Statistical methods in research

Petr Bujok, 2020

What is statistics?

Karl Pearson: 'Statistics is the grammar of science.'

Statistic covers methods used for gathering, organising (transforming and cleansing), analysing, and publishing results from measured data. Data represent specific parts of World which are observed and measured. Necessary note that data are represented by numbers (values) or symbols (strings).

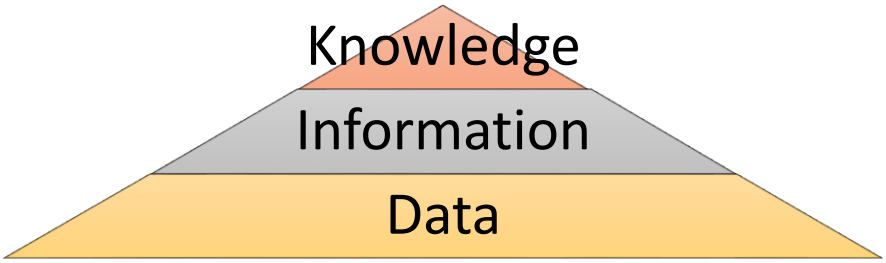


Figure 1 - Relation between data, information and knowledge

Measured data, stored typically in digital form, could be structured in tables, but they do not bring news and facts about measured part of World. On the other hand, if data are used in the (statistical) computational process, provided results could be called information. The information enables typically answer questions 'who', 'where', 'what', etc. Finally, the interpretation of information and providing a comprehensive summary of the observed research area is known as knowledge.

Statistics enables to analyse measured data to extract knowledge and use it in conclusions. More exactly, factors influencing answering questions 'who', 'where', 'what' are detected, predicted and controlled. Statistics typically covers the analytic part where analysis is mentioned as a process to break down whole data to components or answers [1].

Statistical data are usually stored in the **table** where columns represent measured **variables** (temperature, weight, speed, etc.) and rows contain observed **cases** (individuals, objects). All objects of defined (mentioned) part of World are called population, but in many cases, it is not possible to measure the whole population. In these cases, the sample file of sufficient size is extracted from the population, where each object has the same (uniform) probability to be selected.

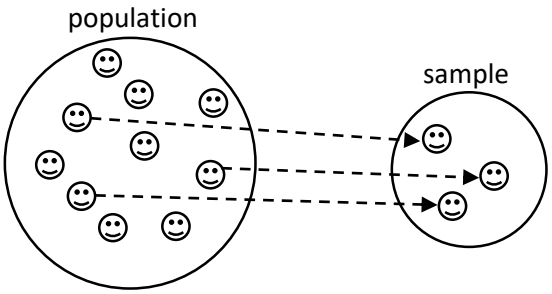


Figure 2 - Population versus sample

Main reasons for using statistics in research

At first, it is important to say that statistics are not necessarily used in all field of research and science. In the field of research where no numerical or categorical data (values) are provided statistics has no usage. On the other hand, experimental research provides such data which should be statistically analysed. In these research areas, statistics are used to provide a numerical or graphical overview of measured data or illustrate the significance of the difference between achieved new results and state-of-the-art results.

Statistics methods – descriptive or inferential?

At first, if the main goal of the analysis is to summarise and describe measured data, the basic characteristics are used (**descriptive statistics**). In the case of analysis, where given facts need to be generalised above and beyond measured data, more sophisticated methods of **inferential statistics** are employed. Inferential statistics is also important in cases where the newly achieved approach has to be compared with the old existing approach. The differences observed between numerical results of both approaches are often very small, and here inferential statistics brings clear decision based on measured data. Deep insight into the introduction to data analysis provides monography [2]

Descriptive statistics

Descriptive statistic describes the relationship between measured variables from a sample or population. Results of descriptive statistics are represented typically as values (tables) or figures (plots).

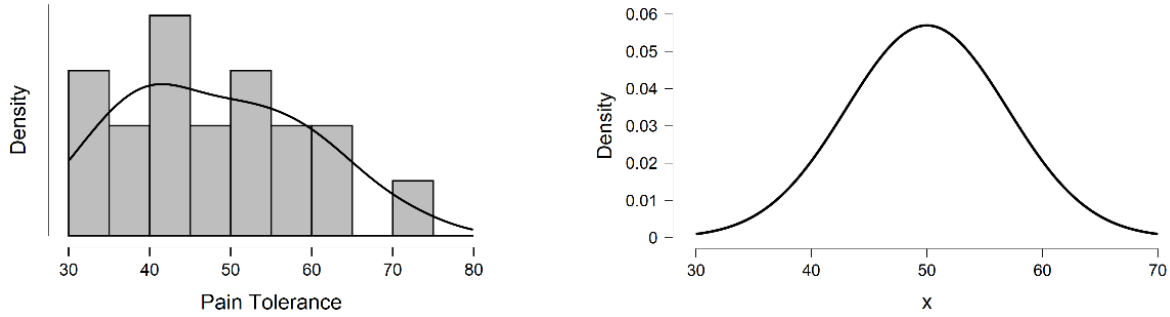


Figure 3 - Empirical histogram and theoretical Gaussian density function

Both forms of results provide an overview of data in the form of basic characteristics. Beside frequencies (*measures of frequency*), the most common numerical characteristics are divided into two groups: *measures of central tendency* (minimum, maximum, arithmetic mean, median, mode, etc.), and *measures of variation* (range, standard deviation, variance, etc.).

In figure 3, the most common descriptive characteristic – *frequency* - is illustrated. The frequencies enable to modelling the distribution of measured variables. The bars of the plot represent numbers of objects with given pain tolerance. The plot is called *histogram* because 'pain tolerance' variable is measured as a quantitative variable. On the right side of this figure, the theoretical probability density function is illustrated to compare achieved distribution with theoretical Gaussian distribution.

We illustrate the aim of descriptive statistics when analysing data of dependency between physical activity (PA) and body mass index (BMI). At first, basic descriptive characteristics are depicted in table 1 for both variables. These values clearly describe both measured variables from example. Both variables are measured completely, values of PA are between 3 and 14, and for BMI between 14 and 35 (it helps to recognise mistakes in data known as *outliers*).

Table 1 - Example of descriptive statistics - basic characteristics

	Physic.Activity	BMI
Valid	100	100
Missing	0	0
Mean	8.6	23.9
Median	8.4	24.5
Std. Deviation	2.3	3.9
Range	11.0	20.9
Minimum	3.2	14.2
Maximum	14.2	35.1
25th percentile	6.8	21.1
50th percentile	8.4	24.5
75th percentile	10.3	26.8

Similarly, the basic characteristics are used to compare the variable in two independent groups of objects (see table 2). Heart rate (beats per minute) was measured on 800 people, and basic descriptive characteristics were computed for male and female independently. From these results, it is clear that male participants have lower average heart rate (see the mean values) and more similar observed values (see standard deviation values). It can indicate that male participants better cope with heart rate in the training exercises compared to female participants.

Table 2 - Example of descriptive statistics: Heart Rate and gender

	Female	Male
Valid	400	400
Missing	0	0
Mean	132	117
Std. Deviation	22.7	19.8
Minimum	78	69
Maximum	196	172

Although the provided values are exact and correct, the graphical form of descriptive statistics is often more absorbable. Figure 4 briefly illustrates the basic characteristics of BMI in one plot know as *box-plot* (or whiskers plot). This plot provides minimum, maximum, median, and 25^{th} and 75^{th} percentiles in the compact view.

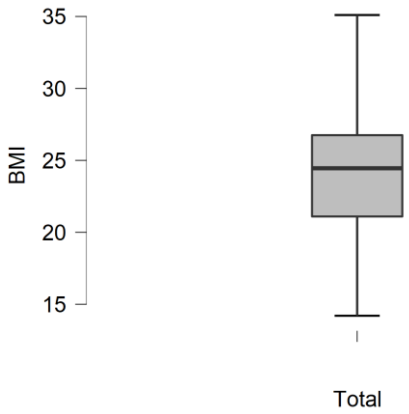


Figure 4 - Box-plot of BMI

Box-plots are very useful in the tasks where the variable is measured in two or more groups of objects. Naturally, the same variable measured in the same units has to be compared in one plot. Figure 5 depicts the relation of the weight of 16 people before and after eight-week calories-intake. Higher weight after the procedure is the obvious and more sophisticated procedure can detect the significance of this weight-gap.

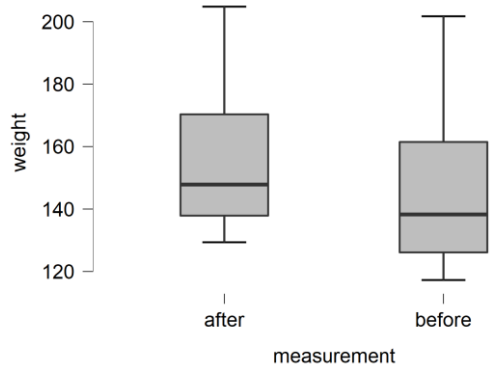


Figure 5 - Box-plot for comparison of the variable in two groups

Necessary note that basic (descriptive) characteristics do not indicate the relation between PA and BMI what was probably the main aim of the research. To illustrate the relationship between two variables, *scatter plots* are widely used. Figure 6 shows an example of a scatter plot for the relation between PA and BMI variables.

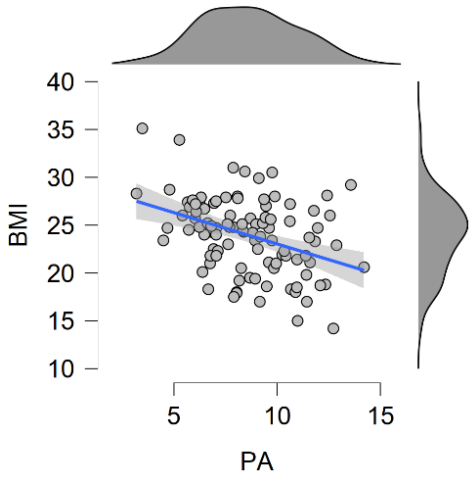


Figure 6 - Relation between BMI and physic.activity

Typically, only the middle part of this plot is provided, and marginal (top and right) parts are hidden (these plots represent achieved distribution curves of variables). In the middle of this plot, points representing each measured object (PA and BMI values) are depicted. In the good cases, these points are formed close to the linear curve (represented as a blue line). In real cases are achieved plots like Figure 4, which are without the estimated auxiliary linear line not very useful. In this case, the blue line indicates a rather negative relation between PA and BMI. It means smaller values of BMI for bigger PA values (people without physical activity have bigger body mass index and vice versa). Further, we can analyse this first-view relation more sophistically computing widely-used correlation coefficient.

The main aim of descriptive statistics is a description of the values measured on the objects. Although these techniques are very useful and important in many fields of research and science, mostly, it is necessary to support descriptive statistics by strict decisions provided by inferential statistics.

Inferential statistics

As was mentioned previously, newly presented numerical (empirical) experiments require unambiguous answers to the research questions. Results of descriptive statistics are often not strongly clear to make a decision without doubts. In these situations, methods of *inferential statistics* can help accept or reject the research hypothesis. The null hypothesis (H_0) is constructed for each research hypothesis as an opposite ('men are higher than women' and H_0 is' men have the same height as women').

A paradigm of inferential statistics typically uses sampled data from a given population. It means that only a portion of objects from the population are selected to describe and make inferences about the whole population [3]. This methodology can be influenced by a wrong decision (the null hypothesis is true, and the results reject it, and vice versa). In these cases, we specified error *type I* (reject the true hypothesis) and *type II* (non-reject the false hypothesis). In this principle, the probability of error type I is corrected by input parameter α called significance level. Unfortunately, decreasing the value of α the probability of error type II is increased. Statisticians recommended using values from $\alpha = 0.05$ in standard cases, to $\alpha = 0.001$ in the cases where the decision is very important.

The α value plays a crucial role in the null hypothesis decision-make process when using statistical software. In these cases, a significance or more often *p-value* (p-level, probability level or p) is computed. Simply said, this p-value can be imagined as 'probability of trust in the null hypothesis'. If the p-value in the test is too small (smaller than given α), then the null hypothesis should be rejected, and vice versa.

To make a correct decision about the given null hypothesis, data and appropriate test are needed. There is a lot of statistical tests for different statistical tasks. At first, the tests are divided into two groups based on the shape of the distribution of the data. When data are distributed normally (Gaussian distribution), parametric tests are recommended to test the null hypothesis. In other cases, nonparametric methods should be employed. To get the information about the distribution of the data, another hypothesis (H_0: 'data are from Gaussian distribution') is tested.

Most often used methods in inferential statistics

Obviously, for different research tasks, different statistical tests are used. In some tasks, it is possible even to use more variants of methods to prove a fact about measured data. Here, only the most common methods used in inferential statistics will be briefly described. For more information and more specific methods, readers should study some more comprehensive articles or books [4].

Dependency of variables

The first group of statistical methods is focused on the evaluation of dependency between two or more variables. These methods strongly vary when different types of variables are measured in experiments – **one qualitative and one quantitative variable**, **both qualitative** variables, and **both quantitative** variables.

If both measured variables are qualitative (typically from categorical scales), the most popular method for evaluating dependency of the variables is **Pearson chi-square test** of independence. We introduce this method when evaluating the dependency of hair length (-1 denotes short hair) on gender (-1 represents male).

Table 3 - Contingency table: hair length and gender

		hair length		
Sex		-1	1	Total
-1	Count	15	1	16
	Expected count	8	8	16
	residuals	2.5	-2.5	0
1	Count	1	15	16
	Expected count	8	8	16
	residuals	-2.5	2.5	0

Table 3 illustrates that 15 male participants have short hair and one long hair, and 15 female participants have long hair and one short hair (rows Count). Expected counts represent theoretical counts achieved if hair length is not dependent on gender. The main aim of the test is to evaluate the difference between the reals counts and expected counts. If the difference (regarding all count cells) is big (using Chi-square distribution), null hypothesis about independency of the variables is rejected. Here (see table 4), the p-value is significantly lower than 0.001. Therefore the length of hair is significantly dependent on the sex of participants. Studying standardised residuals (Table 3), they are a lot of male participants with short hair (2.5 is positive and bigger than quantile 1.96) and a lot of female participants with long hair (the same reason). Negative residuals bigger than quantile -1.96 denote significantly small numbers of truly observed cases (small real counts compared with expected counts).

Table 4 - Chi-square Pearson test and significance level

	Value	df	p
X^2	24.50	1	< .001
N	32		

For evaluate the relationship between one quantitative and one qualitative variable, **one-sample**, **paired,** and **two-sample tests** are employed. These tests are used for data (i.e. quantitative variable) from Gaussian distribution, and they are known as the **t-test**. For other cases, nonparametric variants should be used (**signed**, **Wilcoxon**, **Mann-Whitney** or other tests). The one-sample t-test can be used for comparison of the mean value of a variable from one population with a given value. For example, we need to test if the mean value of IQ is bigger than 100 or not. The null hypothesis is 'Mean value of IQ is equal to 100', and if we reject the null hypothesis, we can prove our research idea.

	N	Mean	SD	SE
IQ	32	115.13	12.16	2.15

Average IQ in a sample of size 32 is 115, and when we compare this sample with a given value 100, parametric t-test rejects null hypothesis (p<0.001). Similarly, the nonparametric one-sample **Wilcoxon** test can be used if the data are not from Gaussian distribution. Although IQ values in our sample are from Gaussian distribution, the Wilcoxon test achieves the same decision as a t-test (p<0.001). Finally, the mean value of IQ in our population is significantly higher than a given value of 100.

	Test	Statistic	df	p
IQ	Student	7.03	31	< .001
	Wilcoxon	457.00		< .001

The paired test enables to compare the difference between the mean values of the variable measured in two moments. For example, we analyse the difference between the weight of 16 participants before and after 8-week calories intake. We can see that the average weight is bigger after this procedure (155 compared to 144 lb before).

	N	Mean	SD
Weight Before	16	144.64	22.70
Weight After	16	155.04	21.44

A parametric paired t-test is applied such that the difference of two observed weight values for each participant is computed, and then one-sample t-test is applied to differences. The null hypothesis is 'difference of weight has zero mean value'. P-value of the t-test is less than 0.001; therefore, weight is significantly increased. Similarly, the nonparametric **Wilcoxon** test is applied (data are from Gauss distribution) with the same decision (p<0.001). Calories intake significantly increases the weight of participants.

Test	Statistic	df	p
Student	-10.84	15	< .001
Wilcoxon	0.00		< .001

Finally, the **two-sample test** is employed when the difference between the mean values of the variable measured in two independent populations is studied. The main idea of this method will be illustrated on example, where the difference between swimming speed (in seconds for 500 meters) and gender (-1 = male).

	Group	N	Mean	SD	SE
swimming	-1	16	87.38	5.19	1.30
	1	16	75.63	3.18	0.80

Because data are from two independent populations, the difference between variances of these populations is tested at first. Because the provided p-value is less than 0.05, we reject equality of variances (it serves for the selection of more appropriate test statistics).

	F	df	p
swimming	5.77	1	0.02

Based on previous results, the null hypothesis ('there is no difference between swimming speed of males and females') is tested. The provided significance level is lower than 0.001; the null hypothesis is rejected. Similarly, the null hypothesis is rejected by the nonparametric two-sample **Mann-Whitney** test, where the same level of significance was achieved ($p<0.001$). There is a significant difference between swimming speed of males and females. The females achieved better results (females have a higher mean value of speed).

	Test	Statistic	df	p
Student		7.72	30	< .001
Mann-Whitney		252.00		< .001

In the cases, where the quantitative variable is not dichotomous (more than two various values are observed), one-way **analysis of variance (ANOVA)** test is employed. Test ANOVA is a generalised variant of the two-sample test, and it is used for example for evaluating dependency of pain tolerance (higher value means higher tolerance) on hair colour (for different types). We can see, people with light blond hair have the highest tolerance, and dark brunette people have the least tolerance.

Hair Color	Mean	SD	N
Dark Blond	51.20	9.28	5
Dark Brunette	37.40	8.32	5
Light Blond	59.20	8.53	5
Light Brunette	42.50	5.45	4

Now, the null hypothesis 'there is equal pain tolerance in all groups of hair colour' will be tested by parametric ANOVA test (data are from Gaussian distribution). The achieved p-value is lower than 0.05, the null hypothesis is rejected, and at least one group of hair colour differs from the remaining groups.

Cases	Sum of Squares	df	Mean Square	F	p
Hair Color	1360.73	3	453.58	6.79	4.11e-3
Residuals	1001.80	15	66.79		

Post-hoc tests are using to detect the difference between individual pairs of groups. There are several types of these tests, here results of Tukey test are presented. Provided p-values of post-hoc tests show two significant differences in pain tolerance: between people with light blond and dark brown hair, and light blond and light brown (as was expected). There is a significant influence of hair colour on pain tolerance; people with light blond hair have the biggest tolerance.

		Mean Difference	SE	t	p_{tukey}
Dark, Blond	Dark, Brunette	13.80	5.17	2.67	0.07
	Light, Blond	-8.00	5.17	-1.55	0.44
	Light, Brunette	8.70	5.48	1.59	0.41
Dark, Brunette	Light, Blond	-21.80	5.17	-4.22	3.71e-3
	Light, Brunette	-5.10	5.48	-0.93	0.79
Light, Blond	Light, Brunette	16.70	5.48	3.05	0.04

Similarly, nonparametric **Kruskal-Wallis** test is employed for test null hypothesis that 'the mean values in all groups of hair colour are equal'. Computed p-value (0.01) enables to reject the null hypothesis, and pain tolerance is influenced by hair colour.

Factor	Statistic	df	p
Hair Color	10.59	3	0.01

Standard one-way ANOVA enables to detect the influence of one qualitative variable on the quantitative variable. In many tasks, where it is necessary to use more than one factor to compare the populations' mean values, more-way ANOVA is used (typically two-way or three-way). **Two-way ANOVA** is used when the dependence of shoe size of people on the country (Scandinavia and Mediterranean), and length of hair (short and long) is studied.

country	hair	Mean	SD	N
mediterr	long	36.71	2.50	7
	short	43.14	1.77	7
scandinavia	long	37.29	1.11	7
	short	42.71	3.59	7

In ANOVA table are evaluated three independent factors influencing shoe size. Computed p-values show that country is not aa significant factor and hair is a significant factor. It means that shoe size differs only for people with different hair length.

Cases	Sum of Squares	df	Mean Square	F	p
country	0.04	1	0.04	9.65e-3	0.92
Residuals	22.21	6	3.70		
hair	246.04	1	246.04	41.92	< .001
Residuals	35.21	6	5.87		
country ✻ hair	1.75	1	1.75	0.37	0.57
Residuals	28.50	6	4.75		

The post-hoc test shows (see p-values of Holm test) that only groups of people with different hair length have significantly different shoe size.

		Mean Difference	SE	t	p_{holm}
scandinavia, short	mediter, short	-0.43	1.10	-0.39	1.00
	scandinavia, long	5.43	1.23	4.41	2.63e-3
	mediter, long	6.00	1.17	5.13	1.46e-3
mediter, short	scandinavia, long	5.86	1.17	5.01	1.46e-3
	mediter, long	6.43	1.23	5.22	1.34e-3
scandinavia, long	mediter, long	0.57	1.10	0.52	1.00

Similarly, nonparametric Friedman test provides a decision of the null hypothesis in ANOVA with more than one factor (the same decision was achieved).

Factor	Chi-Squared	df	p	Kendall's W	F	df num	df den	p_F
country	0.14	1	0.71	-27.27	0.13	3	20	0.94
hair	8.73	1	3.13e-3	-46.26	14.22	3	20	< .001

Correlation coefficient evaluates the relationship between two quantitative variables by values from $\rho \in \langle -1, 1 \rangle$, where values close to 0 mean weak or none dependency. There exist two versions of the coefficient – Pearson and Spearman. Pearson correlation coefficient should be used in cases where both variables are continuous or quantitative at least. On the other hand, if data are not continuous (are typically qualitative), Spearman correlation should be applied.

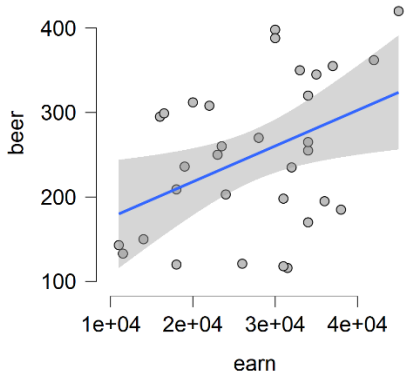


Figure 7 - Scatter plot of dependency between earnings and the amount of beer

We can illustrate this difference in an example where dependency between annual earnings (in €) and the amount of consumed beer (per year in litres) is studied. Figure 7 shows that the relationship is rather positive, i.e. higher consumption of beer means higher earnings (see blue line). Nevertheless, measured objects are not very close to the optimal linear dependency. In table 3, both types of correlation coefficients are computed with p-values.

Table 5 - Comparison of Pearson and Spearman correlation coefficient

Variable	coefficient	earn
beer	Pearson's r	0.42
	p-value	*0.02*
	Spearman's rho	0.34
	p-value	*0.06*

Both coefficients are represented by positive and similar values, but p-values promise different decisions. For each coefficient (estimated from the sample data) the null hypothesis is tested H0: 'Population correlation coefficient equals to zero' (i.e. dependency between variables is not significant). We can see that for the Pearson correlation coefficient, the null hypothesis was rejected (p-value < 0.05), but Spearman coefficient was not evaluated as significant (p>0.05).

Although correlation is successfully used in many real data analysis, **regression** analysis provides more detailed results in cases where more than one variable influence dependent variable. The most widely used kind of regression is called **linear regression model** (LRM) where the dependent variable and mostly independent variables are quantitative. In this model, there is one dependent variable and one or more independent variables (regressors, covariates). The relationship is not both-way as in correlation, here only the dependent variable is described by changes of regressors. Briefly, Least Squares Method is employed to compute parameters of LRM, the significance of each regressor is estimated, and suitability (or quality) of the overall model is evaluated by Index of determination ($R^2 \in \langle 0, 1 \rangle$ where 0 is for poor model and 1 for best model). LRM is used, for example, when studying dependency of earnings on three human factors (shoe size, age and beer consumption).

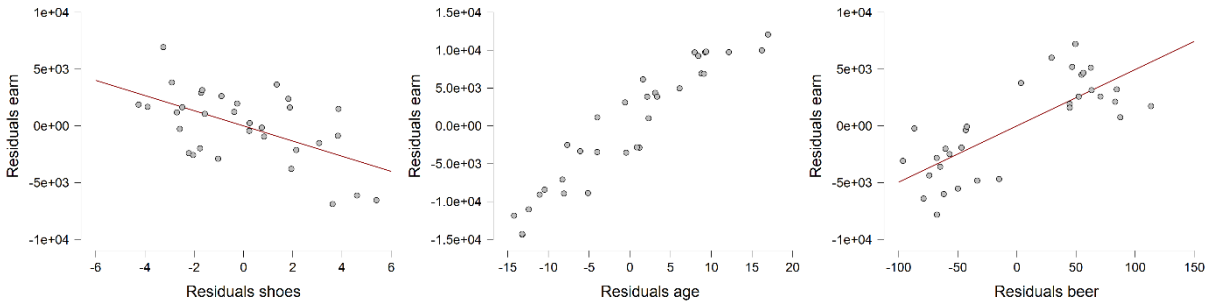
	N	Mean	SD	SE
earnings	32	27437.50	8929.61	1578.55
shoes	32	39.91	3.90	0.69
age	32	34.44	9.52	1.68
beer	32	249.50	90.60	16.02

In the following table, tests of significance for each regressor are depicted. For each regressor, the one-sample t-test is applied to test hypothesis 'parameter of regressor is equal to zero' (regressor is not significant), and if it is rejected (p-value < 0.05) the regressor is significant. Here, all regressors are significant, i.e. influencing dependent variable earnings.

Model		Coeff.	Std Error	t	p
H_1	(Intercept)	12176.99	5865.35	2.08	0.05
	shoes	-668.12	185.51	-3.60	1.21e-3
	age	858.16	54.92	15.63	< .001
	beer	49.58	7.61	6.51	< .001

Quality of this model is estimated in the following table where $R^2 = 0.92$ is close to maximal quality value 1. The earning of people is dependent on shoe size, age and annual consumption of beer. Sign of estimated coefficient (see the previous table) illustrate positive (+) or negative (-) dependency. For example, positive value for beer (49.58) means higher beer consumption higher earnings. For more information about linear regression, more comprehensive papers and books are provided.

Model	R	R^2	Adjusted R^2	RMSE
H_1	0.96	0.92	0.91	2726.00



There exist more other kinds of regression models. Next, **logistic regression** will be briefly introduced. The main difference between linear and logistic regression (LOR) is in the type of dependent variable. The dependent variable in LOR is dichotomous with two possible values (true/false) – i.e. sex or success in the examination. Here, an example of LOR where gender (sex, M-male, F-female) is dependent on the size of shoes (standard EU values). The plot of relation between gender and size of shoes is illustrated in Figure 8 a). It is visible that the points in this plot do not form to the continuous area (compare with Figures 7). Therefore, LRM will provide poor results (blue line estimated rough LRM between these variables). This task is very similar to the two-sample test, where mean values of the quantitative variable (size of shoes) are compared in two groups (gender).

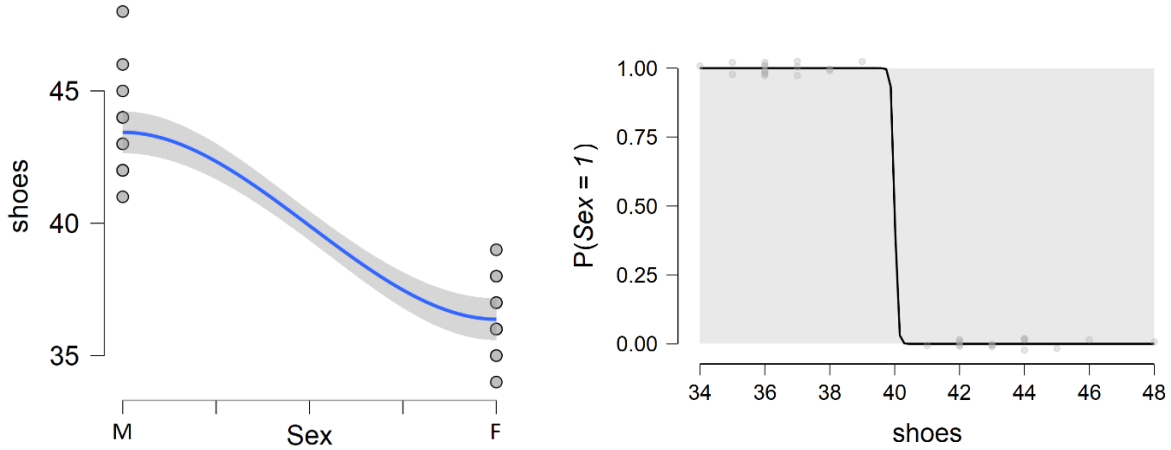


Figure 8-a) relation between the size of shoes and gender, b) LOR model

LOR provides a model of dependency of the dichotomous variable on one or several variables. Similarly, as in LRM, only significant covariates (regressors) are evaluated by a p-value less than 0.05. Moreover, LOR provides a decision-make mechanism enabling distribute data objects into two groups (of gender) based on values of regressors. Finally, if the estimated parameters of the LOR model are

appropriate and accurate (learned model), newly added data objects without the value of dependent (dichotomous) variable are also classified to most appropriate (gender) group. It is clear, the success of classification of the real LOR models is less than 100 %. Estimated values of gender (probability to be Female) are depicted in Figure 8 b). In Table 6, the number of data objects classified to a given gender group and has its own (true) gender are illustrated. Columns 'predicted' denote estimated values of gender (based on the size of shoes) and rows illustrate true gender values (known from data). Here, all data objects are classified successfully, i.e. true male participants were estimated (based on the size of shoes) as male and vice versa.

Table 6 - Confusion matrix of estimated gender from LOR model

Confusion matrix

	Predicted	
Observed	M	F
M	16	0
F	0	16

In Table 7, estimated values of LOR coefficients are provided. These values serve to estimate the gender of newly added data objects only with knowledge of the size of shoes (where true gender is not known). Value of shoe size put into simple linear combination ($869 - 21.2 * shoes\_size$) and resulting value compare with a given limit value (probability), typically set as 0.5 (it classifies data object as male if p<0.5 or female p>0.5).

Table 7 - LOR estimated coefficients

	Estimate	**Standard Error**
(Intercept)	868.76	893053.97
shoes	-21.72	22309.52

Very popular statistical method belonging to multidimensional analysis approaches is **discriminant analysis (DA)**. DA enables (similarly to LOR) to design a rule for classification data objects to independent groups. The optimal classification method is called Linear Discriminant Function (LDF) only if three assumptions have to be valid: 1. differences between group mean values have to be significant, 2. variances of variables in groups have to be statistically equal, and 3. data variables are from Gaussian distribution. A number of groups for classification is not required, but for simplicity, an example with two groups is illustrated. Classification rule is constructed using the size of shoes and speed of swimming where two groups for classification are defined by gender. Typically, a portion of data objects is taken as a training set, and remaining data are testing set (verification). In this case, 32 data objects are divided into 26 for training and 6 for verification.

Test of equality of mean values in groups rejects null hypothesis about group-similarity for both variables. Test of equality covariance matrices (variances) do not reject the null hypothesis, and this assumption is also verified.

Tests of Equality of Class Means				
	F	df1	df2	p
shoes	166.99	1	30	< .001
swimming	59.65	1	30	< .001

Tests of Equality of Covariance Matrices			
	χ^2	df	p
Box's M	4.04	3	0.26

Finally, six data objects (not used in training) are classified by estimated classification rule. We can see that all objects are classified successfully (males as males and vice versa).

Confusion Matrix			
		Predicted	
		F	M
Observed	F	3	0
	M	0	3

The illustration of objects-classification is depicted in Figure 9, where standardised values of measured variables (axes) are used, and LDF is also showed (border between coloured areas). This picture shows that in the cases where data points from two groups are not separated, LDF is not able to be 100 % successful.

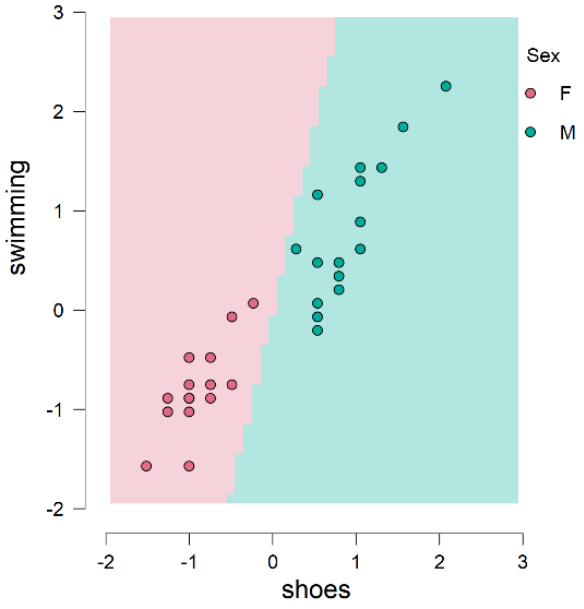


Figure 9 - Decision boundary matrix of DA

In previous paragraphs, methods of classification of data objects into separate groups were introduced (we constructed a rule for classification). Now, the separation of objects into a given number of groups (clusters) will be proposed in **cluster analysis (CA)**. There are two main approaches to cluster data objects in CA – hierarchical and non-hierarchical.

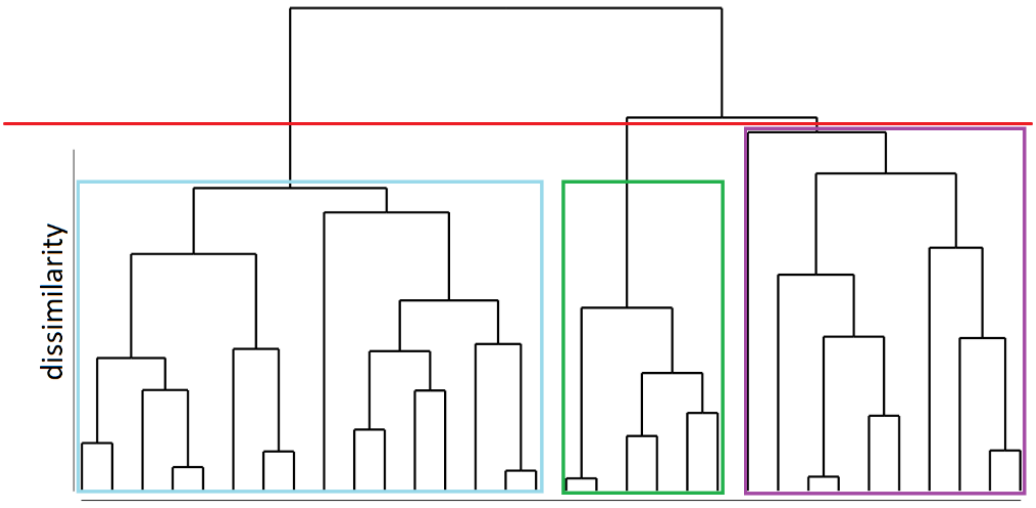


Figure 10 - Example of the dendrogram

At the beginning of hierarchical (connectivity-based) CA, each data object is in one own cluster. Then, the two most similar clusters are joined together. The similarity is typically measured as the reciprocal distance between all pairs of clusters. There exist several various approaches to join the most similar clusters (single linkage, complete linkage, etc.). The joining of pairs of clusters continues until all data objects are in one cluster. The joining process is typically illustrated by the dendrogram. Here, an example of clustering data objects (people) based on earnings, and beer and wine consumption is illustrated in Figure 10. Information, how many clusters are the best for this model can be achieved from the task (two clusters representing gender, etc.). When the appropriate number of clusters is known, the cut of the dendrogram (red line for three clusters) divides data objects into three clusters (blue, green and purple).

The most popular method in the non-hierarchical clustering approach is known as a *k*-means algorithm. The number of clusters (*k*) has to be known before clustering – from the task or numerical *elbow method* (Figure 11). Three different statistical criteria for an estimate the optimal number of clusters are depicted as the least value on curves in the vertical axis (redpoint for BIC criterion). For this example, the optimal number of clusters is k=3.

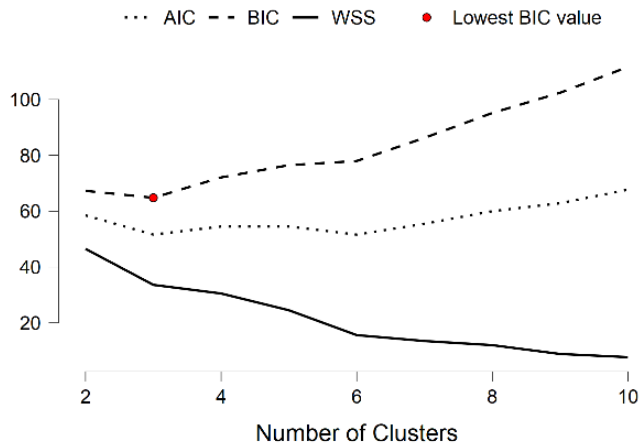


Figure 11 - Results of elbow method for CA

After that, all objects are clustered to the nearest of the three centroids, which were randomly selected from these objects (it exists variant, where centroids are selected as random coordinates in data area). Then, new centroids are re-computed (mean values of the clusters) and if any data object has to be re-

allocated (current centroid is not the nearest one), the model is re-arranged. After this process, data are clustered. Because randomisation in the centroid initialisation, in some tasks, repeated analysis in the same data results in different re-arrangement (data classification).
Here, the elbow method estimated three clusters as the optimal number. In Figure 12, the plot of labelled objects is illustrated with three clusters (some statistical software provides labels in dendrogram).

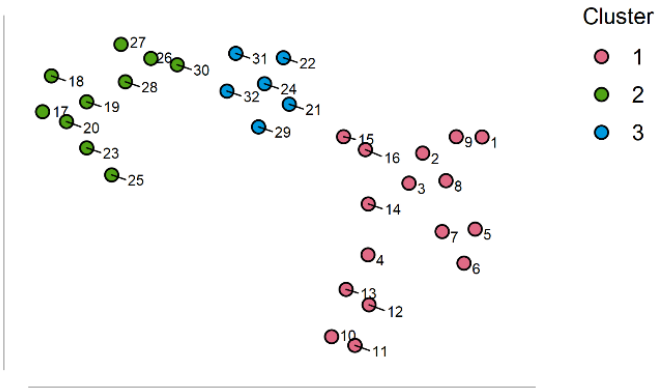


Figure 12 - Illustration of data objects clustering in CA

Interesting information provides a plot of clusters' mean values (Figure 13). All clusters are characterised by all variables, where the positive (bigger) value in the plot denotes bigger values for a given variable in this cluster. Cluster 1 is represented by small earnings, small consumption of beer and big consumption of wine, etc.

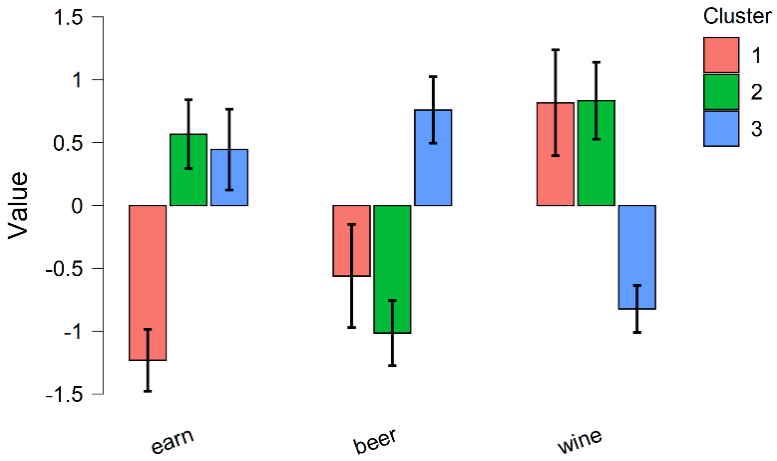


Figure 13 - Mean values of clusters in CA

The last two multidimensional data analysis methods presented here are **Principal Component Analysis (PCA)** and **Factor Analysis (FA)**. Both methods are focused on the reduction of data dimensionality, i.e. the decreasing number of measured variables. The main idea of PCA is to reduce the number of variables based on their variances. The total variance of the model is computed as the sum of variances of variables. The first newly constructed variable is called the first principal component (PC1), and it contains the biggest portion of the total model variance. Then the second principal component contains the biggest portion of the remaining part of the total variance, etc. The number of principal components is equal to the number of original variables, but only some PCs

contains a bigger portion of variability than the original variables. These components are interesting for PCA, and the number of such components typically means the new dimensionality of the task. Notice that each PC is created from all the original variables; therefore, the loss of dimensionality does not mean loss of the original variables. The number of new principal components is given by the task, or it is possible to estimate it using *Eigenvalues*. The Eigenvalues bigger than one denote component with a portion of variance bigger than the original variable.
We apply PCA for data of eight variables measured on 32 people (height, weight, shoes, age, earn, beer, wine, swimming).

Table 8 - Components characteristics in PCA

	Eigenvalue	Proportion var.	Cumulative
PC1	4.82	0.60	0.60
PC2	1.63	0.20	0.81
PC3	1.27	0.16	0.96

In Table 8, component characteristics illustrate the number of PCs and their efficiency (size of Eigenvalues). When we reduce data of eight variables, then the sum of values in the correlation matrix is also eight. Then PC1 has efficiency almost as five original variables (≈ 5), PC2 and PC3 have efficiency more than the original variables. Three first PCs provides information about almost of the original data (sum of the Eigenvalues is 7.8), which is a good result.
Similarly, the Eigenvalues of all PCs are depicted in Figure 14 (numbers are on the horizontal axis), for better illustration of the PCA model. Very good usage of PCA results is in case of scatter plot, where more than two variables. PCA reduces original data into two PCs, and these components are used for the construction of the scatter plot.

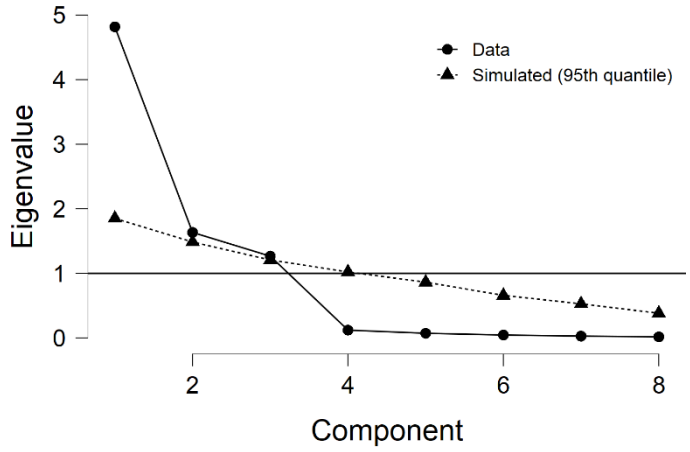


Figure 14 - Plot of Eigenvalues in PCA

Factor analysis also provides a reduction of data dimensionality, where newly constructed *factors* replace the original variable. In FA, the reduction is based on covariance between the original variables. Here, the main role is played by factor loadings (coefficient of a factor in the original data decomposition). Newly constructed factors (their number is known or achieved by PCA) are based on all original data variable, and each factor is loaded (i.e. loadings) by each variable in a given portion. In other words, each original variable divides its loadings among all factors. The factor is composed of variables, which achieves the biggest loadings for this factor.

Table 9 - Factor loadings for FA

	Factor 1	Factor 2	Factor 3	Uniqueness
height	**0.95**			0.03
weight	**0.95**			0.05
shoes	**0.95**			0.02
age		**0.88**		0.00
earn	0.51	**0.81**		0.08
beer	**0.83**		-0.49	0.06
wine			**0.82**	0.23
swimming	**0.92**			0.06

From Table 9 (factor loadings) is clear that factors are loaded relative unambiguously (only earn and beer are located in two factors). If the loadings of the variable are very similar, it is possible to use some rotation (Varimax, etc.). Better insight into factor loadings provides Path diagram (Figure 15). Here, the thickness of the arrow is based on the size of factor loadings (green for positive and red for negative). Factor 1 is for height, weight, size of shoes, beer and swimming. Factor 2 is for age and earnings, and Factor 3 is for wine consumption. Then we can name the factors by some logical labels. PCA and FA are not so exact methods of statistical tests because there is no decision (significance) of the achieved results. Researchers should carefully study the numerical results to achieve the most appropriate results.

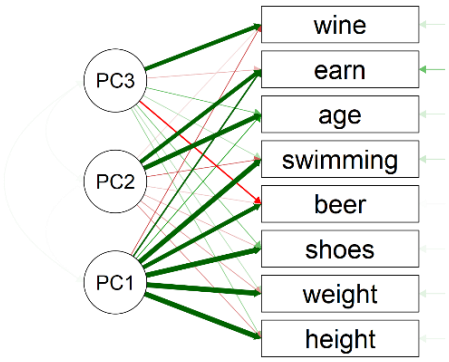


Figure 15 - Path diagram for FA

Design of the experiment, measurement, analysis, conclusion

To achieve appropriate answers for the predefined research questions, researchers have to abide by several rules. Previously introduced statistical methods are efficient when given assumptions are checked.

Preparation of **data measurement** is a very important part of research before analysis because if data are measured in a wrong way or some data are missing, and even the best analytic is not able to construct appropriate conclusion of the analysis. Typically, a deeper consideration or discussion with an expert on data analysis is sufficient. It is possible to measure data in many ways, questionnaire or physical measuring (people, animals, machines, etc.) are the most typical ways.

The steps performed before data analysis in the research are called **experimental design** or design of the experiment (DOE) [5]:

1. Definition of research interest is focused on questions, variables, and objects. The main idea before the measurement is what we want in the research to prove (‘Compare physiological aspects of male and female athletes in a given sport', 'Study the influence of digital devices on children behaviour in school', etc.). Although many researchers often deal with topics which are known for them, it is recommended to write this description (idea) exactly because it results in the next step. Naturally, variables are specified to be measured for support the goal of the research ('gender', 'height', 'weight', etc.). It is important to think about all possible aspects (features, characteristics) which can play a role in the research process.
2. Prepare a research hypothesis. In this phase, the research ideas are defined in more details. The global research idea is divided into several small research hypothesis – each one will be (statistically) tested in the analytical phase. Research hypothesis contains a particular statement regarding the research interest (i.e. 'Men athletes are heavier than women', 'Playing PC games causes worse study results', etc.). For each research hypothesis, the null hypothesis is constructed as a neutral statement ('Men and women athletes have similar weight', 'Playing PC games have no influence on children school results', etc.). The research hypothesis is also called the alternative hypothesis.
3. Prepare the measurement of variables. Some research measurement decided this point because sometimes there is only one possible way to measure a given variable (gender is only Male or Female, etc.). But generally, measured variables are limited in several aspects. Researchers are often able to decide if the variable will be qualitative or quantitative (for example age can be represented as an exact number or category' 20-39', speed of the car can be an exact number or category '100-130 km/h', etc.). If it is possible, it is recommended to measure numerical variables as numbers that are simply transformed to a categorical scale. Unfortunately, the categorical variable is not possible to transform into a number. The accuracy of measurement varies according to the particular variable, i.e. if the variable range is $(0, 0.1)$, accuracy should be at least two decimal places or more, etc.
4. Definition of the experimental settings serves to provide significance of research results. If the research enables to measure an arbitrary number of objects, it is recommended to estimate an appropriate size of the sample. Generally, a higher number of objects means more significant results. Nevertheless, in many research areas, there is not possible to measure a lot of objects (it is expensive, time-consuming, etc.). All objects should be selected to data file randomly to achieve statistically significant results. Unfortunately, in the medical researches, the number of patients is very small to select some sub-groups. Some researches enable to divide objects into the treatment and control group to assess differences of measured variables.

Data in the research are measured variously. It is necessary to guarantee the accuracy of measuring machines (meters) and settings of simulation devices, etc. The calibration is performed in many ways. The simplest way is to measure the known value using the device and compare the measured value with the true value.

Before data analysis, data have to be cleaned, where spurious and outlying values are checked and corrected (third gender value, negative earnings, etc.). If the correction is not possible, the whole object or variable should be removed before analysis. Also, it is possible to transform variables to different units or scales to compare results with other research studies (km/h to mi/h, etc.).

When data are clean, the decision and selection the most proper statistical methods (tests) are performed. Many researchers underestimate this step, and a lot of reports of research are rejected from publication because of the bad analytical part. Authors of study [6] proposed a complex survey

to avoid researchers from making the wrong steps in data analysis. Some methods were described in previous paragraphs. Nevertheless, in the case of more specific researches, more specific methods provide better results and insight into the research problem.

When the analysis of measured data is completed, numerical and graphical results need to be enriched by clear description and interpretation. Here, researchers have to summarise ideas before, during and after performed research. Whether the expected goal was achieved, and if yes, how new ideas were achieved and what are reasons for the newly achieved level of the research area. The conclusion should help the reader to understand the main ideas of the presented part of the research. It should be clear, and it includes all the significant steps of performed research. Standardly, the conclusion contains new possible ways to extend current results.

Statistical software

Although it is possible to perform selected statistical tests manually (using calculator and tables of standard distributions), it is more comfortable and faster to employ some of the offered statistical packages (software). From a price point of view, statistical software is divided into **licensed** (user pay before using) and **free** (available without payment). From a user-interface point of view, software functions are controlled by a graphical **menu** or from the **console**. There are rather small technical (functional) differences between various licensed software, especially in graphical user interface and format of reports (results). The licensed statistical software is popular mainly in the commercial or academic sphere. Mostly, in the individual areas of research, different packages for the researchers are used (it is not a rule, but it is an adaptation the software to a researchers' requirements). Very popular licensed statistical packages with a graphical interface are MiniTab[1], NCSS[2], SPSS[3], Statgraphics[4], Stata[5], Statistica[6], Systat[7], TriloByte[8], etc. Commercial statistical software controlled by console is mainly Matlab[9]. Popular open-source software with a graphical interface is Gretl[10], Jasp [11]or Mystat (student version of Systat). Very popular open-source console statistical software is R project [12]often used with R studio [13]interface. Besides these general statistical packages, researchers often use software focused on particular statistical methods.

[1] http://www.minitab.com/en-us/
[2] https://www.ncss.com/software/ncss/
[3] https://www.ibm.com/products/spss-statistics
[4] https://www.statgraphics.com/
[5] https://www.stata.com/
[6] http://www.statsoft.com/
[7] https://systatsoftware.com/
[8] https://www.trilobyte.cz/
[9] https://www.mathworks.com/products/matlab.html
[10] http://www.learneconometrics.com/gretl/index.html
[11] https://jasp-stats.org/
[12] https://www.r-project.org/
[13] https://rstudio.com/

Reference

[1] T. Ritchey, "Analysis and Synthesis: On Scientific Method - Based on a Study by Bernhard Riemann," *Systems Research,* vol. 8, no. 4, pp. 21-41, 1996.

[2] R. A. Fisher, Statistical Methods for Research Workers, Edinburg: Oliver and Boyd, 1934.

[3] A. a. B. B. S. Zulfiqar, "Basic statistical tools in research and data analysis," *Indian society of anaesthesiologists,* vol. 60, no. 9, pp. 662-669, 2016.

[4] W. B. Michael and S. Hunka, "Research Tools: Statistical Methods," in *Review of Educational Research*, vol. 30(5), 1960, pp. 440-486.

[5] L. Popoola, O. Olorunisola and O. Ademowo, *Data Collection, Management and Analysis in Academic Research,* 2009.

[6] R. M. Khusainova, Z. V. Shilova and O. V. Curteva, "Selection of Appropriate Statistical Methods for Research Results Processing," *Mathematics Education,* vol. 11, no. 1, pp. 303-315, 2016.