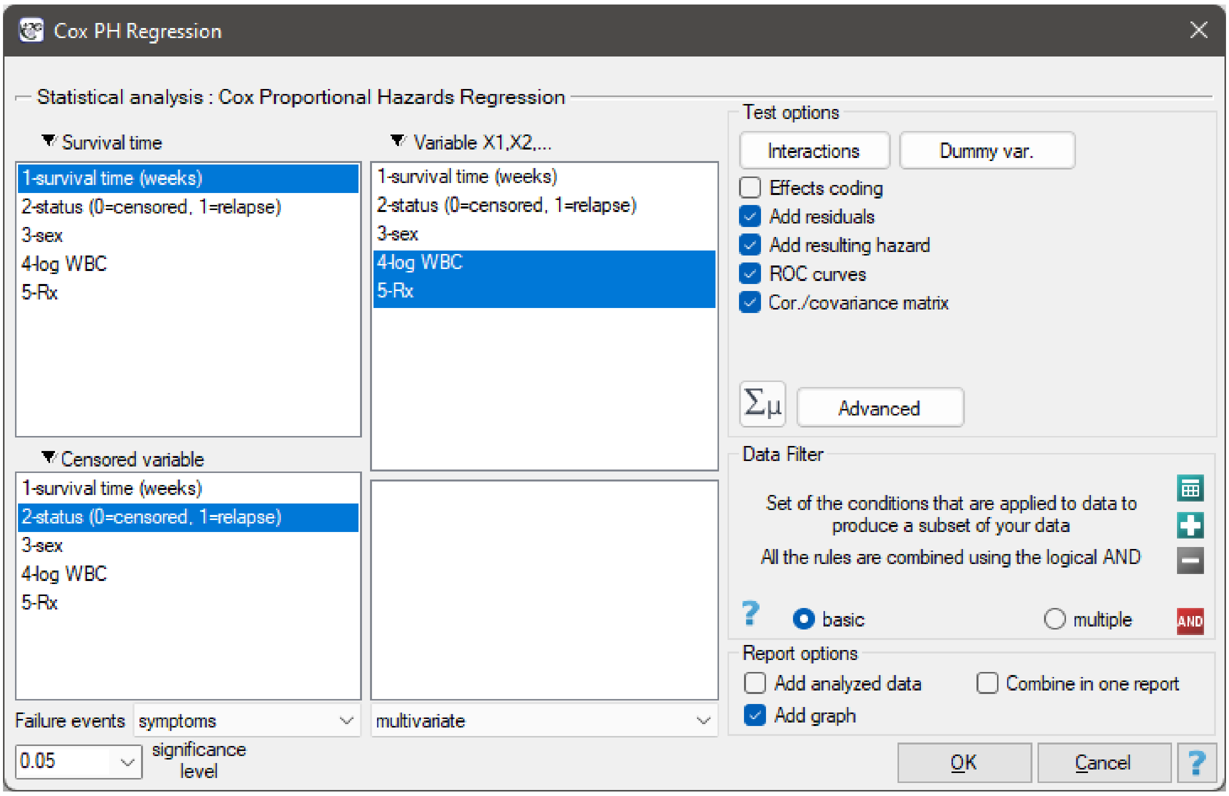


Common for stratas	
Chi-square statistic	2.2412
Degrees of freedom	1
p-value	0.1344

As expected, statistically significant differences only concern the survival curves of the youngest and oldest groups.

27.4 COX PROPORTIONAL HAZARD REGRESSION

The window with settings for Cox regression is accessed via the menu Advanced statistics→Survival analysis→Cox PH regression



Cox PH Regression

Statistical analysis : Cox Proportional Hazards Regression

Survival time: 1-survival time (weeks) (selected); 2-status (0=censored, 1=relapse); 3-sex; 4-log WBC; 5-Rx

Variable X1,X2,...: 1-survival time (weeks); 2-status (0=censored, 1=relapse); 3-sex; 4-log WBC (selected); 5-Rx (selected)

Censored variable: 1-survival time (weeks); 2-status (0=censored, 1=relapse) (selected); 3-sex; 4-log WBC; 5-Rx

Failure events: symptoms

multivariate

0.05 significance level

Test options

☐ Effects coding
☑ Add residuals
☑ Add resulting hazard
☑ ROC curves
☑ Cor./covariance matrix

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

☐ Add analyzed data ☐ Combine in one report
☑ Add graph

Cox regression, also known as the Cox proportional hazard model (Cox D.R. (1972)[47]), is the most popular regressive method for survival analysis. It allows the study of the impact of many independent variables (X_1, X_2, . . ., X_k) on survival rates. The approach is, in a way, non-parametric, and thus encumbered with few assumptions, which is why it is so popular. The nature or shape of the hazard function does not have to be known and the only condition is the assumption which also pertains to most parametric survival models, i.e. hazard proportionality.

The function on which Cox proportional hazard model is based describes the resulting hazard and is the product of two values 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:

$h(t, X_1, X_2, ..., X_k)$ –the resulting hazard describing the risk changing in time and dependent on other factors, e.g. the treatment method,

$h_0(t)$ –the baseline hazard, i.e. the hazard with the assumption that all the explanatory variables are equal to zero,

$\sum_{i=1}^{k} \beta_i X_i$ –a combination (usually linear) of independent variables and model parameters,

$X_1, X_2, . . . X_k$ –explanatory variables independent of time,

$\beta_1, \beta_2, . . . \beta_k$ –parameters.

Dummy variables and interactions in the model

A discussion of the coding of dummy variables and interactions is presented in chapter 24.1 Preparation of the variables for the analysis in multidimensional models.

Correction for ties in Cox regression is based on Breslow's method[27]

The model can be transformed into a the linear form:

$$\ln\left(\frac{h(t, X_1, X_2, ..., X_k)}{h_0(t)}\right) = \sum_{i=1}^{k} \beta_i X_i.$$

In such a case, the solution of the equation is the vector of the estimates of parameters $\beta_0, \beta_1, \ldots, \beta_k$ called **regression coefficients**:

$$b = \left(\begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_k \end{array}\right).$$

The coefficients are estimated by the so-called **partial maximum likelihood estimation**. The method is called "partial" as the search for the maximum of the likelihood function L (the program makes use of the Newton-Raphson iterative algorithm) only takes place for complete data; censored data are taken into account in the algorithm but not directly.

There is a certain error of estimation for each coefficient. The magnitude of that error is estimated from the following formula:

$$SE_b = \sqrt{diag(H^{-1})_b}$$

where:

$diag(H^{-1})$ is the main diagonal of the covariance matrix.

Note
When building a model it ought to be remembered that the number of observations should be ten times greater than or equal to the ratio of the estimated model parameters (k) and the smaller one of the proportions of the censored or complete sizes (p), i.e. ($n \geq 10k/p$) Peduzzi P., et al(1995)[129].

Note
When building the model you need remember that the independent variables should not be multicollinear. In a case of multicollinearity estimation can be uncertain and the obtained error values very high.

Note
The criterion of convergence of the function of the Newton-Raphson iterative algorithm can be controlled with the help of two parameters: the limit of convergence iteration (it gives the maximum number of iterations in which the algorithm should reach convergence) and the convergence criterion (it gives the value below which the received improvement of estimation shall be considered to be insignificant and the algorithm will stop).

27.4.1 Hazard Ratio

An individual hazard ratio (HR) is now calculated for each independent variable :

$$HR_i = e^{\beta_i}.$$

It expresses the change of the risk of a failure event when the independent variable grows by 1 unit. The result is adjusted to the remaining independent variables in the model – it is assumed that they remain stable while the studied independent variable grows by 1 unit.

The HR value is interpreted as follows:

- $HR > 1$ means the stimulating influence of the studied independent variable on the occurrence of the failure event, i.e. it gives information about how much greater the risk of the occurrence of the failure event is when the independent variable grows by 1 unit.
- $HR < 1$ means the destimulating influence of the studied independent variable on the occurrence of the failure event, i.e. it gives information about how much lower the risk is of the occurrence of the failure event when the independent variable grows by 1 unit.
- $HR \approx 1$ means that the studied independent variable has no influence on the occurrence of the failure event (1).

Note

If the analysis is made for a model other than linear or if interaction is taken into account, then, just as in the logistic regression model we can calculate the appropriate HR on the basis of the general formula which is a combination of independent variables.

27.4.2 Model verification

Statistical significance of particular variables in the model (significance of the odds ratio)

On the basis of the coefficient and its error of estimation we can infer if the independent variable for which the coefficient was estimated has a significant effect on the dependent variable. For that purpose we use Wald test.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \beta_i = 0, \\ \mathcal{H}_1: & \beta_i \neq 0. \end{array} \quad \text{or, equivalently:} \quad \begin{array}{cl} \mathcal{H}_0: & OR_i = 1, \\ \mathcal{H}_1: & OR_i \neq 1. \end{array}$$

The Wald test statistics is calculated according to the formula:

$$\chi^2 = \left(\frac{b_i}{SE_{b_i}}\right)^2$$

The statistic asymptotically (for large sizes) has the χ^2 distribution with 1 degree of freedom. On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The quality of the constructed model

A good model should fulfill two basic conditions: it should fit well and be possibly simple. The quality of Cox proportional hazard model can be evaluated with a few general measures based on: L_{FM} –the maximum value of likelihood function of a full model (with all variables),
L_0 –the maximum value of the likelihood function of a model which only contains one free word,
d –the observed number of failure events.

- **Information criteria** are based on the information entropy carried by the model (model insecurity), i.e. they evaluate the lost information when a given model is used to describe the studied phenomenon. We should, then, choose the model with the minimum value of a given information criterion.
  AIC, $AICc$, and BIC is a kind of a compromise between the good fit and complexity. The

second element of the sum in formulas for information criteria (the so-called penalty function) measures the simplicity of the model. That depends on the number of parameters (k) in the model and the number of complete observations (d). In both cases the element grows with the increase of the number of parameters and the growth is the faster the smaller the number of observations.
The information criterion, however, is not an absolute measure, i.e. if all the compared models do not describe reality well, there is no use looking for a warning in the information criterion.

- Akaike information criterion

  $$AIC = -2\ln L_{FM} + 2k,$$

  It is an asymptomatic criterion, appropriate for large sample sizes.
- Corrected Akaike information criterion

  $$AICc = AIC + \frac{2k(k+1)}{d-k-1},$$

  Because the correction of the Akaike information criterion concerns the sample size (the number of failure events) it is the recommended measure (also for smaller sizes).
- Bayesian information criterion or Schwarz criterion

  $$BIC = -2\ln L_{FM} + k\ln(d),$$

  Just like the corrected Akaike criterion it takes into account the sample size (the number of failure events), Volinsky and Raftery (2000)[162].

- **Pseudo R^2** –the so-called McFadden R^2 is a goodness of fit measure of the model (an equivalent of the coefficient of multiple determination R^2 defined for multiple linear regression). The value of that coefficient falls within the range of $<0;1)$, where values close to 1 mean excellent goodness of fit of the model, 0 — a complete lack of fit. Coefficient R^2_{Pseudo} is calculated according to the formula:

  $$R^2_{Pseudo} = 1 - \frac{\ln L_{FM}}{\ln L_0}.$$

  As coefficient R^2_{Pseudo} does not assume value 1 and is sensitive to the amount of variables in the model, its corrected value is calculated:

  $$R^2_{Nagelkerke} = \frac{1-e^{-(2/d)(\ln L_{FM}-\ln L_0)}}{1-e^{(2/d)\ln L_0}} \quad \text{or} \quad R^2_{Cox-Snell} = 1 - e^{\frac{(-2\ln L_0)-(-2\ln L_{FM})}{d}}.$$

- **Statistical significance of all variables in the model**
  The basic tool for the evaluation of the significance of all variables in the model is **the Likelihood Ratio test**. The test verifies the hypothesis:

  $$\begin{array}{cc} \mathcal{H}_0: & \text{all } \beta_i = 0, \\ \mathcal{H}_1: & \text{there is } \beta_i \neq 0. \end{array}$$

  The test statistic has the form presented below:

  $$\chi^2 = -2\ln(L_0/L_{FM}) = -2\ln(L_0) - (-2\ln(L_{FM})).$$

  The statistic asymptotically (for large sizes) has the χ^2 distribution with k degrees of freedom.
  On the basis of test statistics, p value is estimated and then compared with α :

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

- **AUC - area under the ROC curve** –The ROC curve – constructed based on information about the occurrence or absence of an event and the combination of independent variables and model parameters – allows us to assess the ability of the built PH Cox regression model to classify cases into two groups: (1–event) and (0–no event). The resulting curve, and in particular the area under it, illustrates the classification quality of the model. When the ROC curve coincides with the diagonal $y = x$, the decision to assign a case to the selected class (1) or (0) made on the basis of the model is as good as randomly allocating the cases under study to these groups. The classification quality of the model is good when the curve is well above the diagonal $y = x$, that is, when the area under the ROC curve is much larger than the area under the straight line $y = x$, thus larger than 0.5

  Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & AUC = 0.5, \\ \mathcal{H}_1: & AUC \neq 0.5. \end{array}$$

  The test statistic has the form:

$$Z = \frac{AUC - 0.5}{SE_{0.5}},$$

  where:

  $SE_{0.5}$ –field error.

  The statistic Z has asymptotically (for large numbers) a normal distribution.
  On the basis of test statistics, p value is estimated and then compared with α :

$$\begin{array}{lll} \text{if } p \le \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

  In addition, a proposed **cut-off point** value for the combination of independent variables and model parameters is given for the ROC curve.

27.4.3 Analysis of model residuals

The analysis of the of the model residuals allows the verification of its assumptions. The main goal of the analysis in Cox regression is the localization of outliers and the study of hazard proportionality. Typically, in regression models residuals are calculated as the differences of the observed and predicted values of the dependent variable. However, in the case of censored values such a method of determining the residuals is not appropriate. In the program we can analyze residuals described as: Martingale, deviance, and Schoenfeld. The residuals can be drawn with respect to time or independent variables.

Hazard proportionality assumption
A number of graphical methods for evaluating the goodness of fit of the proportional hazard model have been created (Lee and Wang 2003[97]). The most widely used are the methods based on the model residuals. As in the case of other graphical methods of evaluating hazard proportionality this one is a subjective method. For the assumption of proportional hazard to be fulfilled, the residuals should not form any pattern with respect to time but should be randomly distributed around value 0.

Martingale – the residuals can be interpreted as a difference in time $[0, t]$ between the observed number of failure events and their number predicted by the model. The value of the expected residuals is 0 but they have a diagonal distribution which makes it more difficult to interpret the graph (they are in the range of $-\infty$ to 1).

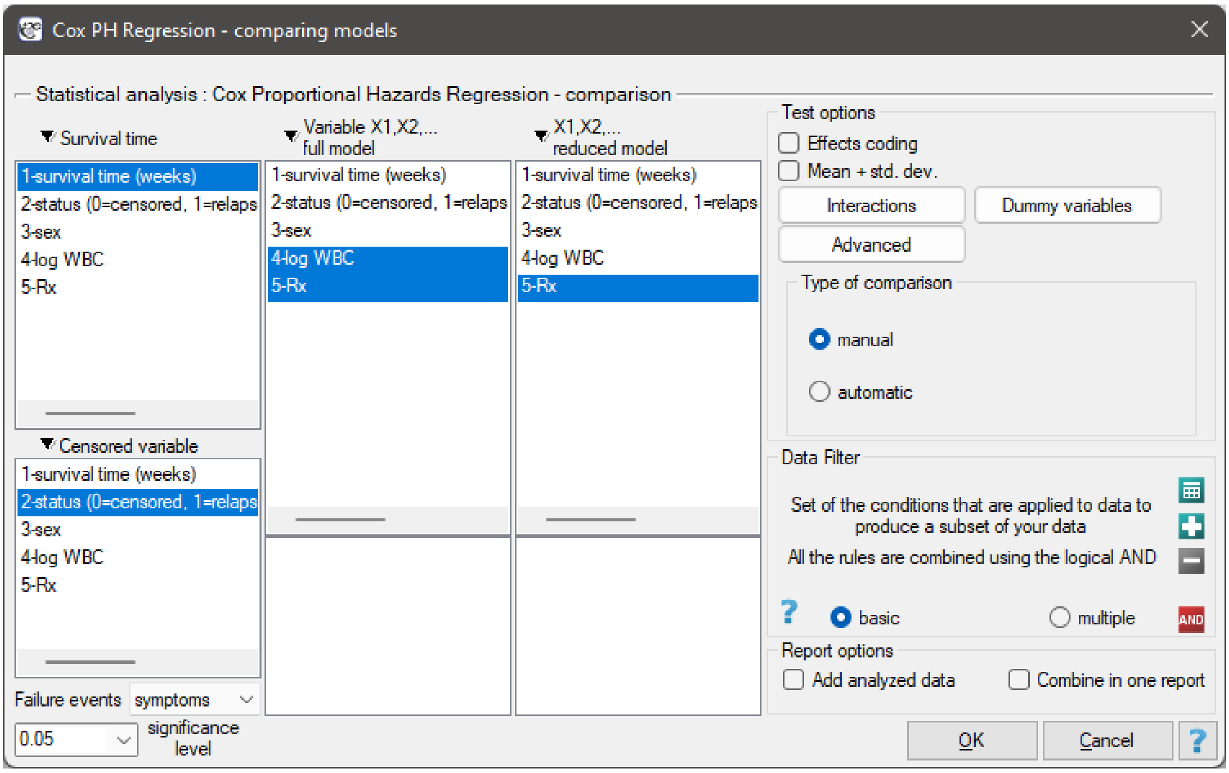
Deviance – similarly to martingale, asymptotically they obtain value 0 but are distributed symmetrically around zero with standard deviation equal to 1 when the model is appropriate. The deviance value is positive when the studied object survives for a shorter period of time than the one expected on the basis of the model, and negative when that period is longer. The analysis of those residuals is used in the study of the proportionality of the hazard but it is mainly a tool for identifying outliers. In the residuals report those of them which are further than 3 standard deviations away from 0 are marked in red.

Schoenfeld – the residuals are calculated separately for each independent variable and only defined for complete observations. For each independent variable the sum of Shoenfeld residuals and their expected value is 0. An advantage of presenting the residuals with respect to time for each variable is the possibility of identifying a variable which does not fulfill, in the model, the assumption of hazard proportionality. That is the variable for which the graph of the residuals forms a systematic pattern (usually the studied area is the linear dependence of the residuals on time). An even distribution of points with respect to value 0 shows the lack of dependence of the residuals on time, i.e. the fulfillment of the assumption of hazard proportionality by a given variable in the model.

If the assumption of hazard proportionality is not fulfilled for any of the variables in Cox model, one possible solution is to make Cox's analyses separately for each level of that variable.

27.5 COMPARISON OF COX PH REGRESSION MODELS

The window with settings for model comparison is accessed via the menu Advanced statistics→Survival analysis→Cox PH Regression – comparing models



Cox PH Regression - comparing models

Statistical analysis : Cox Proportional Hazards Regression - comparison

Survival time: 1-survival time (weeks) [selected]; 2-status (0=censored, 1=relaps; 3-sex; 4-log WBC; 5-Rx

Variable X1,X2,... full model: 1-survival time (weeks); 2-status (0=censored, 1=relaps; 3-sex; 4-log WBC [selected]; 5-Rx [selected]

X1,X2,... reduced model: 1-survival time (weeks); 2-status (0=censored, 1=relaps; 3-sex; 4-log WBC; 5-Rx [selected]

Censored variable: 1-survival time (weeks); 2-status (0=censored, 1=relaps [selected]; 3-sex; 4-log WBC; 5-Rx

Failure events: symptoms

0.05 significance level

Test options

□ Effects coding

□ Mean + std. dev.

Type of comparison

● manual

○ automatic

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options

□ Add analyzed data □ Combine in one report

Due to the possibility of simultaneous analysis of many independent variables in one Cox regression model, there is a problem of selection of an optimum model. When choosing independent variables one has to remember to put into the model variables strongly correlated with the survival time and weakly correlated with one another.

When comparing models with various numbers of independent variables we pay attention to information criteria (AIC, $AICc$, BIC) and to goodness of fit of the model (R^2_{Pseudo}, $R^2_{Nagelkerke}$, $R^2_{Cox-Snell}$). For each model we also calculate the maximum of likelihood function which we later compare with the use of the Likelihood Ratio test.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & L_{FM}=L_{RM},\\ \mathcal{H}_1: & L_{FM}\neq L_{RM},\end{array}$$

where:
L_{FM}, L_{RM} – the maximum of likelihood function in compared models (full and reduced).

The test statistic has the form presented below:

$$\chi^2=-2\ln(L_{RM}/L_{FM})=-2\ln(L_{RM})-(-2\ln(L_{FM}))$$

The statistic asymptotically (for large sizes) has the χ^2 distribution with $df=k_{FM}-k_{RM}$ degrees of freedom, where k_{FM} i k_{RM} is the number of estimated parameters in compared models.

On the basis of test statistics, p value is estimated and then compared with α :

$$\begin{array}{ccl}\text{if } p\leq\alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p>\alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

We make the decision about which model to choose on the basis of the size: AIC, $AICc$, BIC, R^2_{Pseudo}, $R^2_{Nagelkerke}$, $R^2_{Cox-Snell}$ and the result of the Likelihood Ratio test which compares the subsequently created (neighboring) models. If the compared models do not differ significantly, we should select the one with a smaller number of variables. This is because a lack of a difference means that the variables present in the full model but absent in the reduced model do not carry significant information. However, if the difference is statistically significant, it means that one of them (the one with the greater number of variables) is significantly better than the other one.

In the program PQStat the comparison of models can be done manually or automatically.

- **Manual** model comparison – construction of 2 models:
  - a full model – a model with a greater number of variables,
  - a reduced model – a model with a smaller number of variables – such a model is created from the full model by removing those variables which are superfluous from the perspective of studying a given phenomenon.

  The choice of independent variables in the compared models and, subsequently, the choice of a better model on the basis of the results of the comparison, is made by the researcher.

- **Automatic** model comparison is done in several steps:

  step 1 Constructing the model with the use of all variables.

  step 2 Removing one variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.

  step 3 A comparison of the full and the reduced model.

step 4 Removing another variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.

step 5 A comparison of the previous and the newly reduced model.

...

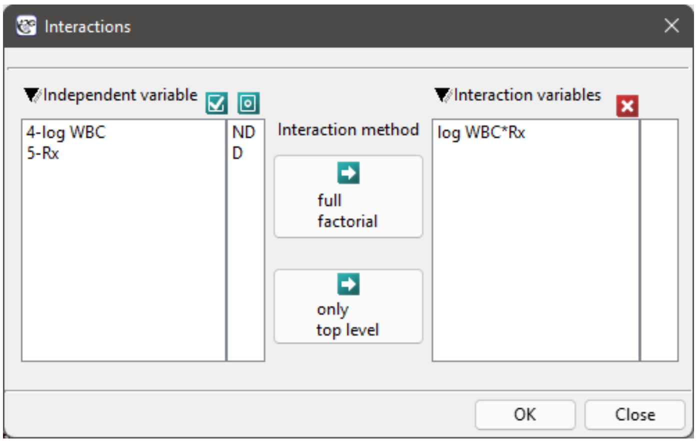
In that way numerous, ever smaller models are created. The last model only contains 1 independent variable.

EXAMPLE 27.2. (remissionLeukemia.pqs file)

The analysis is based on the data about leukemia described in the work of Freirich et al. 1963[63] and further analyzed by many authors, including Kleinbaum and Klein 2005[91]. The data contain information about the time (in weeks) of remission until the moment when a patient was withdrawn from the study because of an end of remission (a return of the symptoms) or of the censorship of the information about the patient. The end of remission is the result of a failure event and is treated as a **complete** observation. An observation is **censored** if a patient remains in the study to the end and remission does not occur or if the patient leaves the study.

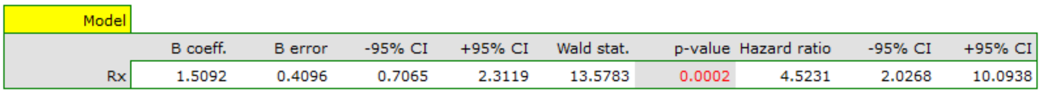
Patients were assigned to one of two groups: a group undergoing traditional treatment (marked as 1 and colled ”placebo group”) and a group with new kind of treatment (marked as 0). The information about the patients’ sex was gathered (1=man, 0=woman) and about the values of the indicator of the number of white cells, marked as ”log WBC”, which is a well-known prognostic factor.

The aim of the study is to determine the influence of kind of treatment on the time of remaining in remission, taking into account possible confounding factors and interactions. In the analysis we will focus on the ”Rx (1=placebo, 0=new treatment)” variable. We will place the ”log WBC” variable in the model as a possible confounding factor (which modifies the effect). In order to evaluate the possible interactions of ”Rx” and ”log WBC” we will also consider a third variable, a ratio of the interacting variables. We will add the variable to the model by selecting, in the analysis window, the Interactions button and by setting appropriate options there.



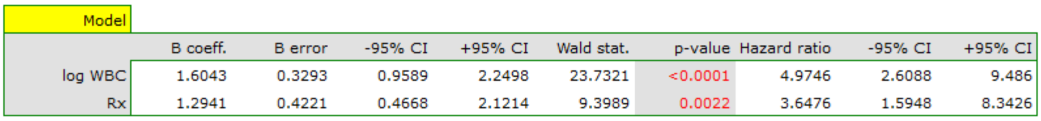
We build three Cox models:

Model A only contains the ”Rx” variable:



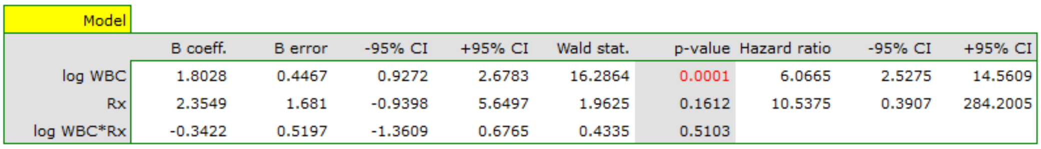
Model	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
Rx	1.5092	0.4096	0.7065	2.3119	13.5783	0.0002	4.5231	2.0268	10.0938

Model B contains the ”Rx” variable and the potentially confounding variable ”log WBC”:



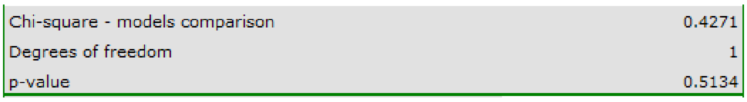
Model	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard ratio	-95% CI	+95% CI
log WBC	1.6043	0.3293	0.9589	2.2498	23.7321	<0.0001	4.9746	2.6088	9.486
Rx	1.2941	0.4221	0.4668	2.1214	9.3989	0.0022	3.6476	1.5948	8.3426

Model C ontains the ”Rx” variable, the ”log WBC” variable, and the potential effect of the interactions of those variables: ”Rx × 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			

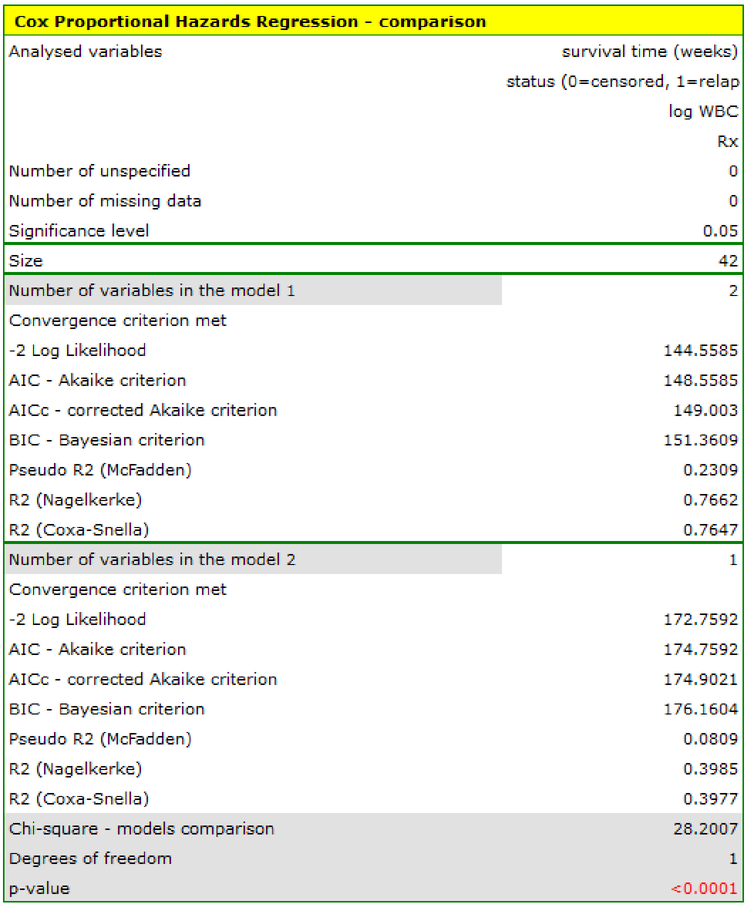
The variable which informs about the interaction of ”Rx” and ”log WBC”, included in model C, is not significant in model C, according to the Wald test. Thus, we can view further consideration of the interactions of the two variables in the model to be unnecessary. We will obtain similar results by comparing, with the use of a likelihood ratio test, model C with model B. We can make the comparison by choosing the Cox PH regression – comparing models menu. We will then obtain a non-significant result (p=0.5134) which means that model C (model with interaction) is NOT significantly better than model B (model without interaction).



Chi-square - models comparison	0.4271
Degrees of freedom	1
p-value	0.5134

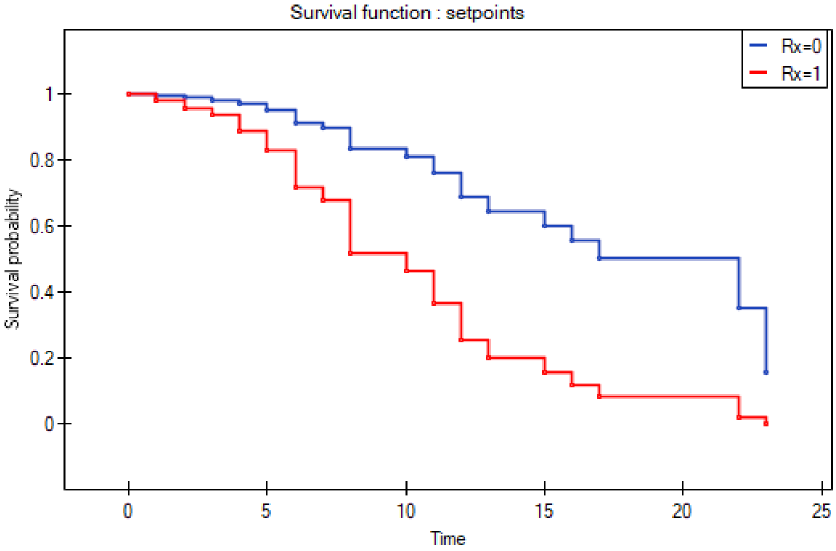
Therefore, we reject model C and move to consider model B and model A.

HR for ”Rx” in model B is 3.65 which means that hazard for the ”placebo group” is about 3.6 greater than for the patients undergoing new treatment. Model A only contains the ”Rx” variable, which is why it is usually called a ”crude” model – it ignores the effect of potential confounding factors. In that model the HR for ”Rx” is 4.52 and is much greater than in model B. However, let us look not only at the point values of the HR estimator but also at the 95% confidence interval for those estimators. The range for ”Rx” in model A is 8.06 (10.09 minus 2.03) wide and is narrower in model B: 6.74 (8.34 minus 1.60). That is why model B gives a more precise HR estimation than model A. In order to make a final decision about which model (A or B) will be better for the evaluation of the effect of treatment (”Rx”) we will once more perform a comparative analysis of the models in the Cox PH pregression – comparing models module. This time the likelihood ratio test yields a significant result (p<0.0001), which is the final confirmation of the superiority of model B. That model has the lowest value of information criteria (AIC=148.6, AICc=149 BIC=151.4) and high values of goodness of fit (Pseudo $R^2_{McFadden} = 0.2309$, $R^2_{Nagelkerke} = 0.7662$, $R^2_{Cox-Snell} = 0.7647$).

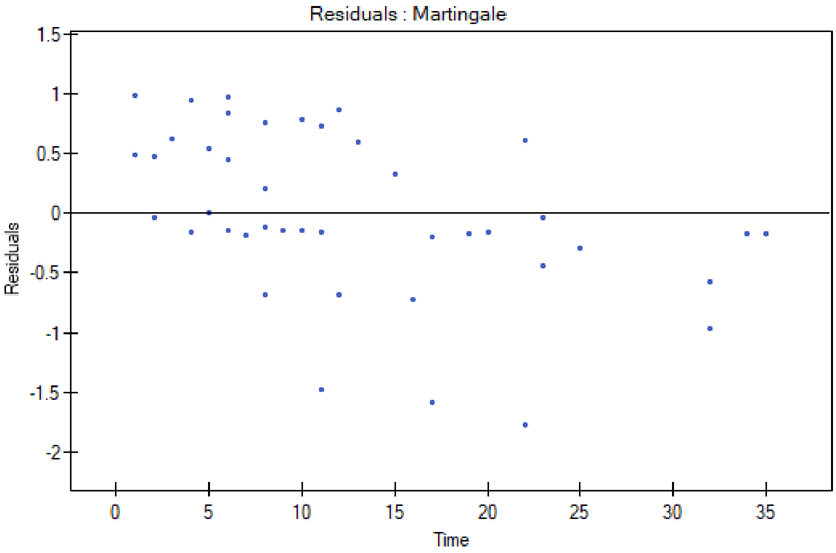


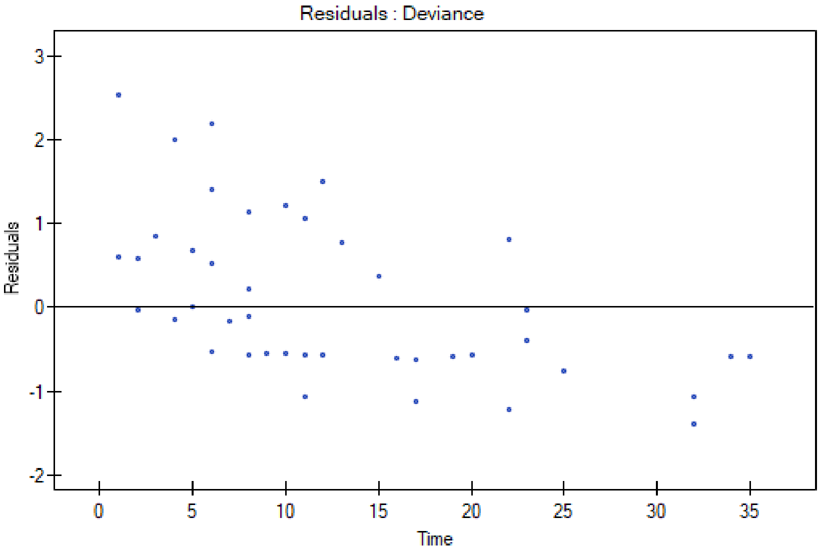
Cox Proportional Hazards Regression - comparison	
Analysed variables	survival time (weeks) status (0=censored, 1=relap log WBC Rx
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Size	42
Number of variables in the model 1	2
Convergence criterion met	
-2 Log Likelihood	144.5585
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

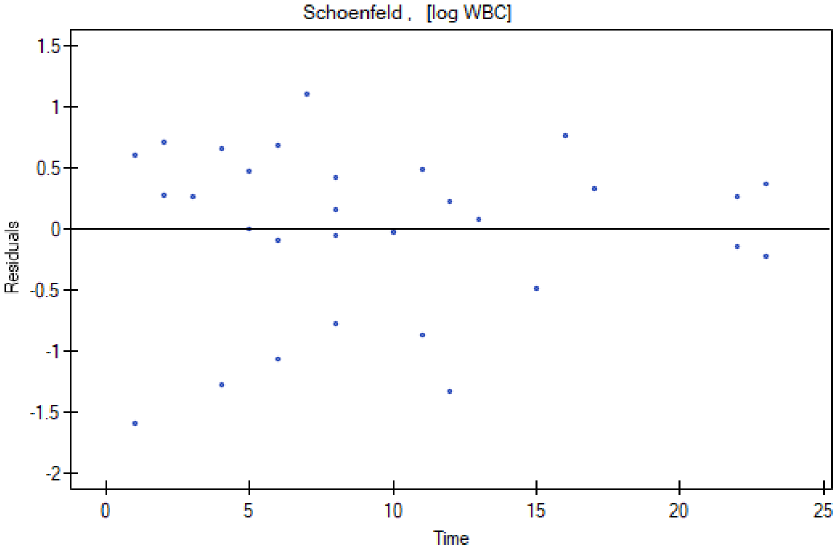
The analysis is complemented with the presentation of the survival curves of both groups, the new treatment one and the traditional treatment one, corrected by the influence of ”log WBC”, for model B. In the graph we observe the differences between the groups, which occur at particular points of survival time. To plot such curves, after selecting Add Graph, we check the Survival Function: in subgroups ... and then, to quickly build a graph of two curves, I choose Quick subgroups and indicate the variable Rx. The Advanced subgroups option allows you to build any number of arbitrarily defined curves.

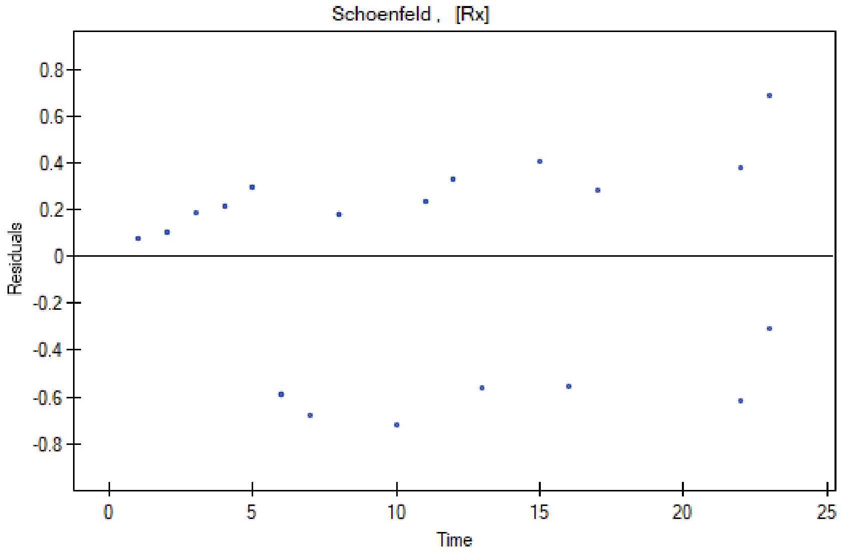


At the end we will evaluate the assumptions of Cox regression by analyzing the model residuals with respect to time.


Residuals : Martingale
Residuals
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
5
10
15
20
25
30
35
Time


Residuals : Deviance
Residuals
3
2
1
0
-1
-2
0
5
10
15
20
25
30
35
Time

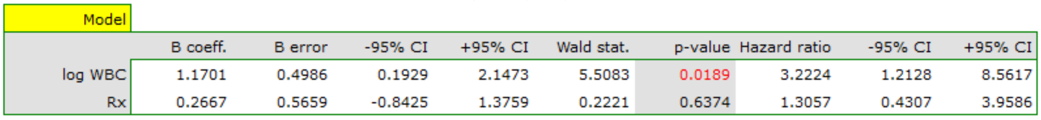

Schoenfeld , [log WBC]
Residuals
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
5
10
15
20
25
Time



We do not observe any outliers, however, the martingale and deviance residuals become lower the longer the time. Shoenfeld residuals have a symmetrical distribution with respect to time. In their case the analysis of the graph can be supported with various tests which can evaluate if the points of the residual graph are distributed in a certain pattern, e.g. a linear dependency. In order to make such an analysis we have to copy Shoenfeld residuals, together with time, into a datasheet, and test the type of the dependence which we are looking for. The result of such a test for each variable signifies if the assumption of hazard proportionality by a variable in the model has been fulfilled. It has been fulfilled if the result is statistically insignificant and it has not been fulfilled if the result is statistically significant. As a result the variable which does not fulfill the regression assumption of the Cox proportional hazard can be excluded from the model. In the case of the ”Log WBC” and ”Rx” variables the symmetrical distribution of the residuals suggests the fulfillment of the assumption of hazard proportionality by those variables. That can be confirmed by checking the correlation, e.g. Pearson’s linear or Spearman’s monotonic, for those residuals and time.

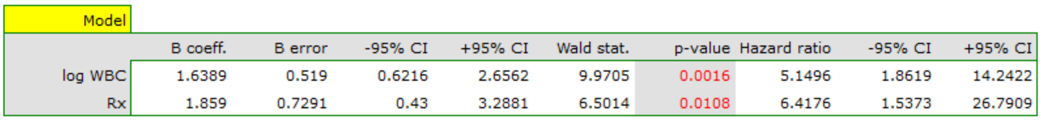
Later we can add the sex variable to the model. However, we have to act with caution because we know, from various sources, that sex can have an influence on the survival function as regards leukemia, in that survival functions can be distributed disproportionately with respect to each other along the time line. That is why we create the Cox model for three variables: ”Sex”, ”Rx”, and ”log WBC”. Before interpreting the coefficients of the model we will check Schonfeld residuals. We will present them in graphs and their results, together with time, will be copied from the report to a new data sheet where we will check the occurrence of Spearman’s monotonic correlation. The obtained values are p=0.0259 (for the time and Shoenfeld residuals correlation for sex), p=0.6192 (for the time and Shoenfeld residuals correlation for log WBC), and p=0,1490 (for the time and Shoenfeld residuals correlation for Rx) which confirms that the assumption of hazard proportionality has not been fulfilled by the sex variable. Therefore, we will build the Cox models separately for women and men. For that purpose we will make the analysis twice, with the data filter switched on. First, the filter will point to the female sex (0), second, to the male sex (1).

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



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

28 META-ANALYSIS

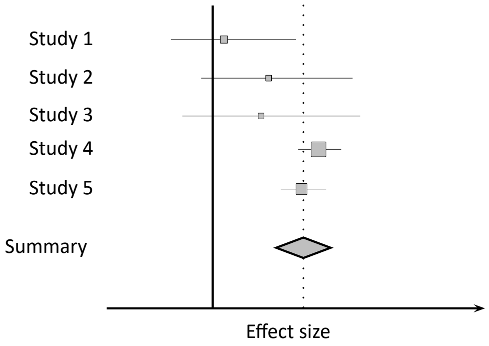
The number of scientific papers being published has increased tremendously in the last decade. This comes with a number of benefits, but it makes it difficult to keep up with the ever-emerging new information. For example, if a doctor were to use a new treatment for his patients based on a scientific paper he had read, he could make a mistake. The error could come from the fact that a whole host of other papers have been published that contradict the effectiveness of that treatment. In order for a doctor's decision to have the least amount of error, he or she should read most of the scientific papers that have been published on the topic. As a result, the constant review of the growing body of literature would take up so much time that there may not be enough time to treat patients. A meta-analysis allows such a review to be done quickly because it is the result of an extensive literature review and a statistical summary of the findings presented therein.

Meta-analysis in PQStat is performed using the following measures:

- **Mean difference**,
- **d Cohen**,
- **g Hedges**,
- **Ratio of two means**,
- **Odds Ratio (OR)**,
- **Relative risk (RR)**,
- **Risk difference (RD)**,
- **Pearson coefficient**,
- **AUC for ROC curve**,
- **Proportion**.

28.1 Introduction

The most familiar image associated with meta-analysis is the forest plot showing the results of each study along with a summary.



In order for the selected literature to be summarized together, it must be consistent in description and the measures given there must be the same.

To be used in a meta-analysis, a scientific paper should describe:

Final result which is some kind of statistical measure indicating the result (effect) obtained in the paper. In fact, these can be different kinds of measures, e.g., difference between means, odds ratio, relative risk, etc.

Standard Error i.e. **SE** allowing one to determine the precision of the study carried out. This precision assigns the **study weight**. The smaller the error (SE), the higher the precision of a given study and the higher the assigned weight will be, making a given study more likely to contribute to the results of a meta-analysis.

Group size is the number of objects on which the study was conducted.

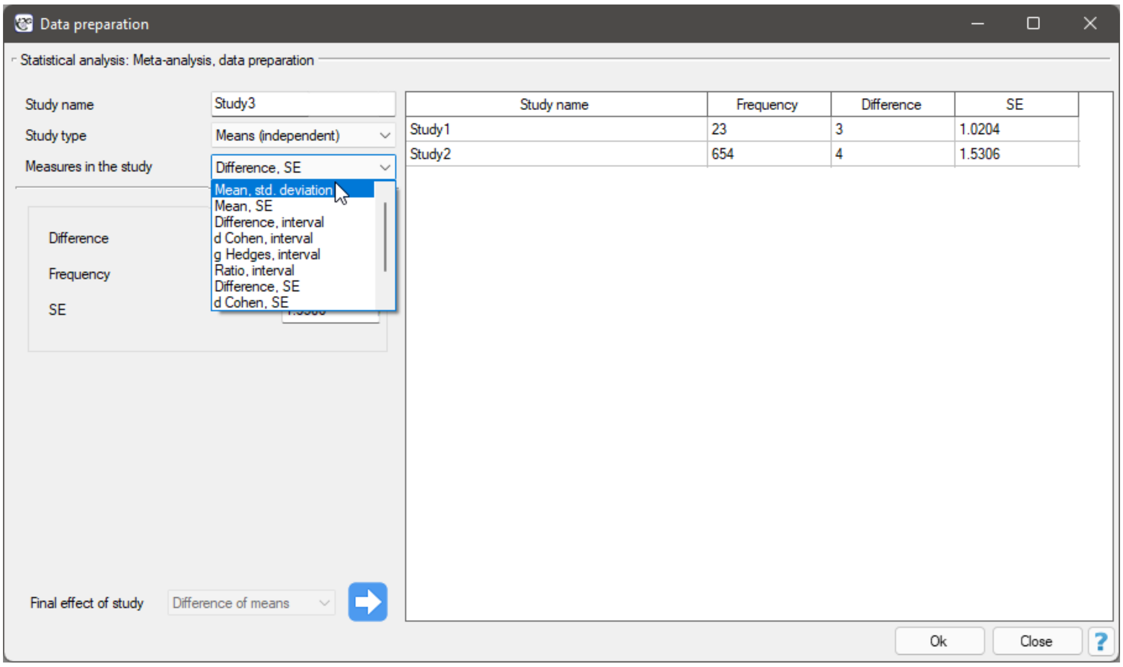
Note!

It often happens, that a scientific paper does not contain all of the elements listed above, in which case you should look for data in the paper from which the calculation of these measures will be possible.

Note!

The PQstat program performs meta-analysis related calculations on data containing: Final Effect, Standard Error, and in some situations Group Size. It is recommended that you enter the data for each publication in the data preparation window before performing the meta-analysis. This is particularly handy when a paper does not explicitly provide these three measures.

The data preparation window is opened via menu:
Advanced Statistics→Meta-analysis→Data preparation.



Data preparation

Statistical analysis: Meta-analysis, data preparation

Study name: Study3
Study type: Means (independent)
Measures in the study: Difference, SE

Difference
Frequency
SE

Final effect of study: Difference of means

Study name	Frequency	Difference	SE
Study1	23	3	1.0204
Study2	654	4	1.5306

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, as it will describe it in all meta-analysis results, including graphs. The desired Final result, Effect Error and Group Size are calculated based on measures extracted from the relevant scientific paper. The measures included in the studies from which the final results can be calculated are shown in the table below:

Study type	Study measures	Final result
Independent means	Means, st. deviations	**a**, **b**, **c**, **e**
	Means, SE	**a**, **b**, **c**, **e**
	Difference, interval	**a**
	d Cohen, interval	**b**
	g Hadges, interval	**c**
	Ratio, interval	**e**
	Difference, SE	**a**
	d Cohen, SE	**b**
	g Hadges, SE	**c**
	Ratio, SE	e
Dependent means	Means, st. deviations	**a**, **b**, **c**
	Means, SE	**a**, **b**, **c**
	Difference, interval	**a**
	d Cohen, interval	**b**
	g Hadges, interval	**c**
	Difference, st. deviation	**a**, **b**, **c**
	Difference, SE	**a**, **b**, **c**
	d Cohen, SE	**b**
	g Hadges, SE	**c**
Mean vs. set	Mean, st. deviation	**a**
	Mean, SE	**a**
Mean - summary	Mean, st. deviation	**d**
	Mean, SE	**d**
Tables	2x2 Tables	**b**, **c**, **f**, **g**, **h**
	OR, interval	**b**, **c**, **f**
	RR, interval	**g**
	RD, interval	**h**
	d Cohen, interval	**b**, **c**, **f**
	g Hadges, interval	**b**, **c**, **f**
	OR, SE	**b**, **c**, **f**
	RR, SE	**g**
	RD, SE	**h**
	d Cohen, SE	**b**, **c**, **f**
	g Hadges, SE	**b**, **c**, **f**
Correlation	Coefficient	**i**
	Coefficient, interval	**i**
	Coefficient, SE	**i**
ROC curve	AUC, SE	**j**
	AUC, interval	**j**
One proportion	Set group size	**k**
	Set proportion	**k**

where the individual end results are:

a - Difference of means

b - d Cohen

c - g Hadges

d - Mean

e - Ratio of means

f - Odds Ratio (OR)

g - Relative risk (RR)

h - Risk differential (RD)

i - Pearson coefficient

j - AUC (ROC curve)

k - Proportion

Note!
In determining the error of coefficients such as OR or RR and others based on tables, when there exist values of zero in the tables or in determining the error of proportions when the proportion is 0 or 1, a continuity correction using an increase factor of 0.5 is applied. The confidence interval for proportions is determined according to the exact Clopper-Pearson method[38].

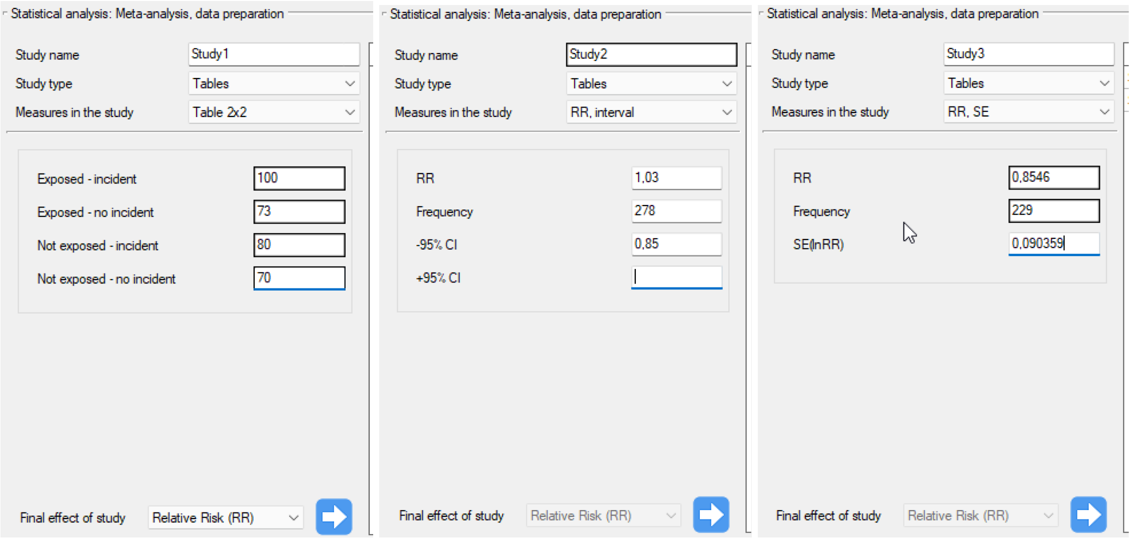
Example 28.1. We are interested in the effect of cigarette smoking on the risk of disease X. We want to conduct a meta-analysis for which the end result will be relative risk (RR). Under this assumption, the papers selected for the meta-analysis must be able to calculate the **RR** and its **error**.

Step 1. Based on the description of the final result (see table above), it was found that the relative risk (described as the score **g**) is possible to determine in the PQStat program in three situations, i.e., if the RR and the confidence interval for it or the RR together with the error are given in the scientific paper, or if the corresponding group sizes in four categories are given, i.e., a 2x2 table.

Step 2. Ten papers were selected for meta-analysis that met the inclusion criteria and had the potential to determine relative risk (see step 1). The needed data included in the selected papers were:

Study 1: group sizes: (smokers and sick)=100, (smokers and non-sick)=73, (non-smokers and sick)=80, (non-smokers and non-sick)=70,
Study 2: group sizes: (smokers and sick)=182, (smokers and non-sick)=172, (non-smokers and sick)=180, (non-smokers and non-sick)=172,
Study 3: group sizes: (smokers and sick)=157, (smokers and non-sick)=132, (non-smokers and sick)=125, (non-smokers and non-sick)=201,
Study 4: group sizes: (smokers and sick)=19, (smokers and non-sick)=15, (non-smokers and sick)=35, (non-smokers and non-sick)=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,

Step 3. KUsing the study preparation window for the meta-analysis, data was input into the datasheet. The first four studies are entered by selecting tables, studies five through nine are entered by selecting RR and range, and the last study provides all the necessary data, i.e., RR and SE. We set Relative risk (RR) as the Final effect of the study:



Statistical analysis: Meta-analysis, data preparation

Study name	Study1
Study type	Tables
Measures in the study	Table 2x2
Exposed - incident	100
Exposed - no incident	73
Not exposed - incident	80
Not exposed - no incident	70
Final effect of study	Relative Risk (RR)

Statistical analysis: Meta-analysis, data preparation

Study name	Study2
Study type	Tables
Measures in the study	RR, interval
RR	1,03
Frequency	278
-95% CI	0,85
+95% CI	
Final effect of study	Relative Risk (RR)

Statistical analysis: Meta-analysis, data preparation

Study name	Study3
Study type	Tables
Measures in the study	RR, SE
RR	0,8546
Frequency	229
SE(lnRR)	0,090359
Final effect of study	Relative Risk (RR)

We move each entered study to the window on the right-hand side ➡. Using the OK button, we transfer the prepared studies to the datasheet. Based on the information about each study in the datasheet, you can proceed to perform a meta-analysis.

P-value, and thus statistical significance is not directly used in meta-analysis. The same effect size may be statistically significant in a large study and insignificant in a study based on small sample size. Moreover, a quite small effect size may be statistically significant in a large study, and a quite large effect size may be insignificant in a small study. This fact is related to the power of statistical tests. When testing for statistical significance, we are testing whether an effect exists at all, i.e., whether it is different from zero, not whether it is large enough to translate into desired effects. For example, the fact that a drug statistically significantly lowers blood pressure by 1mmHg will not result in it being used, because 1mmHg is too small from a clinical perspective. Meta-analysis focuses on the magnitude of individual effects rather than their statistical significance. As a result, it does not matter much whether the papers used in the meta-analysis indicate statistical significance of a particular effect or not.

In PQStat, statistical significance is calculated for each study given the effect ratio and the error of that effect. This is an asymptotic approach, based on a normal distribution and dedicated to large samples. If a different test was used to check statistical significance in the cited study, the results obtained may differ slightly.

28.2 Summary effect

As a result of the meta-analysis, its most desirable element is to summarize the collected studies, i.e., to report the overall effect, M. Such a summary can be done in two ways, by designating a fixed effect or a random effect.

Fixed effect

In calculating the fixed effect, we assume that all studies in the meta-analysis share one common true effect. Thus, if each study involved the same population, e.g., the same country, then to summarize the meta-analysis with a fixed effect we assume that the true (population) effect will be the same in each of these studies. Consequently, all factors that could disturb the size of this effect are the same. For example, if the effect obtained can be affected by the age or gender of the subjects, then these factors are similar in each study. Thus, differences in the obtained effects for individual studies are due only to sampling error (the inherent error of each study) - that is, the size of SE.

The fixed effect estimates the population effect – the true effect for each study.

The confidence interval around the fixed effect (the width of the rhombus in the forest plot) depends only on specific SE.

Random effect

In calculating the random effect, we assume that each study represents a slightly different population, so that the true (population) effect will be different for each population. Thus, if each study involved a different country, then in order to summarize the meta-analysis with a random effect, we assume that some factors that could distort the magnitude of the effect may have different magnitudes across countries. For example, if the effect (e.g., the average increase in fertility) can be affected by the education level of the respondents or the wealth of a country, and these countries differ in these factors, then the true effect (the average increase in fertility) will be slightly different in each of these countries. Thus, the differences in the obtained effects for each study are due to sampling error (the error within each study) – that is, the size of SE, and the differences between the study populations (the variance between the studies - the heterogeneity of the studies) – i.e., T^2. This heterogeneity cannot be too large, too much variance between study populations indicates no basis for a overall summary.

The random effect estimates a weighted mean of the true (population) effects of each study.

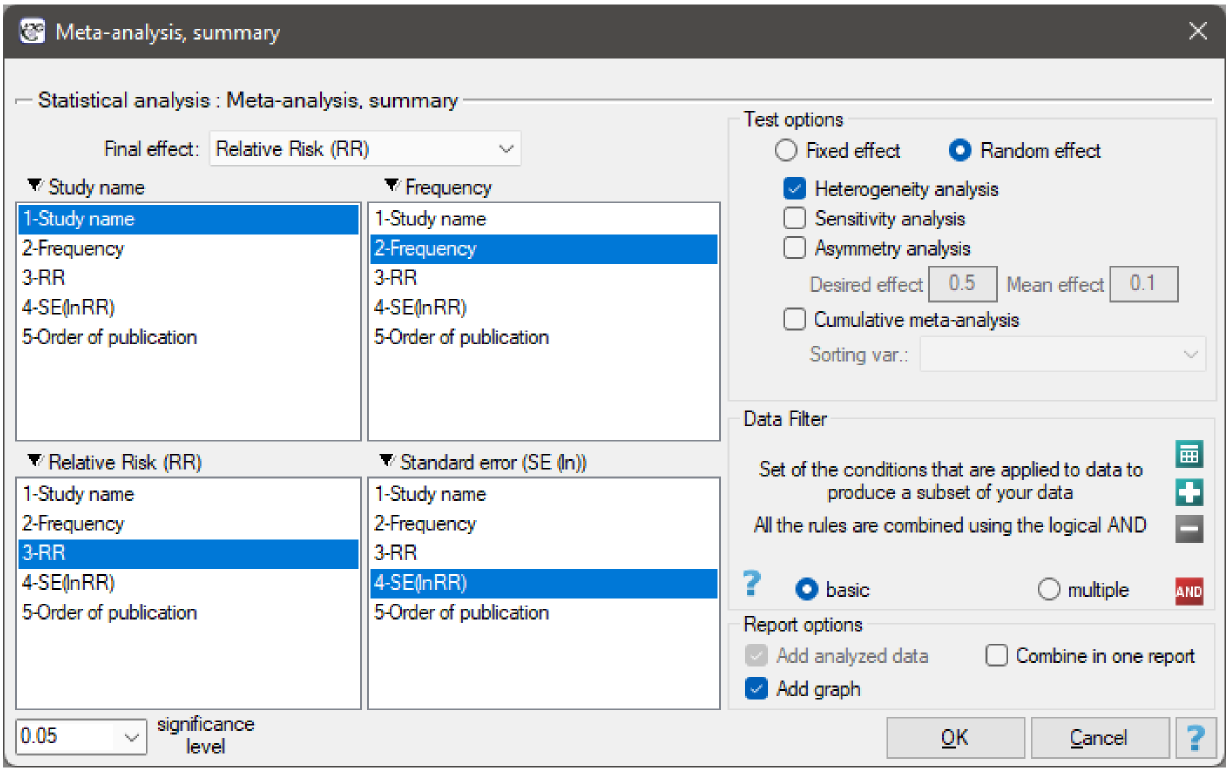
The confidence interval around the variable effect (the width of the rhombus in the forest plot) depends on individual SE and on T^2.

Confidence interval vs. prediction interval

95% confidence interval (width of the rhombus in the forest plot) - means that in 95 percent of cases of such meta-analyses the overall random effect will fall into the interval determined by the rhombus.

95% prediction interval - means that 95 percent of the time the true (population) effect of the new study will fall into the designated interval.

The meta-analysis settings window is opened menu:
Advanced Statistics→Meta-analysis→Meta-analysis, summary.



Meta-analysis, summary

Statistical analysis : Meta-analysis, summary

Final effect: Relative Risk (RR)

Study name: 1-Study name

Frequency: 2-Frequency

Relative Risk (RR): 3-RR

Standard error (SE (ln)): 4-SE(lnRR)

Variables: 1-Study name, 2-Frequency, 3-RR, 4-SE(lnRR), 5-Order of publication

Test options
- ○ Fixed effect ● Random effect
- ☑ Heterogeneity analysis
- □ Sensitivity analysis
- □ Asymmetry analysis
- Desired effect 0.5 Mean effect 0.1
- □ Cumulative meta-analysis
- Sorting var.:

Data Filter

Set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

● basic ○ multiple

Report options
- ☑ Add analyzed data □ Combine in one report
- ☑ Add graph

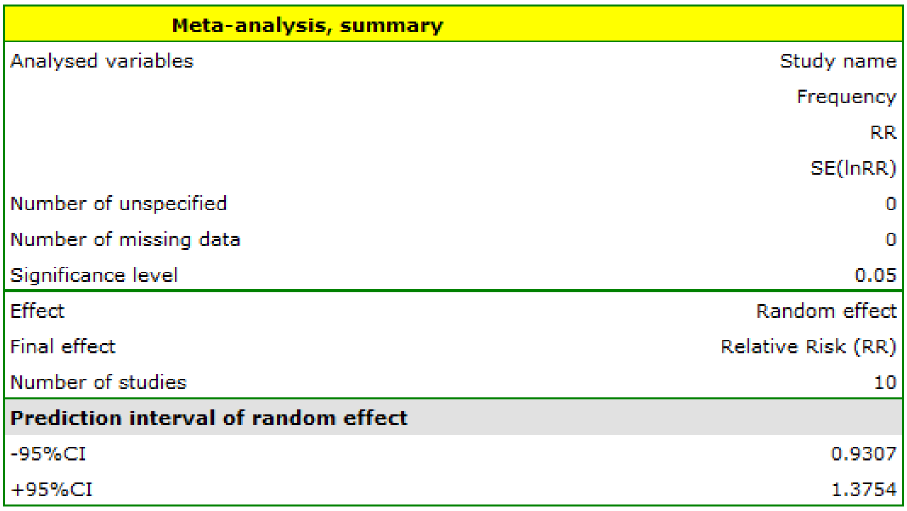
0.05 significance level

In this window, depending on the Final Effect selected, you can summarize the meta-analysis and perform basic analyses to check its assumptions such as heterogeneity, publication bias (sensitivity testing, asymmetry) and perform a cumulative meta-analysis.

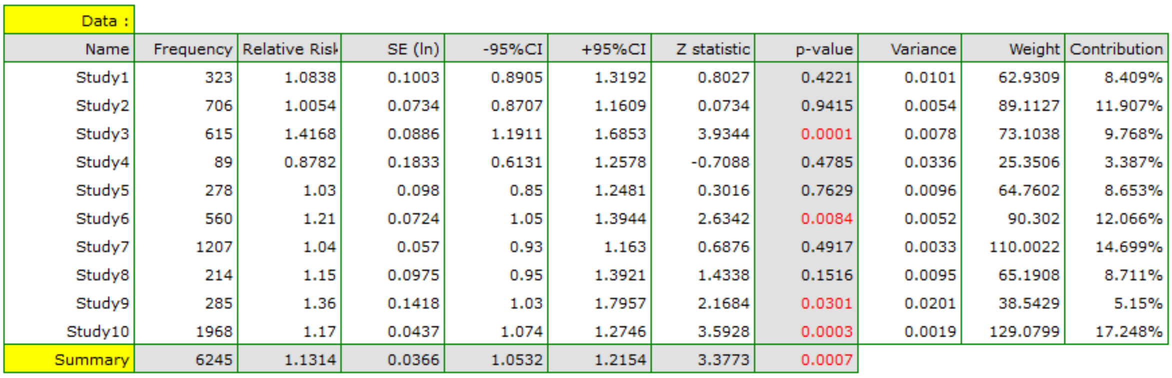
Example 28.2. (MetaAnalysisRR.pqs file)

The risk of disease X for smokers and non-smokers has been studied. Some research papers indicated that the risk of disease X was higher for smokers, while some papers did not prove such a relationship. A meta-analysis was planned to determine whether cigarette smoking affects the occurrence of disease X. A thorough review of the literature on this topic was performed, and based on this, 10 scientific papers were selected for meta-analysis. These papers had data on the basis of which it was possible to calculate the relative risk (i.e. the risk of the disease for smokers in relation to the risk of the disease for non-smokers) and it was possible to establish the error with which the given relative risk is burdened (i.e. the precision of the given study). Data were prepared for meta-analysis and stored in a file.

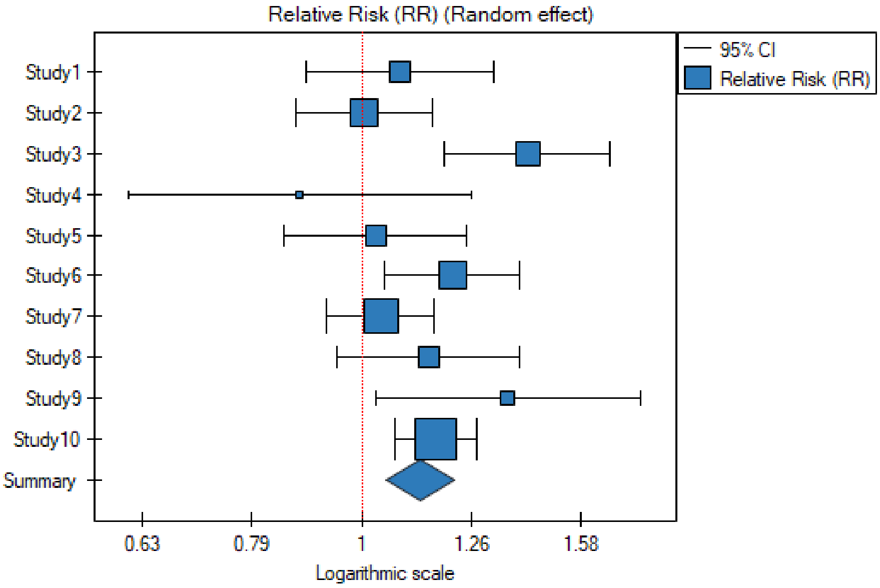
Because the papers included in the meta-analysis were from different locations and included slightly different populations, the summary was chosen using random effect. As the final effect, relative risk was selected and the results were presented on a forest plot.



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 disease for smokers. The overall result of the meta-analysis conducted is also statistically significant and confirms the same effect. The derived relative risk for the overall effect along with the 95 percent confidence interval is above the value of one: RR[95%CI]=1.13[1.05-1.22]. Unfortunately, the prediction interval for the variable effect is wider: [0.93-1.38], which means that in 95% of the cases, the true population relative risk obtained in subsequent studies could be either greater or less than one.

Note!
Before interpreting the results, it is important to check that the assumptions of the meta-analysis are met. In this case, we should consider excluding the third study (see sensitivity analysis , asymmetry analysis, cumulative meta-analysis and the assumption of heterogeneity).

28.3 Weights of individual studies

The weight w_i of the study depends on the observed variability.

For the fixed effect, the variability is due only to sampling error (error within each study) - that is, the size of SE:

$$w_i = \frac{1}{SE^2}$$

For a random effect, variability is due to sampling error (error within each study) - that is, the magnitude of SE, and differences between studies – that is, the observed variance T^2:

$$w_i = \frac{1}{SE^2 + T^2}$$

Based on the weights assigned to each study, the **share** of a given study in the entire analysis is determined. This is the percentage that the weight of a given study represents in relation to the total weight of all included studies.

28.4 Heterogeneity testing

It is difficult to expect every study to end up with exactly the same effect size. Naturally, the results obtained in different papers will be somewhat different. The study of heterogeneity is intended to determine to what extent emerging differences between the results obtained in different papers affect the overall effect constructed in the meta-analysis. **The overall effect summarizes well the results obtained in the different papers if the differences between the different effects are natural i.e. not large**. Large differences in observed effects may indicate heterogeneity of studies and the need to separate more homogeneous subgroups, e.g., divide the collected papers into several subgroups with respect to an additional factor. For example: a given drug has a different effect on younger and older people, so in studies based on data from mainly young people, the effect may differ significantly from studies conducted on older people. Dividing the collected papers into more homogenous subgroups will allow for a good estimation of the overall effect for each of these subgroups separately.

Heterogeneity testing is designed to check whether the variability between studies is equal to zero.

Hypotheses:

$$\begin{array}{cl} \mathcal{H}_0: & \tau^2 = 0, \\ \mathcal{H}_1: & \tau^2 \neq 0, \end{array}$$

where:
τ^2 – is the variance of the true (population) effects of each study.

The test statistic is of the form:

$$Q = T^2 W + k - 1$$

where:
T^2 – is the variance of the observed effects,
W – a factor calculated from the weights assigned to each study,
k – number of studies.

The statistic has asymptotically (for large sample) chi-squared with the degrees of freedom calculated by the formula: $df = k - 1$.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Note

- If the result is statistically significant – this is a strong suggestion to abandon the overall summary of all collected studies.
- If the result obtained is not statistically significant – we can summarize the study with the overall effect. At the same time, it is suggested to summarize with a random effect – according to the following explanation.

  Rationale for choosing a random effect:

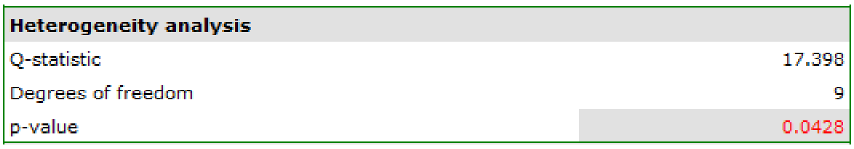
  The overall random effect test takes into account the variability between tests (T^2), while the fixed overall effect does not take this variability into account. However, if T^2 is small, the result of the fixed effect model will be close to that of the random effect model, and when $T^2 = 0$, both models will produce exactly the same result.

Additional measures describing heterogeneity are the coefficients I^2 and H^2:

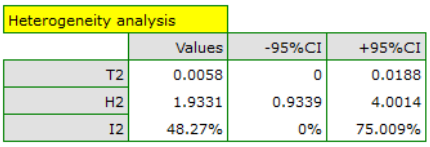
$$I^2 = \frac{H^2 - 1}{H^2}, H^2 = \frac{Q}{k-1}.$$

The I^2 coefficient indicates the percentage of the observed variance that results from the true difference in the magnitude of the effects under study (graphically, it reflects the extent of overlap between the confidence intervals of the individual studies). Because it falls between 0% and 100%, it is subject to simple interpretation and is readily used. If $I^2 = 0$, then all of the observed variance in effect sizes is " false," so if a value of 0 is found in the confidence interval drawn around the I^2 coefficient, the resulting variance can be considered statistically insignificant. On the other hand, the closer the value of I^2 is to 100%, the more one should consider abandoning the overall summary of the study. It is assumed that $I^2 \approx 25\%$ indicates weak, $I^2 \approx 50\%$ moderate, and $I^2 \approx 75\%$ strong heterogeneity among studies. The coefficient H^2, on the other hand, is considered with respect to a value of 1. If the confidence interval for H^2 contains a value of 1, then the variance obtained can be considered statistically insignificant, and the higher the value of H^2, the greater the heterogeneity of the study.

] **Example (28.2) cont.** *(MetaAnalysisRR.pqs file)* When examining the effect of cigarette smoking on the onset of disease X, the heterogeneity assumption of the study was tested. For this purpose, the option Heterogeneity test 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%

A statistically significant result of Q statistic was obtained (p=0.0428). The variance of the observed effects is non-zero (T2=0.0058), and the coefficient I2=48.27%, indicates moderate heterogeneity between studies. Only the confidence interval for the H2 coefficient finds insignificant variability between studies (the range for this coefficient is [0.93-4.00]). With these results in mind, it is important to consider whether the collected papers can be summarized by one overall result (shared relative risk) or whether it is worthwhile to determine a more homogeneous group of papers and perform the analysis again.

28.5 Sensitivity testing

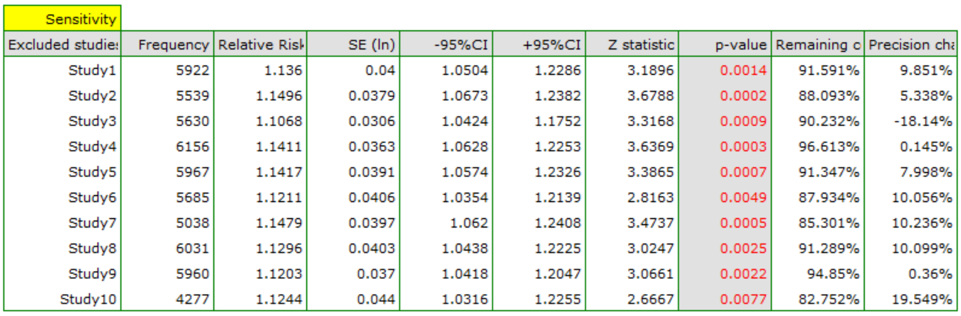
The overall effect of a study may change depending on which studies we include and which we exclude from the analysis. It is the responsibility of the researcher to check how sensitive the analysis is to changes in study selection criteria. Checking for sensitivity helps determine the **changes in overall effect** resulting from removing a particular study. The studies should be close enough that removing one of them does not completely change the interpretation of the overall effect.

The assigned value **remaining contribution** defines the percentage that the total weight of the remaining studies in the analysis represents when a given study is excluded. In contrast, the **precision chenge** indicates how the precision of the overall effect (the width of the confidence interval) will change when a given study is excluded from the analysis.

A good illustration of the sensitivity analysis is a forest plot of the effect size and a plot of the change in precision, with each study excluded.

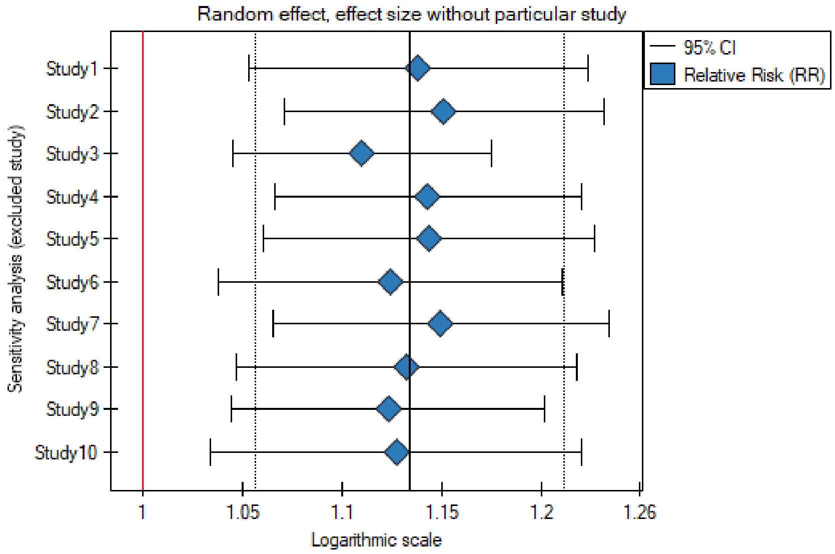
]

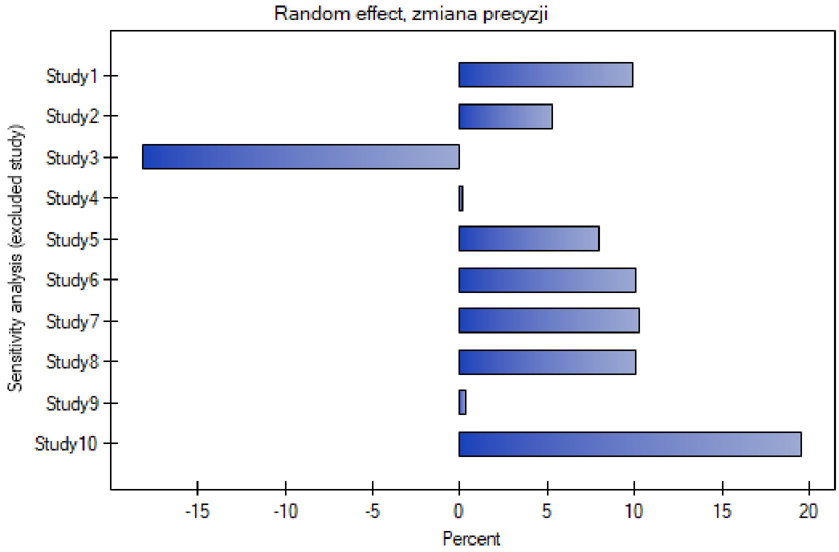
Example (28.2) c.d. *(MetaAnalysisRR.pqs file)* When examining the effect of cigarette smoking on the onset of disease X, the sensitivity of the analysis was checked to exclude individual studies. To do this, the Sensitivity option was selected in the analysis window and forest plot (sensitivity) and bar plot (sensitivity) were selected..



Sensitivity									
Excluded studie:	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 overall relative risk, however, not including particular, indicated studies, still remain statistically significant. The only caveat is study 3. When it is excluded, the precision of the summary obtained increases. The confidence interval for the overall effect is then narrower by about 18%.



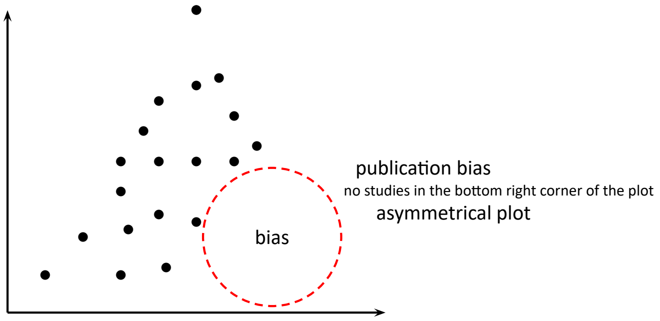


Analysis of the plots leads to the same conclusion. The narrowest interval and the most beneficial change in precision will be obtained when test 3 is excluded.

28.6 Asymmetry testing

Symmetry in the effects obtained is usually indicative of the absence of **publication bias**, but it should be kept in mind that many objective factors can disrupt symmetry, e.g., studies with statistically insignificant effects or small studies are often not published, making it much more difficult to reach such results. At the same time, there are no sufficiently comprehensive and universal statistical tools for asymmetry detection. As a result, a significant part of meta-analyses is published despite the diagnosed asymmetry. Such studies, however, require good justification of such a procedure.

Funnel plot



A standard way to test for publication bias in the form of asymmetry is a funnel plot, showing the relations between study size (Y axis) and summary effect size (X axis). It is assumed that large studies (placed at the top of the graph) in a correctly selected set, are located close together and define the center of the funnel, while smaller studies are located lower and are more diverse and symmetrically distributed. Instead of the study size on the Y-axis, the effect error for a given study can be shown, which is better than showing the study size alone. This is because the effect error is a measure that indicates the precision of the study and also carries information about its size.

Egger's test

Since the interpretation of a funnel plot is always subjective, it may be helpful to use the Egger coefficient (Egger 1997[53]), the interception of the fitted regression line. This coefficient is based on the correlation between the inverse of the standard error and the ratio of the effect size to its error. The further away from 0 the value of the coefficient, the greater the asymmetry. The direction of the coefficient determines the type of asymmetry: a positive value along with a positive confidence interval for it indicates an effect size that is too high in small studies and a negative value along with a negative confidence interval indicates an effect size that is too low in small studies.

Note

Egger's test should only be used when there is a large variation in study sizes and the occurrence of a medium-sized study.

Note

With few studies (small number of k), it is difficult to reach a significant result despite the apparent asymmetry.

Hypotheses:

$$\begin{array}{cl}\mathcal{H}_0: & b=0,\\ \mathcal{H}_1: & b\neq 0,\end{array}$$

where:
b – intercept in Egger's regression equation.

The test statistic is in the form of:

$$t=\frac{b}{SE(b)}$$

where:
$SE(b)$ – standard error of intercept.

The test statistic has t-Student distribution with $k-2$ degrees of freedom.
The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Testing the ”Fail-safe” number

Rosenthal’s Nfs - The ”fail-safe” number described by Rosenthal (1979)[137] specifies the number of papers not indicating an effect (e.g., difference in means equal to 0, odds ratio equal to 1, etc.) that is needed to reduce the overall effect from statistically significant to statistically insignificant.

$$Nfs = \frac{\left(\sum_{i=1}^{k} Z_i\right)^2}{Z_c^2} - k$$

where:
Z_i – the value of the test statistic (with normal distribution) of a given test,
Z_c – the critical value of the normal distribution for a given level of significance,
k – number of studies in the meta-analysis.

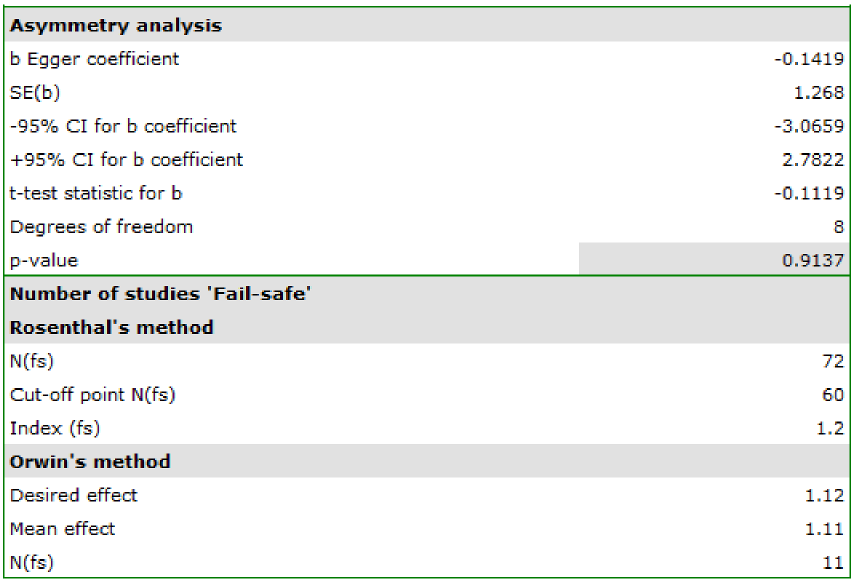
Rosenthal (1984)[126] defined the number of papers being the **cutoff point** as $5k+10$. By determining the quotient of Nfs and the cutoff point, we obtain **coefficient(fs)**. According to Rosenthal’s interpretation, if coefficient(fs) is greater than 1, the probability of publication bias is minimal.

Orwin’s Nfs - the ”fail-safe” number described by Orwin (1983) determines the number of papers with the average effect indicated by the researcher $\overline{M_{fs}}$ that is needed to reduce the overall effect to the desired size M_d indicated by the researcher.

$$Nfs = k\frac{M - M_d}{M_d - \overline{M_{fs}}}$$

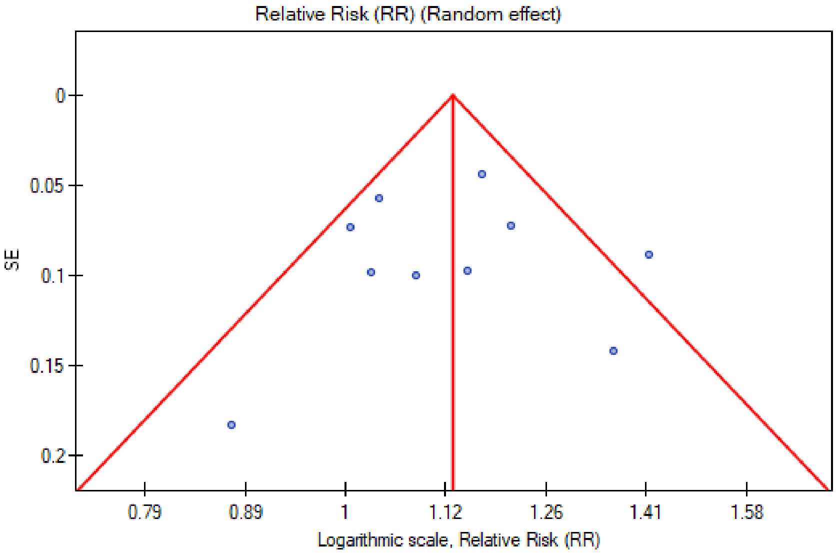
where:
M – the overall effect obtained in the meta-analysis.

Example (28.2) c.d. *(MetaAnalysisRR.pqs file)* When examining the effect of cigarette smoking on the onset of disease X, the assumption of study asymmetry, and therefore publication bias, was checked. To do this, the option Asymmetry was selected in the analysis window and Funnel plot was selected.



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 result are not statistically significant (p=0.9137), indicating no publication bias.



The points representing each study are symmetrically distributed in the funnel plot. Admittedly, one study is outside the boundary of the triangle (Study 3), but it is close to its edges. On the basis of the diagram we also have no fundamental objections to the choice of studies, the only concern being the third study.

The number of ”fail-safe” papers determined by Rosenthal’s method is large and is at 72. Thus, if the overall effect (relative risk shared by all studies) were to be statistically insignificant (cigarette smoking would have no effect on the risk of disease X), 72 more papers with a relative risk of one would have to be included in the pooled papers. The obtained effect can be therefore considered stable, as it will not be easy (with a small number of papers) to undermine the obtained effect.

The resulting overall relative risk is RR=1.13. Using Orwin’s method it was checked how many papers with relative risk equal to 1.11 it would take for the overall relative risk to fall to 1.12. The result was 11 papers. On the other hand, by reducing the size of the relative risk from 1.11 to 1.10 only 5 papers are needed for the overall relative risk to be 1.12.

28.7 Cumulative meta-analysis

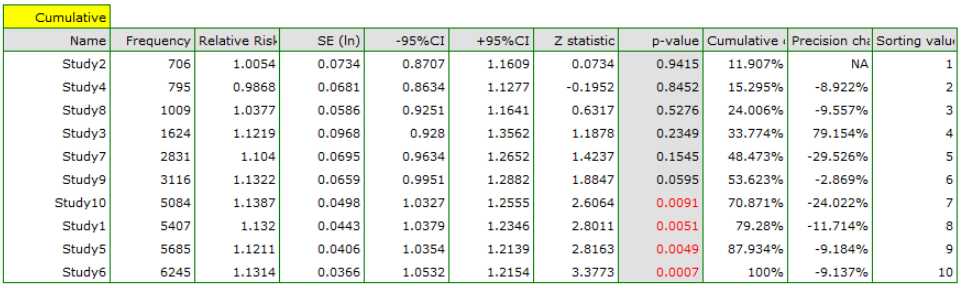
The typical purpose of conducting a cumulative meta-analysis is to show how the effect has changed since the last meta-analysis on a topic was conducted/published, or how it has changed over the years. Then chronologically (according to the timeline) more studies are added and the overall effect is calculated each time. Equally important is the cumulative analysis in a study of how the overall effect changes depending on the magnitude of the impact of a selected additional factor. The studies are then sorted according to the magnitude of that factor and the **for successively added studies, a cumulative overall effect is calculated**.

Depending on the purpose of the cumulation, the variable by which the individual studies will be sorted should be chosen, i.e., the order in which the studies are added to the meta-analysis summary. This can be any numerical variable.

The assigned value of **Cumulative contribution** defines the percentage that is represented by the total weight of the included studies in the analysis i.e., the given study and the studies preceding it. In contrast, **Precision change** indicates how the precision of the overall effect (the width of the confidence interval) will change when a given study is included with the studies preceding it.

A good illustration of the cumulative analysis is a forest plot of the effect size and a plot of the change in precision, with each study included.

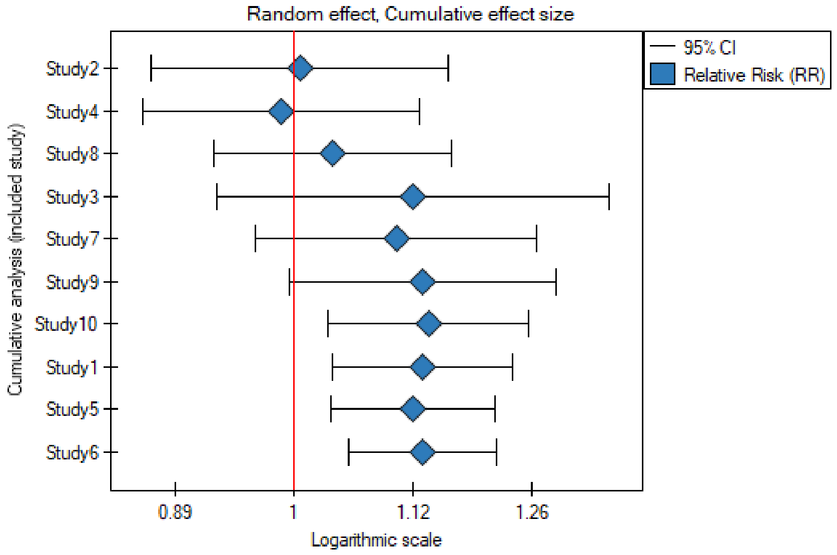
Example (28.2) cont. *(MetaAnalyzisRR.pqs file)* By investigating the effect of cigarette smoking on the onset of disease X, we examined how the results evolved over time. To do this, the cumulative meta-analysis was selected in the analysis window as well as the variable by which subsequent papers would be included in the meta-analysis, and forest plot (cumulative) and bar plot (cumulative) were specified.

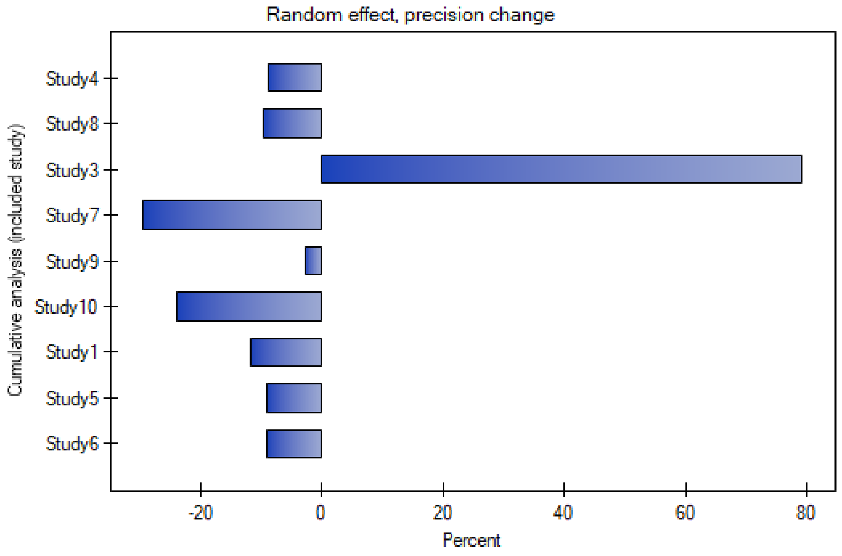


Cumulative

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 were added, the resulting overall effect gained strength, and its significance was obtained by adding Study 10 to the earlier papers, and then subsequent studies as well. In general, the addition of more papers increased the precision of the derived relative risk, except when Study 3 was added. The confidence interval of the overall relative risk then increased by 79.15%. We see this effect in the table and in the accompanying charts. As a result, one should consider excluding Study 3 from the meta-analysis.





28.8 Group comparison

There are situations in which the data collected are of the same effect, performed on the same population, but under slightly different conditions. Suppose that part of the study was performed under condition A and part under condition B. Then it may be interesting to compare the overall effects obtained for each group. Demonstrating differences between overall effects may be the main goal of a meta-analysis, and then it is inadvisable to compound both subgroups simultaneously with one overall effect. However, if the researcher realizes that the studies were conducted under different conditions, but it seems appropriate to summarize all the studies together, then showing the absence of statistically (or clinically) significant differences, the researcher can make a joint summary taking into account this division into subgroups A and B, i.e. determine a overall summary adjusted for different conditions of the experiment. For example, Country A has a slightly different climate than Country B. We have a number of studies from country A and a number of studies from country B. If our study population is the vegetation of these two countries, we can test whether the climatic conditions affect the obtained study effects for each country. A comparative analysis of the subgroups thus determined will allow us to assess whether climate has a major influence on the results obtained or not, and whether the results of the studies covering these two countries can indeed be summarized in one overall effect, or whether we should determine separate summaries for each country. Another example can be a situation where some of the studies are studies in which randomization was performed, but in some of them we do not have full randomization, then we can divide the studies into subgroups to then check whether the studies without randomization give similar results to the studies with randomization in order to include them in further, combined analysis.

Examination of group heterogeneity

We can compare groups by choosing as overall effect: fixed effect, random effect – separate T^2 or random effect – pooled T^2, where T^2 is the variance of the observed effects.

- **Fixed effect** is chosen when we assume that 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 the groups differ in 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.

The main goal is to compare groups, that is, to determine whether the groups being compared differ in their true (i.e., population) overall effect. In practice, this is to test whether the variance of group overall effects is zero, i.e., to test the heterogeneity of the groups. For a description and interpretation of the results of heterogeneity analysis, see chapter Heterogeneity testing, except that in the case of group comparisons, heterogeneity refers to the compound effects of the groups being compared, not the individual studies, and the outcome depends on the overall effect chosen.

Hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & \tau^2 = 0,\\ \mathcal{H}_1: & \tau^2 \neq 0,\end{array}$$

where:
τ^2 – is the variance of the true (population) summary effects of the groups being compared.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lcl}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

If the result is statistically significant (a score of **Q-statistic**, I^2 **coefficient** or H^2 **coefficient**), this is a strong suggestion to drop the overall summary of the groups being compared.

Examining heterogeneity in groups

An additional option of the analysis is the possibility to analyze each group separately for heterogeneity, as described in Heterogeneity testing. The results obtained (in particular, the variance T^2) make it easier to decide how to compare the groups, i.e., whether to choose a random effect (separate T^2) or a random effect (pooled T^2).

Joint summary of the groups

In a situation where, based on the results of the group comparison, the differences obtained between the overall effects of the groups are small and insignificant, a joint summary of the groups can be performed. The summation is done in correction for the division into the indicated groups. For example, if we split the study based on the different conditions of the experiment conducted, then the joint summary will be done in correction for the different conditions of the experiment. The result of joint summation (overall efect of both grous) depends on the observed differences (on the variation between studies and between groups) i.e. on the choice of ovearall effect (whether it is fixed or random (separate T^2) or random (pooled T^2)).

A good illustration of the joint (ovearall) summary of the groups in the meta-analysis is a forest plot showing the results of each study with each group's summary and the joint summary of the groups.

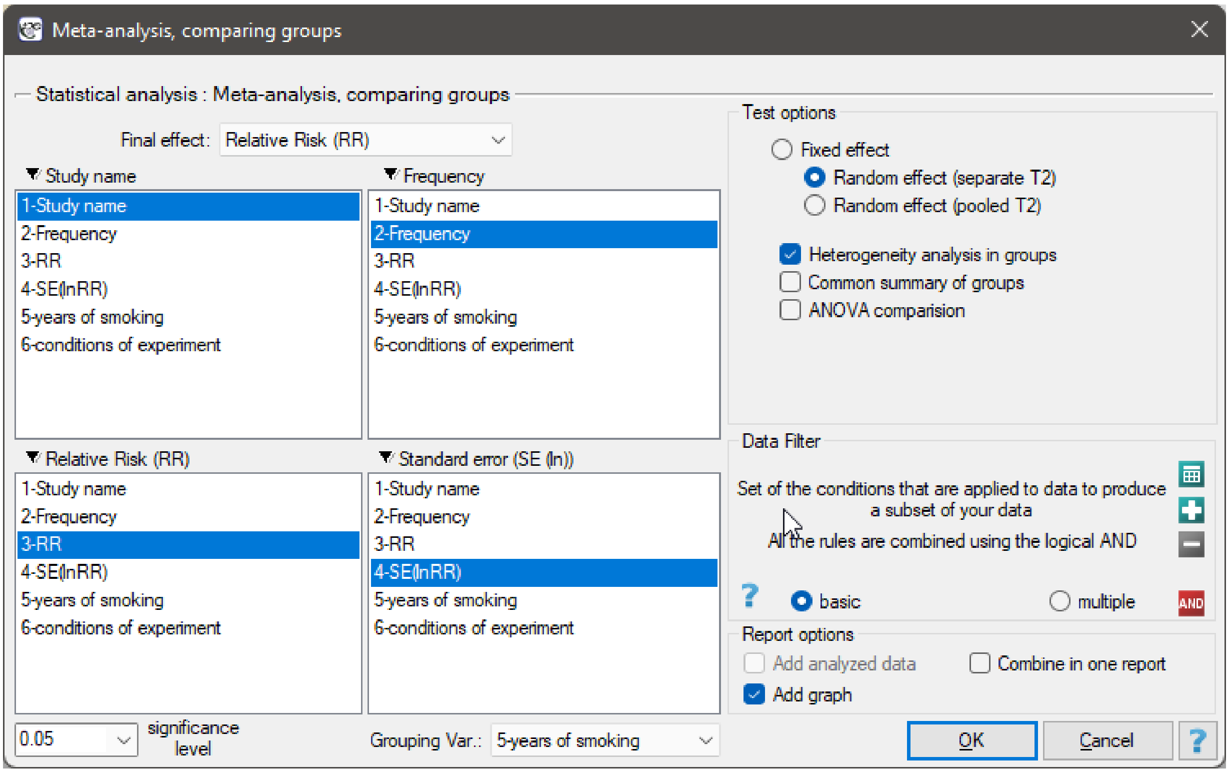
ANOVA comparison

ANOVA comparison is an additional option for comparing groups. It is a slightly different method of comparison than comparison by testing heterogeneity of groups (based on a different mathematical model). Both methods, however, give overlapping results as to the comparison of groups. In case of comparison of groups by ANOVA method the observed variance is broken down into between-group variance and within-group variance. The within-group variance is then broken down into the variance of each group separately. As a result, the following Q statistics are determined:

- The Q statistic (group 1) – examines that part of the total variance that relates to group one, i.e., the variance between studies located within group one,
- The Q statistic (group 2) – examines that part of the total variance that relates to the second group, i.e. the variance between studies within the second group,
- ...
- The Q statistic (group g) – examines that part of the total variance that relates to the last group, i.e., the variance between studies within the last group,
- The Q statistic (within groups) = Q (group 1) + Q (group 2) + ... + Q (group g) - examines that part of the total variance that relates to the inside of the individual groups, i.e., the variance of the within-group tests,
- **The Q statistic (between groups)** - examines that part of the total variance that relates to differences between groups, i.e., the between-group variance (**same result as examining the heterogeneity of groups**) ,
- The 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 window with settings of group comparison for meta-analysis is opened via menu: Advanced Statistics→Meta-analysis→Meta-analysis, comparing groups.



Meta-analysis, comparing groups

Statistical analysis : Meta-analysis, comparing groups

Final effect: Relative Risk (RR)

Study name: **1-Study name**, 2-Frequency, 3-RR, 4-SE(lnRR), 5-years of smoking, 6-conditions of experiment

Frequency: 1-Study name, **2-Frequency**, 3-RR, 4-SE(lnRR), 5-years of smoking, 6-conditions of experiment

Relative Risk (RR): 1-Study name, 2-Frequency, **3-RR**, 4-SE(lnRR), 5-years of smoking, 6-conditions of experiment

Standard error (SE (ln)): 1-Study name, 2-Frequency, 3-RR, **4-SE(lnRR)**, 5-years of smoking, 6-conditions of experiment

0.05 significance level

Grouping Var.: 5-years of smoking

Test options
○ Fixed effect
● Random effect (separate T2)
○ Random effect (pooled T2)
☑ Heterogeneity analysis in groups
☐ Common summary of groups
☐ ANOVA comparision

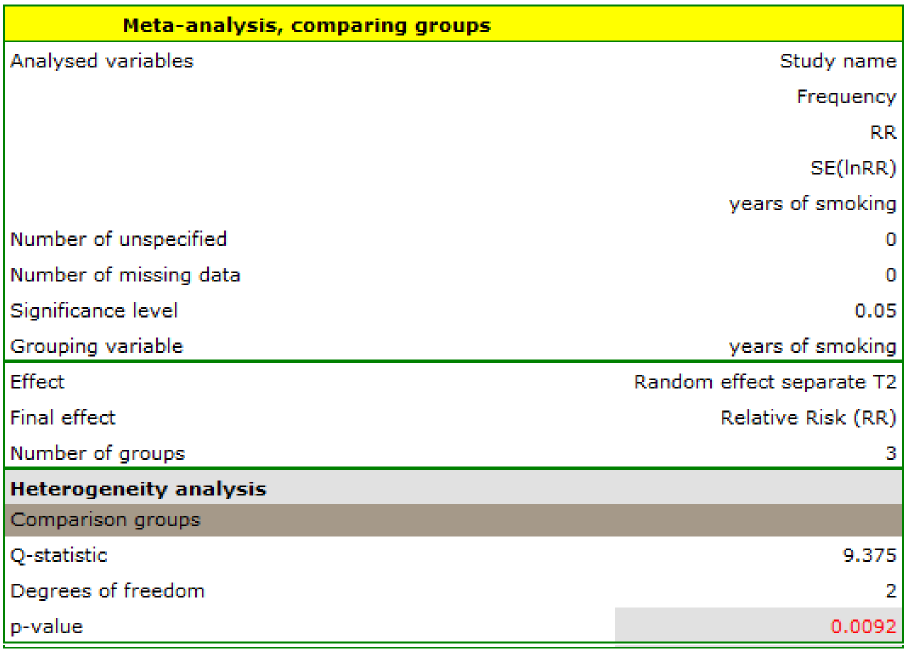
Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

Report options
☐ Add analyzed data ☐ Combine in one report
☑ Add graph

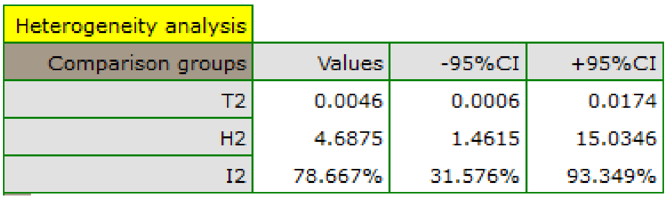
EXAMPLE 28.3. (MetaAnalysisRR.pqs file)

The risk of disease X was examined for smokers and non-smokers. A meta-analysis was conducted to determine whether smoking duration affects the onset of disease X. A thorough review of the literature on this topic was carried out, and 17 studies were identified that had a description of the relative risk and its error (i.e. the precision of the study). Because the studies involved different smoking times, 3 groups of studies were identified:
(1) studies on people who have been smoking for more than 10 years,
(2) studies on people who have been smoking for 5 to 10 years,
(3) studies on people who have been smoking for less than 5 years.
In addition, a subdivision was made between the two different conditions of the studies (different inclusion/exclusion criteria of subjects). Data were prepared for meta-analysis and stored in a file.
The purpose of conducting the meta-analysis was to compare age groups. In addition, it was examined whether the different conditions of the experiment translated into differences in the relative risk obtained.

Because the papers included in the meta-analysis were from different locations and included slightly different populations, the summary was made by selecting random effect (separate T2). As the final effect, relative risk was selected and the results were presented on a forest plot.

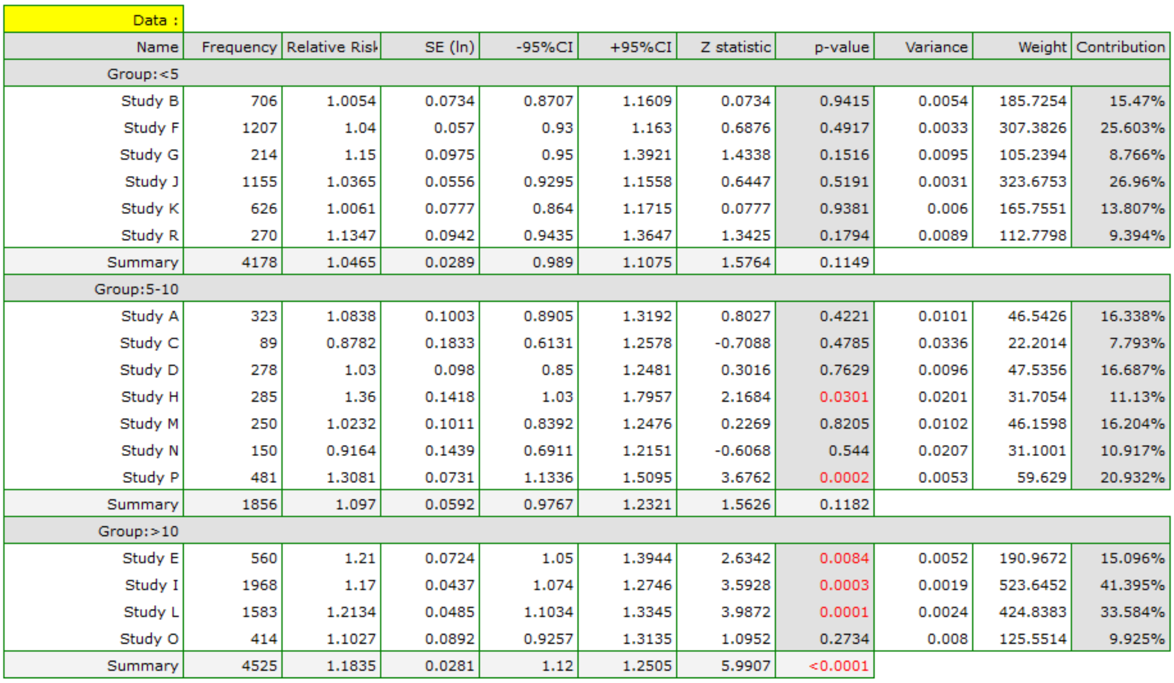


Meta-analysis, comparing groups	
Analysed variables	Study name
	Frequency
	RR
	SE(lnRR)
	years of smoking
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Grouping variable	years of smoking
Effect	Random effect separate T2
Final effect	Relative Risk (RR)
Number of groups	3
Heterogeneity analysis	
Comparison groups	
Q-statistic	9.375
Degrees of freedom	2
p-value	0.0092



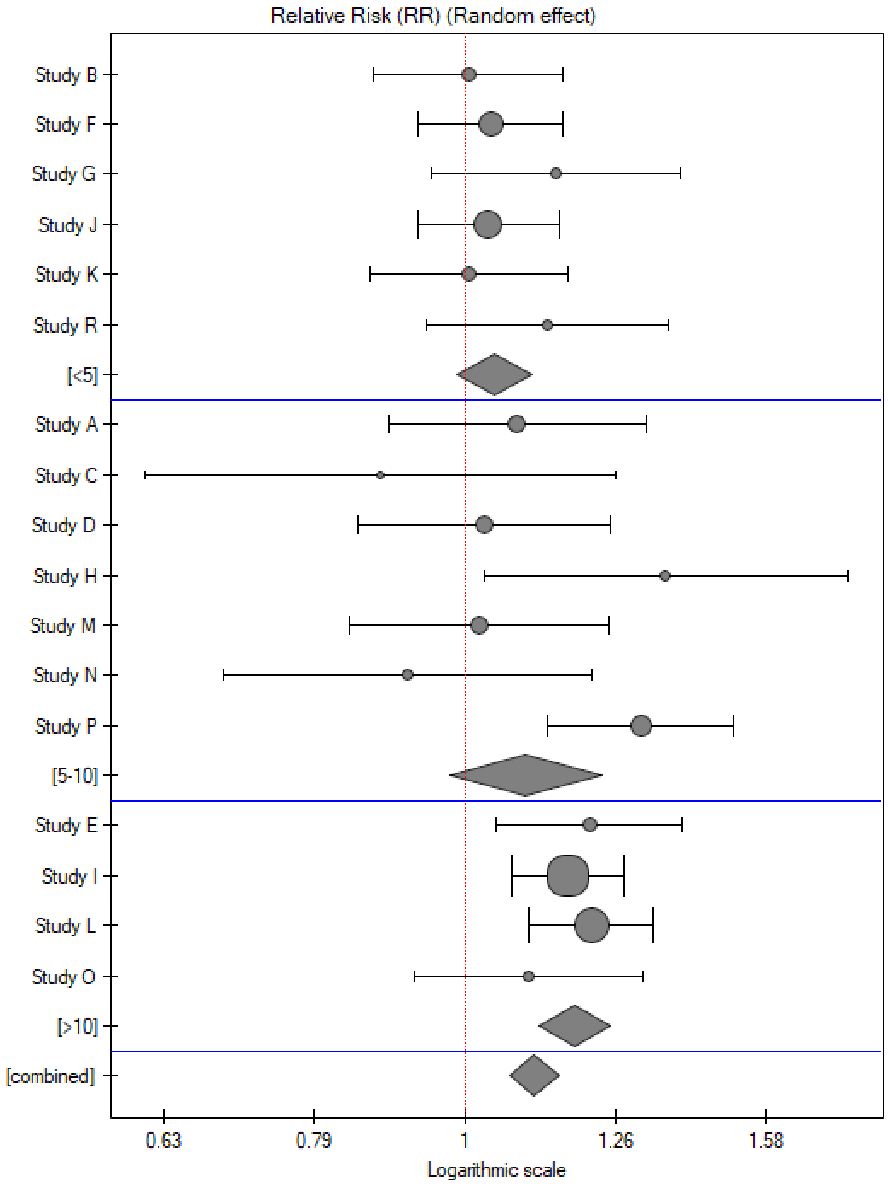
Heterogeneity analysis			
Comparison groups	Values	-95%CI	+95%CI
T2	0.0046	0.0006	0.0174
H2	4.6875	1.4615	15.0346
I2	78.667%	31.576%	93.349%

The groups are statistically significantly different (p=0.0092), which we observe not only based on the test of heterogeneity, but also on the coefficient of H2 (the coefficient along with the confidence interval is above the value of one) and I2 (78 is high heterogeneity). Therefore, the collected papers will not be summarized by a overall effect but only by a separate summary of each group.

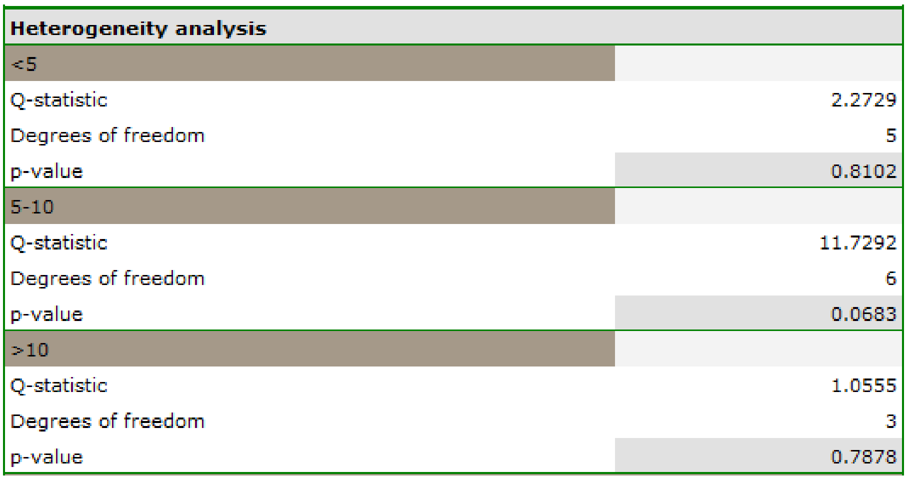


Data :										
Name	Frequency	Relative Risk	SE (ln)	-95%CI	+95%CI	Z statistic	p-value	Variance	Weight	Contribution
Group:<5										
Study B	706	1.0054	0.0734	0.8707	1.1609	0.0734	0.9415	0.0054	185.7254	15.47%
Study F	1207	1.04	0.057	0.93	1.163	0.6876	0.4917	0.0033	307.3826	25.603%
Study G	214	1.15	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			

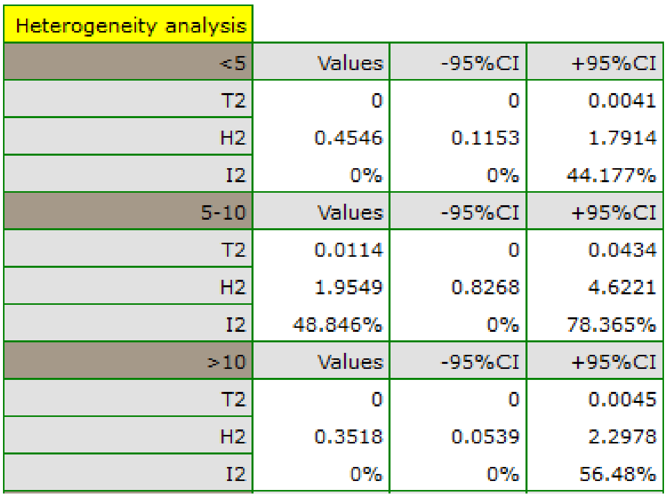
The forest plot also shows a summary of each group and does not contain a joint summary of the groups.



In addition, homogeneity within each group was checked to ascertain the feasibility of summarizing them separately.

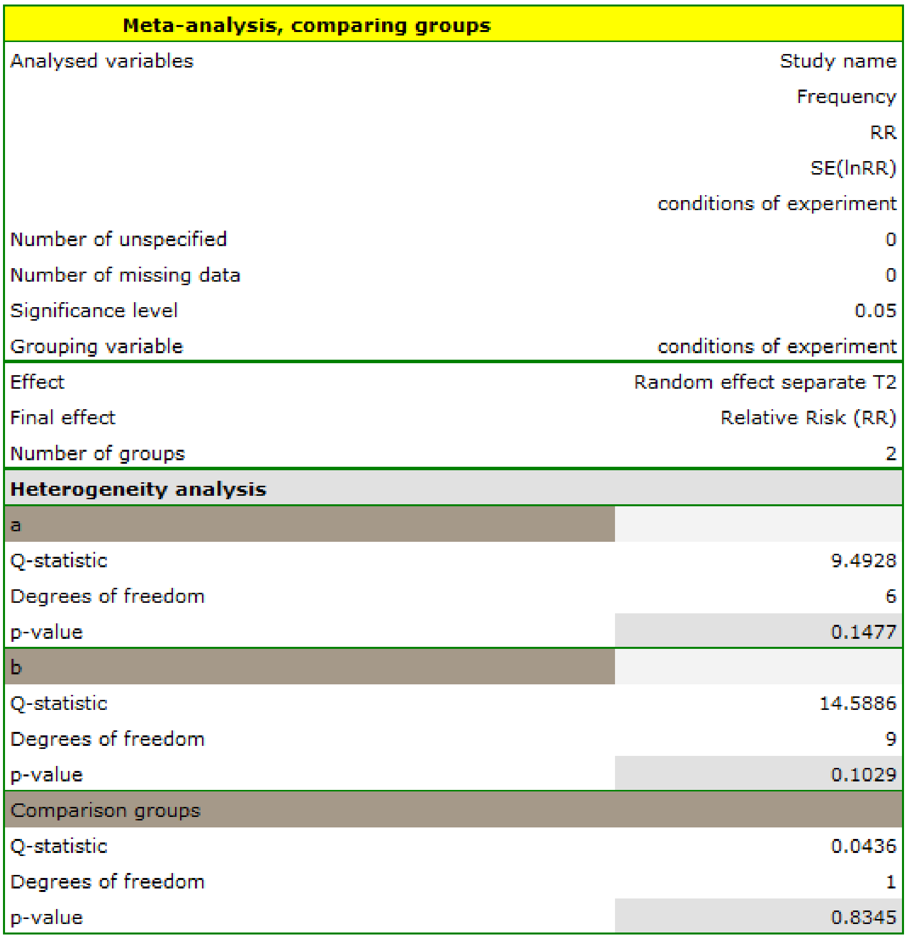


Heterogeneity analysis	
<5	
Q-statistic	2.2729
Degrees of freedom	5
p-value	0.8102
5-10	
Q-statistic	11.7292
Degrees of freedom	6
p-value	0.0683
>10	
Q-statistic	1.0555
Degrees of freedom	3
p-value	0.7878

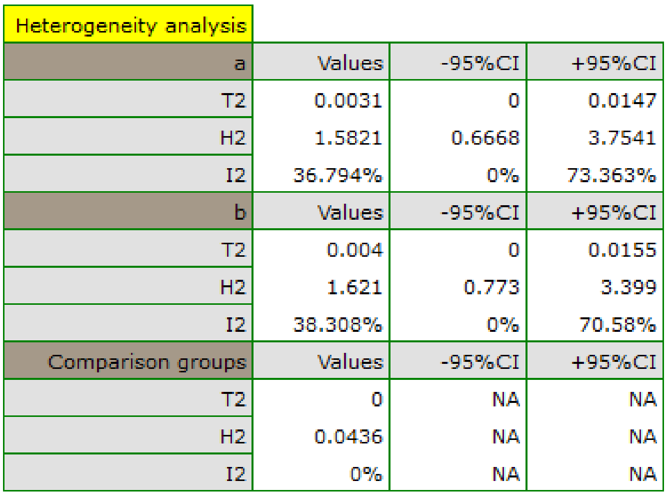


Heterogeneity analysis			
<5	Values	-95%CI	+95%CI
T2	0	0	0.0041
H2	0.4546	0.1153	1.7914
I2	0%	0%	44.177%
5-10	Values	-95%CI	+95%CI
T2	0.0114	0	0.0434
H2	1.9549	0.8268	4.6221
I2	48.846%	0%	78.365%
>10	Values	-95%CI	+95%CI
T2	0	0	0.0045
H2	0.3518	0.0539	2.2978
I2	0%	0%	56.48%

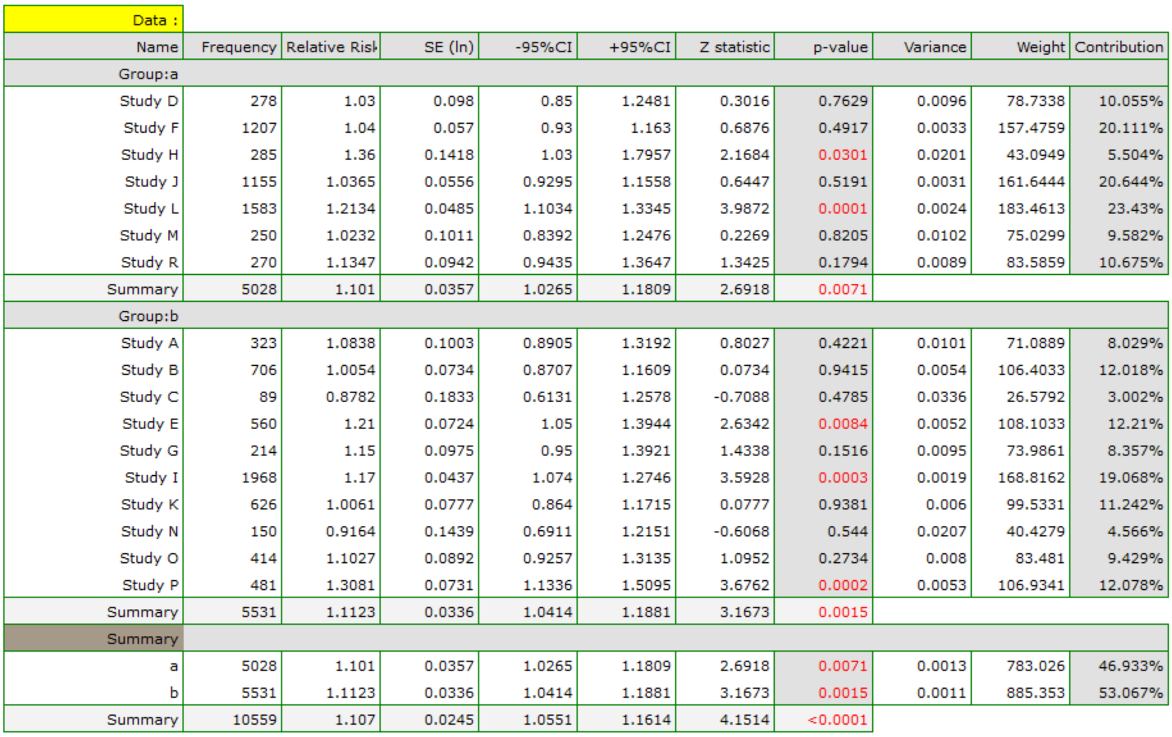
In contrast, the results of the comparison concerning the different study conditions indicate that there is no significant effect of these conditions on the overall effect. In this case, it is possible to calculate a joint overall effect when correcting for the different test conditions, i.e., a joint summary of the two groups.



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

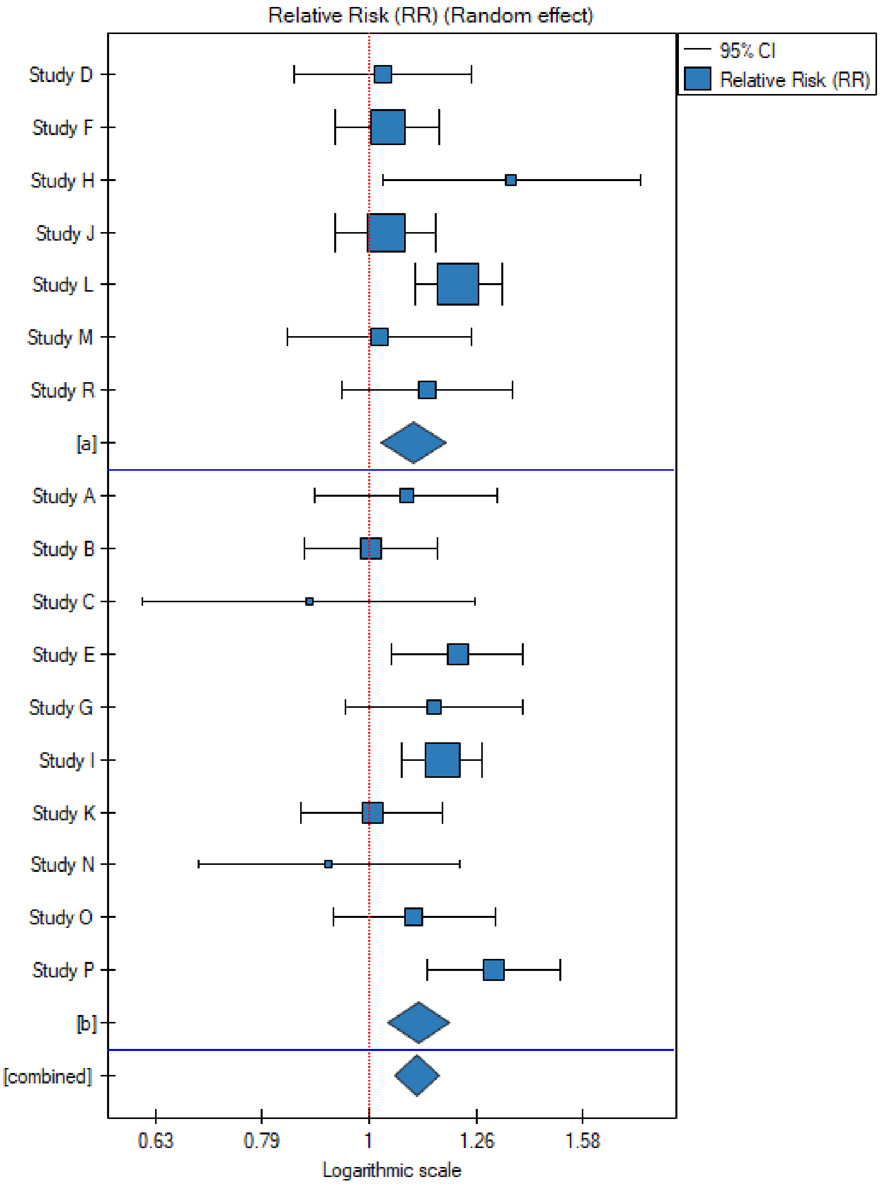


Heterogeneity analysis			
a	Values	-95%CI	+95%CI
T2	0.0031	0	0.0147
H2	1.5821	0.6668	3.7541
I2	36.794%	0%	73.363%
b	Values	-95%CI	+95%CI
T2	0.004	0	0.0155
H2	1.621	0.773	3.399
I2	38.308%	0%	70.58%
Comparison groups	Values	-95%CI	+95%CI
T2	0	NA	NA
H2	0.0436	NA	NA
I2	0%	NA	NA



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			



28.9 Meta-regression

Meta-regression analysis is conducted in an analogous manner to the regression analysis described in the section Multiple Regression. In the case of meta-regression, the study objects are the individual studies, their results (e.g., odds ratios, relative risks, differences in means) constitute the dependent variable Y i.e., the explained variable, and the additional conditions for conducting these studies constitute the independent variables (X_1, X_2, . . ., X_k) i.e., the explanatory variables. As in traditional regression models, the independent variables may interact and those described by a nominal scale may be subject to special coding (for more information, see Preparation of the variables for analysis in multivariate models). The number of independent variables should be small, less than the number of papers on which the study is based on ($n \geq k + 1$).

We can perform meta-regression by choosing a fixed effect or a random effect.

- **Fixed effect** is chosen when we assume that the studies represent one common true effect such that all factors that could perturb the magnitude of this effect are the same except for the factors tested as independent variables in the model (X_1, X_2, $\ldots$, X_k). This is a situation that occurs very rarely in real research because it requires fully controlled conditions, which is almost impossible in different studies, conducted in different locations and by different researchers. The use of fixed effect would be justified, for example, in a situation when all the tests are carried out at the same location, on the same population, changing only those conditions that are described by the characteristic being tested. For example, if we wanted to test the effect of changing temperature on changing the relative risk of disease described in each study, then all studies should be conducted on the same population under exactly the same conditions except for the change in temperature, which is the independent variable X in the model.

- **Random effect** is chosen when we assume that studies may represent slightly different populations, i.e., factors that could perturb the magnitude of the effect under study are not described in all papers (they can be assumed to be similar, but not necessarily exactly the same). Each paper provides the magnitudes of the factors we are interested in, which are involved in model building as independent variables (X_1, X_2, $\ldots$, X_k). The use of a random effect is common because individual studies are usually conducted at different locations under slightly different conditions, the variability of interest is only in the conditions that describe the factors given in the study, e.g., temperature, which will be the independent variable X in the model.

Model verification

- **Statistical significance of individual variables in the model**.

  Based on the coefficient and its error, we can conclude whether the independent variable for which this coefficient was estimated has a significant effect on the final effect. For this purpose, we test the hypotheses:

$$\begin{array}{ll}\mathcal{H}_0: & \beta_i = 0,\\ \mathcal{H}_1: & \beta_i \neq 0.\end{array}$$

  Calculate the test statistic using the formula:

$$Z = \frac{b_i}{SE_{b_i}}$$

  Test statistics has the normal distribution.
  The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{lll}\text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,\\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0.\end{array}$$

- **Quality of the built model** of a linear multivariate regression can be assessed by several measures.

  - **Coefficient** R^2 – is a measure of model fit. It expresses the percentage of variability between study effects explained by the model.
    The value of this coefficient is in the range $<0;1>$, where 1 means a perfect fit of the model, 0 – a complete lack of fit. In determining it we use the following equation:

$$R^2 = T^2_{(modelu)} + T^2_{(total)},$$

    where:
    $T^2_{(modelu)}$ – variance between studies explained by the model,
    $T^2_{(total)}$ – total variance between studies.

- **Coefficient** I^2 – determines the percentage of the observed variance that results from the true difference in the magnitude of the effects under study.

Note
For a detailed representation of the variance described by the coefficients, see chapter Testing heterogeneity

- **Statistical significance of all variables in the model**
  The primary tool for estimating the significance of all variables in the model is an ANOVA that determines Q (of the model).

$$\begin{array}{cl} \mathcal{H}_0: & \text{all } \beta_i = 0, \\ \mathcal{H}_1: & \text{exists } \beta_i \neq 0. \end{array}$$

Using the ANOVA approach, the observed variance between tests is broken into the variance explained by the model and the variance of the residual (not explained by the model). As a result, the following Q statistics are determined:

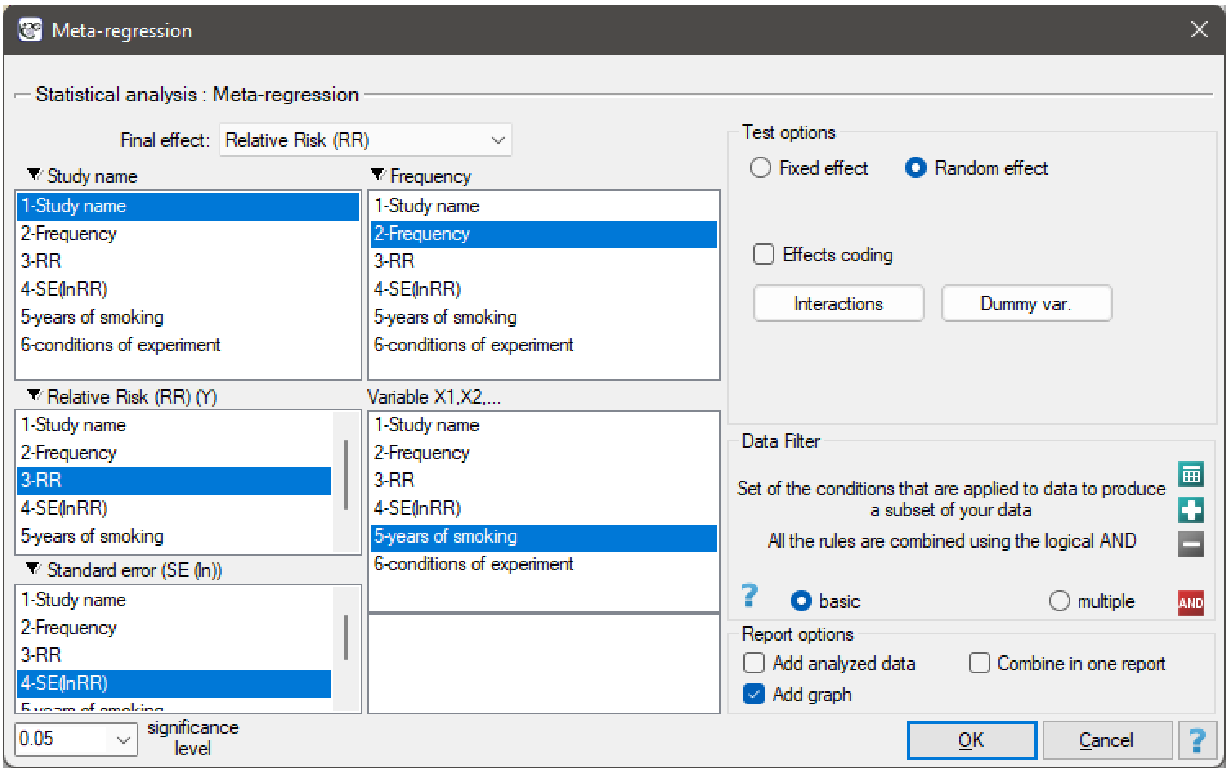
- The Q statistic (of the residuals) - examines the portion of the total variance that is not explained by the model,
- The Q statistic (of the model) - examines the portion of the total variance that is explained by the model,
- The Q statistic (total) - examines the variance between all studies.

Each of the above Q statistics has χ^2 distribution with the appropriate number of degrees of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α:

$$\begin{array}{ccl} \text{if } p \le \alpha & \Longrightarrow & \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The window with settings of group comparison for meta-analysis is opened via menu: Advanced Statistics→Meta-analysis→Meta-regression.



Meta-regression

Statistical analysis : Meta-regression

Final effect: Relative Risk (RR)

Study name: 1-Study name

Frequency: 2-Frequency

Relative Risk (RR) (Y): 3-RR

Variable X1,X2,...: 5-years of smoking

Standard error (SE (ln)): 4-SE(lnRR)

Variables listed: 1-Study name, 2-Frequency, 3-RR, 4-SE(lnRR), 5-years of smoking, 6-conditions of experiment

0.05 significance level

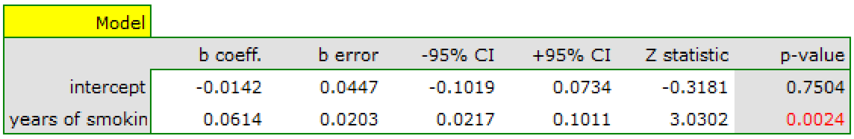
Test options
○ Fixed effect ● Random effect
□ Effects coding

Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
● basic ○ multiple

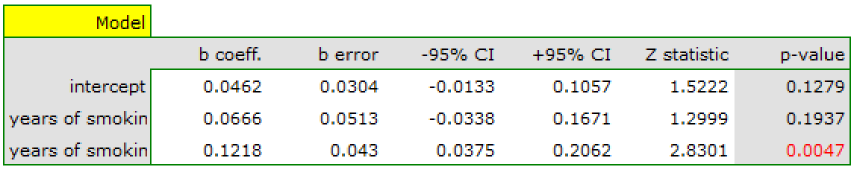
Report options
□ Add analyzed data □ Combine in one report
☑ Add graph

Example (28.3) cont. *(MetaAnalysisRR.pqs file)* The risk of disease X was examined for smokers and non-smokers. A meta-analysis comparing groups of studies was conducted to determine whether the number of years of smoking affected the onset of disease X and whether different conditions of the experiment resulted in different relative risks. On the basis of the comparison of the groups of studies, it was possible to establish that the last group (the group of smokers who have been smoking the longest, i.e. for more than 10 years) shows an association between smoking and the onset of disease X. On the other hand, for the groups with shorter smoking duration, no significant effect could be obtained. However, it was observed that the effect systematically increased with increasing 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 constructed. In the first model, the grouping variable Years of smoking was treated as a continuous variable. In the second model, it was determined that the variable Years of smoking would be treated as a categorical (dummy) variable with the reference group smoking less than 5 years. Data were prepared for meta-regression and stored in a file.

Because the papers included in the meta-analysis were from different locations and included slightly different populations, the meta-regression was performed by selecting random effect. The relative risk was selected as the final effect, and the results were presented in the graph.



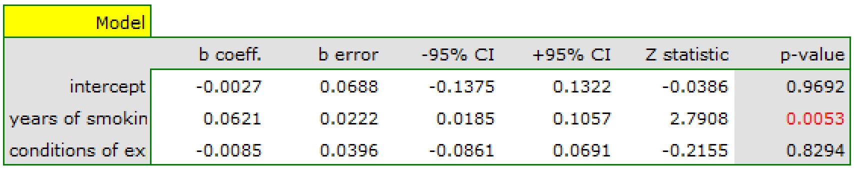
Model	b coeff.	b error	-95% CI	+95% CI	Z statistic	p-value
intercept	-0.0142	0.0447	-0.1019	0.0734	-0.3181	0.7504
years of smokin	0.0614	0.0203	0.0217	0.1011	3.0302	0.0024



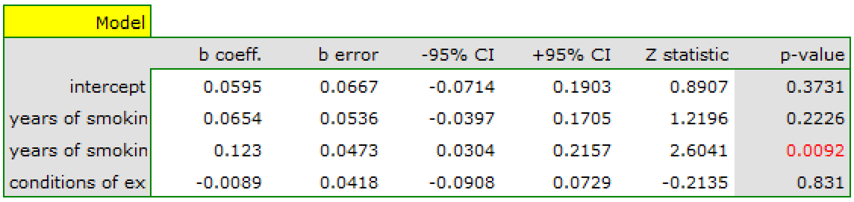
Model	b coeff.	b error	-95% CI	+95% CI	Z statistic	p-value
intercept	0.0462	0.0304	-0.0133	0.1057	1.5222	0.1279
years of smokin	0.0666	0.0513	-0.0338	0.1671	1.2999	0.1937
years of smokin	0.1218	0.043	0.0375	0.2062	2.8301	0.0047

Both models confirmed a significant association between the duration of smoking and the magnitude of the relative risk of disease X. In the first model, the logarithm of the relative risk of disease X increased by 0.0614 with increasing time of smoking (moving to the subsequent group of years of smoking). Analysis of the results of the second model leads to similar conclusions. In this case, the results are considered for the group of smokers smoking less than 5 years. The logarithm of relative risk for smokers between 5 and 10 years increases by 0.0666 (relative to smokers younger than 5 years), and for smokers older than 10 years it increases by 0.1218 (relative to smokers younger than 5 years).

Since part of the study was conducted according to other criteria (under different conditions) the obtained results of both models were corrected for different conditions of the study.

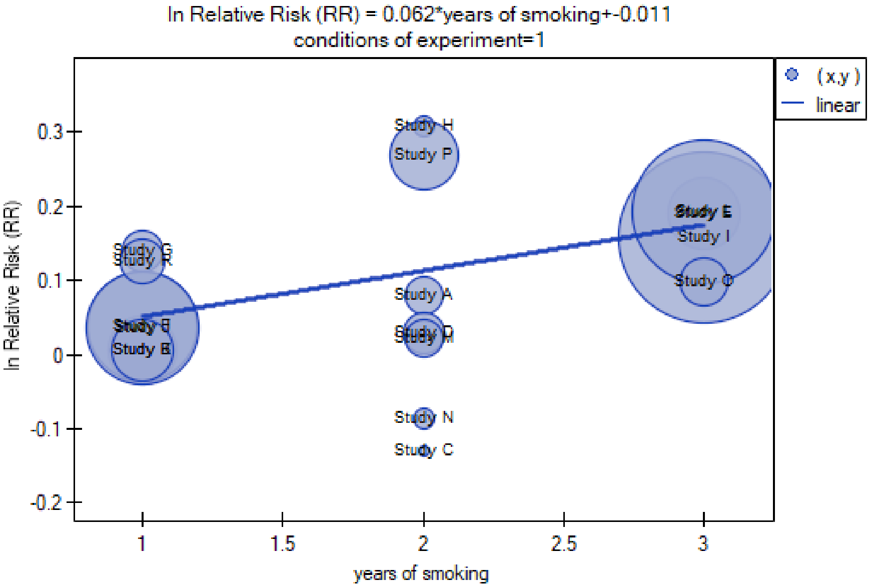


Model						
	b coeff.	b error	-95% CI	+95% CI	Z statistic	p-value
intercept	-0.0027	0.0688	-0.1375	0.1322	-0.0386	0.9692
years of smokin	0.0621	0.0222	0.0185	0.1057	2.7908	0.0053
conditions of ex	-0.0085	0.0396	-0.0861	0.0691	-0.2155	0.8294



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, and thus it can be concluded that the risk of disease X increases with years of smoking regardless of what methodology (inclusion/exclusion criteria of subjects) was used to conduct the study. The resulting relation for the first model, assuming that the study was conducted under condition ”a” (indicated as first conditions) is shown in the graph.



29 RELIABILITY ANALYSIS

Reliability analysis is usually associated with the complex scale construction, in particular summary scales (these consist of many individual items). Reliability analysis, associated as its internal consistency, informs us to what extent a particular scale measures what it should measure. In other words, to what extend the scale items measure the things that are measured by the whole scale.

When every scale item measures the same construct (the correlation between the items should be high) we can call it reliable scale. This assumption can be checked by calculating the matrix of the Pearson's correlation coefficient. Many measures of concordance can be used in reliability analysis. However, the most popular technique is the α-Cronbach coefficient and so-called split-half reliability.

Cronbach's α coefficient was named for the first time in 1951[49], by Cronbach. It measures the proportion of single item variances a and the whole scale variance (items sum). It is calculated according to the following formula:

$$\alpha_C = \frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k} sd_i^2}{sd_t^2}\right),$$

where:
k – number of scale items,
sd_i^2 – variance of i item,
sd_t^2 – variance of items sum.

Standardised reliability coefficient $\alpha_{standard}$ is calculated according to the following formula:

$$\alpha_{standard} = \frac{k\overline{r_p}}{1+(k-1)\overline{r_p}},$$

where:
$\overline{r_p}$ – mean of all the Pearson's correlation coefficients for ($k(k-1)/2$) scale items.

Alpha can take on any value less than or equal to 1, including negative values, although only positive values make sense. If all scale items are reliable, the reliability coefficient is 1.

There are some values that help in an assessesment of particular scale items usefulness:

- the value of α_C coefficient calculated after removing a particular scale item,
- the value of standard deviation of a scale calculated after removing a particular scale item,
- mean value of a scale calculated after removing a particular scale item,
- the Pearson's correlation coefficients between a particular item and the sum of other items.

Split-half reliability

Split-half reliability is a random scale item division into 2 halves and an analysis of the halves correlation. It is carried out by the Spearman-Brown split-half reliability coefficient, published independently by Spearman (1910)[154] and Brown (1910)[33]:

$$r_{SH} = \frac{2r_p^*}{1+r_p^*},$$

where:
r_p^* – the Pearson's correlation coefficient between halves of a scale.

If two halves, randomly selected, are ideally correlated: $r_{SH} = 1$.

A formula for the split-half reliability coefficient proposed by Guttman (1945)[71]:

$$r_{SHG} = 2\left(1 - \frac{sd_{t1}^2 + sd_{t2}^2}{sd_t^2}\right),$$

where:
sd_{t1}^2, sd_{t2}^2 – variance of the first and the second half of a scale,
sd_t^2 – variance of the sum of all scales items.

Note
The scale is realiable if the scales reliability coefficients (α_C, $\alpha_{standard}$, r_{SH} , r_{SHG}) are larger than 0.6 and smaller than 1.

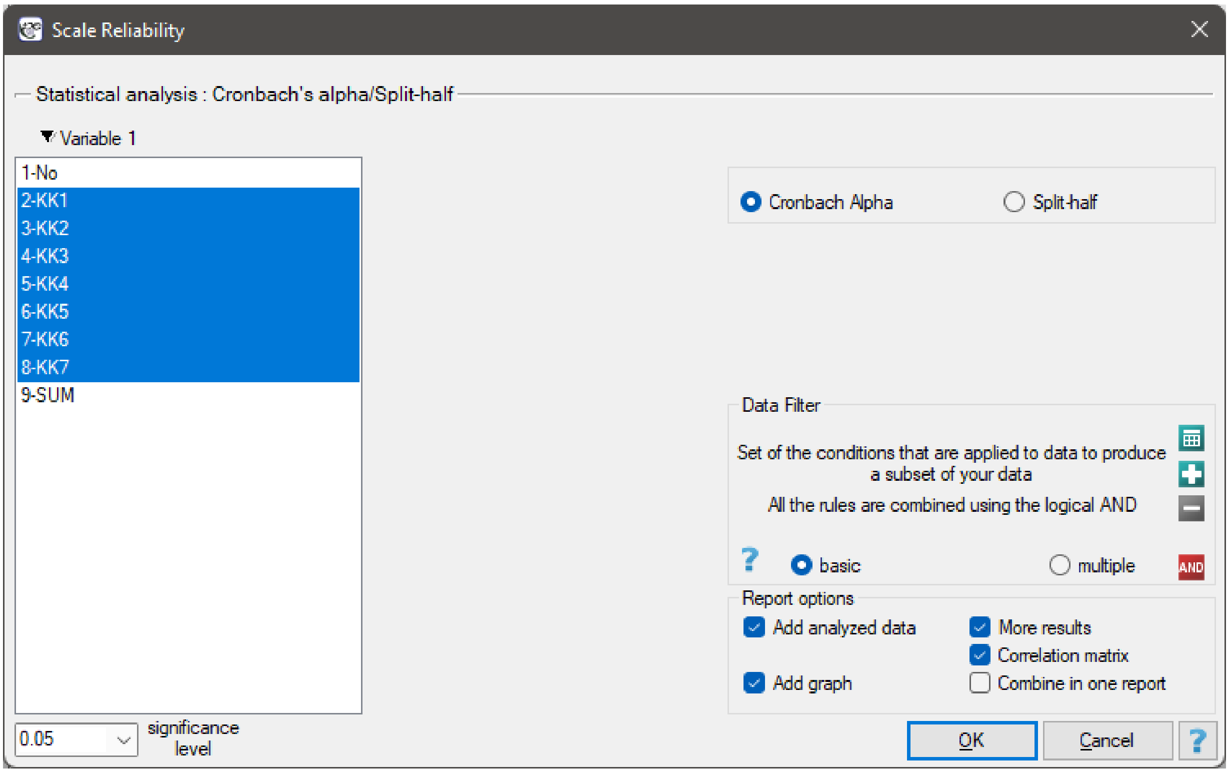
Standard error of measurement is calculated for the reliable scale, according to the following formula:

$$SEM = sd_t\sqrt{1-\alpha_C}$$ – for the Cronbach's alpha coefficient of reliability

or

$$SEM = sd_t\sqrt{1-r_{SH}}$$ – for the split-half reliability coefficient

The settings window with the Cronbach's alpha/Split-half can be opened in Statistics menu →Scale reliability.

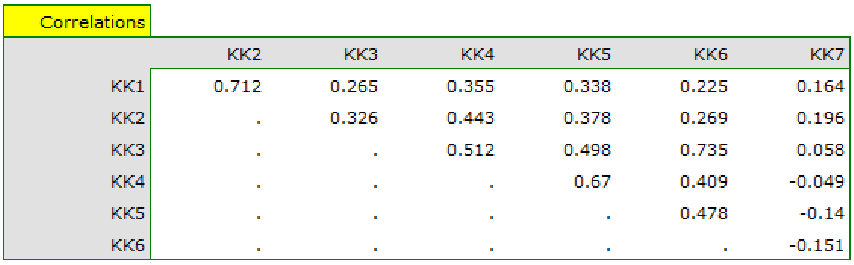


EXAMPLE 29.1. (scale.pqs file)
A ”competence scale”, created in some company, enables an assessment of the usefulness of future employees. Apart from participation in a job interview, candidates fill in the questionnaire that includes the ”competence scale” questions. There are 7 questions in the scale. For each question, one can get 1 - 5 points, where 1 - the lowest mark, 5 - the highest mark. The maximum score of the questionnaire is 35. In the table, there are scores obtained by 24 candidates.

Lp	KK1	KK2	KK3	KK4	KK5	KK6	KK7	SUMA
1	3	3	5	5	5	5	1	27
2	5	4	4	3	3	5	1	25
3	5	5	3	5	3	2	1	24
4	1	2	5	5	5	5	2	25
5	4	5	5	5	5	5	1	30
6	4	4	5	5	5	5	3	31
7	1	1	5	5	5	5	2	24
8	5	5	5	5	3	5	3	31
9	3	2	2	5	4	2	1	19
10	3	4	3	4	4	2	1	21
11	4	4	3	4	4	4	4	27
12	1	1	3	4	1	1	3	16
13	3	3	4	5	5	5	1	26
14	4	5	5	5	5	5	2	31
15	1	4	4	4	1	4	4	22
16	1	4	5	5	5	5	1	26
17	5	5	5	5	5	5	2	32
18	5	3	5	5	3	5	4	30
19	1	1	2	2	2	1	4	13
20	5	5	5	5	5	5	5	35
21	5	3	5	5	5	5	1	29
22	5	5	5	5	5	1	5	31
23	2	1	5	3	2	4	1	18
24	5	5	5	5	5	5	5	35

For checking the accuracy of the ”competence scale”, the reliability should be analysed.

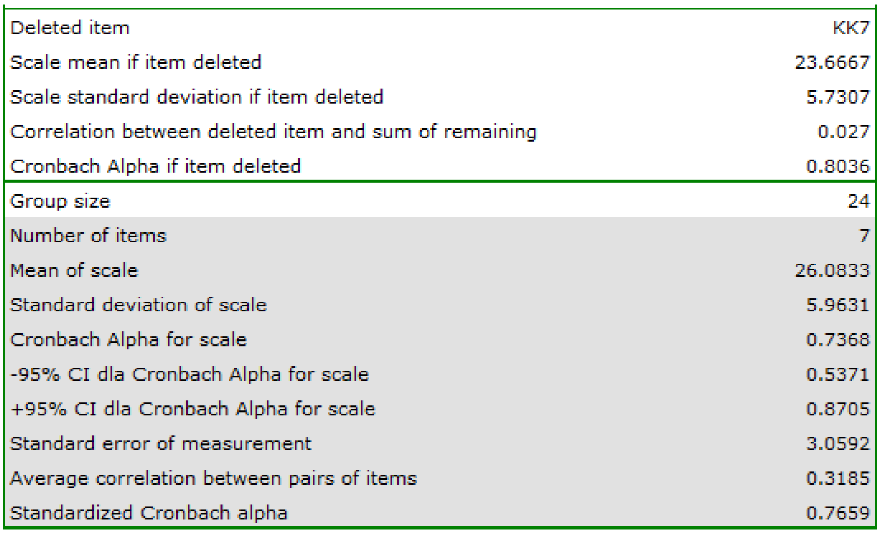
The correlation matrix indicates that the last item is least correlated with the other items. Thus, it is suspected that the item does not measure the same construct as the others.



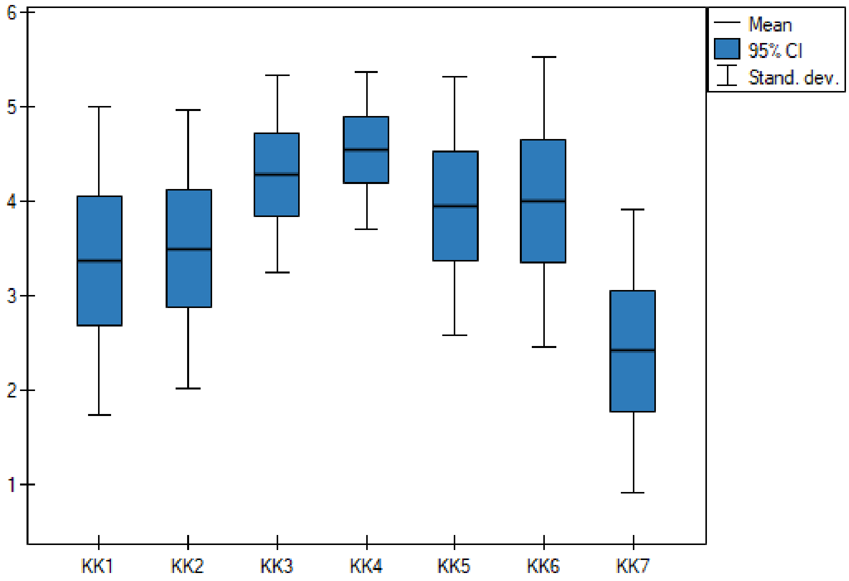
Correlations

	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

The competence scale turned out to be a reliable scale. Cronbach alpha coefficient is 0.7368, and mean of all the Pearson’s correlation coefficients is 0.3185.

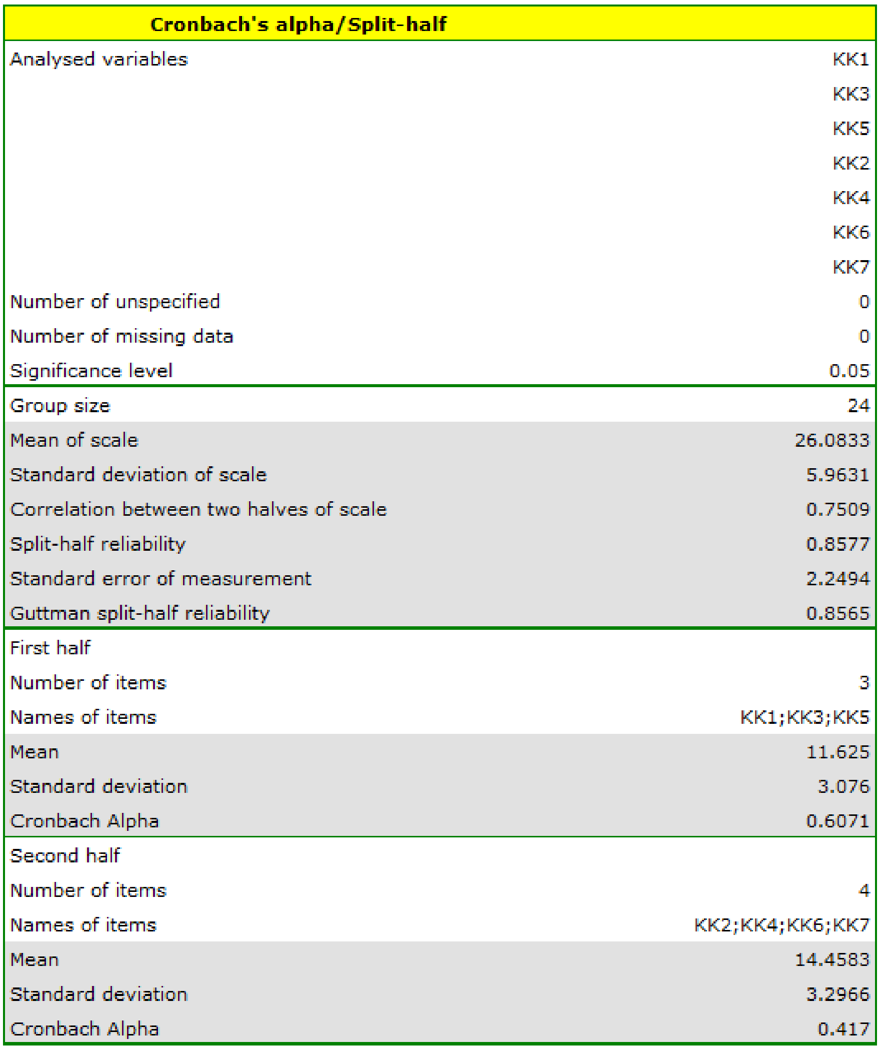


Deleted item	KK7
Scale mean if item deleted	23.6667
Scale standard deviation if item deleted	5.7307
Correlation between deleted item and sum of remaining	0.027
Cronbach Alpha if item deleted	0.8036
Group size	24
Number of items	7
Mean of scale	26.0833
Standard deviation of scale	5.9631
Cronbach Alpha 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



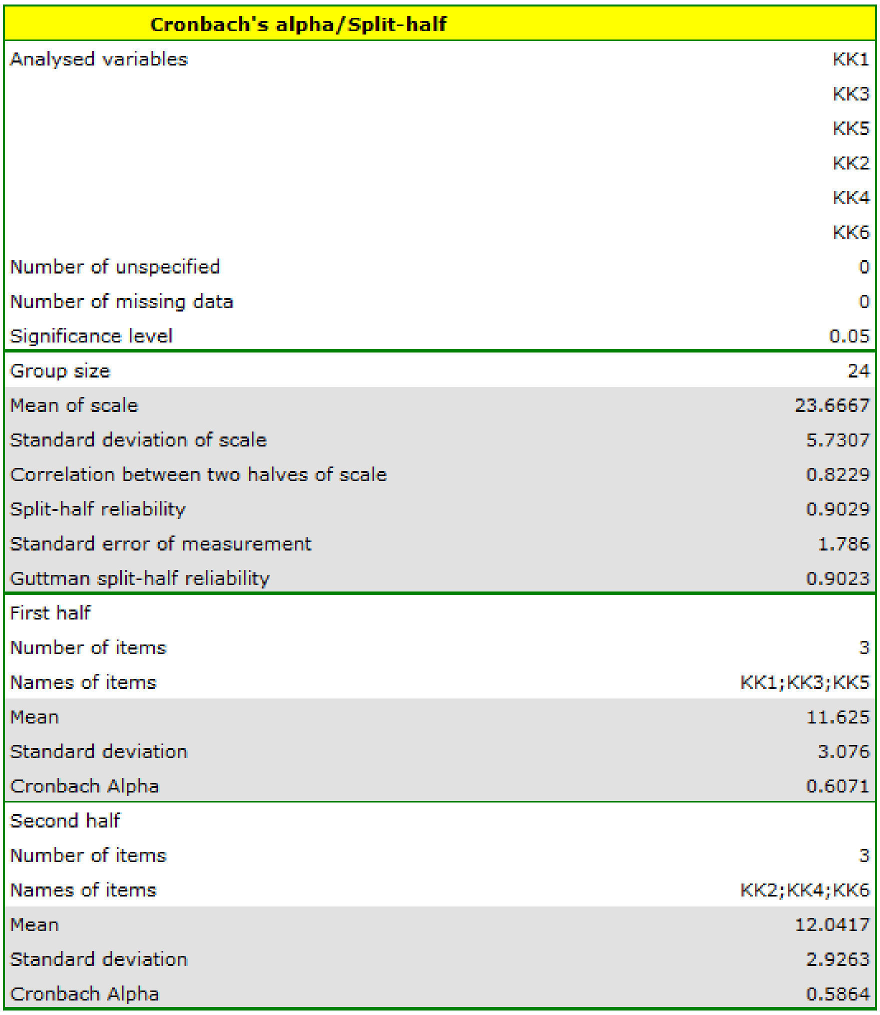
A more precised analysis of each item indicates that, except the last one, they all influence scale reliability in a similar way. Correlation between the KK7 item and the other scales items, is the weakest: 0.0270. Removing the KK7 item from the scale, the Cronbach alpha coefficient would increase to 0.8036.

Similar conclusion can be drawn on the basis of split-half reliability analysis, carried out on the items randomly divided into 2 halves (KK1, KK3, KK5) (KK2, KK4, KK6, KK7).



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

Spearman-Brown split-half reliability Coefficient is 0.8578. Guttman split-half reliability coefficient is 0.8565. The halves are well correlated – the correlation coefficient is 0.7509. However, the value of Cronbach alpha coefficient is too low for the second half (0.416958). This half includes the KK7 item, which shows a weak correlation with the other scale items. Removing the item and repeating the analysis, all the items are really high and reliable.

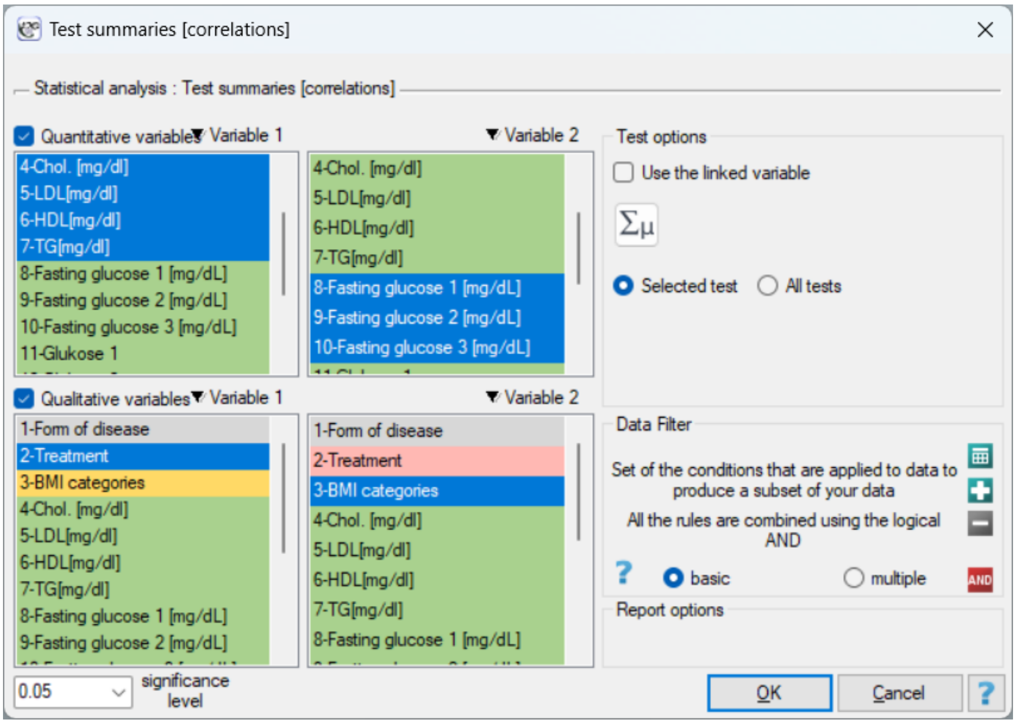


Cronbach's alpha/Split-half	
Analysed variables	KK1
	KK3
	KK5
	KK2
	KK4
	KK6
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Group size	24
Mean of scale	23.6667
Standard deviation of scale	5.7307
Correlation between two halves of scale	0.8229
Split-half reliability	0.9029
Standard 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

30 Test summaries

To speed up the work, we can perform individual tests in sets. The quantitative data will be able to be further described by means, medians, etc., and the qualitative data by counts and percentages.

The settings window with the Test summaries can be opened in Stistics→Summaries→Test summaries and then the selected group of analyses.



Test summaries [correlations]
Statistical analysis : Test summaries [correlations]
☑ Quantitative variables
Variable 1
Variable 2
4-Chol. [mg/dl]
5-LDL[mg/dl]
6-HDL[mg/dl]
7-TG[mg/dl]
8-Fasting glucose 1 [mg/dL]
9-Fasting glucose 2 [mg/dL]
10-Fasting glucose 3 [mg/dL]
11-Glukose 1
☑ Qualitative variables
Variable 1
Variable 2
1-Form of disease
2-Treatment
3-BMI categories
4-Chol. [mg/dl]
5-LDL[mg/dl]
6-HDL[mg/dl]
7-TG[mg/dl]
8-Fasting glucose 1 [mg/dL]
9-Fasting glucose 2 [mg/dL]
Test options
□ Use the linked variable
Σμ
● Selected test ○ All tests
Data Filter
Set of the conditions that are applied to data to produce a subset of your data
All the rules are combined using the logical AND
? ● basic ○ multiple
AND
Report options
0.05 significance level
OK
Cancel
?

At our disposal we have:

1. Comparison of two dependent groups:
   - The t-test for dependent groups
   - The Wilcoxon test (matched-pairs)
   - The Bowker-McNemar test
   - Normality of distribution test Kołmogorov-Smirnov (or another one suggested by the user)
   - and others ...
2. Comparison of two independent groups:
   - The t-test for independent groups
   - The Mann-Whitney U test
   - The Chi-square tests, Fisher exact, OR/RR
   - Normality of distribution test Kołmogorov-Smirnov (or another one suggested by the user)
   - and others ...
3. Korelację:
   - Pearson linear correlation
   - Spearman's monotonic correlation

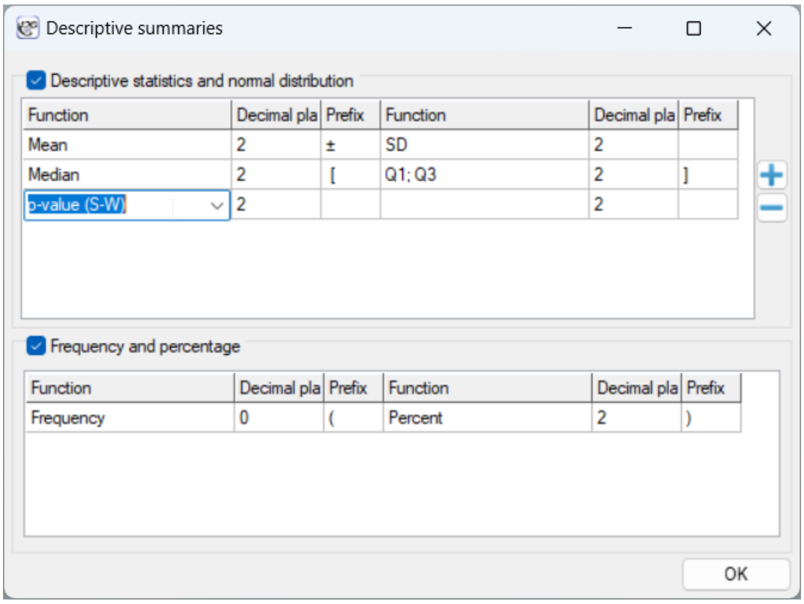
- The Chi-square tests, Fisher exact, correlation co.
- Normality of distribution test Kołmogorov-Smirnov (or another one suggested by the user)
- and others ...

In the program, for each of the analyzed variables, depending on whether they are quantitative or qualitative, we can return results:

- **selected tests** - automatically according to the rule described below returned in the table report;
- **all tests** and accompanying coefficients and measures regardless of whether the minimum conditions for their use are met.

Notes on the program's automatic test selection
Note 1!
If the user does not describe each variable with the appropriate scales before analysis, the quantitative data will be treated as an interval scale, and the qualitative data as a nominal scale.

Note 2!
Testing the normality of the distribution is based on the results of the normality test selected by the user when setting the descriptive statistics.



☑ Descriptive statistics and normal distribution

Function	Decimal pla	Prefix	Function	Decimal pla	Prefix
Mean	2	±	SD	2	
Median	2	[	Q1; Q3	2	]
p-value (S-W)	2			2	

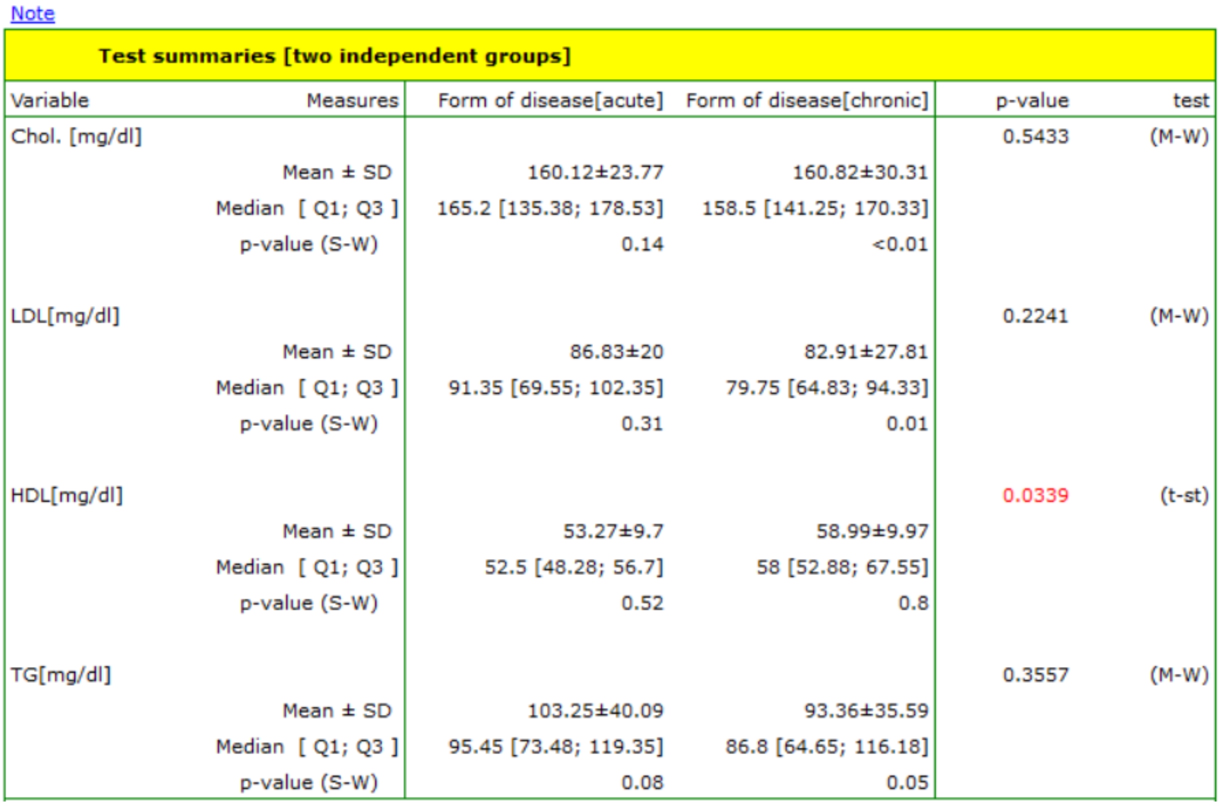
☑ Frequency and percentage

Function	Decimal pla	Prefix	Function	Decimal pla	Prefix
Frequency	0	(	Percent	2	)

Note 3!
If the user chooses not to indicate tests that assess the normality of the distribution in the window of these statistics, then it will be checked based on the Kolmogorov-Smirnov test. The analyses we propose are robust to small deviations from the normal distribution, and the Kolmogorov-Smirnov test is the most conservative among the available tests, by which we show the non-normality of the distribution only when the tested distribution differs greatly from the normal distribution. In this situation, we test the normality of the distribution (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 in measurements, (3) for correlation based on the model residuals.

Example 30.1. (Summaries.pqs file)
We want to make an automatic comparison between two independent groups: chronic patients and acute patients. We make the comparison based on the data described on the interval scale: Chol, LDL,

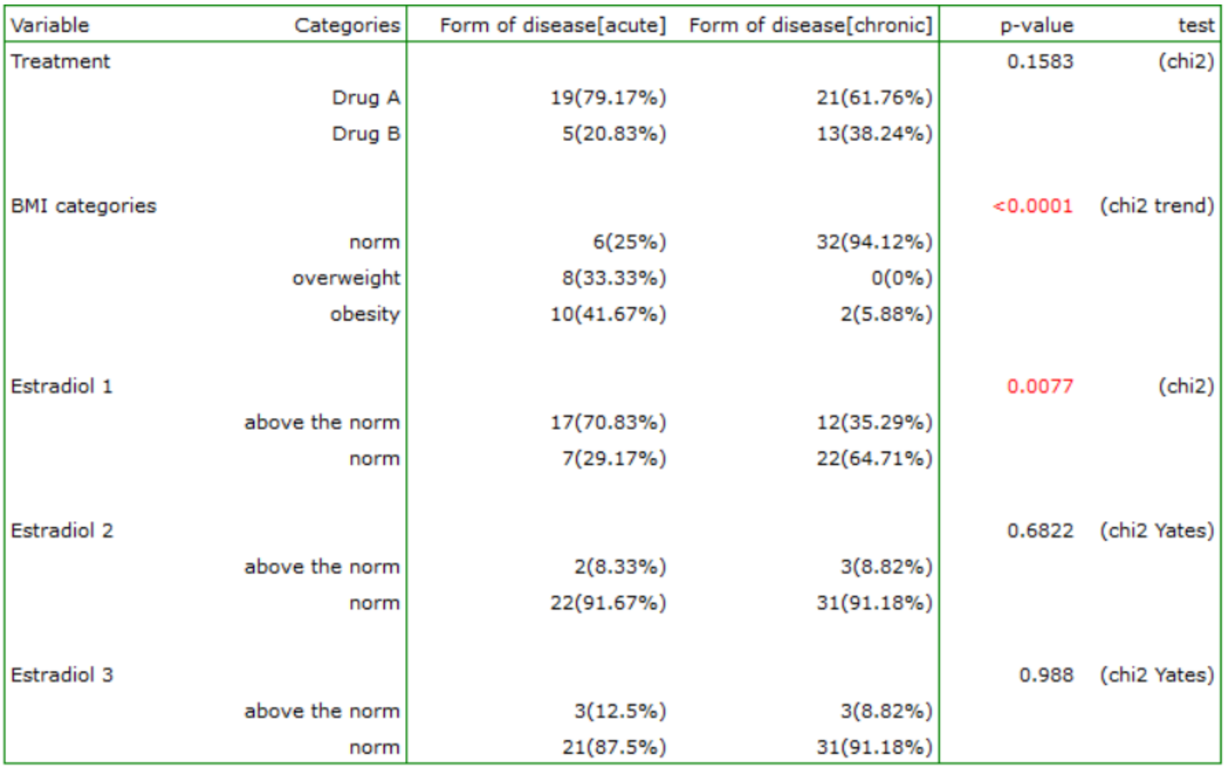
HDL, TG, based on the ordinal feature BMI, and nominal data: Treatment, Estradiol 1, Estradiol 2, Estradiol 3. To do this, we choose the menu Summaries → Test summaries [two independent groups] and select the grouping variable: Treatment, then we select interval data (in the quantitative variables section) and nominal and ordinal data (in the qualitative variables section). We select the option Selected test and perform the analysis.



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%)		

The result is both a description of each group and the statistical test selected for comparison.

In the Note at the top of the report, there is a description informing the principle of selecting a statistical test suitable for the analysis:
”**Quantitative variables**:
For the interval scale with normality of distribution, the unpaired Student’s t-test (t-st) or its Cochran-Cox correction (C-C) was determined when the variances of the groups differed. For the interval scale, when the condition of normality of distribution was not met, as for the ordinal scale, the Mann-Whitney (M-W) test was determined. Normality of the data was tested with the Shapiro-Wilk and equality of variance was tested with the Fisher-Snedecor (F-S) 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 ordinal scale, the chi-square test for trend was determined. If the scale was not marked for the variables analyzed, it was assumed that the data came from the nominal scale.”

31 THE WIZARD

The Wizard is a tool which makes the navigation easier to go, through the basic statistics included in an application, especially for a novice user. It includes suggestions of assumptions which should be checked before the choice of a particular statistic test. The last step of the wizard is to select an appropriate statistic test and to open the window with the settings of the test options.

The Wizard may be launched by:
- Statistics→Wizard,
- button on a toolbar.

A launched wizard window includes the possibility to choose the kind of an analysis that a user wants to carry out. A user may choose:

Comparison — 1 group - to compare values of measurements coming from a 1 population with the specific value given by the user. This population is represented by raw data gathered in a 1 column or cumulated to the form of a frequency table.

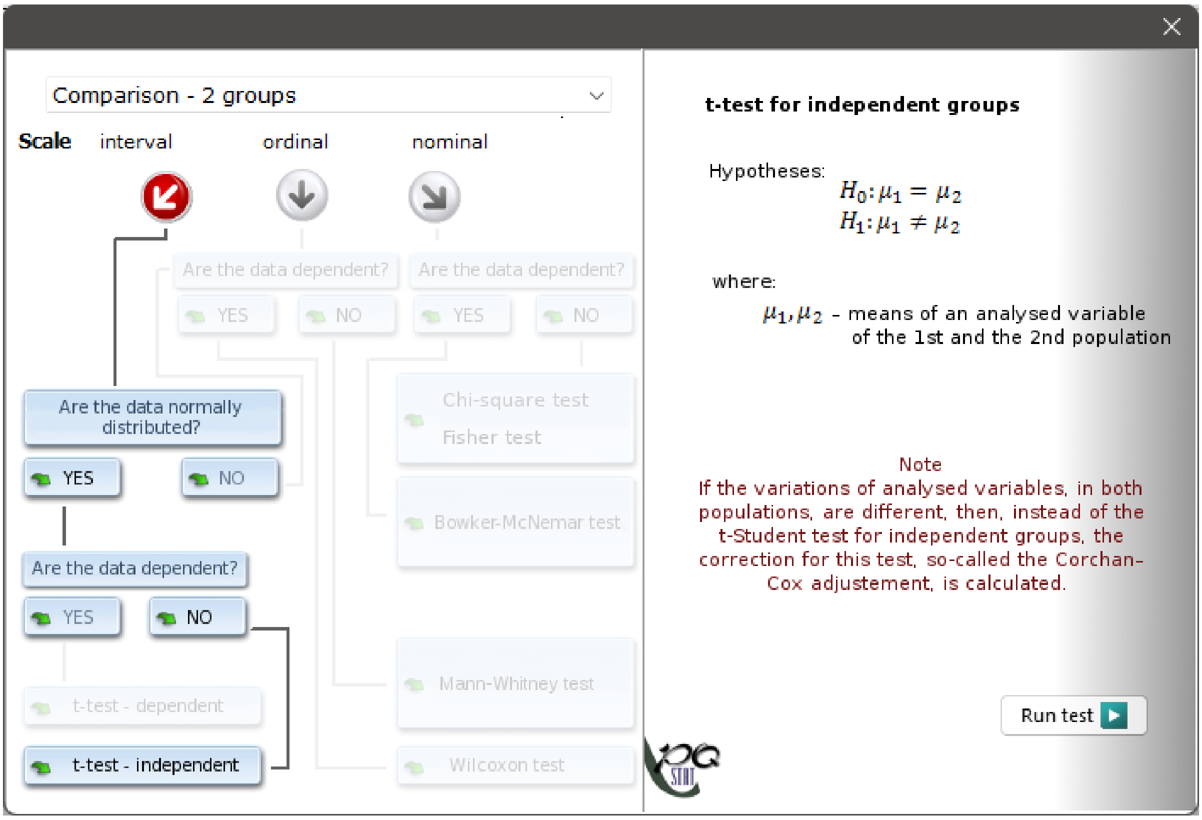
Comparison — 2 groups - to compare values of measurements coming from 2 populations. These populations are represented by raw data gathered in 2 columns or cumulated to the form of a contingency table.

Comparison — more than 2 groups - to compare values of measurements coming from several populations. The populations are represented by data collected in the form of raw data, in several columns.

Correlation - to check the occurrence of dependence between 2 parameters coming from a 1 population. These features are represented by raw data gathered in 2 columns or cumulated to the form of a contingency table.

Agreement - to check the concordance of obtained measurements. These features are represented by raw data gathered in several columns or cumulated to the form of a contingency table.

When the user chooses the kind of an analysis, a graph will occur. The graph is divided according to a scale, on which the measurement of the analysed features was done (interval scale, ordinal scale, nominal scale).



t-test for independent groups

Hypotheses:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

where:

μ_1, μ_2 - means of an analysed variable of the 1st and the 2nd population

Note
If the variations of analysed variables, in both populations, are different, then, instead of the t-Student test for independent groups, the correction for this test, so-called the Corchan-Cox adjustement, is calculated.

The user moves on the graph by selecting the adequate answers to the asked questions. After the user gets through the way on the graph, chosen by himself, he is able to perform this test, which – according to the replies – is an appropriate one to solve the determined statistical problem.

32 OTHER NOTES

32.1 FILES FORMAT

PQS - default file format for PQStat files; is used for representing all objects created with ℙℚStat (project,datasheet,report,graph);

PQX - XML file for PQStat, is used for representing all objects created with ℙℚStat; PQX files are stored in Unicode text format (support UTF-8 character encoding); recommended for use on computers with a small amount of memory.

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*

[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] Bland J.M., Altman D.G. (1986), *Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327 (8476): 307–10*

[22] Bland J.M., Altman D.G. (1999), *Measuring agreement in method comparison studies. Statistical Methods in Medical Research 8 (2): 135–60*

[23] Bowker A.H. (1948), *Test for symmetry in contingency tables. Journal of the American Statistical Association, 43, 572-574*

[24] Box G. E. , Cox D. R. (1964), *An analysis of transformations. Journal of the Royal Statistical Society, Series B 26: 211–252*

[25] 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*

[26] Breslow N.E. (1996), *Statistics in epidemiology: the case-control study', Journal of the American Statistical Association, 91, 14—28*

[27] Breslow N.E. (1974), *Covariance analysis of censored survival data. Biometrics, 30(1):89–99*

[28] Brookmeyer R. and Crowley J. (1982a), *A confidence interval for the median survival time. Biometrics 38, 29-41*

[29] Brown L.D., Cai T.T., DasGupta A. (2001), *Interval Estimation for a Binomial Proportion. Statistical Science, Vol. 16, no. 2, 101-133*

[30] Brown M.B., Forsythe A. B. (1974a), *Robust tests for equality of variances. Journal of the American Statistical Association, 69,364-367*

[31] Brown M. B., Forsythe A. B. (1974), *The ANOVA and multiple comparisons for data with heterogeneous variances. Biometrics, 30, 719-724*

[32] 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*

[33] Brown W. (1910), *Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296-322*

[34] Hochberg Y. (1988), *A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–803*

[35] Chow S.C., Shao J., and Wang H. (2008). *Sample Size Calculations in Clinical Research, Second Edition. Chapman and Hall/CRC. Boca Raton, Florida.*

[36] Cicchetti D. and Allison T. (1971),*A new procedure for assessing reliability of scoring eeg sleep recordings. American Journal EEG Technology, 11, 101-109*

[37] Cleveland, W. S. (1979),*Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association. 74*

[38] Clopper C. and Pearson S. (1934), *The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26: 404-413*

[39] Cochran W.G. (1950), *The comparison ofpercentages in matched samples. Biometrika, 37, 256-266*

[40] Cochran W.G. (1952), *The chi-square goodness-of-fit test. Annals of Mathematical Statistics, 23, 315-345*,

[41] Cochran W. G. (1954), *Some methods for strengthening the common chi-square tests. Biometrics, 10(4) 17-45 1*

[42] Cochran W.G. and Cox G.M. (1957), *Experimental designs (2nd 4.). New York: John Wiley and Sons.*

[43] Cohen J. (1960), *A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 10,3746*

[44] Cohen J. (1968), *Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220*

[45] Cohen J. (1988), *Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, Hillsdale, New Jersey*

[46] Conover W. J. (1999), *Practical nonparametric statistics (3rd ed). John Wiley and Sons, New York*

[47] Cox D.R. (1972), *Regression models and life tables. Journal of the Royal Statistical Society, B34:187-220*

[48] Cramkr H. (1946), *Mathematical models of statistics. Princeton, NJ: Princeton University Press.*

[49] Cronbach L.J. (1951), *Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334*

[50] 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*

[51] Dunn O. J. (1964), *Multiple comparisons using rank sums. Technometrics, 6: 241–252*

[52] Durbin J. (1951), *Incomplete blocks in ranking experiments. British Journal of Statistical Psychology, 4: 85–90*

[53] 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*

[54] Epps T.W., Pulley L.B. (1983), *A test for normality based on the empirical characteristic function. Biometrika. 1983;70:723–726*

[55] 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.*

[56] Fisher R.A. (1934), *Statistical methods for research workers (5th ed.). Edinburgh: Oliver and Boyd*

[57] Fisher R.A. (1935), *The logic of inductive inference. Journal of the Royal Statistical Society, Series A, 98,39-54*

[58] Fisher R.A. (1936), *The use of multiple measurements in taxonomic problems. Annals of Eugenics 7 (2): 179–188*

[59] Fleiss J.L. (1971), *Measuring nominal scale agreement among many raters. Psychological Bulletin, 76 (5): 378–382*

[60] Fleiss J.L., Cohen J. (1973), *The equivalence of weighted kappa and the intraclass correlation coef-fcient as measure of reliability. Educational and Psychological Measurement, 33, 613-619*

[61] Fleiss J.L., Levin B., Paik M.C. (2003), *Statistical methods for rates and proportions. 3rd ed. (New York: John Wiley) 598-626*

[62] 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*

[63] 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*

[64] 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*

[65] 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*

[66] Gehan E. A. (1965a), *A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. Biometrika, 52:203—223*

[67] Gehan E. A. (1965b), *A Generalized Two-Sample Wilcoxon Test for Doubly-Censored Data. Biometrika, 52:650—653*

[68] Goodman L. A. (1960), *On the exact variance of products. Journal of the American Statistical Association, 55, 708-713*

[69] Greenhouse S. W., Geisser S. (1959), *On methods in the analysis of profile data. Psychometrika, 24, 95–112*

[70] Green S.B. (1991), *How many subjects does it take to do a regression analysis? Multivariate Behavioral Research, 26, 499-510*

[71] Guttman L. (1945), *A basic for analyzing test-retest reliabilit. Psychometrika, 10, 255-282*

[72] 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*

[73] 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*

[74] 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*

[75] Hanusz Z., Tarasińska J. (2014), *On multivariate normality tests using skewness and kurtosis, Colloquium Biometricum 44, 139-148*

[76] Henderson, R. (1916), *Note on graduation by adjusted average.Transactionsof the Actuarial Society of America, 17:43–48*

[77] Henze N., Zirkler B. (1990), *A class of invariant consistent tests for multivariate normality. Comm. Statist. Theory Methods. 1990;19:3595–3617*

[78] Hochberg Y. (1988), *A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika 75 (4): 800–802*

[79] Holm S. (1979), *A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70*

[80] Hotelling H. (1931), *The generalization of Student's ratio. Annals of Mathematical Statistics 2 (3): 360–378*

[81] 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*

[82] 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*

[83] 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*

[84] Iman R. L., Davenport J. M. (1980), *Approximations of the critical region of the friedman statistic, Communications in Statistics 9, 571–595*

[85] Jarque C. M., Bera A. K., (1987)., *A test for normality of Observations and Regression Residuals, International Statistical Review 55, 163-172*

[86] 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*

[87] Jonckheere A. R. (1954), *A distribution-free k-sample test against ordered alternatives. Biometrika, 41: 133–145*

[88] Kaplan E.L., Meier P. (1958), *Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53:457-481*

[89] Kendall M.G. (1938), *A new measure of rank correlation. Biometrika, 30, 81-93.*

[90] Kendall M.G., Babington-Smith B. (1939), *The problem of m rankings. Annals of Mathematical Statistics, 10, 275-287*

[91] Kleinbaum D. G., Klein M. (2005), *Survival Analysis: A Self-Learning Text, Second Edition (Statistics for Biology and Health)*

[92] Kolmogorov A.N. (1933), *Sulla deterrninazione empirica di una legge di distribuzione. Giorn-de1l'Inst. Ital. degli. Art., 4, 89-91*

[93] Kruskal W.H. (1952), *A nonparametric test for the several sample problem. Annals of Mathematical Statistics, 23, 525-540*

[94] Kruskal W.H., Wallis W.A. (1952), *Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, 583-621*

[95] Lancaster H.O. (1961), *Significance tests in discrete distributions. Journal of the American Statistical Association 56:223-234*

[96] Lawley D. N. (1938), *A generalization of Fisher's z-test. Biometrika 30: 180–187*

[97] Lee E. T., Wang J. W. (2003), *Statistical Methods for Survival Data Analysis, ed. third, Wiley*

[98] Lenth, R. V., (2001), *Some Practical Guidelines for Effective Sample Size Determination. The American Statistician, 55(3), 187-193*

[99] 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*

[100] 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.*

[101] 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*

[102] 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*

[103] 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*

[104] Lloyd S. P. (1982), *Least squares quantization in PCM. IEEE Transactions on Information Theory 28 (2): 129–137*

[105] Lund R.E., Lund J.R. (1983), *Algorithm AS 190, Probabilities and Upper Quantiles for the Studentized Range. Applied Statistics; 34*

[106] Mahalanobis P. C. (1930), *On tests and measures of group divergence. Journal of the Asiatic Society of Bengal 26: 541–588*

[107] Mahalanobis P. C. (1936), *On the generalized distance in statistics. National Institute of Science of India 12: 49–55*

[108] 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*

[109] 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*

[110] Mantel N. (1963), *Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. J. Am. Statist. Assoc., 58, 690-700*

[111] Mantel N. (1966), *Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. Cancer Chemotherapy Reports, 50:163—170*

[112] Marascuilo L.A. and McSweeney M. (1977), *Nonparametric and distribution-free method for the social sciences. Monterey, CA: Brooks/Cole Publishing Company*

[113] Mardia K. V. (1970), *Measures of multivariate skewness and kurtosis with applications, Biometrica 57, 519-530*

[114] Mardia K. V. (1974), *Applications of some measuresof multivariate skewness and kurtosis for testing normality and robustness studies, Sankhay B 36, 115-128*

[115] Mauchly J. W. (1940), *Significance test for sphericity of n-variate normal population. Annals of Mathematical Statistics, 11, 204-209.*

[116] McNemar Q. (1947), *Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157*

[117] 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*

[118] Miettinen O.S. (1985), *Theoretical Epidemiology: Principles of Occurrence Research in Medicine. John Wiley and Sons, New York*

[119] Miettinen O.S. and Nurminen M. (1985), *Comparative analysis of two rates. Statistics in Medicine 4: 213-226*

[120] 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*

[121] Nadaraya, E. A. (1964), *On Estimating Regression. Theory of Probability and Its Applications. 9 (1): 141–2*

[122] Newcombe R.G. (1998), *Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods. Statistics in Medicine 17: 873-890*

[123] Newman S.C.(2001), *Biostatistical Methods in Epidemiology. 2nd ed. New York: John Wiley*

[124] 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.*

[125] Ogilvie J. C. (1965), *Paired comparison models with tests for interaction. Biometrics 21(3): 651-64*

[126] Orwin R. G. (1983), *A Fail-SafeN for Effect Size in Meta-Analysis. J Educ Behav Stat, 8(2):157-159*

[127] 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*

[128] 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*

[129] 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*

[130] 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*

[131] Pillai K. C. (1955), *Some new test criteria in multivariate analysis. Annals of Mathematical Statistics 26: 117–121*

[132] Plackett R.L. (1984), *Discussion of Yates' "Tests of significance for 2x2 contingency tables". Journal of Royal Statistical Society Series A 147:426-463*

[133] Pratt J.W. and Gibbons J.D. (1981), *Concepts of Nonparametric Theory. Springer-Verlag, New York*

[134] 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*

[135] 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*

[136] Rosenbaum P.R., Rubin D.B. (1983a), *The central role of the propensity score in observational studies for causal effects. Biometrika; 70:41–55*

[137] Rosenthal R. (1979), *The "file drawer problem" and tolerance for null results. Psychological Bulletin, 5, 638-641*

[138] Rothman K.J., Greenland S., Lash T.L. (2008), *Modern Epidemiology, 3rd ed. (Lippincott Williams and Wilkins) 221−225*

[139] Roy S. N. (1939), *p-statistics or some generalizations in analysis of variance appropriate to multivariate problems. Sankhya 4: 381–396*

[140] Royston P. (1992), *Approximating the Shapiro–Wilk W-test for non-normality". Statistics and Computing 2 (3): 117–119*

[141] Royston P. (1993b), *A toolkit for testing for non-normality in complete and censored samples. Statistician 42: 37–43*

[142] Rufibach K. (2010), *Assessment of paired binary data; Skeletal Radiology volume 40, pages1–4*

[143] Satterthwaite F.E. (1946), *An approximate distribution of estimates of variance components. Biometrics Bulletin, 2, 1 10-1 14*

[144] 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*

[145] 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*

[146] Scott D. W., (1992), *Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley.*

[147] Shapiro S.S. and Wilk M.B. (1965), *An analysis of variance test for normality (complete samples). Biometrika 52 (3–4): 591–611*

[148] Sheather S.J. (2009), *A modern approach to regression with R. New York, NY: Springer*

[149] Shrout P.E., and Fleiss J.L (1979), *Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86, 420-428*

[150] Šidák Z. K. (1967), *Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association, 62 (318): 626–633*

[151] Silverman B. W., (1986), *Density estimation for statistics and data analysis, London: Chapman and Hall*

[152] 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*

[153] Sobel M. E. (1982). *Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology 13: 290–312*

[154] Spearman C. (1910), *Correlation calculated from faulty data. British Journal of Psychology, 3, 271-295*

[155] Tamhane A. C. (1977), *Multiple comparisons in model I One-Way ANOVA with unequal variances. Communications in Statistics, A6 (1), 15-32*

[156] Tarone R. E., Ware J. (1977), *On distribution-free tests for equality of survival distributions. Biometrica, 64(1):156-160*

[157] Tarone R.E. (1985), *On heterogeneity tests based on efficient scores. Biometrika 72, 91–95*

[158] 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*

[159] Terrell G. R. (1990), *The maximal smoothing principle in density estimation. Journal of the American Statistical Association 85, 470–477*

[160] Terrell G.R., Scott D. W. (1985), *Oversmoothed nonparametric density estimates. Journal of the American Statistical Association 80, 209-214*

[161] Thode H. C. (2002), *Testing For Normality. CRC Press; 2002. 506 s.*

[162] Volinsky C.T., Raftery A.E. (2000) , *Bayesian information criterion for censored survival models. Biometrics, 56(1):256–262*

[163] Wallenstein S. (1997), *A non-iterative accurate asymptotic confidence interval for the difference between two Proportions. Statistics in Medicine 16: 1329-1336*

[164] Wallis W.A. (1939), *The correlation ratio for ranked data. Journal of the American Statistical Association, 34,533-538*

[165] Watson, G. S. (1964), *Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A. 26 (4): 359–372*

[166] Welch B. L. (1951), *On the comparison of several mean values: an alternative approach. Biometrika 38: 330–336*

[167] Wilcoxon F. (1945), *Individual comparisons by ranking methods. Biometries, 1,80-83*

[168] Wilcoxon F. (1945), *Individual comparisons by ranking methods. Biometries, 1, 80-83*

[169] Wilcoxon F. (1949), *Some rapid approximate statistical procedures. Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation*

[170] Wilcoxon F. (1949), *Some rapid approximate statistical procedures. Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation*

[171] Wilcoxon F. (1949), *Some rapid approximate statistical procedures. Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation*

[172] Wilson E.B. (1927), *Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association: 22(158):209-212*

[173] Wilks S.S. (1932), *Certain generalizations in the analysis of variance. Biometrika 24: 471–494*

[174] Yates F. (1934), *Contingency tables involving small numbers and the chi-square test. Journal of the Royal Statistical Society, 1,2 17-235*

[175] Youden W.J. (1950), *Index for rating diagnostic tests. Cancer. 3: 32–35*

[176] 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*

[177] Zar J. H., (2010), *Biostatistical Analysis (Fifth Edition). Pearson Educational*

[178] Zweig M.H., Campbell G. (1993), *Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry 39:561-577*