Classes as part of the biostatistics module

April 20–24, 2026

Agnieszka Strzelecka

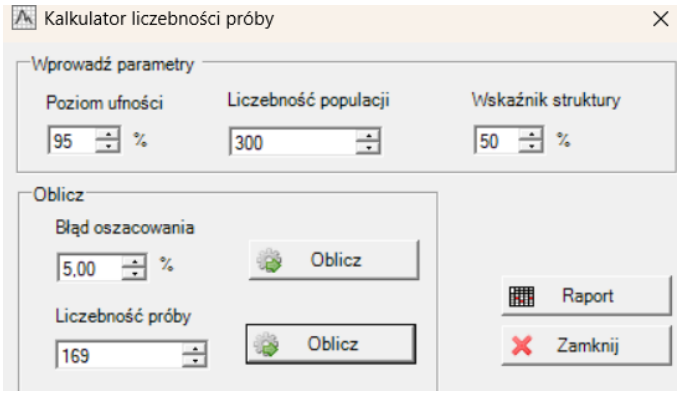
„Edukacja bez granic dla kierunków medycznych w UJK"
projekt dofinansowany przez Narodową Agencję Wymiany Akademickiej

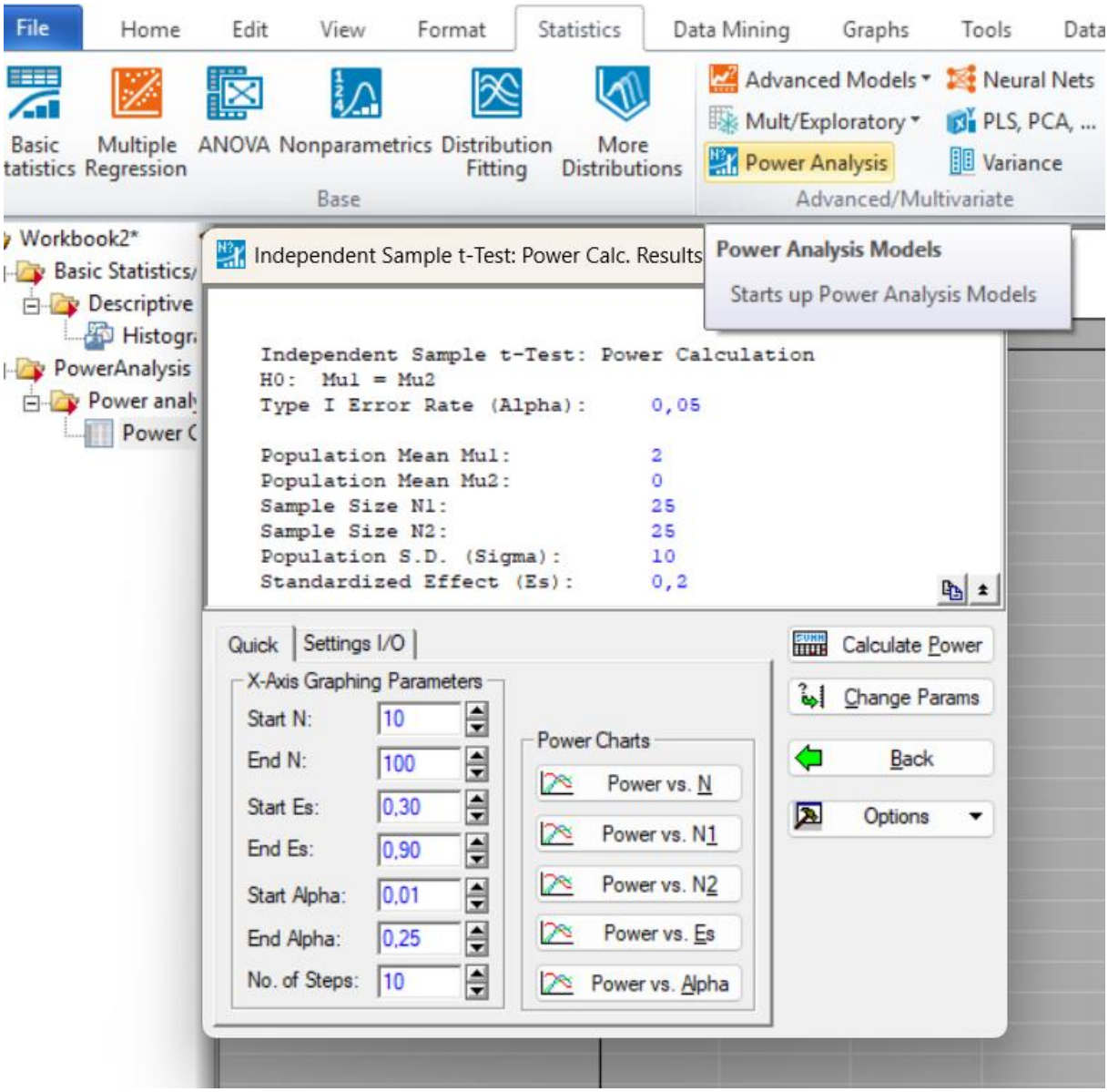
Using TIBC Statistica 13.3 for preliminary data analysis. Creating and managing a database (data management). Descriptive statistics—selecting, calculating, and interpreting statistical measures; graphical presentation of data according to its type and the measurement scale used. Validation of the tool. Factor analysis.

Introduction to Statistical Concepts

One of the fundamental goals of science is to explain and predict the outcomes of observed events and the causal relationships between them. Statistics is a valuable tool in this regard. Based on knowledge of the characteristics of a suitably selected subset (sample) of a given population, we will draw conclusions regarding the characteristics under consideration for the remaining, unknown elements of that population (the general population). To this end, we will use the main branches of mathematical statistics: statistical hypothesis testing and the estimation of parameters or functions.

The first step we must take is to define the population about which we want to formulate research questions and specify the purpose of our research. We must therefore clarify what we mean by the terms population and sample. A POPULATION is the set of all elements that are subject to study from the perspective of various research criteria, and a SAMPLE is a subset selected from the population for the purpose of drawing conclusions about the population. We must ensure that the sample is representative of the entire population. We can then generalize the results of the study conducted on the sample to all elements of the population that were not studied. Examples of sample selection for research:





It is also important to have a sufficiently large sample size to ensure that the analyses have adequate statistical power. The characteristics that distinguish the units within the population under study are called statistical characteristics. A population has many characteristics. However, for a specific study, we select only those most relevant to the problem being analyzed. After identifying the characteristics that interest us, we must decide how we will measure the values of these characteristics during the observation. In a specific study, the researcher must determine what type of scale to use for the characteristic

under study. The choice of a specific scale determines the applicability of appropriate statistical tools for analyzing the collected data.

When conducting a statistical study, we should consider the following stages:

I. Preliminary stage – defining the purpose of the study, the type of statistical population, and the statistical characteristics to be examined.

II. Data collection stage – observation of units, test results, or interviews

III. Processing of raw statistical data – verification of collected data, calculation of descriptive statistics, and graphical presentation of the processed data.

IV. Analysis of the processed data – point and interval estimation, hypothesis testing, and meta-analysis.

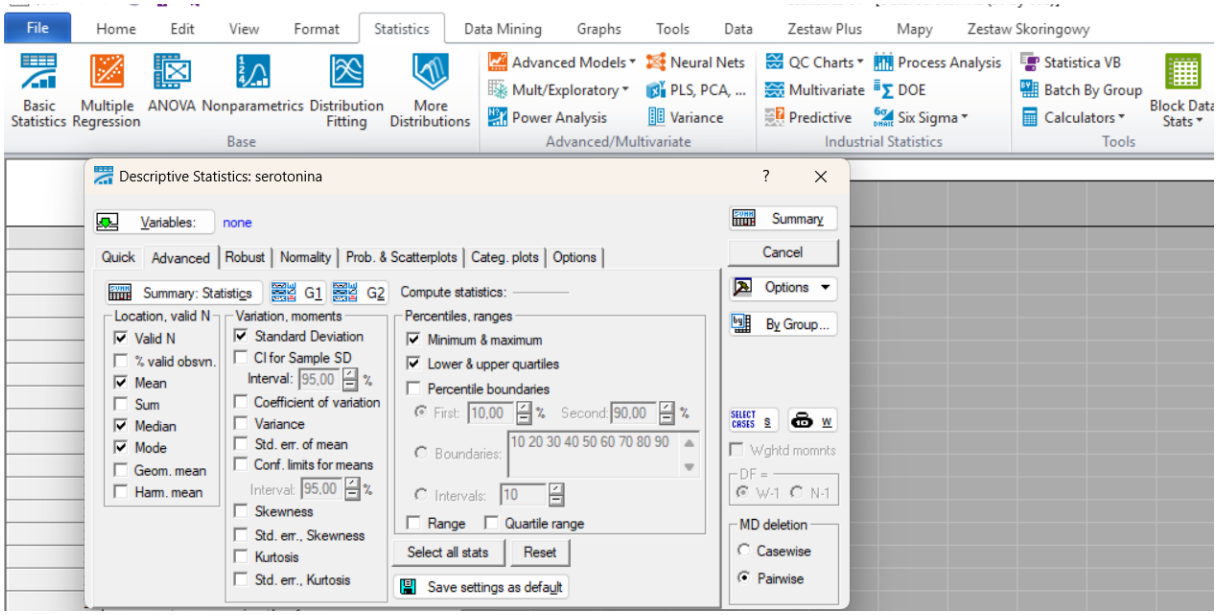
DESCRIPTIVE STATISTICS

Once the observation phase is complete—that is, once all relevant results have been collected—we obtain "raw" statistical data (a large number of reports, individual data points, etc.). The collected data should be organized and appropriately described statistically. Of course, we can assess the differences between distributions or groups of results "by eye," but it is definitely a better idea to use certain measures that help us describe the characteristics of a distribution or group of results. A statistical description involves calculating certain numerical characteristics (called statistics) of the observed features. It serves as the starting point for drawing conclusions when working with a random sample. Statistics characterize a population in such a way that comparing different statistical populations can be reduced to comparing these statistics.

Access path to the analysis

Statistics/Basic statistics/ Descriptive statistics/OK/Variables/Advanced/Summary

If you want to analyze a selected parameter based on a grouping variable, select the Groups tab, choose Grouping Variable (categorical variable), and then select Collect results in a single table/OK in the current dialog box.

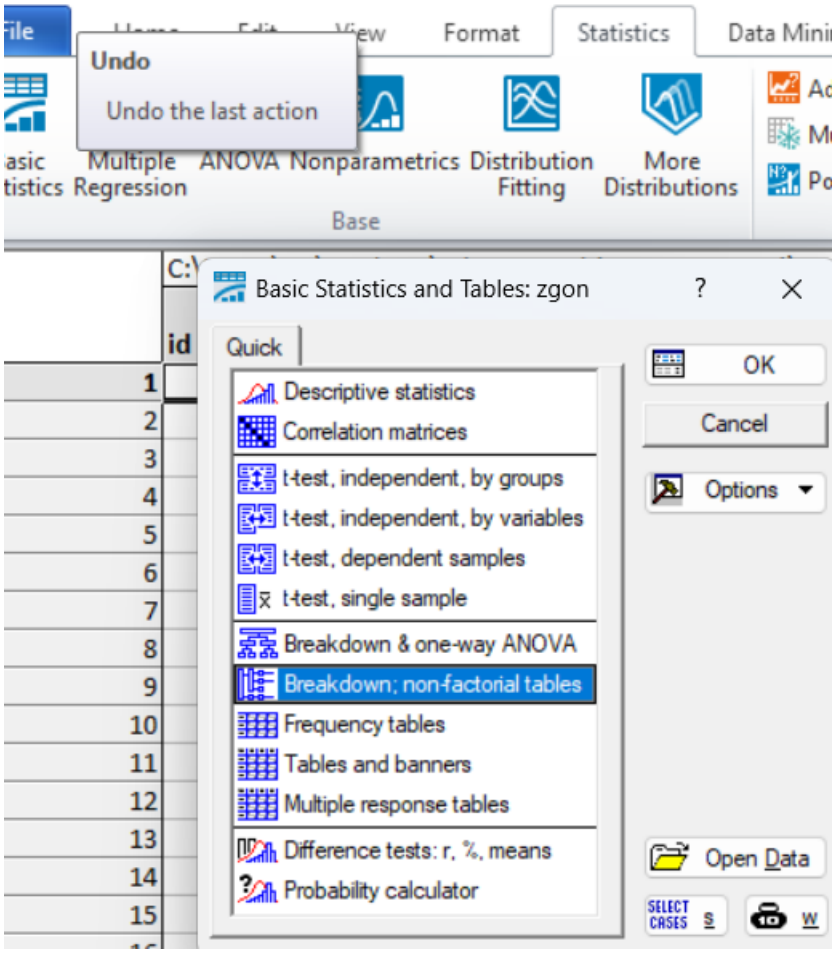


Variable	Descriptive Statistics									
	Valid N	Mean (M)	Median (Me)	Mode (Mo)	Frequency of Mode	Minimum (Min)	Maximum (Max)	Lower Quartile (Q1)	Upper Quartile (Q3)	Std.Dev. (SD)
	150	35,0893	35,50000	30,30000	4	15,40000	51,20000	30,50000	39,60000	6,138241

Below is a description of the collected data

The study included 150 women. The lowest observed value of the measured parameter “Body Fat” among the women in the study was 15.40, and the highest was 51.20 (these are the minimum and maximum values). The mean of the studied parameter “Body Fat” was 35.089, with half of the women in the study having a value of 35.50 for this parameter (this refers to the median value). The typical “middle” 50% of the women observed had body fat levels between 30.50 and 39.60 (these are the Q1 and Q3 values). The standard deviation of the “Body Fat” parameter value among the women studied was 6.138, and its relative variation in values was small. Typical observations of the “Body Fat” parameter can be considered those with values between 28.95 and 41.227 (the first value is the difference between the mean and the standard deviation, and the second value is the sum of the mean and the standard deviation).

If we are analyzing a continuous quantitative variable by a grouping variable, we can use simplified cross-tabs. Go to Statistics/ Basic Statistics/ Breakdown Table of Descriptive Statistics, enter the data, and select the appropriate statistical measures.



A pivot table in Excel is a powerful tool that summarizes, analyzes, and organizes large datasets, allowing you to uncover patterns, trends, and insights quickly.

A pivot table is a feature in Excel that reorganizes and summarizes data from raw datasets into a more digestible format. It allows you to sort, filter, group, and calculate data, making it easier to analyze totals, averages, counts, and other metrics across different categories or time periods. Pivot tables are especially useful for comparing sales, tracking performance, or analyzing large datasets without manually creating complex formulas. Before creating a pivot table, ensure your data is properly structured: Use a tabular format: each row represents a single record, and each column represents a field (e.g., date, product,

amount).Include unique headers in a single row for each column.Avoid empty rows, merged cells, or mixed data types in a column. Optionally, format your data as an Excel table (Ctrl+T) to simplify updates and dynamic ranges. Creating a Pivot Table

Click any cell within your dataset. Go to the Insert tab and select PivotTable. Choose the data range and decide whether to place the pivot table in a new worksheet or an existing one. Click OK to open the PivotTable Fields pane.

Building and Using a Pivot Table

Pivot tables have four main areas: Rows: categories you want to display as rows (e.g., products, regions). Columns: categories displayed as columns (optional). Values: numerical data to summarize (e.g., sum of sales, count of orders). Filters: fields to filter the data dynamically (e.g., by country or date),.

You can drag and drop fields into these areas to customize the table. Excel automatically summarizes data, usually by sum or count, but you can change the calculation type to average, max, min, or percentage.

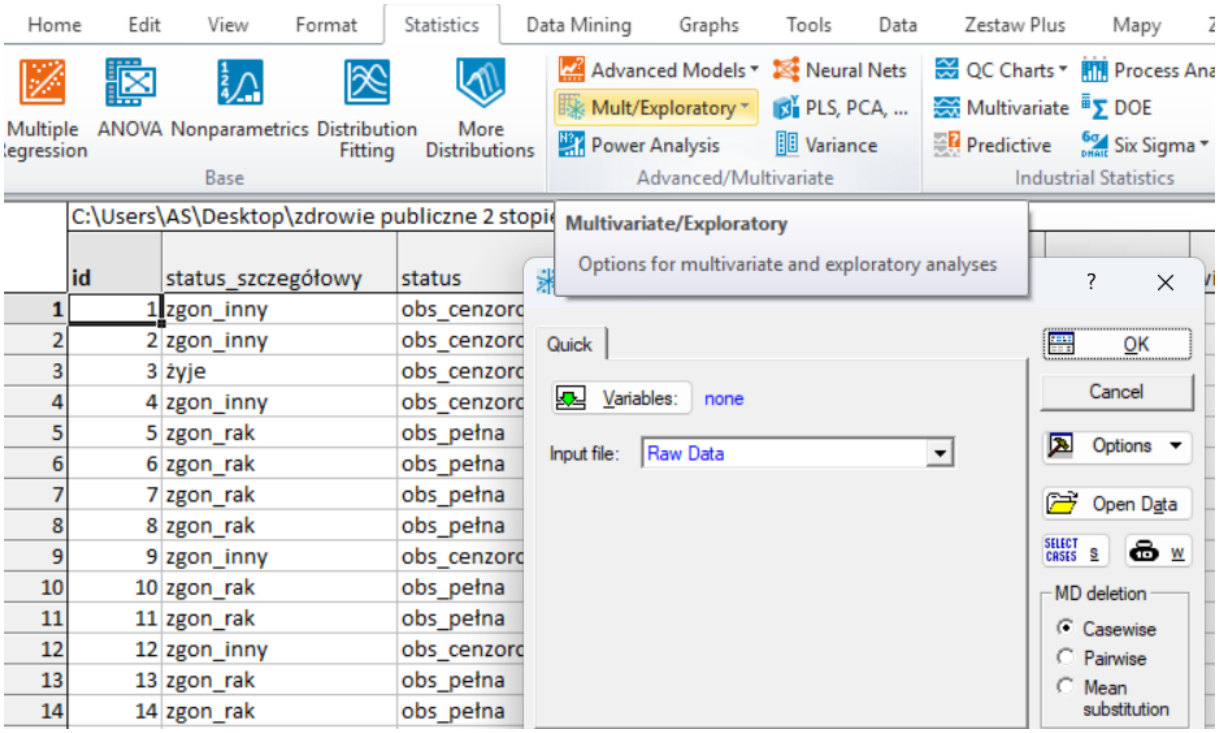
Factor Analysis is a statistical technique used in data analysis to identify hidden patterns or underlying relationships among a large set of variables. It helps reduce data complexity by grouping correlated variables into smaller sets called factors which represent shared characteristics or dimensions within the data.

Factor analysis serves several purposes and objectives in statistical analysis:

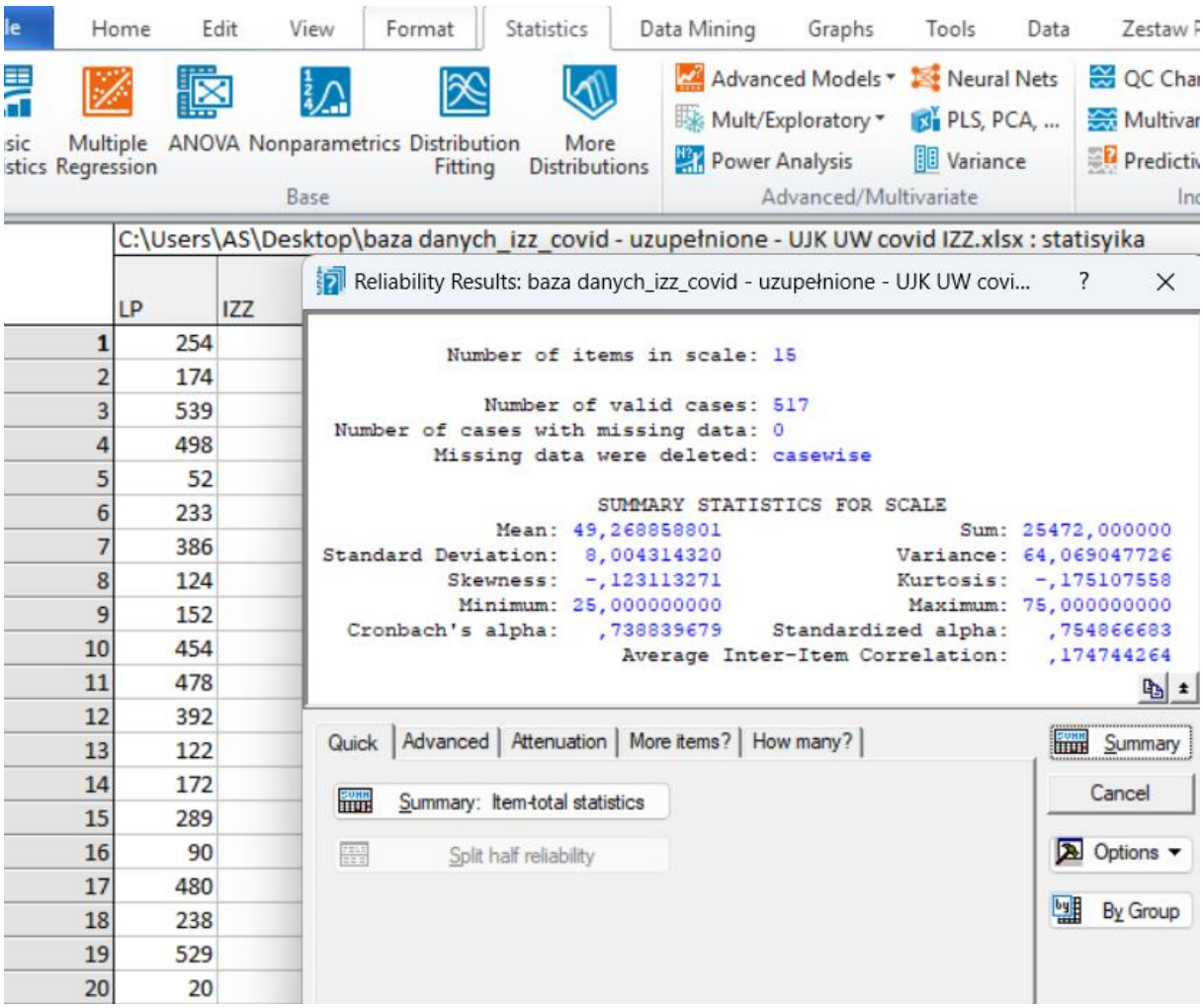
- Dimensionality Reduction: Simplifies datasets by grouping correlated variables into fewer factors making data easier to interpret.
- Identifying Latent Constructs: Reveals hidden variables like traits, attitudes that explain observed patterns.
- Data Summarization: Summarizes many variables into a concise set of factors while retaining key information.
- Hypothesis Testing: Evaluates whether the data support expected relationships among variables.
- Variable Selection: Highlights the most relevant variables for analysis or modelling.

- Improving Models: Reduces multicollinearity and enhances predictive model performance.

Using factor analysis, we can identify the variables that have the greatest influence on the variables under study.



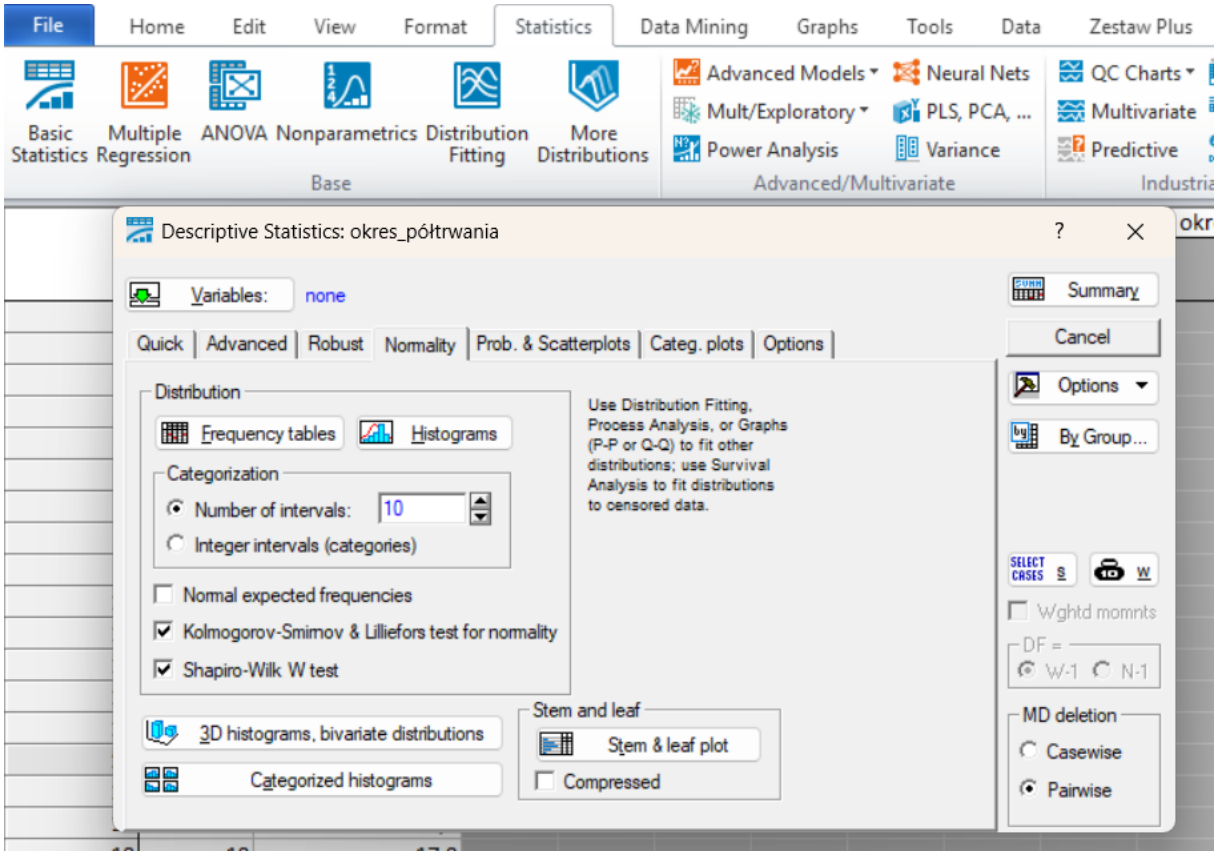
Cronbach's Alpha (α) measures the internal consistency of your scale. It assesses how well your items work together to measure the same underlying concept. For example, if you're measuring customer satisfaction, the coefficient indicates whether all your questions consistently tap into that same construct. The Alpha coefficient ranges from 0 to 1, where higher values indicate better internal consistency. It serves as a reliability check to determine whether your scale items measure the same construct or diverge into different dimensions.

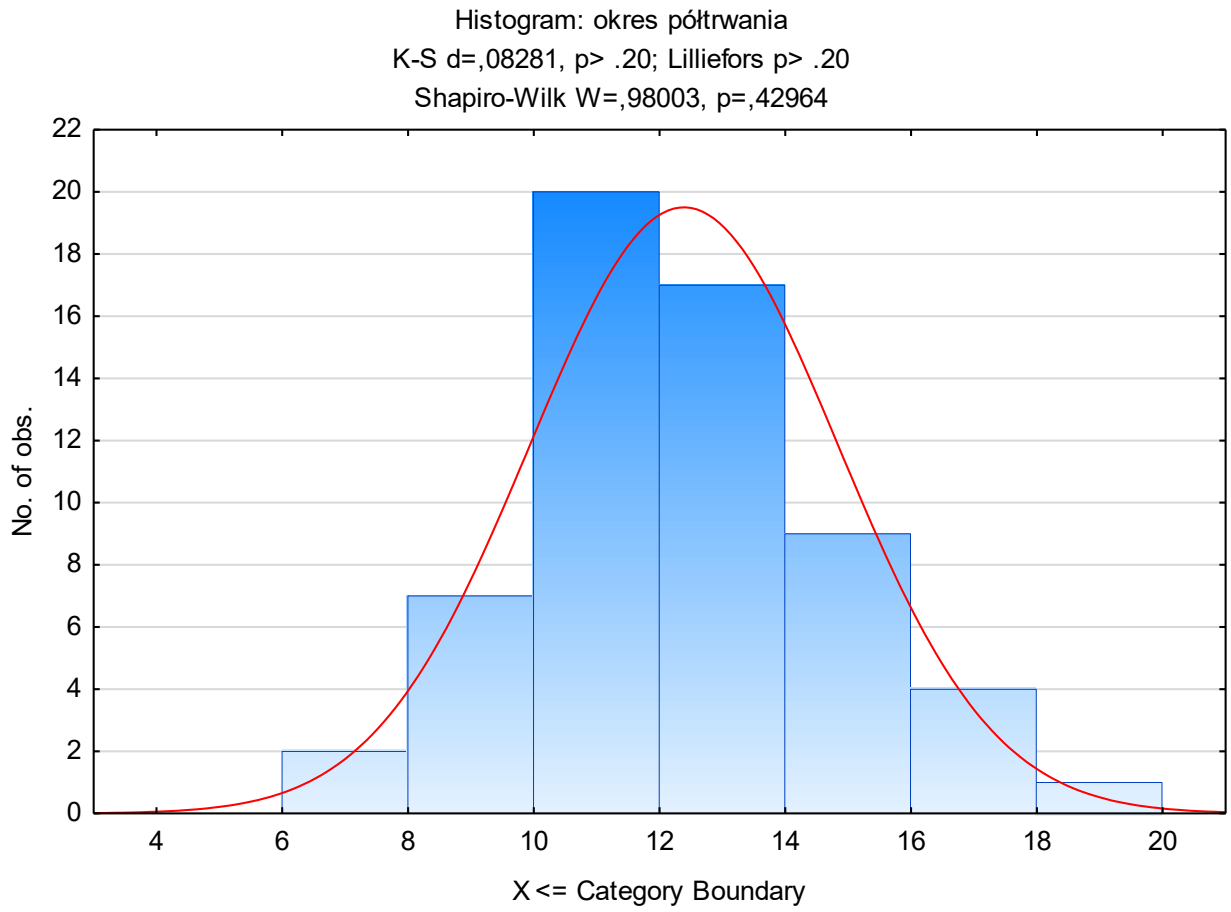

Home
Edit
View
Format
Statistics
Data Mining
Graphs
Tools
Data
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
More Distributions
Base
Advanced Models
Mult/Exploratory
Power Analysis
Neural Nets
PLS, PCA, ...
Variance
Advanced/Multivariate
C:\Users\AS\Desktop\baza danych_izz_covid - uzupełnione - UJK UW covid IZZ.xlsx : statisyika
LP
IZZ
1 254
2 174
3 539
4 498
5 52
6 233
7 386
8 124
9 152
10 454
11 478
12 392
13 122
14 172
15 289
16 90
17 480
18 238
19 529
20 20
Reliability Results: baza danych_izz_covid - uzupełnione - UJK UW covi...
Number of items in scale: 15
Number of valid cases: 517
Number of cases with missing data: 0
Missing data were deleted: casewise
SUMMARY STATISTICS FOR SCALE
Mean: 49,268858801 Sum: 25472,000000
Standard Deviation: 8,004314320 Variance: 64,069047726
Skewness: -,123113271 Kurtosis: -,175107558
Minimum: 25,000000000 Maximum: 75,000000000
Cronbach's alpha: ,738839679 Standardized alpha: ,754866683
Average Inter-Item Correlation: ,174744264
Quick
Advanced
Attenuation
More items?
How many?
Summary: Item-total statistics
Split half reliability
Summary
Cancel
Options
By Group

Point and interval estimation, hypothesis testing. Testing hypotheses about the parameters of one or two populations (Student's t-tests, nonparametric tests). Analysis of variance (ANOVA) (parametric and nonparametric tests).

NORMALITY OF THE DISTRIBUTION

We check the assumptions for the applicability of parametric tests—specifically, the normality of the distribution. Statistics / Basic Statistics / Descriptive Statistics In the Descriptive Statistics tool window that appears, select the Normality, and then choose: Shapiro–Wilk Test / Histograms. If the variable(s) whose normality is to be tested have not yet been specified, a variable selection window will appear. This window can also be opened at any time using the Variables button. From the generated histogram, we read the p-values of the performed tests.





These values form the basis for verifying competing statistical hypotheses:

H_o: the sample comes from a normal distribution (it is normally distributed)

H_1: the sample does not come from a normal distribution (it is not normally distributed) Let us summarize the results in the form of a table:

test	$p - value$	poz. istot. α	porównanie	decyzja
Kołmogorowa-Smirnowa	> 0,2	0,05	$p > \alpha$	H_0
Lillieforsa	> 0,2	0,05	$p > \alpha$	H_0
Shapiro-Wilka	0,42964	0,05	$p > \alpha$	H_0

And we note that, for each of the tests considered—at the adopted significance level $\alpha = 0.05$—there are no grounds to reject the null hypothesis H_0. Therefore, we conclude that the sample comes from a normal distribution."

STUDENT'S T-TEST, VERIFICATION OF THE EQUALITY OF MEANS – TWO INDEPENDENT SAMPLES

Comparing means in two populations based on independent samples.

General plan for solving the task:

1. Preliminary analysis: determining descriptive statistics in the studied groups (with particular emphasis on means and standard deviations), graphical analysis of distributions .
2. Checking the assumptions of the Student's t-test (normality of distributions, homogeneity of variances).
3. Conducting the Student's t-test if these assumptions are met (it is possible to use a modification of the Student's t-test if the assumption of normality is met but the assumption of homogeneity of variances is violated).
4. Applying another statistical test if the assumptions mentioned in point 2 are not met, in particular, if the assumption of normality of distribution is violated.

PRELIMINARY ANALYSIS – DETERMINING DESCRIPTIVE STATISTICS IN THE STUDIED GROUPS:

Statistics tab → *Basic Statistics* → *Breakdown & one-way ANOVA (dependent variable: serotonin, grouping: group)* → *Descriptive statistics (select statistics by checking the boxes)* → *Summary*

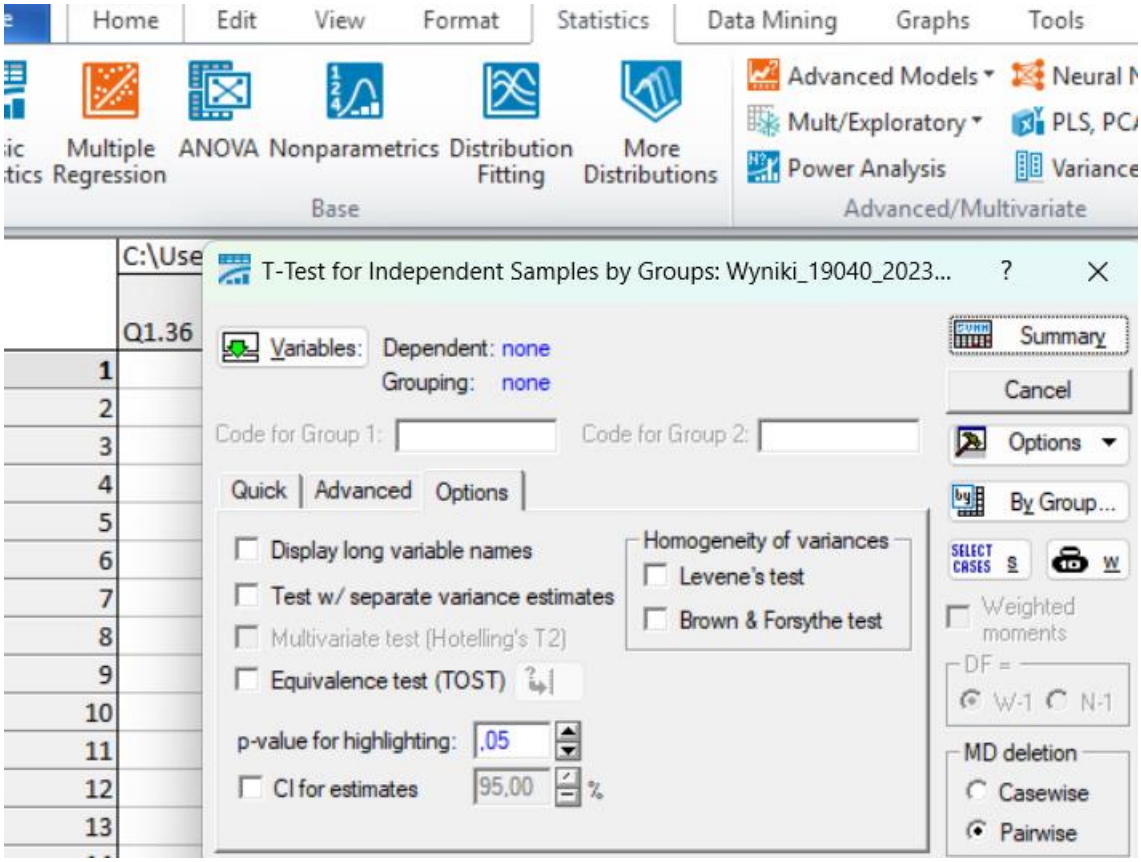
Before applying the Student's t-test, the following must be checked:

1. normality of distributions (this was verified above using the Shapiro-Wilk test),
2. homogeneity of variances.

We formulate the sets of hypotheses regarding the normality of distributions (separately for patients and healthy subjects) and conclusions regarding the fulfillment of the normality assumption: (We perform the Shapiro-Wilk test).

If the assumption of normality is met, we can compare the mean serotonin concentrations using the Student's t-test for independent samples (also known as unpaired).

In Statistica, the term "for independent samples" is used.



Verification of the hypothesis in Statistica: t-test for independent samples

(Statistics → Basic Statistics → t-test for independent samples (by groups))

ASSESSMENT OF THE HOMOGENEITY OF VARIANCES ASSUMPTION

Set of hypotheses using the notation from the statistical model formulated above:

$H_0: \sigma_1^2 = \sigma_2^2$

VS.

$H_1: \sigma_1^2 \neq \sigma_2^2$

STUDENT'S T-TEST FOR INDEPENDENT SAMPLES

Set of hypotheses using the introduced notation:

$H_0: \mu_1 = \mu_2$

VS.

$H_1: \mu_1 \neq \mu_2$

Since we have concluded that the variances are not equal, we should apply the t-test with separate variance estimates:

Statistics → Basic Statistics → t-test for independent samples (by groups): dependent variable = serotonin, grouping variable = group. Go to the Options tab and check "Test with separate variance estimates" → Summary button

FINAL CONCLUSION (was the null hypothesis rejected?):

At the significance level of 0.05, we reject the null hypothesis of the equality of means...

STUDENT'S T-TEST, VERIFICATION OF EQUALITY OF MEAN LEVELS – PAIRED SAMPLES

Comparing means in two populations based on paired samples.

The analyzed problem concerns the comparison of means in two related (paired) groups, as each patient had their blood pressure measured before starting the therapy and at a specific time after the therapy began. The subject of interest is the difference of "after" minus "before".

Therefore, we should use the Student's t-test (for paired samples) as the test of first choice. Consequently, it will be necessary to verify whether the assumption for the applicability of this test is met.

General plan for solving the task:

1. Preliminary analysis: determining descriptive statistics in the studied groups (with particular emphasis on means and standard deviations), graphical analysis of the distribution of differences.
2. Checking the assumption of the Student's t-test for paired samples (normality of the distribution of differences).
3. Conducting the Student's t-test if this assumption is met.
4. Applying another statistical test in case the assumption of the normality of the distribution of differences is violated.

Solution:

PRELIMINARY ANALYSIS – DETERMINING DESCRIPTIVE STATISTICS

The assessment of differences is the most important, but we can determine descriptive statistics not only for the "difference" variable, but also for the "before" and "after" variables. Evaluating the "before" variable can be useful, as the result of the conducted study will refer to...

Statistics →Basic Statistics →Descriptive Statistics: in the variable selection window, select variables.

The main subject of interest is the "difference" variable, and the normality of this specific variable must be assessed.

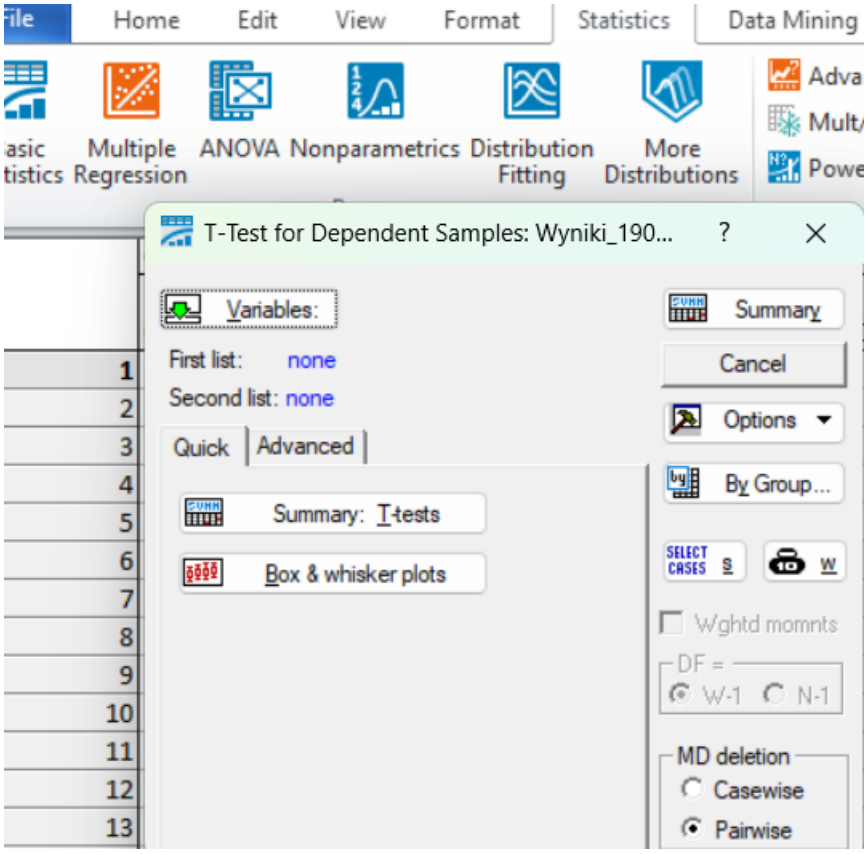
Hypothesis framework

$H_0: \mu_1 = \mu_2$

vs.

$H_1: \mu_1 \neq \mu_2$

Since the assumption of the normality of the distribution of differences is met, we perform the Student's t-test:



Statistics tab →Basic Statistics →t-test for dependent samples: Variables (first list: "after" variable, second list: "before" variable).

In this case, the p-value is found in the field named "p". Please note that since the hypothesis was one-tailed, the read value must be divided by two. Ultimately, the p-value for the system of hypotheses of the form:

$H_0: \mu_{\text{after}} - \mu_{\text{before}} = 0$

vs.

$H_1: \mu_{\text{after}} - \mu_{\text{before}} < 0$

FINAL CONCLUSION

We formulate one [appropriate] of the following two final conclusions:

We succeeded in rejecting the null hypothesis (and thus confirming the working/alternative hypothesis), the p-value...

NONPARAMETRIC TESTS FOR TWO RELATED SAMPLES

Wilcoxon Signed-Rank and Sign Tests – Two Related Samples

- Objective: To compare the distributions of two populations based on related (paired) samples.
- Note: In all statistical tests used below, we assume a significance level of $\alpha = 0.05$.

The analyzed problem concerns the comparison of distributions (of a continuous quantitative variable) in two related groups, as the level of indicator X was measured twice for each patient.

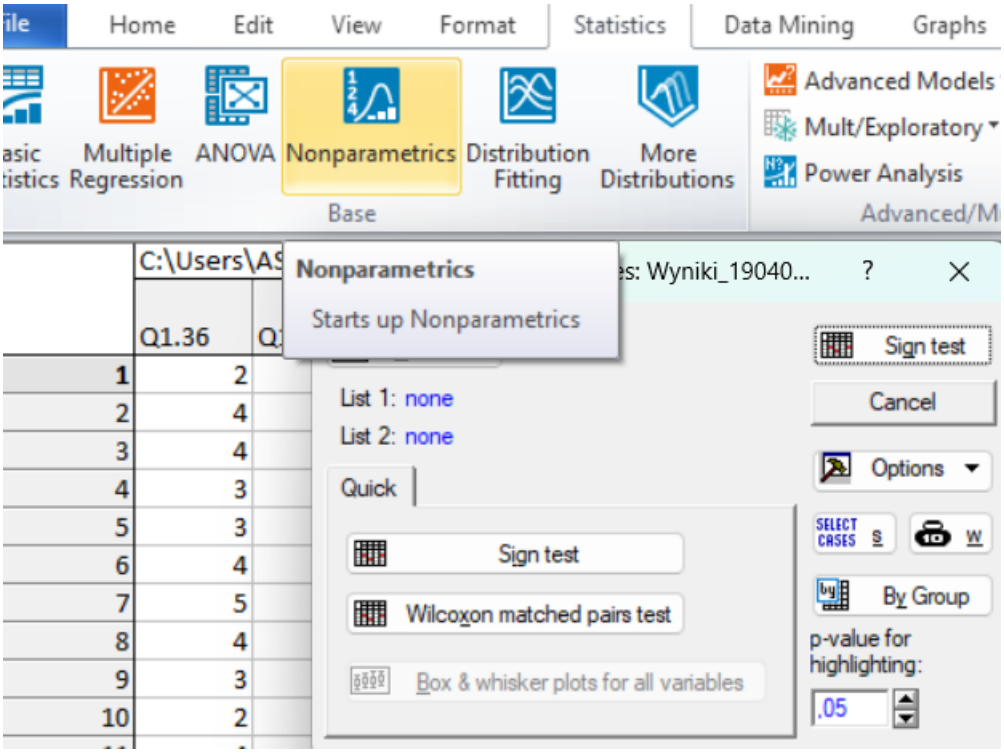
Therefore, our first-choice test should be the paired t-test. Consequently, it is necessary to check whether the assumptions for the applicability of this test are met. We proceed analogously to the aforementioned paired t-test.

General Plan for Solving the Task:

1. Preliminary analysis: Determining descriptive statistics for the studied groups and performing a graphical analysis of the distributions (boxplot, Q-Q plot, histograms).
2. Checking the assumptions of the paired t-test (normality of distributions, equality of variances).

3. Conducting the appropriate version of the paired t-test, if these assumptions are met.
4. Applying an alternative statistical test if the assumptions mentioned in step 2 are violated—specifically, the violation of the assumption of normality of distribution.

PRELIMINARY ANALYSIS – DETERMINING DESCRIPTIVE STATISTICS IN THE STUDY GROUPS:



The most crucial step is evaluating the differences, but descriptive statistics can be calculated not only for the "difference" variable but also for the "before" and "after" variables. A preliminary analysis must be performed (generating descriptive statistics for the analyzed variables):

Statistics → Basic Statistics/Tables → Descriptive Statistics (variables: *before*, *after*, and *difference*) → More tab (select statistics by checking them) → Summary

To expand the information about the samples themselves, boxplots can additionally be generated (although this is not strictly necessary from the perspective of the conducted statistical analysis).

Statistics → Nonparametrics → Comparing two dependent samples (variables) → Box & whisker plot for all variables: select variables [*before* and *after*]

GRAPHICAL ANALYSIS AND VERIFICATION OF NORMALITY

We create a histogram with the Shapiro-Wilk test value for the *difference* variable:

Statistics → Basic Statistics/Tables → Descriptive Statistics (variable: *difference*) → Normality tab → check: Shapiro-Wilk W test → Click: Histograms

- Conclusion: The assumption of normality for the distribution of differences is violated ($p\text{-value} = 0.0015$). Therefore, the paired t-test cannot be applied.
- Instead, we must use nonparametric tests.

In nonparametric tests, we do not compare mean levels, but rather either the entire distributions with each other or certain measures of location of these distributions.

In Statistica, we have two built-in tests of this type (concerning two related samples):

- A. Sign test
- B. Wilcoxon signed-rank test

Hypotheses for both tests:

$H_0: F_1(x) = F_2(x)$

vs.

$H_1: F_1(x) \neq F_2(x)$

Both of these tests can be found under:

Statistics → Nonparametrics → Comparing two dependent samples (variables) → click: Variables: select variables [*before* and *after*]

where we immediately input the variables and then, using the respective buttons, run:

- → Sign test
- → Wilcoxon matched pairs test

If it is possible to perform both tests, the Wilcoxon signed-rank test is more highly recommended because it has greater statistical power (it rejects a false null hypothesis more frequently) since it utilizes more information from the sample.

FINAL CONCLUSION (example of whether we succeeded in rejecting the null hypothesis):

In light of the conducted Sign and Wilcoxon signed-rank tests, at a significance level of 0.05, we do not reject the hypotheses regarding the similarities of the compared distributions.

Here is the professional translation of your text into English, formatted clearly for readability:

NONPARAMETRIC TESTS FOR TWO INDEPENDENT SAMPLES

Wald-Wolfowitz Runs Test, Kolmogorov-Smirnov Test, Mann-Whitney U Test – Two Independent Samples

Objective: To compare the distributions in two populations based on independent samples.

Note: In all statistical tests used below, we assume a significance level of 0.05.

The analyzed problem involves comparing the distributions of a continuous quantitative variable in two independent groups, as no individual can simultaneously belong to both the Asperger's syndrome group and the clinically healthy group. Therefore, our first-choice test should be the independent samples t-test. Consequently, it is necessary to check whether the assumptions for applying this test are met.

General Step-by-Step Plan:

1. Preliminary analysis: Calculating descriptive statistics for the studied groups (with a particular focus on means and standard deviations) and graphical analysis of the distributions.
2. Checking the assumptions of the independent t-test (normality of distributions, homogeneity of variances).
3. Conducting the appropriate version of the t-test, provided these assumptions are met.
4. Applying an alternative statistical test if the assumptions mentioned in step 2 are violated—particularly the assumption of normality.

PRELIMINARY ANALYSIS – CALCULATING DESCRIPTIVE STATISTICS FOR THE STUDIED GROUPS:

In Statistica: Statistics tab → Basic Statistics/Tables → Descriptive Statistics → 'More' tab (select statistics by checking the boxes) → 'By Group' button (select the grouping variable

[Asperger], options: check *"Categorized representation"*, uncheck *"Analyze without grouping"*) → OK → Summary.

Generating a box plot: * Graphs → 2D Graphs → Box Plots (Graph type: select dependent variable [time] and grouping variable [Asperger], middle point: *line*).

Alternatively (especially since—as it will turn out—the normality assumption is violated), we can also use:

- Statistics → Nonparametrics → Comparing two independent samples (groups) → Box-whisker plot for groups (select dependent variable [time] and grouping variable [Asperger]).

ASSESSING THE NORMALITY ASSUMPTION FOR TASK COMPLETION TIME DISTRIBUTIONS

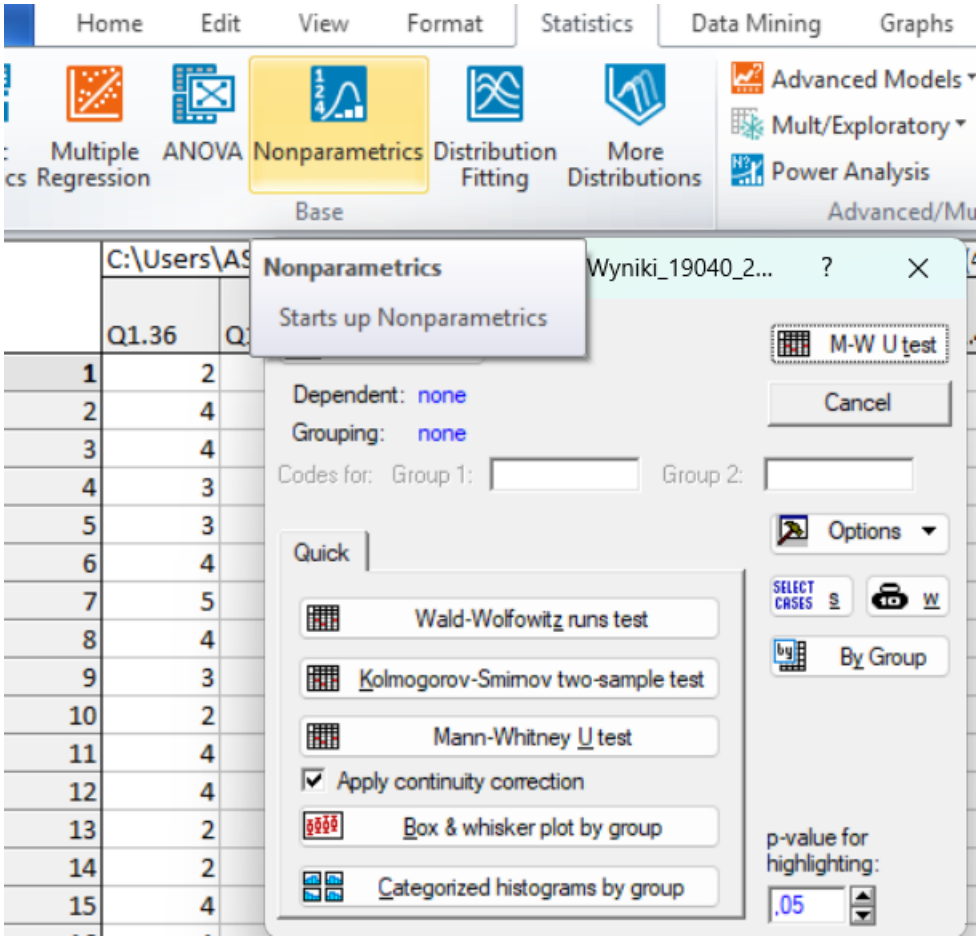
We perform the Shapiro-Wilk test (the instructions for this test can be found in the section on parametric t-tests).

Final conclusion regarding normality: The assumption of normal distribution is violated. Therefore, we cannot use the independent samples t-test. Instead, we must turn to nonparametric tests.

Nonparametric tests do not compare mean levels; instead, they compare either the entire distributions or certain measures of location of these distributions.

In Statistica, three such tests are available:

- A. Wald-Wolfowitz runs test
- B. Kolmogorov-Smirnov test
- C. Mann-Whitney U test



All of these allow us to verify the consistency of the distributions from which the samples were drawn. These tests do not require the assumption of normality to be met.

Hypothesis setup:

$$\mathrm{H}_0 : F_1 = F_2$$

vs.

$$\mathrm{H}_1 : F_1 \neq F_2$$

All of the mentioned tests can be found in:

- Statistics → Nonparametrics → Comparing two independent samples (groups) → 'Variables' button (dependent variable: [time], grouping variable: [Asperger]).

Once the variables are selected, you can run the tests individually using the respective buttons:

- Wald-Wolfowitz runs test
- Kolmogorov-Smirnov two-sample test
- Mann-Whitney U test

The Mann-Whitney U test is considered the most powerful nonparametric alternative to the independent samples t-test. Therefore, in the event of conflicting results among the three tests, we should primarily rely on the findings of the Mann-Whitney U test, while additionally reporting that another test yields the opposite result.

FINAL CONCLUSION (Did we manage to reject the null hypothesis?):

Based on the Mann-Whitney U test, at a significance level of 0.05, we do not reject the null hypothesis that the task completion time distributions are identical in the population of individuals with Asperger's syndrome and the clinically healthy population. Nevertheless, the Kolmogorov-Smirnov test yields the opposite result.

One-Way Analysis of Variance (One-Way ANOVA)

Objective: To compare means across more than two populations based on independent samples.

The problem under analysis involves comparing the means of four independent groups, as each individual was treated with only one of the drugs: A, B, C, or D. Therefore, one-way analysis of variance (ANOVA) should be selected as the statistical technique of first choice. Consequently, it is necessary to verify whether the assumptions for applying ANOVA are met.

General Plan for Solving the Problem:

1. Preliminary analysis: determining descriptive statistics for the analyzed groups (with a particular focus on means and standard deviations) and graphical analysis of the distributions.
2. Adopting a statistical model for the analyzed phenomenon and verifying the assumptions of the ANOVA test (normality of distributions, homogeneity of variances).
3. Conducting the analysis of variance, provided these assumptions are met.
4. Applying an alternative statistical test if the assumptions mentioned in point 2 are not met.

Solution:

PRELIMINARY ANALYSIS – DETERMINING DESCRIPTIVE MEASURES IN THE ANALYZED GROUPS

Statistica path: > Statistics ➔ Basic Statistics/Tables ➔ Breakdown & one-way ANOVA (dependent variable: *Enzym* [Enzyme], grouping variable: *Lek* [Drug]) ➔ Descriptive statistics (select specific statistics by checking the boxes) ➔ Summary

Box Plot:

Statistica path: > Graphs ➔ 2D Graphs ➔ Box Plots [dependent: *Enzym* [Enzyme], grouping: *Lek* [Drug]]

We assume that:

- The response variable (dependent variable) is the enzyme level (quantitative variable).
- The experimental factor (grouping variable) is the drug (qualitative/categorical variable with 4 levels).

MAIN TEST

Hypothesis verification tool: One-Way Analysis of Variance

(In Statistica: Statistics ➔ ANOVA ➔ One-Way ANOVA)

Hypothesis setup using the established notation:

$H_0: \mu_A = \mu_B = \mu_C = \mu_D$

vs.

H_1:not all means $\mu_A, \mu_B, \mu_C, \mu_D$ are equal

Before applying the ANOVA test, the following assumptions must be verified:

1. Normality of distributions,
2. Homogeneity (equality) of variances.

VERIFYING HOMOGENEITY OF VARIANCE

To check the equality of variances (also referred to as homogeneity of variance or homoscedasticity), various tests can be used in Statistica, including Levene's test, Cochran's C, Hartley's, and Bartlett's tests. For each of these tests, the null hypothesis states that the variances of the enzyme level (E) distribution are equal, while the alternative hypothesis is the negation of the null hypothesis:

$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$

vs.

H_1:not all variances $\sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2$ are equal

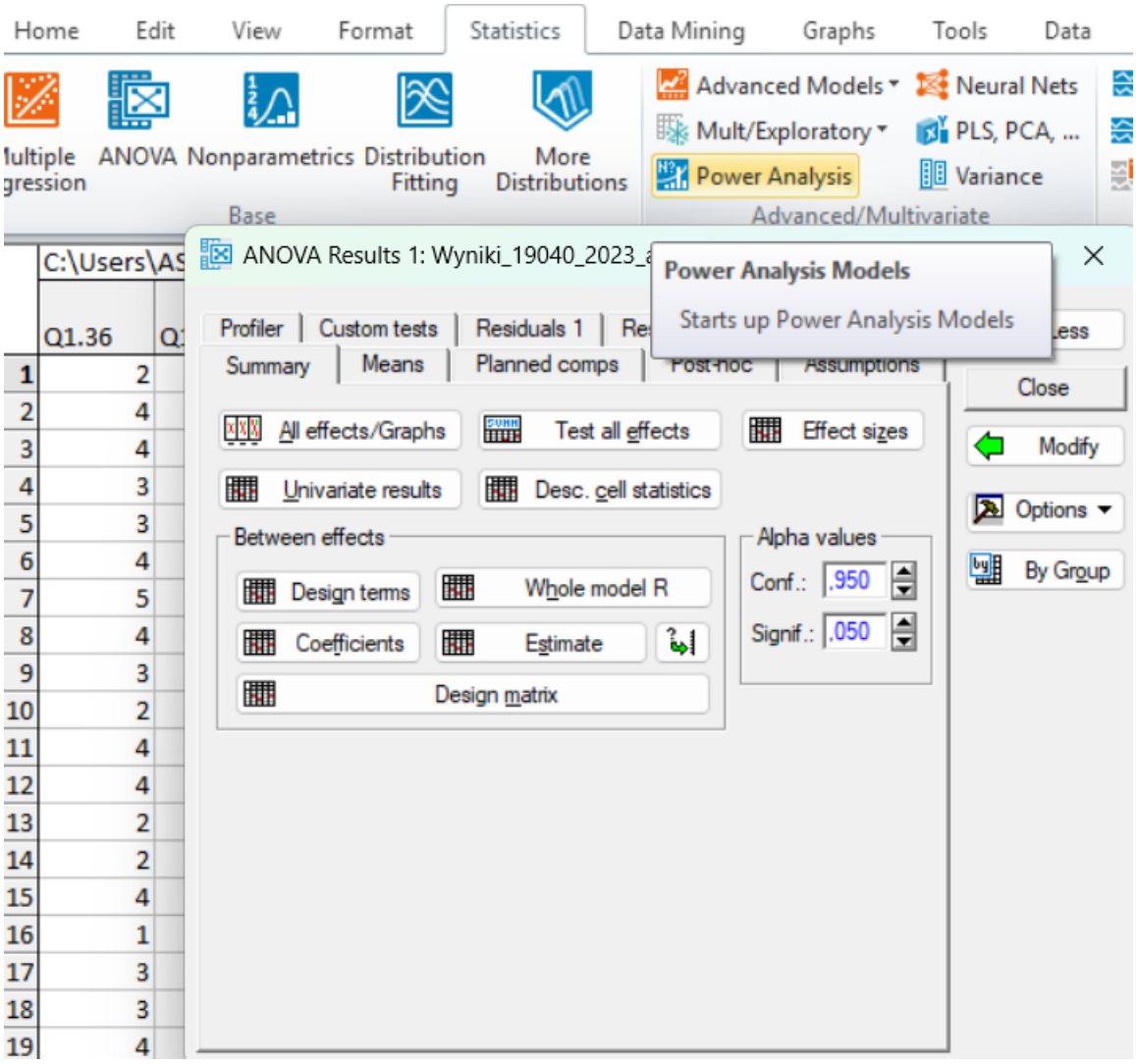
Statistica path: > Statistics ➔ ANOVA ➔ One-Way ANOVA ➔ Dependent variables: [Enzyme], qualitative factor: [Drug] ➔ All effects

We find the values in the row corresponding to the factor name, i.e., in the row labeled as "Lek" (Drug) (this is the value highlighted on a gray background in the table above; in practice, this value is equal to 0, or more precisely, it is 0.00001).

Since the values are < 0.05, we reject the null hypothesis and accept the alternative hypothesis, which states that not all means are equal.

We can view the plots of the means:

In the ANOVA results, go to the "Podstawowe" (Quick/Basic) tab -> click the "Średnie/Wykresy" (Means/Plots) button.



POST-HOC TESTS

If the null hypothesis in the analysis of variance was successfully rejected, post-hoc tests should now be conducted to determine:

1. between which drugs there are statistically significant differences between the means,
2. how to identify homogeneous groups (i.e., groups within which there are no significant differences between the means).

Post-hoc tests are sometimes referred to as multiple comparison tests.

There are many post-hoc tests. These tests differ in their so-called power, which in this context is understood as the ability to reject a false null hypothesis of equal means.

A given test is considered conservative if it rejects a false null hypothesis less frequently than another test.

In post-hoc tests, either all pairs of means are compared, or a pre-selected mean is compared with the remaining means.

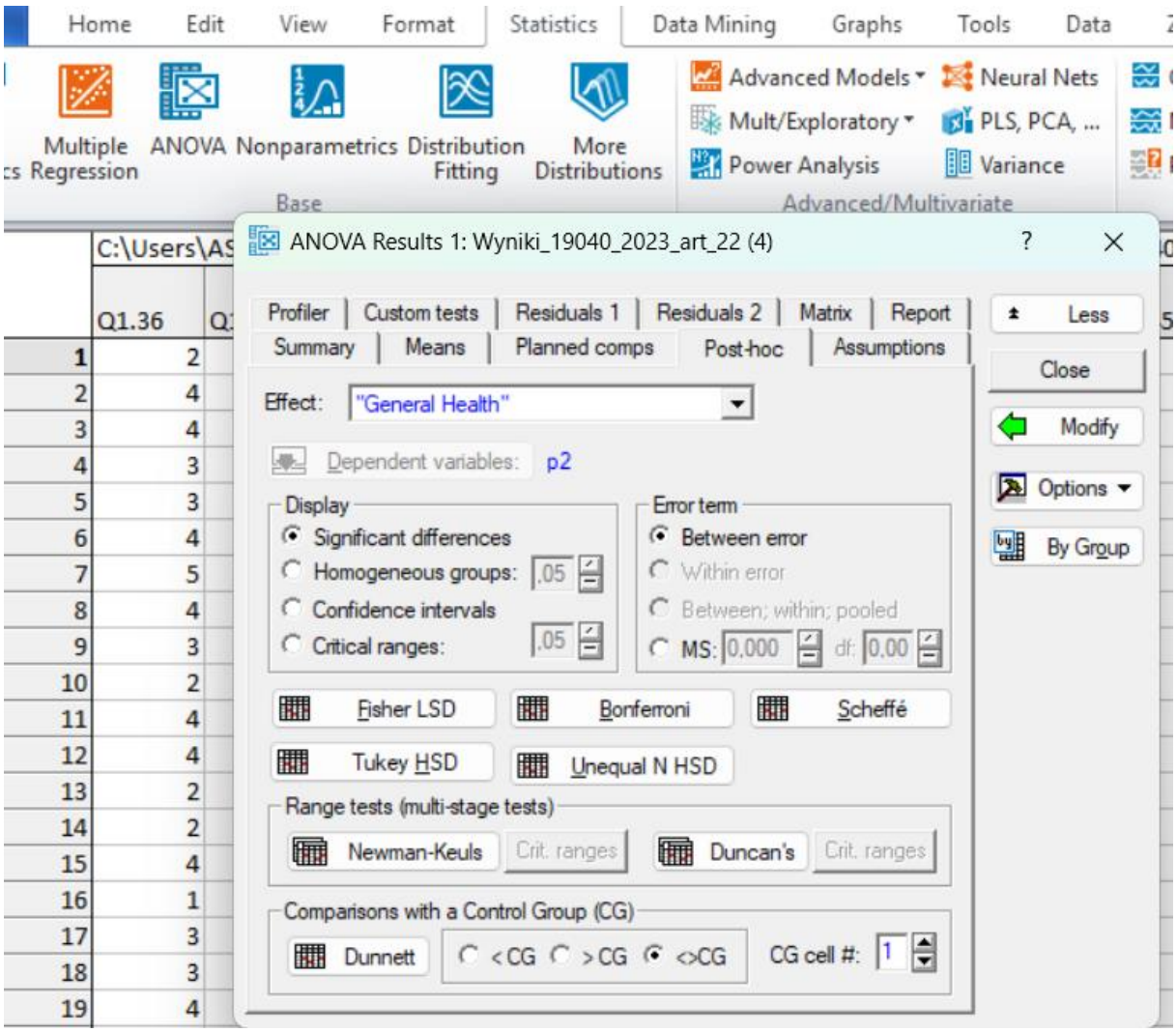
Post-hoc tests comparing all pairs of means, ordered from most to least conservative:

1. Scheffé's test
2. Tukey's test (the most commonly used in practice)
3. Newman-Keuls test
4. Duncan's test
5. Fisher's LSD (Least Significant Difference) test

Scheffé's test ensures the assumed family-wise significance level for all tested pairs.

Tukey's test comes in two variants:

- for balanced designs (equal number of replications for factor levels) – the Honestly Significant Difference test (in Statistica: Tukey's HSD test);
- for unequal sample sizes (in Statistica: Tukey's test for unequal N).



We will use Tukey's test to evaluate which pairs of means significantly differ from each other.

In Statistica: In the ANOVA results, open the "post-hoc" tab -> in the "Pokaż" (Show) section, select "Istotne różnice" (Significant differences), and then click the "Test Tukey'a (HSD)" (Tukey HSD) button.

ID	Test HSD Tukeya;				
	DRUG	{1} 85,247	{2} 79,527	{3} 73,573	{4} 79,613
1	A		0,003869	0,000159	0,004535
2	B	0,003869		0,002505	0,999949
3	C	0,000159	0,002505		0,002133
4	D	0,004535	0,999949	0,002133	

NONPARAMETRIC TESTS FOR MULTIPLE INDEPENDENT SAMPLES

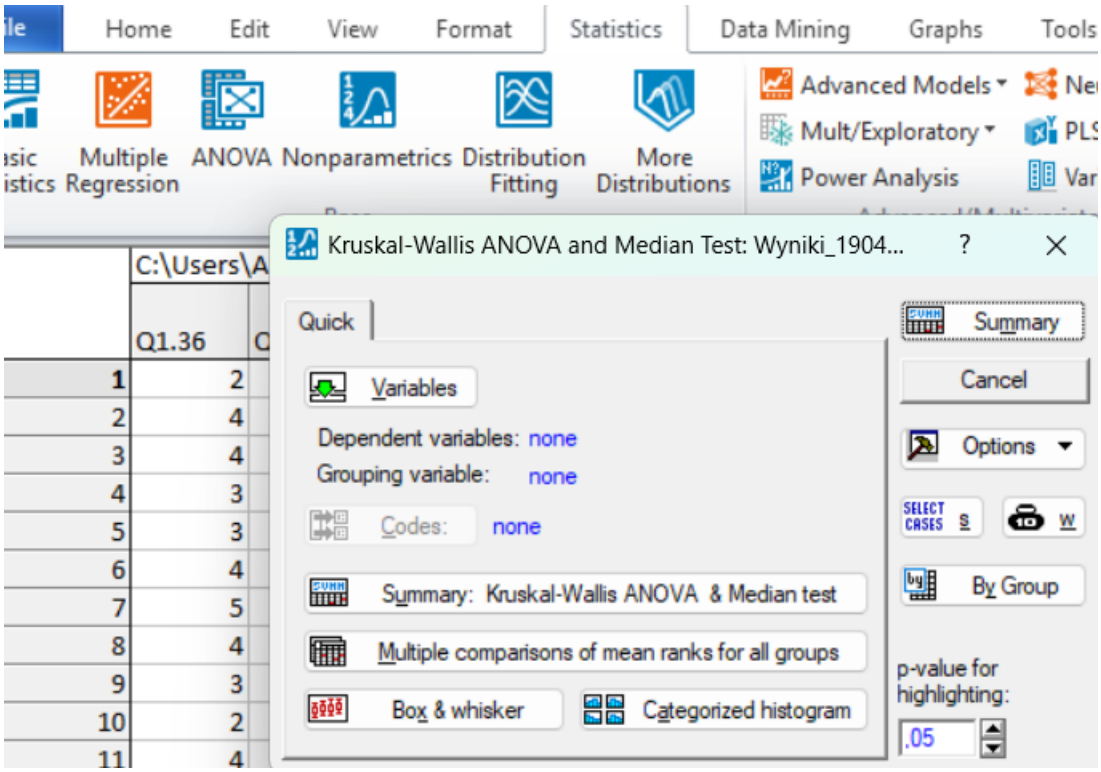
In nonparametric tests, instead of comparing mean levels, we compare either the entire distributions with each other or certain measures of location of these distributions.

Statistica provides two tests of this type:

- **A.** the Kruskal-Wallis test,
- **B.** the median test.

The first test is used to verify the consistency of distributions across individual groups, while the second test is used to verify the equality of medians across individual groups.

These tests only require the assumption of continuity of the distributions of the analyzed variable.



Both of these tests can be found under: Statistics / Nonparametric / Comparison of Multiple Independent Samples (Groups) / button: Variables: Select the variables [dependent: parameter, grouping: group] where you enter the variables immediately, and then click the button: / Summary

Testing for Independence of Variables (chi-square test). Statistical Inference in Correlation and Regression Analysis (linear).

CHI-SQUARE (χ^2) TEST OF INDEPENDENCE

General framework of hypotheses:

Null hypothesis (H0): Characteristics A and B are independent

vs.

Alternative hypothesis (H1): Characteristics A and B are dependent

The χ^2 independence test is based on a comparison of observed counts with expected counts (expected counts are calculated under the assumption that the null hypothesis is true).

When we accept H_o—if the p-value of the test is greater than the chosen significance level α ($p > \alpha$)

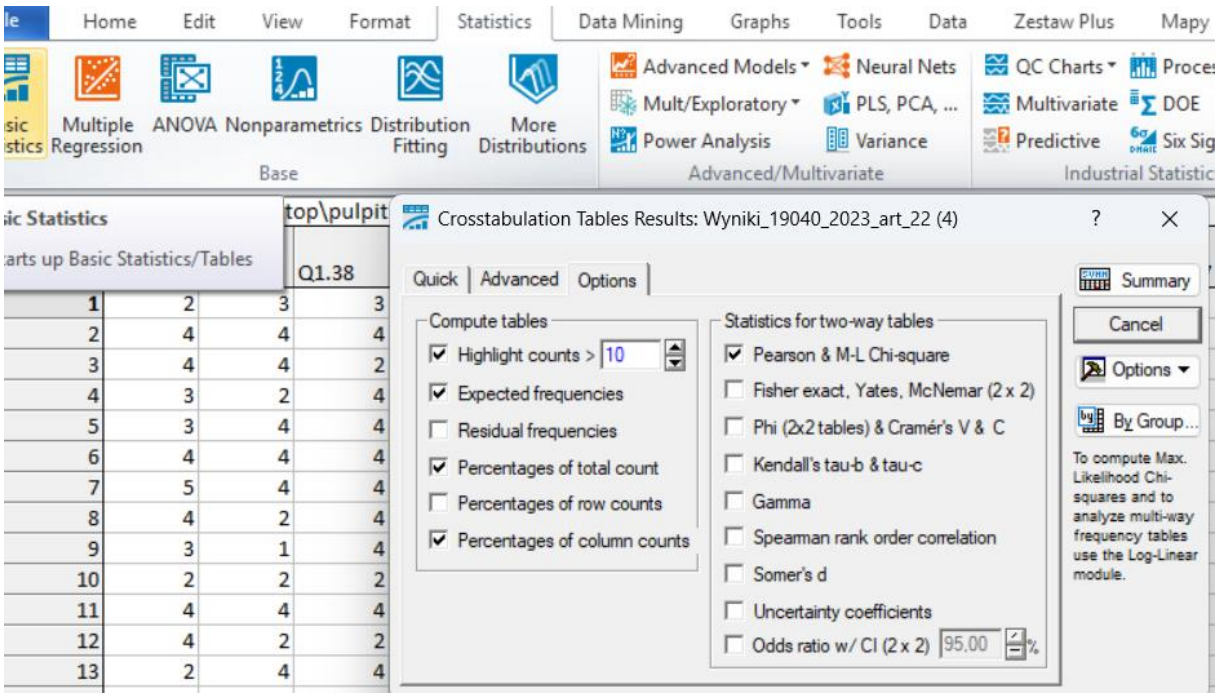
When we reject H0 in favor of H1—if the p-value of the test is less than the chosen significance level α ($p < \alpha$)

Since we are analyzing two categorical variables, we should use the chi-square test of independence for this purpose

Path to the tool:

Statistics/Basic Statistics/Multivariate Tables/OK/Summary/ (List 1: physical activity, List 2: gender/OK/OK/Options (tab at the top)/Add to selection: expected frequencies, Pearson's chi-square, and NW/More/Exact two-way tables.

As a result, three tables will appear in the workbook, which constitute the solution to the problem: a table showing the distribution of the variables under study, a table with expected frequencies, and the value of the χ^2 test statistic—all of these elements are visible in the workbook (the access path is described below in the description). Keep in mind that before reading the value of the chi-square statistic, you should check whether the expected values are at least 5.



Summary Frequency Table (Marked cells have counts > 10 (Marginal summaries are not marked)			
General Health	SEX 1	SEX 2	Row Totals
Problematic	33	46	79
Excellent	34	45	79
Sufficient	46	72	118
Inadequate	13	23	36
All Grps	126	186	312

	2-Way Summary Table: Expected Frequencies Marked cells have counts > 10		
General Health	SEX 1	SEX 2	Row Totals
Problematic	31,9038	47,0962	79,0000
Excellent	31,9038	47,0962	79,0000
Sufficient	47,6538	70,3462	118,0000
Inadequate	14,5385	21,4615	36,0000
Totals	126,0000	186,0000	312,0000

	Statistics: General Health(4) x SEX(2)		
Statistic	Chi-square	df	p
Pearson Chi-square	,6635557	df=3	p=,88174
M-L Chi-square	,6657238	df=3	p=,88123

We use the Pearson chi-square test if the expected frequencies are at least 5 or greater than 5. We use the non-parametric chi-square test if the expected frequencies are less than 5.

Assumptions: The expected counts must be at least 5; in our case, they are less than 5. Test value p = 0.88124, chi-square test (Pearsona)

Final conclusion: (Can the null hypothesis be rejected?) At a significance level of $\alpha = 0.05$, we have no basis for rejecting H_o regarding the independence of the variables under study ($p = 0.881$), and we therefore accept that the characteristics under study are independent

Or

At a significance level of $\alpha = 0.05$, we have grounds to reject H_o regarding the independence of the variables under study in favor of hypothesis H_1 ($p = 0.002$), and we therefore conclude

CORRELATION AND REGRESSION

Objective:

To measure correlation in a sample. To estimate correlation in a population. To test the significance of correlation in a population.

Note:

In all statistical tests used below, we assume a significance level of $\alpha = 0.05$.

A correlation plot is used to visually assess the type, direction, and strength of the correlation between a pair of variables.

LINEAR CORRELATION

Based on the conclusions drawn from a visual assessment of the correlation plot, we begin by calculating Pearson's linear correlation coefficient:

Statistics Basic Statistics/ Correlation Matrices/button: Two Lists of Variables [set: First List – first variable, Second List – second variable + OK] button: Correlations

Its value is 0.960452 ≅ 0.960. This means that there is indeed a strong, positive, and linear correlation. In this sense, we treat the calculated correlation coefficient as a sample estimate (since the analysis concerns only what is happening within the studied group of patients).

If we treat the sample as a random sample drawn from the population it represents, then the calculated sample correlation coefficient (let's denote it as r, or $\hat{\rho}$ [the Greek letter "rho" with a diaeresis]) is an estimate (the value of the estimator) of the unknown linear correlation in the population (denoted by ρ [the Greek letter "rho"]).In this case, a significance test can be performed on the linear correlation coefficient in the population.

Statistical model

We are interested in the linear correlation coefficient ρ in the population.

To describe the relationship between a pair of variables in the population, we use a two-dimensional random variable (X, Y) with a joint two-dimensional normal distribution, in which the unknown parameter is the correlation coefficient ρ, where $-1 \leq \rho \leq 1$.

We have a single sample of observation pairs: 〖(X)〗_1,Y_1), (X_2,Y_2), … , (X_n,Y_n), of size n, drawn from the same distribution as the random variable (X,Y).

Hypothesis framework:

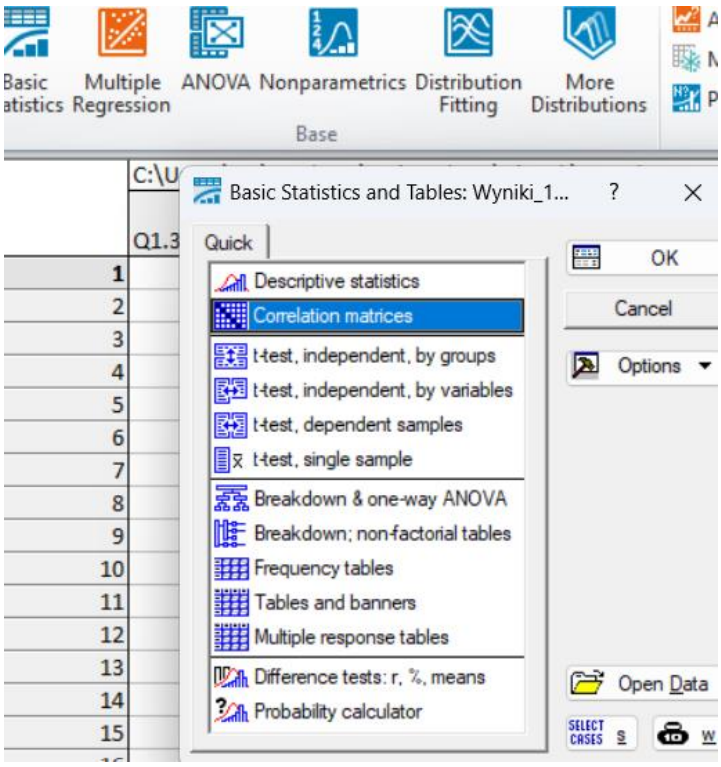
H_0 : the correlation of interest in the population is not positive

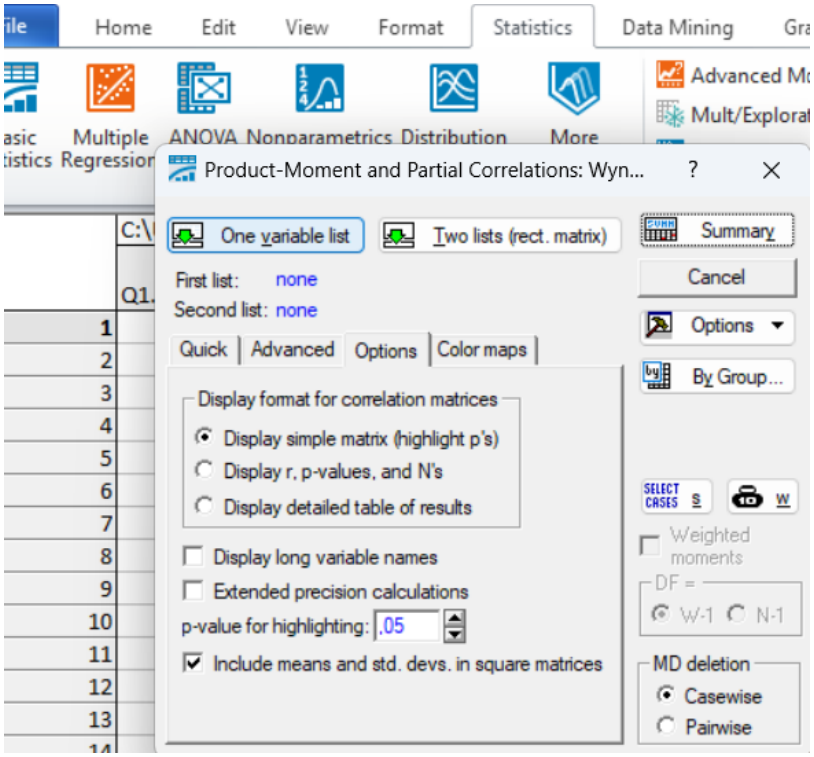
vs.

H_1 : the correlation of interest in the population is positive

A necessary assumption for the test that must be verified is that the distribution from which the sample is drawn is normally distributed in two dimensions. Statistica does not include any tools that would allow this to be done. The p-values (0.3824 and 0.2362) from the Shapiro-Wilk test obtained for both samples are greater than the significance level of $\alpha = 0.05$. We therefore conclude that both (one-dimensional) samples come from normal distributions.

We assume that the sample of observation pairs comes from a two-dimensional normal distribution. We therefore return to the basic significance test for Pearson's linear correlation coefficient. In Statistica, we will perform this test while calculating the correlation coefficient from the sample, that is, by following the steps described earlier: Statistics/Basic Statistics/Correlation Matrices/button: Two Lists of Variables [set: First List – first variable, Second List – second variable + OK]. Options [select: Display p-value and N] /button: Summary





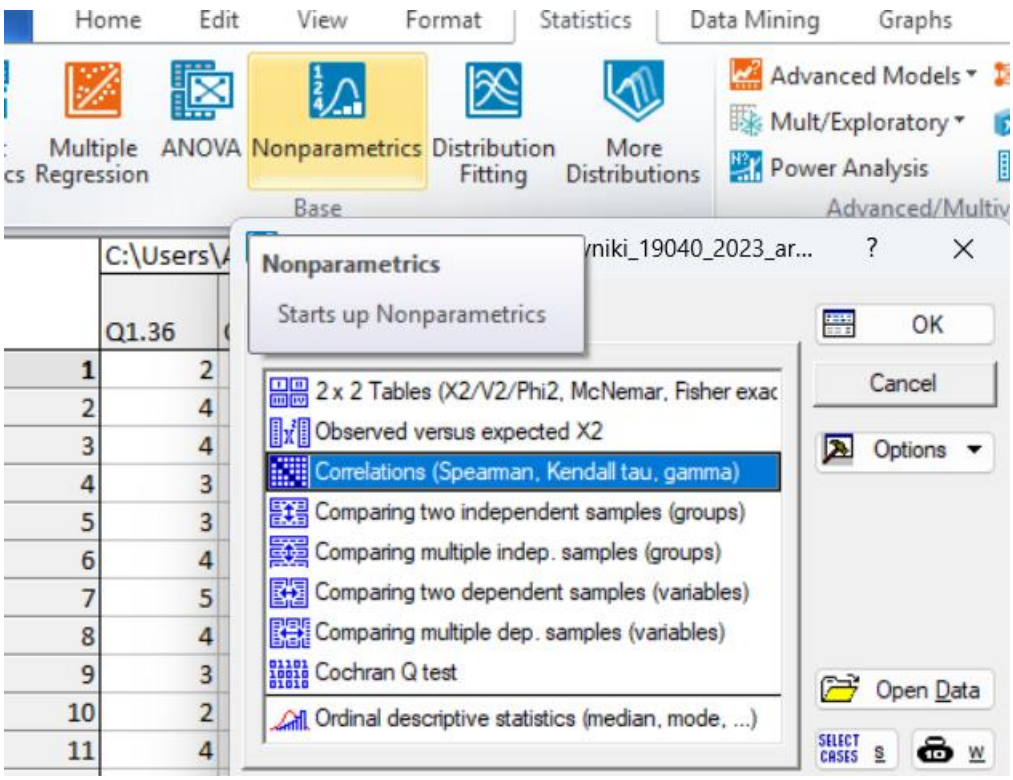
In addition to displaying the correlation coefficient from the sample again, we are also given the p-value ≅ 0.000 for the two-tailed significance test of the correlation coefficient in the population.

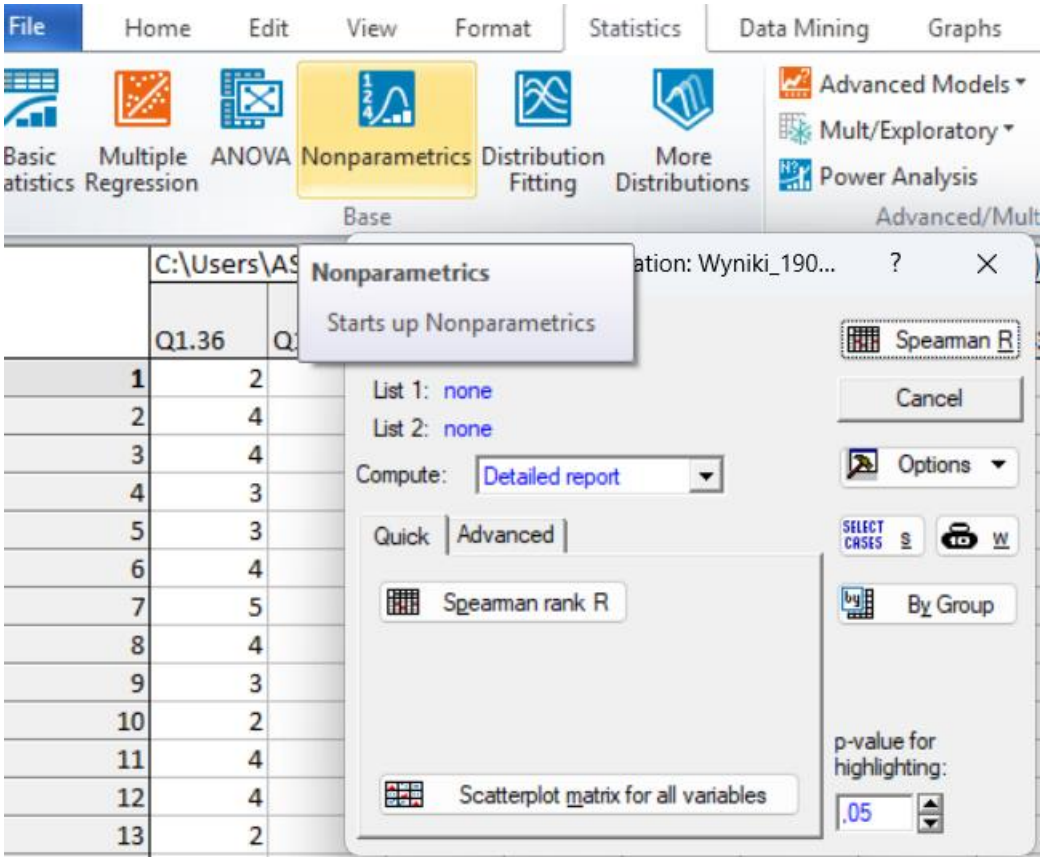
The direction of the inequality in hypothesis H_1 (where $\rho > 0$) is consistent with the sign of the sample correlation coefficient: $\hat{\rho} = 0.9605 > 0$; therefore, the p-value for the one-tailed test is $p/2 \cong 0.000$.

NONLINEAR CORRELATION

Generally, we examine a type of correlation other than linear in two cases: for quantitative variables, if the shape of the correlation differs from a linear one; for qualitative variables expressed on an ordinal scale (since it is not possible to consider a linear correlation in such cases). It is also worth using nonlinear correlation coefficients when the resolution of the analyzed problem is based on a significance test of the linear correlation coefficient, and that test encounters a problem of non-fulfillment or lack of decisiveness regarding the assumption of two-dimensional normality of the sample. In Statistica, we have tools available for calculating Sperman's rank correlation coefficient and Kendall's τ coefficients (in two variants: "gamma"—Kendall's τ without accounting for repeated values in the sample, and "Kendall's tau"—with such repetitions taken into account).

Statistics / Nonparametric / Correlations (Sperman, Kendall's tau, gamma) / button: Variables [select both], confirm: OK / tab: More / buttons [all in sequence]: Spearman's R; Gamma; Kendall's tau; Matrix of overall variable dispersion.


Home
Edit
View
Format
Statistics
Data Mining
Graphs
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
More Distributions
Advanced Models
Mult/Exploratory
Power Analysis
Base
Advanced/Multiv
C:\Users\A
Q1.36
Nonparametrics
Starts up Nonparametrics
OK
Cancel
Options
2 x 2 Tables (X2/V2/Phi2, McNemar, Fisher exac
Observed versus expected X2
Correlations (Spearman, Kendall tau, gamma)
Comparing two independent samples (groups)
Comparing multiple indep. samples (groups)
Comparing two dependent samples (variables)
Comparing multiple dep. samples (variables)
Cochran Q test
Ordinal descriptive statistics (median, mode, ...)
Open Data


File
Home
Edit
View
Format
Statistics
Data Mining
Graphs
Basic Statistics
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
More Distributions
Advanced Models
Mult/Exploratory
Power Analysis
Base
C:\Users\AS
Q1.36
Nonparametrics
Starts up Nonparametrics
Spearman R
List 1: none
List 2: none
Cancel
Compute: Detailed report
Options
Quick
Advanced
Spearman rank R
By Group
p-value for highlighting:
.05
Scatterplot matrix for all variables

We specify the following regression model:

$\hat{y} = \beta_0 + \beta_1 \cdot x$

$y = \hat{y} + \varepsilon = \beta_0 + \beta_1 \cdot x + \varepsilon$

Where y is the dependent variable, x is the independent variable, ε is the random component, and β_0 and β_1 are the parameters of this model.

Statistics / Multiple Regression / button: Variables [we set: Dependent Variables – first variable, List of Independent Variables – second variable] / OK / OK

We can view the detailed results on the "More" tab, under the "Summary" button: Regression Results:

Note that the regression obtained from the sample is of the form:

(Second variable)̂ = -125.575 + 1.347 × first variable

or, symbolically, in one of two forms:

ŷ_i = -125.575 + 1.347 × x_i

y_i = ŷ_i + e_i = -125.575 + 1.347·x_i + e_i

where ŷ_i denotes the predicted value (by the estimated linear regression) of the dependent variable y () for the i-th individual, for whom a value of the second variable equal to y_i was observed, while the value of the independent variable x was observed as x_i.

Furthermore, e_i denote the deviations of the observed values of the variable y from their predicted values, i.e., the so-called residuals of the linear regression model (e_i = y_i − ŷ_i).

The coefficient of determination R^2 (also known as the goodness-of-fit coefficient; referred to in Statistica as "Multiple R^2") is 0.922468923 ≅ 0.922. This means that 92.2% of the variation in the second variable is explained by the estimated linear regression.

Significance Tests for Linear Regression Coefficients

We formulate a set of hypotheses for each regression parameter in the population:

H_0: the parameter is equal to zero

vs.

H_1: the parameter is not equal to zero

Symbolic representation of the hypothesis for (free term): $H_0 : \beta_0 = 0$

vs.

$H_1 : \beta_0 \neq 0$

Symbolic representation of the hypothesis for (the coefficient of the variable):

$$H_0 : \beta_1 = 0$$

vs.

$$H_1 : \beta_1 \neq 0$$

In the last column of the table (Summary of Regression for the Dependent Variable), we find two p-values (we have two parameters: the intercept and the slope associated with the single variable x). Both are practically equal to 0.00. Conclusion regarding the significance of the linear regression model parameters: Overall conclusion regarding the estimated linear regression model The estimated linear regression model fits the empirical data very well (coefficient of determination $R^2 \cong 0.922$). It describes a statistically significant linear relationship between the second variable and the first (both p-values $\cong 0.00$ in the significance test of the linear regression coefficients).

	Regression summary for the dependent variable: second variableR= ,96045246 R^2= ,92246892 Popraw. R2= ,91859237 F(1,20)=237,96 p<,00000 Błąd std. estymacji: 7,8945					
	b*	Bł. std. z b*	b	Bł. std. z b	t(20)	p
Free term			-125,575	12,92649	-9,71451	0,000000
First variable	0,960452	0,062262	1,347	0,08735	15,42599	0,000000

Construction and interpretation of the receiver operating characteristic (ROC) curve. Application of the logistic regression model. Interpretation of the parameters of the logistic regression equation; calculation and interpretation of the odds ratio. A case study based on an article by Agnieszka Strzelecka et al.

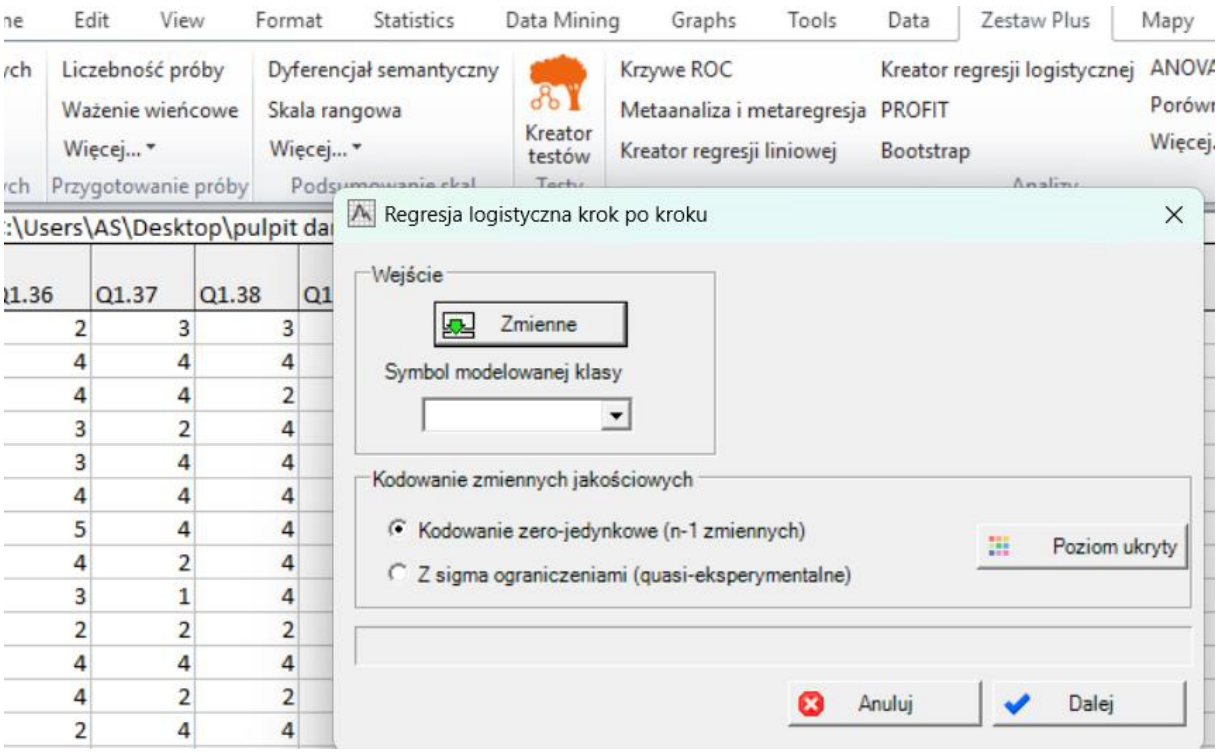
The tool for building a logistic regression model is located in the "Set Plus / Logistic Regression Model" tab. When analyzing data using logistic regression, we can include any set of variables (categorical and continuous) in the model, provided they have a justified influence on the dependent variable. The dependent variable in the model can take two values (yes–no, improvement–no improvement, etc.) or another format for representing the values of the variable in question.

ROC Curve

One of the outputs of a logistic regression model is the probability that individual cases belong to the modeled class. Every decision rule is designed to deliver the best results—the fewest misclassifications. To plot the ROC curve, go to the regression results tab

Results Residuals1 tab ROC Curve

The principles of model building and the ROC curve will be discussed using the article as an example. Determinants of primary healthcare patients' dissatisfaction with the quality of provided medical services https://www.aaem.pl/Determinants-of-primary-healthcare-patients-dissatisfaction-with-the-quality-of-provided,132783,0,2.html



To discover the determinants of primary health care patients' dissatisfaction with the quality of medical services.

The final form of the logistic regression model with qualitative explanatory variables pointed to the determinants of primary health care patients' dissatisfaction with the quality of provided medical services. The model's goodness of fit was verified using the Hosmer-Lemeshow (HL) test. Furthermore, a ROC curve was built, using which the conformity of the indications of patients' satisfaction and dissatisfaction resulting from the model was evaluated against the actual indications. The area under the plot of the ROC curve was calculated, marked as AUC (area under curve), which is a measure of the goodness of the model.

The assessment of the quality of health services (a dependent variable in the logistic regression model) was defined as a dichotomous variable which assumed the following two variants: a patient's satisfaction or dissatisfaction with the medical service received. The construction of the model was preceded by a preliminary selection of predictors through an assessment of their quality using Cramér's V coefficient. A number of predictors were discarded at that stage. The remaining predictors were included in the sequential construction of the logistic regression model. To achieve this, the forward stepwise

regression was used, and the significance of the difference between the subsequent models that were built was evaluated using an LR (likelihood ratio) test. In the final step, another group of variables was discarded, as they proved to be insignificant – see Tab. 2. The remaining variables, however, were included in the final version of the model – see Tab. 3. The predictors' statistical significance was verified using the Wald test.

Tab. 2. The statistically insignificant predictors which were dismissed

Predictors	Value of the Wald statistic	p-value
Type of visit	2.178102	0.139987
Patient's consent to a given treatment	0.001434	0.969788
The doctor taking the patient's monetary situation into account	2.393783	0.302132
Assessment of house visits	3.662625	0.160203
Waiting time for an appointment	1.589945	0.207334
Manner of registering to a GP	4.768181	0.189581
The ability to freely choose a doctor	1.549588	0.213196
The manner in which a doctor welcomes the patient during a visit	0.778833	0.677452
The course of a visit	3.799975	0.149571
Waiting time for a visit	0.160168	0.689002
Transfer of the patient's medical data between healthcare institutions	3.317207	0.345255
Standardised (uniform) medical data of a patient (for example test results, or orders)	3.352502	0.340402
Access to medical data 24h a day, seven days a week	0.674903	0.879091
Security of the patient's sensitive data	1.801599	0.614587
A website of the primary health care institution	0.577075	0.901660
Providing a flow of information about a patient between institutions using an information system in healthcare	5.145542	0.076324
Utilising e-prescriptions	3.949213	0.138816
Utilising e-referrals	2.167056	0.338400
Utilising the Internet to exchange or obtain patients' medical data from other medical institutions	3.394166	0.183217
Assessment of the inconvenience of waiting	1.510083	0.219126

Using logistic regression, the determinants which in the patients' opinions influence the dissatisfaction with medical services were established (Tab. 3).

Tab. 3. The final predictors which were included in the (logistic regression) model

Variable – reference variant	Estimate of the logistic regression parameter	OR (95% CI)	p-value
Constant term	-5.199	0.006 (0.002-0.014)	0.000
The attitude of a doctor towards the patient – the	2.170	8.754 (3.625-21.140)	0.000

doctor is rude			
A GP refuses admittance in an emergency	1.895	6.656 (3.062-14.469)	0.000
A GP sends the patient to a different primary care physician	1.655	5.232 (2.598-10.535)	0.000
The attitude of the remaining personnel – the remaining personnel are rude	1.282	3.603 (1.745-7.439)	0.001
Understanding information about health – the information about the health condition is not understandable to the patient, lack of full explanation of the health condition and the further process of treatment	1.428	4.170 (1.598-10.877)	0.004
Clear identification of the patient in the system – lack of clear identification	1.126	3.084 (1.311-7.255)	0.010
Availability of information about health – difficult access to information regarding the health condition (medical documentation)	0.677	1.968 (1.125-3.443)	0.018

Based on the estimated logistic regression one may state that the chance of occurrence of a patient's dissatisfaction with a medical service is 8.7 times greater when a doctor is rude during a visit, compared to the situation when they are polite (OR = 8.754; 95% Cl: 3.625-21.140; p 0.000). Patients expect politeness and understanding from the doctor's side during a doctor's visit. The next determinant of dissatisfaction with the quality of services is the fact that a doctor does not admit patients in an emergency (OR = 6.656; 95% Cl: 3.062-14.469; p = 0.000) or when a GP sends the patient to a different GP (OR = 5.232; 95% Cl: 2.598-10.535; p = 0.000). A patient is often forced in such a situation to use paid medical advice. The described situation happens more often in smaller primary health care institutions, where 1-2 GPs admit patients.

Another determinant is the lack of clear patient identification in the healthcare system (OR = 3.084; 95% Cl: 1.311-7.255; p = 0.010), which may lead to mistakes and a lack of complete medical documentation. Patients do not have full knowledge regarding primary health care institutions sharing medical documentation. This issue is regulated, amongst others, in the Act defining the principles of the functioning of primary health care and on Patients' Rights and on the Ombudsman for Patient Rights. This state causes the chance of dissatisfaction with the quality of medical services to increase almost two-fold, compared to the case when patients have full information about and access to their medical documentation (OR = 1.968; 95% Cl: 1.125-3.443; p = 0.018) (Tab. 3).

The value of the statistic HS = 7.9319, with the value of p= 0.16003, indicated a significant model fit of the logistic regression model. Based on the analysis of the area under the ROC curve, one can also state that the model fits the data well (the area is AUC = 0.834) (Fig. 1) and is characterised by a good predictive ability resulting from the obtained sensitivity and specificity graphs for different levels of probability (Fig. 2).

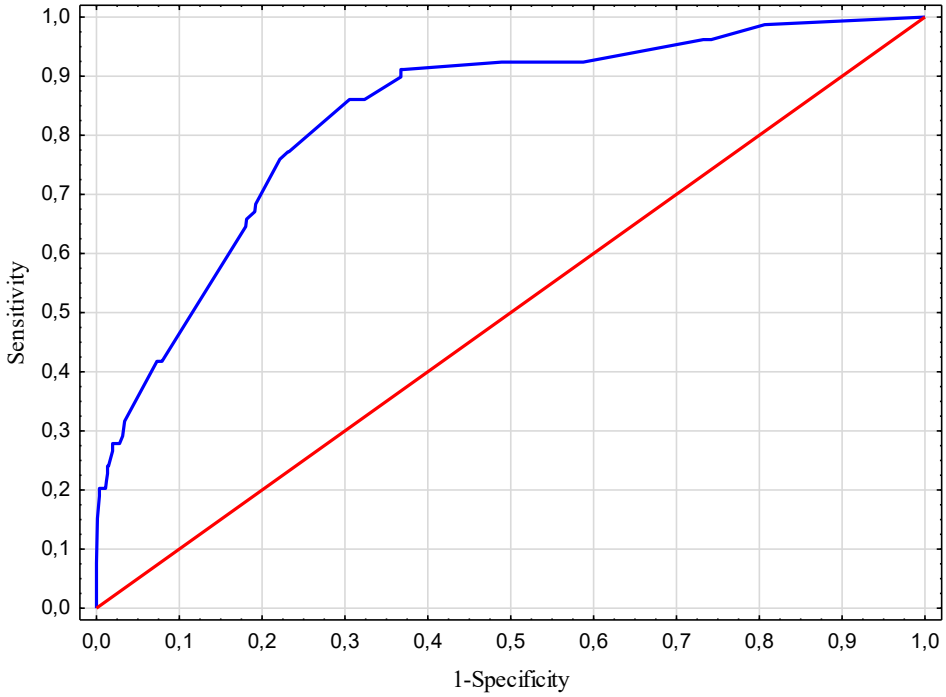


Fig. 1. The ROC curve graph

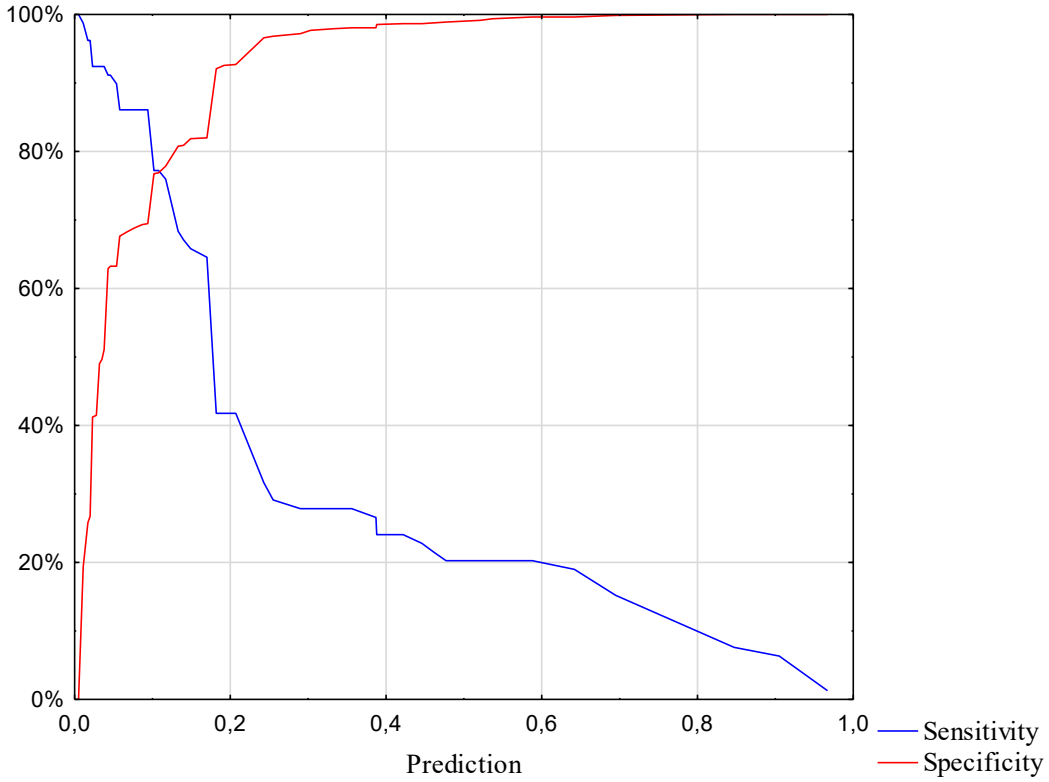


Fig. 2. The sensitivity and specificity graph

Plotting a survival curve. Testing the null hypothesis that there are no differences between survival functions. Survival analysis based on the Cox proportional hazards model.

Survival Analysis

Objective

To conduct a survival analysis using descriptive methods and survival analysis techniques, such as the Kaplan-Meier estimator, the log-rank test, and the Cox proportional hazards model.

Note: In all statistical tests performed below, the significance level is set at $\alpha = 0.05$.

Task Description and Data Preparation

Perform a survival analysis for patients undergoing surgery due to cancer N, assuming that the endpoint of the study is death caused by cancer N.

1. Defining the Endpoint and Censoring

- Endpoint (event of interest): Death due to cancer N.

- Observation status: In survival analysis, every patient requires a variable indicating whether their survival time observation is complete (event) or censored (truncated):
    - Complete observations (status = 1): Death due to cancer N.
    - Censored observations (status = 2 or 3): Patients who are still alive at the end of the study or who died from a cause other than cancer N.

2. Preparing and Loading Data into Statistica

- Before importing the data into Statistica, it is recommended to pre-process it in Excel and create a sheet named dane_przekodowane (*recoded_data*).
- While not strictly necessary, it is highly recommended to replace numeric codes of qualitative variables with text descriptions (e.g., change 1 to present, 0 to absent). This way, you will not have to memorize what each code represents, and chart legends will be generated with clear text labels automatically.
- It is best to preserve the original status variable (which includes information about deaths from causes other than cancer N) under a name like detailed_status, and create a new variable named status representing the "proper" status of the patient for this specific survival analysis.
- As always, verify that all other variables are correct and do not contain suspicious entries or missing values.

Research Question

We will analyze a mini-research problem regarding the relationship between the survival time of patients operated on for cancer N and the occurrence (or absence) of symptom X.

Research Question: Does survival time depend on the presence (or absence) of symptom X?

Part I: Preliminary Data Analysis (Descriptive Survival Measures)

We begin with a preliminary data review related to our research question, including:

1. Distribution of the symptom_X variable:
    - *Path:* Statistics →Basic Statistics →Frequency tables (select symptom_X →click Summary).
2. Number of complete and censored observations:
    - *Path:* Statistics →Basic Statistics →Frequency tables (select status →click Summary).

3. Contingency table (crosstabulation) for symptom X and survival status:
    - *Path:* Statistics →Basic Statistics →Tables and banners →Click Specify table (List 1 = status, List 2 = symptom_X) →OK →OK.
    - *Optional:* In the Options tab, check Column percentages (to calculate percentages immediately), go to the More tab, and click Detailed 2-way tables.
4. Descriptive statistics for survival times based on symptom X:
    - *Path:* Statistics →Basic Statistics →Breakdown & one-way ANOVA (select time as the dependent variable and symptom_X as the grouping variable; check the appropriate boxes in the Descriptive statistics tab).
5. Survival times for complete vs. censored observations:
    - *Path:* Statistics →Basic Statistics →Breakdown & one-way ANOVA (select time as the dependent variable and status as the grouping variable).
6. Survival times by both status and symptom X simultaneously:
    - *Path:* Statistics →Basic Statistics →Breakdown & one-way ANOVA (select time as the dependent variable and both symptom_X and status as grouping variables).

Descriptive Survival Measures (Without Hypothesis Testing)

- Mean survival time: Calculated as the sum of survival times divided by the total number of patients.
    - *Result:* The mean survival time in the group with symptom X is approx. 1818 days, while in the group without this symptom, it is approx. 2415 days.
    - *Preliminary conclusion:* The occurrence of symptom X is associated with a shorter mean survival time.
- Average hazard rates (which tend to overestimate the actual risk): Calculated as the number of complete observations divided by the sum of survival times.
    - *Result:* The average hazard rate in the group with symptom X is approx. 0.000251, whereas in the group without the symptom, it is approx. 0.000058.
    - *Preliminary conclusion:* The average hazard rate in the group with symptom X is approx. 4.3 times higher than in the group without this symptom.

Part II: Comparing Survival Curves (Log-Rank Test)

We will now analyze survival times using formal statistical tests, starting with comparing the survival curves using the log-rank test.

Hypothesis Setup

H_0:Survival curves in the group with symptom X and the group without symptom X are identical

H_1:Survival curves in the group with symptom X and the group without symptom X differ

Execution in Statistica

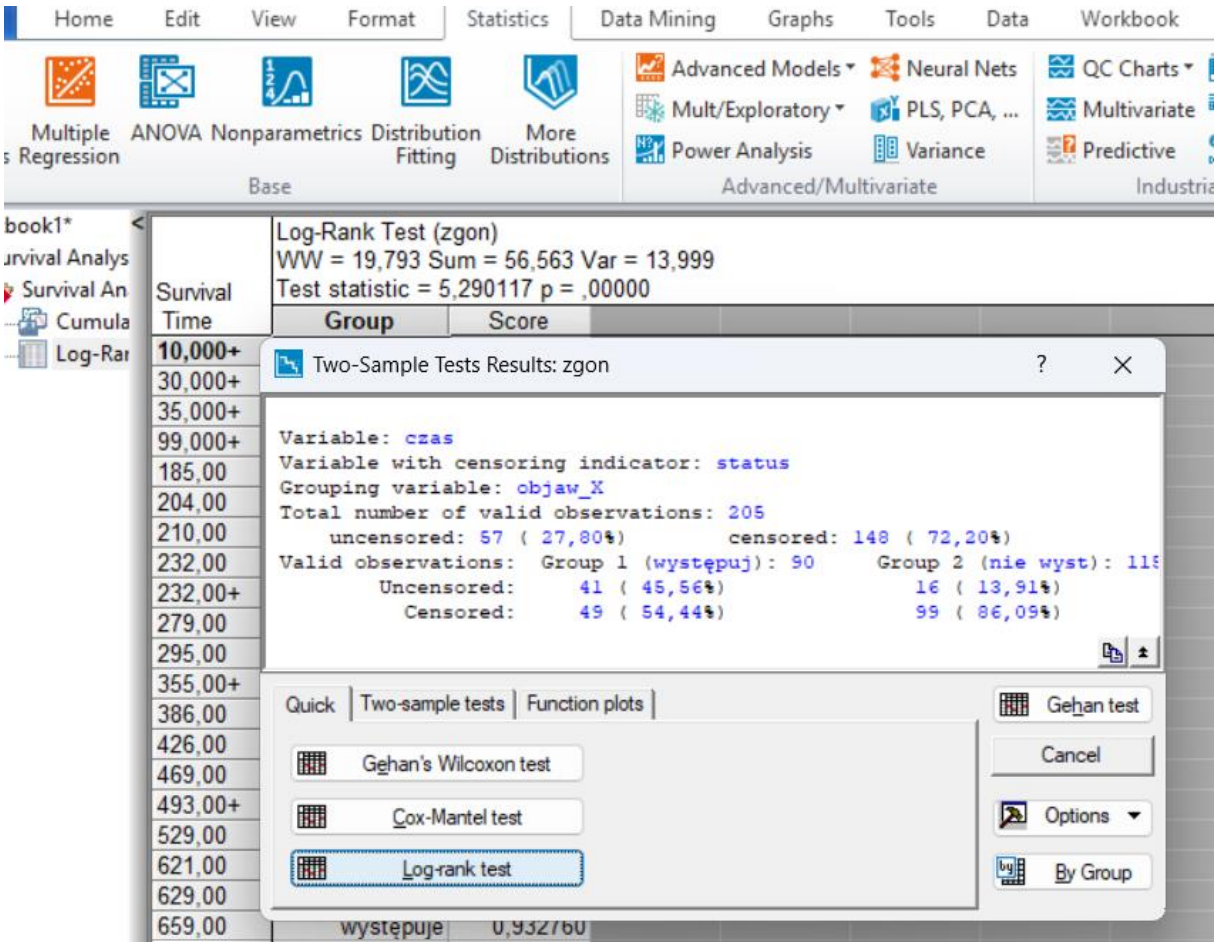
1. Navigate to: Statistics →Advanced Models →Survival Analysis.
2. Since the variable symptom_X takes only two values, select Comparing two samples and click OK.
3. In the dialog box, specify the variables for the survival analysis:

Field Name in Statistica	Variable to select from the data file	Description
Survival time	time	Contains patient survival times
Censoring indicator	status	Contains information on whether survival time is complete or censored
Grouping variable	symptom_X	Contains group membership info (with vs. without symptom)

4. After clicking OK, proceed to the next dialog box where you must specify which values of the symptom_X variable represent complete observations and which represent censored observations. You can type the codes manually (e.g., "complete", "censored") or, more conveniently, double-click the input fields to select the codes from the drop-down list.
5. Click OK, then click the Log-rank test button to obtain the test output.

Results and Interpretation

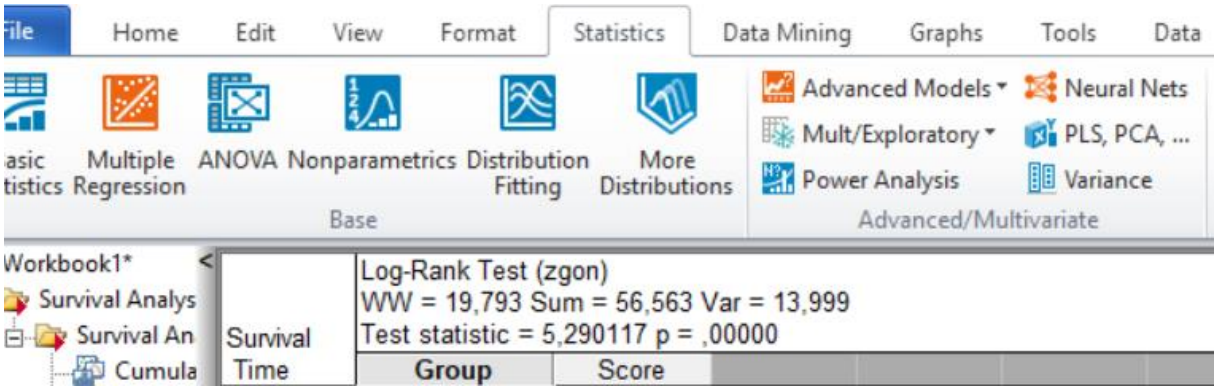
The first column of the resulting table contains the survival times of individual patients, marked with a "+" sign if the time is censored (uncensored values are shown without a "+"). The p-value for the log-rank test is highlighted in yellow.



Conclusion Template:

- If $p < 0.05$: Since the p-value in the log-rank test is [...], we reject H_0regarding the identity of survival curves.
- If $p \geq 0.05$: Since the p-value in the log-rank test is [...], there is no basis to reject H_0regarding the identity of survival curves.

Conclusion for our study: Since the p-value in the log-rank test is less than the significance level $\alpha = 0.05$(specifically, the p-value is nearly 0, and certainly $p < 0.00001$), we reject H_0regarding the identity of the survival curves.



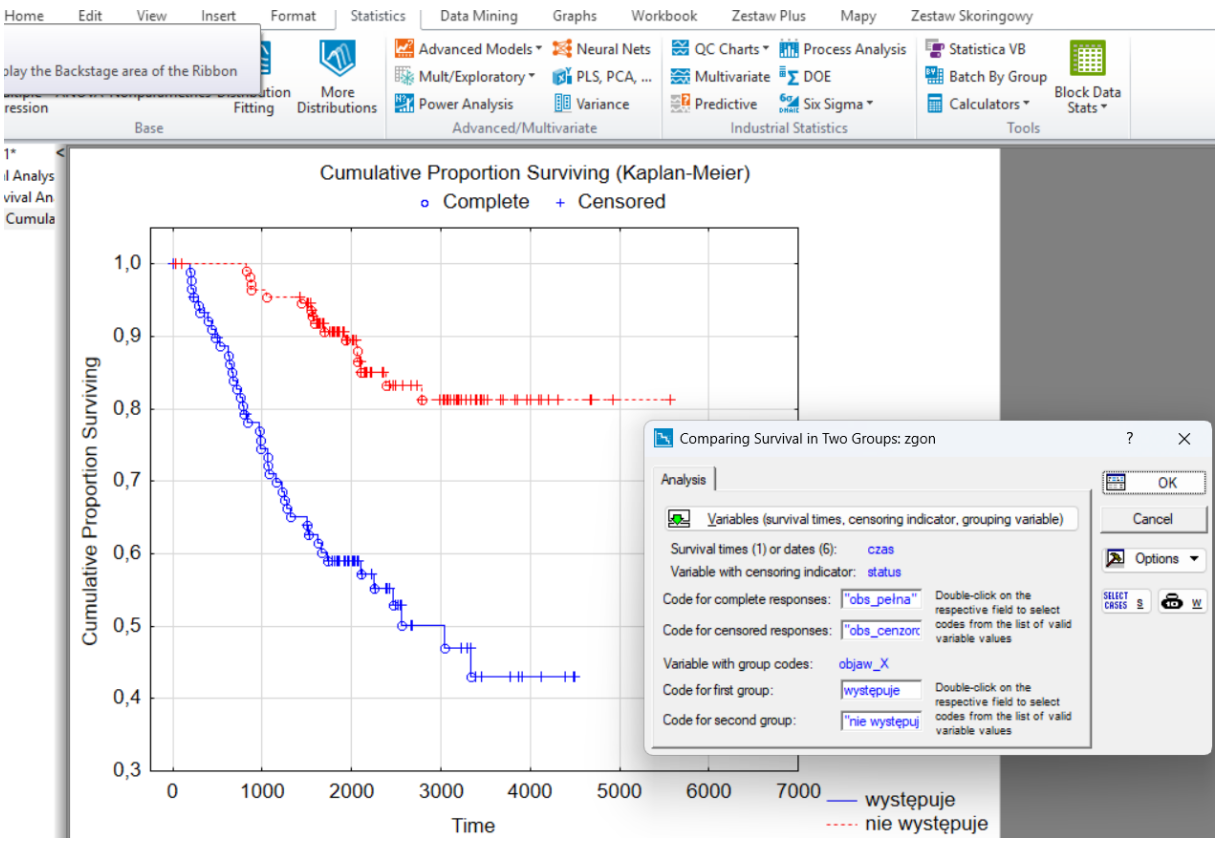
Part III: Kaplan-Meier Survival Curves

To visualize the survival probability over the entire follow-up period, we can generate a Kaplan-Meier plot in Statistica.

1. On the results dialog for the two-sample test, go to the Function plots tab.
2. Select Survival probability for groups.

Plot Interpretation

The survival curves visualize the probability of survival over time. As shown, at any given time starting from the day of surgery, patients without symptom X (upper curve) have a higher probability of survival than patients with symptom X. While this visual relationship applies to our sample, the statistical significance of this difference was confirmed by the log-rank test.

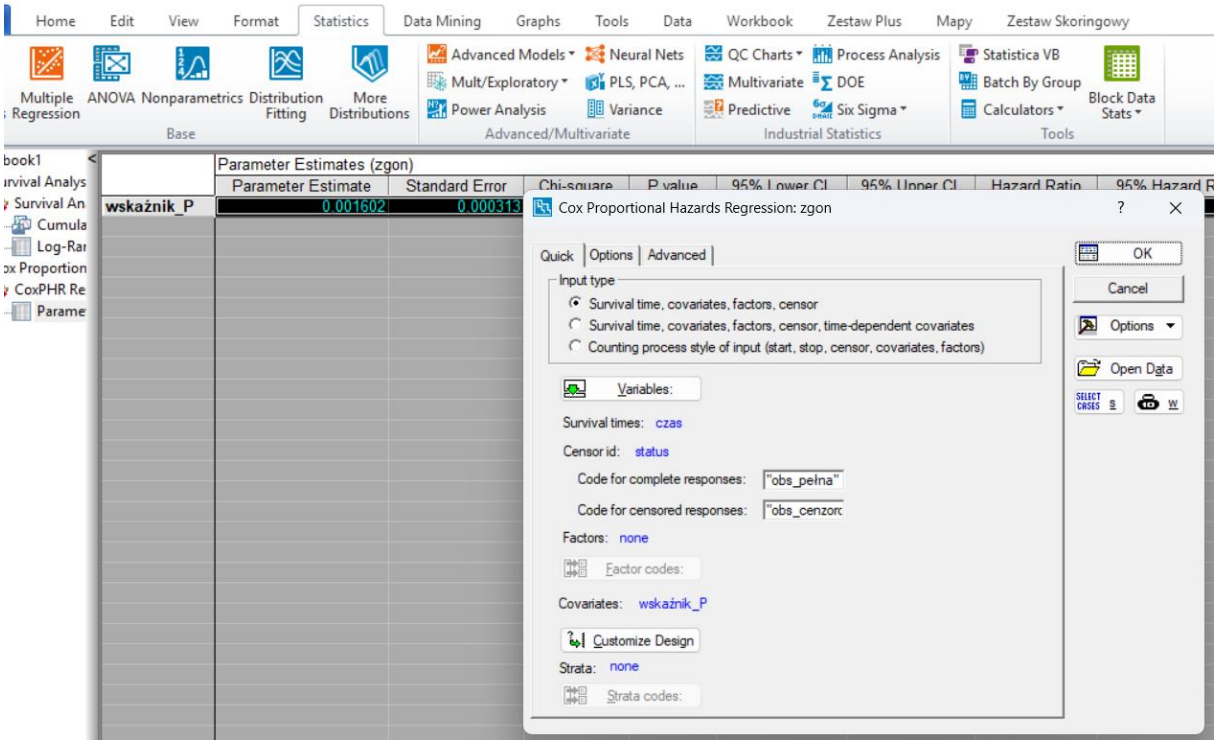


Part IV: Cox Proportional Hazards Model

Further survival analysis can be conducted using the Cox proportional hazards regression model, which allows us to quantify the risk.

1. Univariate Model (Crude Hazard Ratio - Crude HR)

1. Go to: Statistics →Advanced Models →Cox Proportional Hazards Models.
2. Select the survival variables and the qualitative predictor (symptom_X). Specify the codes for complete and censored observations.
3. Go to the Options tab, and under Coding of qualitative variables, set the reference class (typically the group with the lower hypothesized risk). In our case, this will be the group where symptom X is absent.
4. Click OK to run the analysis, and on the results dialog, click Parameter estimates.



Hypotheses for the Cox Model Parameter:

H_0:HR $= 1$ (equal risk of death in both groups)

H_1:HR $\neq 1$ (different risk of death in both groups)

Alternatively:

$H_0: \beta = 0$ vs. $H_1: \beta \neq 0$

(where β is the model coefficient for the predictor symptom_X).

Results:

- Hazard Ratio (HR): ≈ 4.4
- 95% Confidence Interval (95% CI): $2.4 - 7.8$
- p-value: 0.00001

Conclusion: Patients with symptom X have approximately 4.4 times higher risk of death compared to patients without symptom X. This association is statistically significant ($p = 0.00001$), which is also reflected by the fact that the 95% confidence interval for the HR does not include 1 (which would indicate equal risk).

Note: The HR calculated above is a "crude" HR, meaning it was determined without adjusting for other potential confounding variables.

2. Multivariate Model (Adjusted Hazard Ratio - Adjusted HR)

If we include additional variables in the model, we obtain the adjusted HR. Let us calculate the HR for symptom X while adjusting for the patients' sex.

1. Set up the Cox proportional hazards model as before, but include both symptom_X and sex as predictors.
2. Generate the parameter estimates table.

Results (Adjusted for Sex):

- Adjusted HR for symptom X: ≈ 4.1
- 95% Confidence Interval (95% CI): $2.3 - 7.4$
- p-value: 0.00002

Conclusion: After adjusting for sex, patients with symptom X still have approximately 4.1 times higher risk of death compared to those without symptom X. This relationship remains statistically significant ($p = 0.00002$, and the 95% CI does not include 1).

3. Cox Model with Continuous Variables (Indicator P)

The Cox proportional hazards model can also evaluate whether continuous (quantitative) variables are associated with survival time. For example, let's assess whether indicator P is associated with survival.

Before using formal survival analysis methods, we can perform a preliminary check of the range of indicator P values among patients (e.g., using Breakdown & one-way ANOVA in Basic Statistics):

The descriptive results might suggest a relationship (e.g., higher values of indicator P are observed in censored cases compared to complete cases in our sample). However, these values do not account for time-to-event dynamics. We must use Cox regression to formally assess this.

1. In the variable selection window for the Cox model, select the appropriate variables. Ensure that indicator_P is selected in the Continuous predictors (or *Independent covariates*) section.
2. Specify the complete/censored codes and go to the results table (the qualitative coding step is skipped here since we are only evaluating a continuous variable).

Results:

- Hazard Ratio (HR): 1.0016

- 95% Confidence Interval (95% CI): $1.0010 - 1.0022$
- p-value: < 0.00001

Conclusion: An increase in indicator P is statistically significantly associated with an increase in the risk of death (since $\mathrm{HR} > 1$ and $p < 0.00001$). The $\mathrm{HR} = 1.0016$ can be interpreted as follows: a one-unit increase in indicator P is associated with an approximate 0.16% increase in the risk of death (calculated as: $(1.0016 - 1) \times 100\%$).

	Parameter Estimates								
	Parameter Estimate	Standard Error	Chi-square	P value	95% Lower CL	95% Upper CL	Hazard Ratio	95% Hazard Ratio Lower CL	95% Hazard Ratio Upper CL
wskaźnik_P	0,001602	0,000313	26,27547	0,000000	0,000990	0,002215	1,001604	1,000990	1,002218