# Výuka statistiky s Excelem
# Statistics Lectures in Excel

Oldřich Kříž, Jiří Neubauer, Marek Sedlačík

## 1. Introduction

An elementary course of Statistics belongs to basic subjects taught at most of the economic faculties. This course provides a fundamental economic education. In spite of the fact that it is focused on non-statisticians, it is necessary to pay attention to this education because it contributes in a specific way to the creation of general economic thinking.

In this contribution the authors are thinking of the way of teaching in which students would be able to understand the basis of statistics easily and naturally, would succeed in getting into principles of statistics, and would not be afraid of using statistical tools effectively and reasonably. It is not a good feedback for the teacher if almost all students are afraid of Statistics even if they consider some parts of Statistics as Mathematics.

## 2. Statistical software and teaching

Nowadays, it is not any novelty that the use of modern computer technique in teaching Statistics is almost obvious. Scientific calculators were used for numerical calculation not a long time ago and they are a helpful tool even at the present time. However, the statistical analysis via computer provides more useful outputs. Besides numerical calculation we can mention tables and graphs which describe the range of monitored variable characteristics. These outputs help to the students understand philosophy of Statistics. The computers which entered this area several years ago, are essential for many contemporary statistical methods.

We can find a lot of statistical software on the market, such as Statistica, SPSS, Statgraphics, Unistat, QCExpert, Adstat, Matlab (statistics toolbox) and so on. These products differ for example in a range of offered methods and analysis, in graphical interfaces, in user's accessibility, or in universality. The statisticians usually use the software which they are familiar with, or which is available in a workplace. The aim of this contribution is not any description or comparison of the statistical software. We would like to analyse how it can be used in teaching basic Statistics, which comprises construction of statistical graphs, computation of descriptive measures of data sets, normality tests, confidence intervals, hypothesis testing, analysis of variance and linear regression. Such dataset analysis can be made in any statistical software.

Provided the teacher of Statistics asks a question which of the mentioned software is to be used in the lessons, two basic problems should be solved. The first problem is the accessibility of the chosen software at school since the price of required multi-licence is not often low. The second one is the accessibility of this software for the students because they should be able to use it not only at school but also at their homes or dormitories if we want to make their study more effective.

The authors of the contribution have tried technical and didactical advantages of Unistat, QCExpert and Matlab. The first two software products are not very difficult to operate as well as orientate in the offered structure of methods or particular techniques. The graphical interface is also sufficient for the teaching in a basic course of Statistics. Moreover, both of them are explained in Czech language. They can be used as an aid during the course. Matlab is a well-known and widely dispersed product with a huge range of application in different exact disciplines. However, special toolboxes are required and knowledge of Matlab language is necessary. It can be recommended for advanced parts of Statistics. The common disadvantage of the given products is their bounded use of individual students' work at their homes. They are not accessible for students at our faculty which causes that the computer aided teaching ends out of the computer classroom.

## 3. Statistics in Excel

This handicap of statistical software could be solved at least partially by a familiar product Excel, the part of MS Office. An indisputable advantage is that it is mass-used among students and also classrooms are equipped with Excel. It is not special statistical software. However, the tabular nature of Excel enables to utilize several implemented tools and interesting attributes.

In the first instance, there is the extensity of using it for own numerical computation via equations defined by the user. Excel is capable of dealing with matrix calculus; furthermore, we can clearly organize different numerical outputs into tables which are widely used in statistics. Another asset is that the user – in our case a student – works interactively and that he/she sorts out the outputs according to his/her requirements. Graphical abilities of Excel allow us to choose from 14 types of graphs, where each one could be demonstrated in different modifications.

Implemented procedures from various parts of statistics (known as analytic tools) could be used in many fields of statistics.  It is possible to apply packages of descriptive statistics, frequency distribution, one or two sample tests, analysis of variance, regression, correlation, etc. These procedures still give the same type of outputs and hence it is possible to comment on results easily. The widest offer of Excel is in category of implemented statistical functions. It is suitable to incorporate the functions into the statistical lectures gradually and with didactical intention to use their basic property, which is instantaneous conversion of outputs on the change of inputs. On that account, the students could combine different functions, test the behaviour of data sets or models and penetrate into the statistical philosophy.

On the other hand, there are some drawbacks of Excel. Statistical functions have a bit vague and sometimes misguided terminology of functions description and of the help too.  Except the terminology, there is another problem that we face us from a didactic point of view. Above all, that concerns divisions during finding inverse functions of some probability distributions. This disunity of entering parameters of functions is illusory and it could cause troubles for students and sometimes also computing errors. More inconveniences of this sort can be found.

Primarily, the authors did not conceal the ambition to put together an application in Excel, which could eliminate the problems mentioned above and which could sufficiently support statistics lectures, especially in case of describing data sets and practical use of estimates and tests with regard to concrete problems. The ground of this application is formerly used worksheet [3] which is extended by a data list and some additional statistical procedures – see later. The data list allows us to use the application not only for own data sets, but also for chosen data from the offered list of variables. Moreover, there is a possibility to apply selected procedures either separately to the data or separately to inserted characteristics, which is very important considering the didactics of statistics.

## 3.1 Worksheet STAT1

To begin with, let us introduce the working principle and individual lists of prepared aid STAT1. In each computational list is the item menu of variables which could undergo the processing; furthermore, we have to select required parameters of a given assignment (highlighted in red colour), for example the significance level, various constants etc. Statistical outputs and results are in green cells. The first sheet “*data*” contains not only data sets from a certain statistical exercise and textbook but of course, it is possible to paste your own values with an appellation. The choice of concrete data for given processing is available separately in each sheet. The second sheet “*descriptive measures*” offers all standard descriptive characteristics, including Q-Qplot and tests of normality, see figure 1. Next two sheets “*discrete data*”and “*continuous data*” (see figure 2) allow us to explore date with respect to the data type. You can find there frequency distribution table, descriptive measures and parallel graphs. The following sheets pursue estimation and hypothesis testing for one or two populations with reference to the data type, see figures 3, 4 and 5. The last sheet “*tables*” provides us with quantiles and functions of commonly used distributions. An additional asset is the fact that all outputs are commented in short verbal sentences and the user could choose an appropriate conclusion regarding the problem.

After a brief introduction of the application STAT1, we can proceed to the description of a real data processing. More precisely, let us consider the following representative problem.

Assume a minor road in the neighbourhood of an elementary school. On account of several traffic accidents and several complaints of parents, there was conducted a speed measurement in the mentioned place. A measured data set of 50 observations is in the table 1 – *speed 1*. According to a head teacher’s announcement, it is inevitable to investigate whether drivers respect the speed limit and possibly what is the proportion of drivers who exceed the limit of 50 km per hour. Moreover, a new speed bump was installed as a consequence and a speed measurement of 60 observations was realized repeatedly (see table 1 – *speed 2*). Perform an exhaustive statistical data processing with respect to the problem, utilize the aid STAT1.

| ***speed 1*** | | | | ***speed 2*** | | | |
|---|---|---|---|---|---|---|---|
| 54.1 | 55.3 | 50.4 | 48.3 | 46.7 | 51 | 43.6 | 49.9 |
| 48.1 | 54.3 | 63.9 | 36.1 | 46.8 | 39.1 | 45 | 54.1 |
| 46.4 | 45.3 | 47.5 | 52.9 | 44.8 | 46.6 | 42.5 | 45.4 |
| 44 | 52.6 | 45.9 | 48.7 | 52.1 | 51.6 | 48.7 | 44.1 |
| 49.2 | 45.6 | 46.8 | 45.8 | 39.8 | 43.2 | 47.7 | 48.6 |
| 50.3 | 40.8 | 61.3 | | 49.8 | 42.4 | 45.9 | 43 |
| 52.6 | 56.7 | 39.9 | | 43.3 | 50.1 | 48.6 | 39.6 |
| 53.3 | 59.6 | 49.6 | | 46.1 | 43.7 | 53.5 | 46.3 |
| 39 | 51.5 | 48 | | 46.2 | 46.8 | 43.7 | 44.7 |
| 57.8 | 44.9 | 43.4 | | 46.3 | 48.7 | 53.5 | 41 |
| 47.6 | 43.4 | 47.6 | | 47.1 | 51.4 | 48.7 | 45.3 |
| 57 | 57.8 | 49.7 | | 50.1 | 42.4 | 43.1 | 45.6 |
| 49.5 | 50 | 50.4 | | 55.9 | 41.2 | 48.3 | 42.5 |
| 41.3 | 54.9 | 58.1 | | 37.7 | 46.9 | 47.7 | 46.6 |
| 38.9 | 49.4 | 44 | | 52.6 | 44.4 | 46.2 | 42.7 |

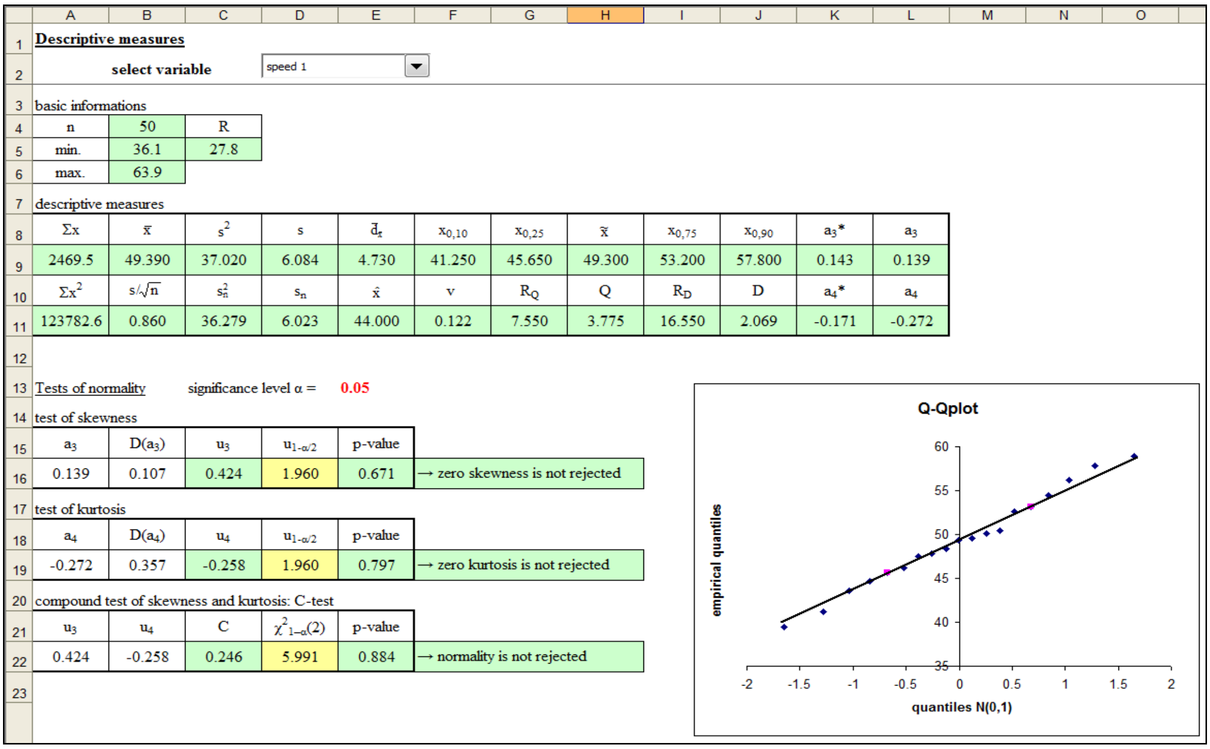
Tab. 1: Data sets

### 3.1.1 Organizing data

The first step is to paste the data from the table 1 into the first sheet “*data*” of STAT1. We use corresponding appellations *speed 1* and *speed 2*.

### 3.1.2 Descriptive statistics, frequency tables and graphs

The sheet “*descriptive measures*” enables us to glance the structure of the data set *speed 1* which represents the behaviour of drivers before the installation of the speed bump. Standard descriptive characteristics, the Q-Qplot and tests of normality for a given significance level are available.



**Descriptive measures**

**select variable** speed 1

basic informations

| | | |
|---|---|---|
| n | 50 | R |
| min. | 36.1 | 27.8 |
| max. | 63.9 | |

descriptive measures

| $\Sigma x$ | $\bar{x}$ | $s^2$ | $s$ | $\bar{d}_x$ | $x_{0,10}$ | $x_{0,25}$ | $\tilde{x}$ | $x_{0,75}$ | $x_{0,90}$ | $a_3^*$ | $a_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2469.5 | 49.390 | 37.020 | 6.084 | 4.730 | 41.250 | 45.650 | 49.300 | 53.200 | 57.800 | 0.143 | 0.139 |
| $\Sigma x^2$ | $s/\sqrt{n}$ | $s_n^2$ | $s_n$ | $\hat{x}$ | v | $R_Q$ | Q | $R_D$ | D | $a_4^*$ | $a_4$ |
| 123782.6 | 0.860 | 36.279 | 6.023 | 44.000 | 0.122 | 7.550 | 3.775 | 16.550 | 2.069 | -0.171 | -0.272 |

Tests of normality significance level $\alpha$ = 0.05

test of skewness

| $a_3$ | $D(a_3)$ | $u_3$ | $u_{1-\alpha/2}$ | p-value | |
|---|---|---|---|---|---|
| 0.139 | 0.107 | 0.424 | 1.960 | 0.671 | → zero skewness is not rejected |

test of kurtosis

| $a_4$ | $D(a_4)$ | $u_4$ | $u_{1-\alpha/2}$ | p-value | |
|---|---|---|---|---|---|
| -0.272 | 0.357 | -0.258 | 1.960 | 0.797 | → zero kurtosis is not rejected |

compound test of skewness and kurtosis: C-test

| $u_3$ | $u_4$ | C | $\chi^2_{1-\alpha}(2)$ | p-value | |
|---|---|---|---|---|---|
| 0.424 | -0.258 | 0.246 | 5.991 | 0.884 | → normality is not rejected |

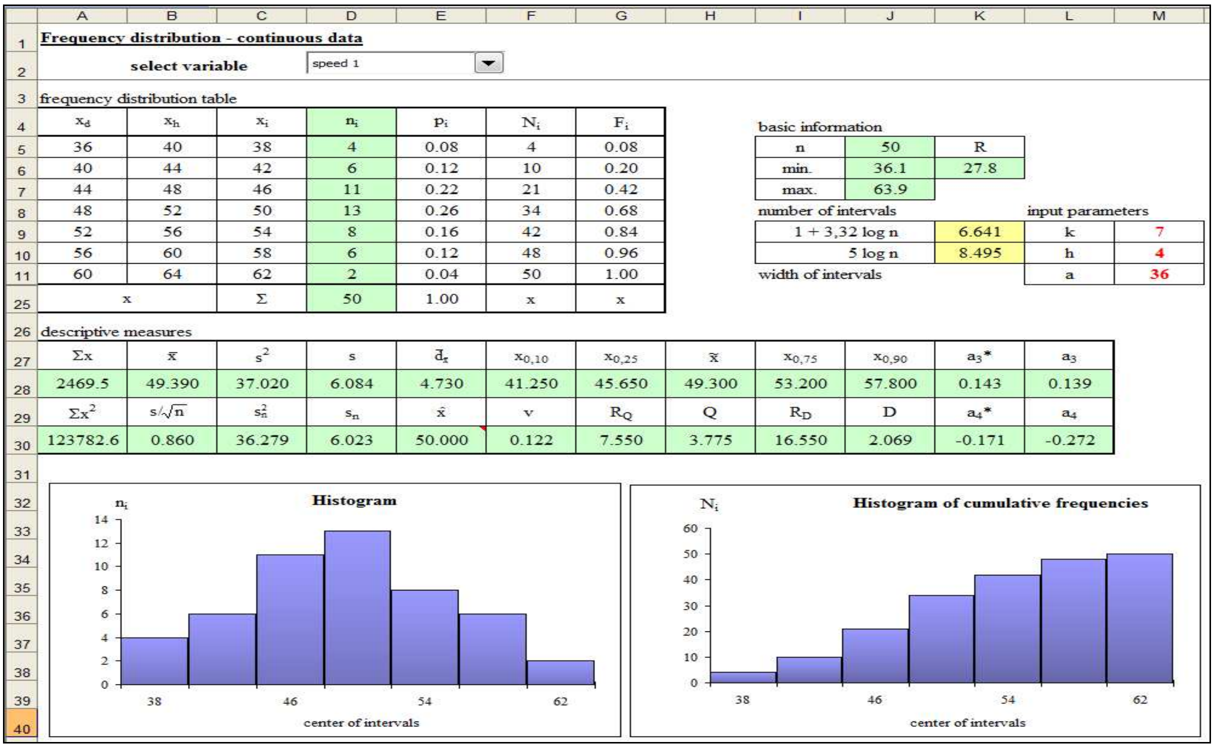
Fig. 1: Descriptive measures based on the sample

Because of the type of the data set *speed 1*, we select the sheet “*continuous data*” as a way to investigate the frequency distribution. The frequency distribution table, descriptive measures and histograms for chosen input parameters are produced.

### 3.1.3 Point estimators and confidence intervals

In addition to the descriptive statistics, the worksheet STAT1 provides point and interval estimates of the population mean, the variance and the standard deviation of the normal distribution, the estimates of the mean based on large samples are also included. These estimates are computed in the sheets “*one sample – normal*” and “*one large sample*”. It is possible to calculate the given estimates using datasets (sheet “*data*”), or just insert values of necessary point estimates, such as the sample mean, the sample variance or the sample standard deviation and, of course, the sample size.

We illustrate the possible using of STAT1 on the data set *speed 1* which contains the speed velocity of 50 cars on the road near the school before the speed bump was installed, see the figure 3. After the selecting of the proper dataset and choosing the confidence level $\alpha$, all necessary outputs are automatically computed. The two-sided 95% confidence interval of the mean of the *speed 1* is (47.661; 51.119) km per hour.



Frequency distribution - continuous data

select variable: speed 1

frequency distribution table

| $x_d$ | $x_h$ | $x_i$ | $n_i$ | $p_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|---|---|
| 36 | 40 | 38 | 4 | 0.08 | 4 | 0.08 |
| 40 | 44 | 42 | 6 | 0.12 | 10 | 0.20 |
| 44 | 48 | 46 | 11 | 0.22 | 21 | 0.42 |
| 48 | 52 | 50 | 13 | 0.26 | 34 | 0.68 |
| 52 | 56 | 54 | 8 | 0.16 | 42 | 0.84 |
| 56 | 60 | 58 | 6 | 0.12 | 48 | 0.96 |
| 60 | 64 | 62 | 2 | 0.04 | 50 | 1.00 |
| x | | Σ | 50 | 1.00 | x | x |

basic information

| | | |
|---|---|---|
| n | 50 | R |
| min. | 36.1 | 27.8 |
| max. | 63.9 | |

number of intervals

| | |
|---|---|
| 1 + 3,32 log n | 6.641 |
| 5 log n | 8.495 |

width of intervals

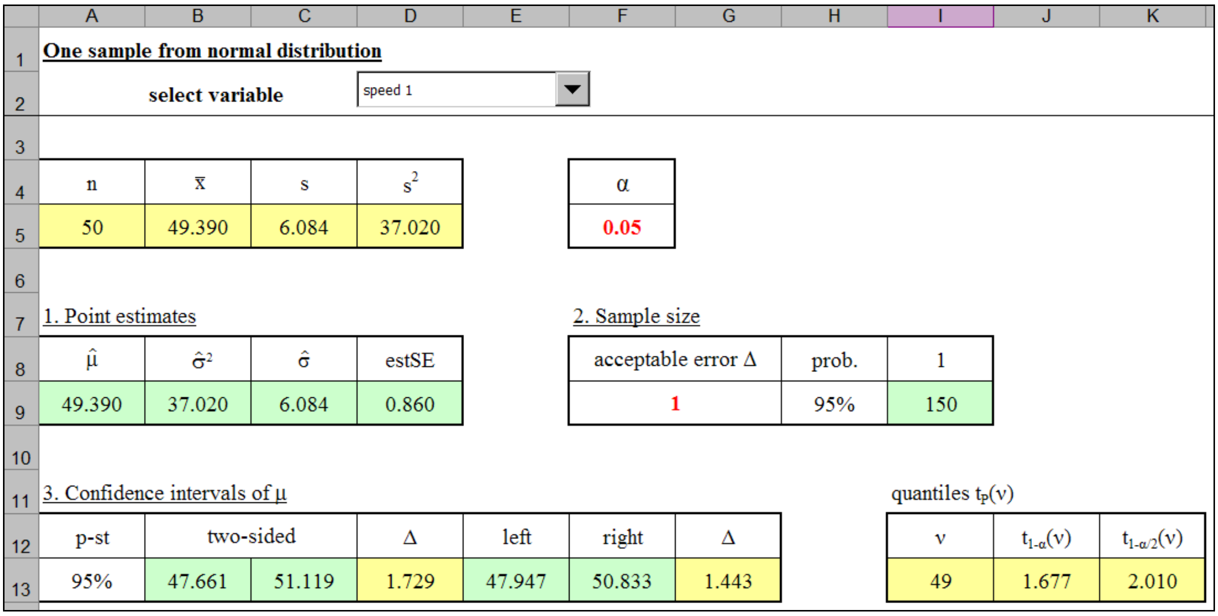
input parameters

| | |
|---|---|
| k | 7 |
| h | 4 |
| a | 36 |

descriptive measures

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Sigma x$ | $\bar{x}$ | $s^2$ | s | $\bar{d}_{\bar{x}}$ | $x_{0,10}$ | $x_{0,25}$ | $\tilde{x}$ | $x_{0,75}$ | $x_{0,90}$ | $a_3^*$ | $a_3$ |
| 2469.5 | 49.390 | 37.020 | 6.084 | 4.730 | 41.250 | 45.650 | 49.300 | 53.200 | 57.800 | 0.143 | 0.139 |
| $\Sigma x^2$ | $s/\sqrt{n}$ | $s_n^2$ | $s_n$ | $\hat{x}$ | v | $R_Q$ | Q | $R_D$ | D | $a_4^*$ | $a_4$ |
| 123782.6 | 0.860 | 36.279 | 6.023 | 50.000 | 0.122 | 7.550 | 3.775 | 16.550 | 2.069 | -0.171 | -0.272 |

Fig. 2: Frequency distribution based on the sample



One sample from normal distribution

select variable: speed 1

| n | $\bar{x}$ | s | $s^2$ |
|---|---|---|---|
| 50 | 49.390 | 6.084 | 37.020 |

| α |
|---|
| 0.05 |

1. Point estimates

| $\hat{\mu}$ | $\hat{\sigma}^2$ | $\hat{\sigma}$ | estSE |
|---|---|---|---|
| 49.390 | 37.020 | 6.084 | 0.860 |

2. Sample size

| acceptable error Δ | prob. | 1 |
|---|---|---|
| 1 | 95% | 150 |

3. Confidence intervals of μ

| p-st | two-sided | | Δ | left | right | Δ |
|---|---|---|---|---|---|---|
| 95% | 47.661 | 51.119 | 1.729 | 47.947 | 50.833 | 1.443 |

quantiles $t_p(\nu)$

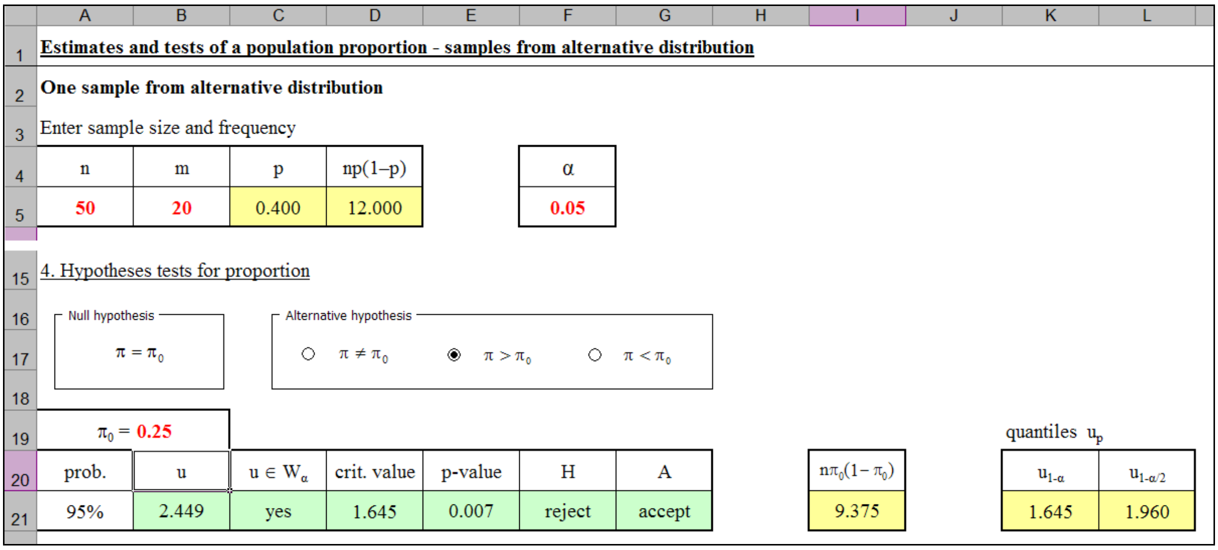
| ν | $t_{1-\alpha}(\nu)$ | $t_{1-\alpha/2}(\nu)$ |
|---|---|---|
| 49 | 1.677 | 2.010 |

Fig. 3: Point estimates and confidence intervals based on the sample from the normal distribution

### 3.1.4 Hypothesis testing

The parents claim that at least 25 % of drivers do not obey the speed limit of 50 km per hour. Using the test of proportion we can decide whether they are right or not. The number of cars, whose measured speed was above the given limit, is 20. The point estimator of the probability

$\pi$ that the speed of the car exceeds the limit of 50 km per hour is 20/50 = 0.4. We can test the hypothesis that the probability is 0.25 against the alternative hypothesis that the probability is larger than 0.25 which describes the fact the parents say. We choose the significance level 0.05. All results are summarized in the figure 4. We can select another alternative hypothesis by switching among them using the mouse. The test statistic is 2.449, the critical value of the test is equal to 1.645, p-value is 0.007. According to these results we can reject the null hypothesis that the probability is 0.25 and accept the hypothesis that the probability is larger on the given significance level. We must admit that the parents were right.



**Estimates and tests of a population proportion - samples from alternative distribution**

**One sample from alternative distribution**

Enter sample size and frequency

| n | m | p | np(1–p) |
|---|---|---|---|
| 50 | 20 | 0.400 | 12.000 |

| α |
|---|
| 0.05 |

4. Hypotheses tests for proportion

Null hypothesis: $\pi = \pi_0$

Alternative hypothesis: ○ $\pi \neq \pi_0$ ● $\pi > \pi_0$ ○ $\pi < \pi_0$

$\pi_0$ = 0.25

| prob. | u | u ∈ $W_\alpha$ | crit. value | p-value | H | A |
|---|---|---|---|---|---|---|
| 95% | 2.449 | yes | 1.645 | 0.007 | reject | accept |

| $n\pi_0(1-\pi_0)$ |
|---|
| 9.375 |

quantiles $u_p$

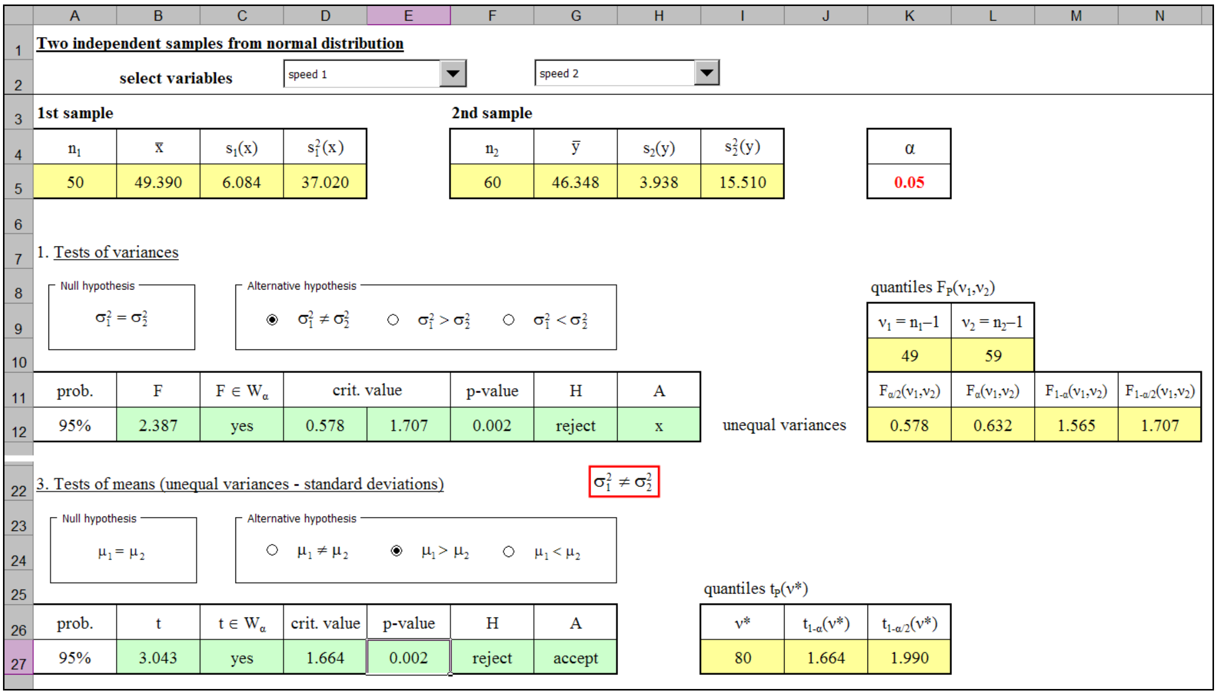
| $u_{1-\alpha}$ | $u_{1-\alpha/2}$ |
|---|---|
| 1.645 | 1.960 |

Fig. 4: The test of the population proportion

The new speed bump was installed and we obtained 60 measurements of the car speed. One can expect that this new obstacle on the road causes the decrease in the car speed near the school. Using the worksheet STAT1 the testing is very easy and user-friendly. First of all, we have to select datasets – *speed 1*, *speed 2* – and determine the significance level $\alpha$. Before we start to test the hypothesis that the installed bump reduces the speed of the cars significantly, we should find out if the population variances are equal or not. Using the F-test, we get the result that they are unequal, see the figure 5. When the bump has a significant impact, the population mean before installation must be larger than the population mean after it. Under assumption of unequal variances we obtain the value of the test statistic 3.043, the critical value is equal to 1.664, p-value is 0.002, see the figure 5. According to these results we can claim that the installed speed bumps reduces the speed of the cars near the school significantly (significance level is 0.05).

## 4. Conclusion

The contribution deals with the introducing of one teaching tool based on Microsoft Excel which helps students with study of Statistics not only in the lessons with the teacher at school but also at home during their individual preparation for the lessons or the exam. The crucial point of teaching is not just repeating numerical calculations, but using the whole teacher's and student's potential in order to understand statistical procedures and methods. Provided lessons are led in the way in which a student is able to continue with the study at home on his or her own computer and the acquired knowledge can be verified and deepened. The author's main aim of this tool was to make basic statistical computation easy and intuitive. The described worksheet STAT1 covers descriptive statistics, point and interval estimates and hypothesis testing in the content of lectures in Statistics usually taught at economic faculties.



**Two independent samples from normal distribution**

**select variables** speed 1 ▼ speed 2 ▼

**1st sample**

| $n_1$ | $\bar{x}$ | $s_1(x)$ | $s_1^2(x)$ |
|---|---|---|---|
| 50 | 49.390 | 6.084 | 37.020 |

**2nd sample**

| $n_2$ | $\bar{y}$ | $s_2(y)$ | $s_2^2(y)$ |
|---|---|---|---|
| 60 | 46.348 | 3.938 | 15.510 |

| α |
|---|
| 0.05 |

1. Tests of variances

Null hypothesis: $\sigma_1^2 = \sigma_2^2$

Alternative hypothesis: ● $\sigma_1^2 \neq \sigma_2^2$ ○ $\sigma_1^2 > \sigma_2^2$ ○ $\sigma_1^2 < \sigma_2^2$

| prob. | F | $F \in W_\alpha$ | crit. value | | p-value | H | A |
|---|---|---|---|---|---|---|---|
| 95% | 2.387 | yes | 0.578 | 1.707 | 0.002 | reject | x |

unequal variances

quantiles $F_P(\nu_1,\nu_2)$

| $\nu_1 = n_1-1$ | $\nu_2 = n_2-1$ |
|---|---|
| 49 | 59 |

| $F_{\alpha/2}(\nu_1,\nu_2)$ | $F_\alpha(\nu_1,\nu_2)$ | $F_{1-\alpha}(\nu_1,\nu_2)$ | $F_{1-\alpha/2}(\nu_1,\nu_2)$ |
|---|---|---|---|
| 0.578 | 0.632 | 1.565 | 1.707 |

3. Tests of means (unequal variances - standard deviations) $\sigma_1^2 \neq \sigma_2^2$

Null hypothesis: $\mu_1 = \mu_2$

Alternative hypothesis: ○ $\mu_1 \neq \mu_2$ ● $\mu_1 > \mu_2$ ○ $\mu_1 < \mu_2$

| prob. | t | $t \in W_\alpha$ | crit. value | p-value | H | A |
|---|---|---|---|---|---|---|
| 95% | 3.043 | yes | 1.664 | 0.002 | reject | accept |

quantiles $t_P(\nu^*)$

| $\nu^*$ | $t_{1-\alpha}(\nu^*)$ | $t_{1-\alpha/2}(\nu^*)$ |
|---|---|---|
| 80 | 1.664 | 1.990 |

Fig. 5: The test of means based on the samples from normal distribution

According to the results of questionnaire which was given to the students of Statistics at the Faculty of Economy and Management, University of Defence, the mentioned worksheet seems to be a helpful and useful tool for the students. More than 87 % of the students consider it to be a very good tool, 57 % of them used it actively during preparation for the lessons or doing their homework. 94 % of respondents think that the operating of this application is easy and natural, for 74 % of the students outputs are intuitive and logical. More than 90 % of respondents appreciate the possibility of using it for basic statistical data analysis of their own measurements. The authors of the described tool believe on the basis of their gained experience that the using of the tool in the lessons of Statistics is well founded above all for two reasons. It is known and widely-spread among students at our faculty and students are familiar with it. It also provides for individual student's work at school even at home background.

## References

1. KŘÍŽ, O., NEUBAUER, J., SEDLAČÍK, M. *Pravděpodobnost a náhodná veličina.* [Skripta]. 1. vyd. Brno : Univerzita obrany, 2007. 144 s. ISBN 978-80-7231-488-1.
2. KŘÍŽ, O., NEUBAUER, J., SEDLAČÍK, M. *Popisná statistika a výběrová šetření* [Skripta]. 1. vyd. Brno : UO, 2009. 181 s. ISBN 978-80-7231-707-3.
3. KŘÍŽ, O., NEUBAUER, J. Výuka statistiky podporovaná Excelem. In *Sborník XXV. mezinárodního kolokvia řízení osvojovacího procesu: sborník abstraktů a elektronických verzí příspěvků na CD-ROMu* [CD-ROM]. Brno, UO, 2007. Adresář: 6clanky/1krizo.pdf. ISBN 978-80-7231-228-3.
4. STINSON, C., DODGE, M. *Mistrovství v Microsoft Office Excel 2003*. Brno: CP Books, 2005. 890 s. ISBN 80-251-0669-1.

RNDr. Oldřich  Kříž
Department of Econometrics FEM, University of Defence
Kounicova 65, 612 00, Brno, Czech Republic
E-mail: Oldrich.Kriz@unob.cz
Telefon: + 420 973 443 334

Mgr. Jiří Neubauer, Ph.D.
Department of Econometrics FEM, University of Defence
Kounicova 65, 612 00, Brno, Czech Republic
E-mail: Jiri.Neubauer@unob.cz
Telefon: + 420 973 442 029

Mgr. Marek Sedlačík, Ph.D.
Department of Econometrics FEM, University of Defence
Kounicova 65, 612 00, Brno, Czech Republic
E-mail: Marek.Sedlacik@unob.cz
Telefon: + 420 973 443 591