Péter Kovács, PhD – Éva Kuruczleki

Statistics II

Learning guide

2019

Methodological expert: Edit Gyáfrás

This teaching material has been made at the University of Szeged, and supported by the European Union. Project identity number: EFOP-3.4.3-16-2016-00014

Contents

[Preface 3](#_Toc491019954)

[1. Sampling and introduction to the usage of SPSS 4](#_Toc491019955)

[2. Estimation 11](#_Toc491019956)

[3. Hypothesis testing 20](#_Toc491019957)

[3.1. One sample tests 21](#_Toc491019958)

[3.2. Two samples tests 25](#_Toc491019959)

[Review Section (Topic 1-3) 33](#_Toc491019960)

[4. Relationships, causal models 41](#_Toc491019961)

[4.1. Crosstabs analysis 42](#_Toc491019962)

[4.2. Analysis of variance 48](#_Toc491019963)

[4.3. Correlation and regression analysis 57](#_Toc491019964)

[Review Section (Topic 4) 64](#_Toc491019965)

.

# Preface

In order to understand news, social and business phenomena and our environment, interpret the relationships among social and business data correctly we need statistical literacy, reasoning and thinking which can include knowledge of basic statistical key figures, understanding concepts describing basic information about research methods, about visualization (about both visualization and interpretation) and the knowledge about data sources and the ability to evaluate the used data sources.

Very often we can’t examine the whole population directly, we have only sample(s) and we should infer the population properties from sample(s). After the first introductory semester, during this semester we concentrate on inferential statistics and the potential biases and errors, sampling methods, point and interval estimation of the population mean, proportion, standard deviation; hypothesis testing and relationship analysis among variables: crosstabs analysis, analyze of variance, correlation; casual models among scale variables: regression models. During the semester, we will use several data sources and IBM SPSS software to do the statistical analysis.

In the second semester probability and chance will have an important role, so it is crucial to be familiar with the terms of probability, expected value, density function, cumulative probability function and normal distribution.

The literature of the semester is Lind-Marchal-Mason: Statistical Techniques in Business & Economics (Eleventh Edition), McGraw Hill). Moreover, PowerPoint files and videos are available to support your learning. This document is a learning guide which contains the key terms, the source of the materials, suggested learning activities, sample exercises and solutions in each topic.

In order to review the previous elements a sample paper and an SPSS based test are available after the third and the fourth topics.

# Sampling and introduction to the usage of SPSS

Goals

**This chapter introduces the basic terms of statistical samples and the usage of the SPSS software at a basic level.** Learning of this chapter is successful if the Reader is able to

* distinguish among the population and a sample,
* explain the meaning of descriptive and inferential statistics,
* identify the types of samples
* use some basic menus (variable settings, data transformation, descriptive statistics) in SPSS.

**Knowledge obtained by reading this chapter:** basic terms of statistical sampling, basics of SPSS

**Skills obtained by reading this chapter:**

* statistical communication – basic terminology, making connections between statistical and everyday terms,
* organization – design, plan and carry out simple analyses following the necessary steps of statistical analyses using a new statistical software.

**Attitudes developed** **by reading this chapter**: openness towards the different forms of statistics, i.e. inferential statistics.

**This chapter makes the Reader to be autonomous in**: differentiation samples from the population, identifying variables and some SPSS analysis methods.

Definitions

**Population:** a collection of all possible individuals, objects, or measurements of interest.

**Sample:** a portion or part of the population of interest

**Descriptive statistics:** Describe the observed elements.

**Statistical inference:** Inferences to the populations which are based on the sample

(estimation, hypothesis testing).

**Representativeness:** A term used to describe the extent to which

different characteristics of a sample accurately represent the

characteristics of the population from which the sample was

selected.

**Representative sample:** A sample that is similar in terms of characteristics of the population to which the findings of a study are being generalized. A representative sample is not biased and therefore does not display any patterns or trends that are different from those displayed by the population from which it is drawn. It is rather difficult and often impossible to obtain a representative sample. Nonrandom samples usually tend to have a kind of bias. The use of a random sample usually leads to a representative sample.

**Probability Samples:** each member of the population has a known non-zero probability of being selected. Methods include random sampling, systematic sampling and stratified sampling.

**Random sampling** is the purest form of probability sampling. Types: simple random sample with replacement (each member of the population has an equal and known chance of being selected) and simple random sample without replacement.

**Stratified sampling** is a commonly used probability method that is superior to random sampling because it reduces sampling error. Stratum is a subset of the population that share at least one common characteristic, such as males and females.

**Cluster Sample**: a probability sample in which each sampling unit is a collection of elements.

**Nonprobability Samples:** members are selected from the population in some nonrandom manner. Methods include convenience sampling, judgment sampling, quota sampling, and snowball sampling.

**Snowball sampling** is a special nonprobability method used when the desired sample characteristic is rare.

Learning activities

In order to learn the basic terms

1. Read Chapter 8.1-8.4 from the book (Page 266-275)!
2. Open and explore 1\_sampling.ppt!
3. Read Chapter 8.5-8.6 from the book (Page 275-295)!
4. Explore Sampling distribution
5. Check your knowledge: solve the chapter exercises in the book!
6. Explore and solve the sample tasks!

Sample tasks

1. A part of a questionnaire is known below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *What is your gender?*

|  |  |
| --- | --- |
|  | 1-Male |
|  | 2-Female |

 |

 |
|

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *What is your training programme?*

|  |  |
| --- | --- |
|  | 1-Business Administration and Management |
|  | 2-Commerce and marketing |
|  | 3-Finance and accounting |
|  | 4-Other:  |

 |

 |
|

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Evaluate the lecturer based on the lectures! (9: do not know)*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Preparedness |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 1 | 2 | 3 | 4 | 5 | 9 |

 |

 |

 |

The answers comes from an online questionnaire system. The data can be found in the survey.xls file.

* 1. Import the data into SPSS!
	2. Set the properties of the variables!
	3. Prepare a frequency table by gender!
	4. Set values for the gender variable at the following way: 1 – male, 2 – female. Prepare the frequency table by gender again!
	5. Prepare descriptive statistics about the preparedness of the lecturer (frequency, mean, mode, median, standard deviation, skewness). Do the results make sense? What can be the reason?
	6. Set the ‘9’ value as a Missing value! Prepare the descriptive statistics again!
	7. Delete one data from each column randomly! Solve task e) again with the gender and training program together!
	8. Recode the training program variable by Automatic recode!
	9. Create faculty variable from the training program variable!
	10. Export the results into a Word document!

Sample tasks solutions

The solutions contain the created SPSS output tables. (Task a, b and j does not require output table as a solution.)

1. Import the data into SPSS!
2. Set the properties of variables!
3. Prepare a frequency table by gender!

|  |
| --- |
| **Gender** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 1 | 33 | 38,8 | 38,8 | 38,8 |
| 2 | 52 | 61,2 | 61,2 | 100,0 |
| Total | 85 | 100,0 | 100,0 |  |

1. Set values for gender variable at the following way: 1 – male, 2 – female. Prepare the frequency table by gender again!

|  |
| --- |
| **Gender** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | male | 33 | 38,8 | 38,8 | 38,8 |
| female | 52 | 61,2 | 61,2 | 100,0 |
| Total | 85 | 100,0 | 100,0 |  |

1. Prepare descriptive statistics about the preparedness of lecturer (frequency, mean, mode, median, standard deviation, skewness). Do the results make sense? What can be the reason?

|  |
| --- |
| **Statistics** |
| Preparedness of lecturer  |
| N | Valid | 85 |
| Missing | 0 |
| Mean | 5,04 |
| Median | 5,00 |
| Mode | 5 |
| Std. Deviation | ,663 |
| Skewness | 4,988 |
| Std. Error of Skewness | ,261 |

|  |
| --- |
| **Preparedness of lecturer** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 4 | 5 | 5,9 | 5,9 | 5,9 |
| 5 | 78 | 91,8 | 91,8 | 97,6 |
| 9 | 2 | 2,4 | 2,4 | 100,0 |
| Total | 85 | 100,0 | 100,0 |  |

1. Set the ‘9’ value as a Missing value! Prepare the descriptive statistics again!

|  |
| --- |
| **Statistics** |
| Preparedness of lecturer  |
| N | Valid | 83 |
| Missing | 2 |
| Mean | 4,94 |
| Median | 5,00 |
| Mode | 5 |
| Std. Deviation | ,239 |
| Skewness | -3,765 |
| Std. Error of Skewness | ,264 |

|  |
| --- |
| **Preparedness of lecturer** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 4 | 5 | 5,9 | 6,0 | 6,0 |
| 5 | 78 | 91,8 | 94,0 | 100,0 |
| Total | 83 | 97,6 | 100,0 |  |
| Missing | 9 | 2 | 2,4 |  |  |
| Total | 85 | 100,0 |  |  |

1. Delete one data from each column randomly! Solve task e) again with the gender and training programme together!

(The values from the first row were deleted.)

|  |
| --- |
| **Statistics** |
|  | Preparedness of lecturer | Gender | Training programme |
| N | Valid | 82 | 84 | 85 |
| Missing | 3 | 1 | 0 |
| Mean | 4,94 | 1,61 |  |
| Median | 5,00 | 2,00 |  |
| Mode | 5 | 2 |  |
| Std. Deviation | ,241 | ,491 |  |
| Skewness | -3,738 | -,447 |  |
| Std. Error of Skewness | ,266 | ,263 |  |

|  |
| --- |
| **Preparedness of lecturer** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 4 | 5 | 5,9 | 6,1 | 6,1 |
| 5 | 77 | 90,6 | 93,9 | 100,0 |
| Total | 82 | 96,5 | 100,0 |  |
| Missing | 9 | 2 | 2,4 |  |  |
| System | 1 | 1,2 |  |  |
| Total | 3 | 3,5 |  |  |
| Total | 85 | 100,0 |  |  |

|  |
| --- |
| **Gender** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | male | 33 | 38,8 | 39,3 | 39,3 |
| female | 51 | 60,0 | 60,7 | 100,0 |
| Total | 84 | 98,8 | 100,0 |  |
| Missing | System | 1 | 1,2 |  |  |
| Total | 85 | 100,0 |  |  |

|  |
| --- |
| **Training programme** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid |  | 1 | 1,2 | 1,2 | 1,2 |
| business administration and management | 24 | 28,2 | 28,2 | 29,4 |
| commerce and marketing | 20 | 23,5 | 23,5 | 52,9 |
| finance and accounting | 39 | 45,9 | 45,9 | 98,8 |
| other | 1 | 1,2 | 1,2 | 100,0 |
| Total | 85 | 100,0 | 100,0 |  |

1. Recode the training programme variable by Automatic recode!

After recoding:

|  |
| --- |
| **Training programme** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | business administration and management | 24 | 28,2 | 28,6 | 28,6 |
| commerce and marketing | 20 | 23,5 | 23,8 | 52,4 |
| finance and accounting | 39 | 45,9 | 46,4 | 98,8 |
| other | 1 | 1,2 | 1,2 | 100,0 |
| Total | 84 | 98,8 | 100,0 |  |
| Missing | 5 | 1 | 1,2 |  |  |
| Total | 85 | 100,0 |  |  |

1. Create faculty variable from the training programme variable!

After recoding:

|  |
| --- |
| **Faculty** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Faculty of Economics | 83 | 97,6 | 98,8 | 98,8 |
| Other faculty | 1 | 1,2 | 1,2 | 100,0 |
| Total | 84 | 98,8 | 100,0 |  |
| Missing | 3,00 | 1 | 1,2 |  |  |
| Total | 85 | 100,0 |  |  |

1. Export the results into a Word document!

When on the Output window, go to File/Export and set the desired file format (.doc format in this case)

# Estimation

Goals

**This chapter introduces the theoretical background and application of estimation**. Learning of this chapter is successful if the Reader is able to do the followings:

* calculate point estimations (for mean, proportion and standard deviation) and interpret the result
* calculate interval estimations (for mean, proportion and standard deviation) and interpret the result
* apply SPSS for interval estimation (for the mean).

**Knowledge obtained by reading this chapter:** calculation of point and interval estimations of population parameter (mean, standard deviation, proportion) both paper and SPSS-based

**Skills obtained by reading this chapter:**

* statistical communication – estimating population parameters with the help of sample data
* logical skills – identifying which formula is needed in certain situations (i.e. differentiating between interval estimations for the mean depending on sampling methods).

**Attitudes developed** **by reading this chapter**: confidence in the application of different estimation methods.

**This chapter makes the Reader to be autonomous in**: differentiating sample and population properties, giving estimation for population parameter based on sample characteristics.

Definitions

**Estimation:** Estimate the population parameter from a sample. Types: point- and interval estimations.

**Point estimation:** The statistic is computed from sample to estimate the population parameter

**Standard error of the estimation**: The difference on average between the sample statistics

and the population parameter with a given sample size

**Standard error of the mean**: The difference on average between

the sample means and the population mean

**Maximum error (error bound):** with a given probability the maximum error of the estimation

**Proportion** is the fraction or percentage that indicates the part of the population or sample having a particular trait of interest.

Learning activities

In order to learn how to calculate and interpret estimations

1. Read Chapter 9 from the book (Page 296-327)!
2. Open and explore 2\_estimation.ppt!
3. Explore and solve the sample tasks!
4. Check your knowledge: solve the chapter exercises in the book!

Sample tasks

**1.** We observe machines which fill bottles with coffee. We have a random sample with replacement. In the sample we know the weight of the coffee in a bottle (gram). We assume the weights’ normal distribution with 1g standard deviation.

55, 54, 54, 56, 57, 56, 55, 57, 54, 56, 55, 54, 57, 54, 56, 50.

1. Compute the point estimate of the population mean! Calculate the standard error of the estimation!
2. Develop a 95 % confidence interval for the population mean in line with the given condition!
3. Compute the point estimate of the population standard deviation!
4. Develop a 95 % confidence interval for the population mean if we know nothing about the population standard deviation!

**2.** We examine the number of borrowed books of a library’s borrowers. We have a random sample with replacement.

|  |  |
| --- | --- |
| Books | Number of borrowers |
| 12345 |  40120200 100 40 |
| Total | 500 |

1. What is the size of the population? What about the population

mean?

1. What is the size of the sample?
2. Compute the point estimate of the population mean! Calculate the standard error of the estimation!
3. Develop a 90 % confidence interval for the population mean!
4. Compute the point estimate of the population standard deviation!
5. Develop a 95 % confidence interval for the population standard deviation!
6. Estimate the proportion of those who borrowed at least 4 books! Calculate the standard error of the estimation!
7. Develop a 99 % confidence interval for proportion of those who borrowed at least 4 books!

**3.** A survey is to be conducted to determine the mean family income in Southern Illinois. The sponsor of the survey wants the estimate to be within $100 with a 95 percent level of confidence. The standard deviation of the incomes is estimated to be $400. How large a sample is required?

**4.** A sample of 80 Chief Financial Officers revealed 20 had at one time been dismissed from a job. Develop a 94 percent confidence interval for the proportion that has been dismissed from a job.

**5.** A random sample of 20 retired Florida residents revealed they listened to the radio an average (mean) of 40 minutes per day with a standard deviation of 8.6 minutes. Develop a 95 percent confidence interval for the population mean listening time.

**6.** The survey2.sav file contains data about a course evaluation. Develop a 95% confidence interval for the mean of age!

Sample tasks solutions

**1.** We observe machines which fill bottles with coffee. We have a random sample with replacement. In the sample we know the weight of the coffee in the bottle (gram). We assume the weights’ normal distribution with 1g standard deviation.

55, 54, 54, 56, 57, 56, 55, 57, 54, 56, 55, 54, 57, 54, 56, 50.

1. Compute the point estimate of the population mean! Calculate the standard error of the estimation!





When we estimate the population mean, the error of the estimation

is on average 0.25 g.

1. Develop a 95 % confidence interval for the population mean in line with the given condition!



1-α=0.95 1-(α/2)=0.975



 (54.51;55.49) g

With 95% probability, the population mean is between 54.51 and 55.49 g.

1. Compute the point estimate of the population standard deviation!



1. Develop a 95 % confidence interval for the population mean if we know nothing about the population standard deviation!







 (54.07;55.93)g

With 95% probability, the population mean is between 54.07 and 55.93 g.

**2.** We examine the number of borrowed books of a library’s borrowers. We have a random sample with replacement.

|  |  |
| --- | --- |
| Books | Number of borrowers |
| 12345 |  40120200 100 40 |
| Total | 500 |

1. What is the size of the population? What about the population mean?

Do not know the size of the population

Do not know the population mean

1. What is the size of the sample?

 (with replacement)

1. Compute the point estimate of the population mean! Calculate the standard error of the estimation!

books

The value of the sample mean is 2.96 books. The point estimation of the population is mean is 2.96 books.

Standard error: first estimate “s”! (task E)

 book

When we estimate the population mean, the error of the estimation is on average 0.047 books.

1. Develop a 90 % confidence interval for the population mean!



 (2.844;3.036) books

With 90% probability, the population mean is between 2.844 and 3.036 books.

1. Compute the point estimate of the population standard deviation!

 books

1. Develop a 95 % confidence interval for the population standard deviation!







(0.979;1.109) books

With 95% probability, the population standard deviation is between 0.979 and 1.109 books.

1. Estimate the proportion of those who borrowed at least 4 books! Calculate the standard error of the estimation!



When we estimate the population proportion, the error of the estimation is on average 2 percentage points.

1. Develop a 99 % confidence interval for proportion of those who borrowed at least 4 books!

Condition: n\*p;n\*q>10 → 140; 360>10

 (22.8;33.2)%



With 99% probability, the proportion of those who borrowed at least 4 books is between 22.8 and 33.2 percent.

**3.** A survey is to be conducted to determine the mean family income in Southern Illinois. The sponsor of the survey wants the estimate to be within $100 with a 95 percent level of confidence. The standard deviation of the incomes is estimated to be $400. How large a sample is required?







The minimum required sample size is 62 elements.

(If we replace back 61: delta won’t be within 100, if we replace 62, delta will be within 100.)

**4.** A sample of 80 Chief Financial Officers revealed 20 had at one time been dismissed from a job. Develop a 94 percent confidence interval for the proportion that has been dismissed from a job.



Condition: n\*p;n\*q>10 → 20; 60>10



 (16;34)%



With 94% probability, the proportion that has been dismissed from a job is between 16 and 34 percent.

**5.** A random sample of 20 retired Florida residents revealed they listened to the radio an average (mean) of 40 minutes per day with a standard deviation of 8.6 minutes. Develop a 95 percent confidence interval for the population mean listening time.







 (35.98;44.02) min

With 95% probability, the population mean is between 35.98 and 44.02 minutes.

**6.** The survey2.sav file contains data about a course evaluation. Develop a 95% confidence interval for the mean of age!

|  |
| --- |
| **Case Processing Summary** |
|  | Cases |
| Valid | Missing | Total |
| N | Percent | N | Percent | N | Percent |
| Age (year) | 85 | 100,0% | 0 | 0,0% | 85 | 100,0% |

|  |
| --- |
| **Descriptives** |
|  | Statistic | Std. Error |
| Age (year) | Mean | 21,82 | ,136 |
| 95% Confidence Interval for Mean | Lower Bound | 21,55 |  |
| Upper Bound | 22,09 |  |
| 5% Trimmed Mean | 21,79 |  |
| Median | 22,00 |  |
| Variance | 1,576 |  |
| Std. Deviation | 1,255 |  |
| Minimum | 19 |  |
| Maximum | 25 |  |
| Range | 6 |  |
| Interquartile Range | 2 |  |
| Skewness | ,527 | ,261 |
| Kurtosis | -,069 | ,517 |

With a 95% probability, the mean of age is between 21.55 and 22.09 years.

# Hypothesis testing

Goals

This chapter introduces the theoretical background and application of hypothesis testing. Learning of this chapter is successful if the Reader is able to do the followings:

* choose and apply the appropriate test statistics to a given problem
* interpret the results
* apply SPSS for hypothesis testing and interpret the results.

**Knowledge obtained by reading this chapter:** basics of hypothesis testing (one and two-sample tests) both paper (mean, proportion, standard deviation) and SPSS-based (mean).

**Skills obtained by reading this chapter:**

* statistical reasoning – using one and two sample tests to compare population parameters to a hypothetic value (one sample tests) or comparing two population parameters (two sample tests).
* logical skills – identifying which formula is needed in certain situations (i.e. differentiating between paired and independent samples tests for the population mean depending on the sample characteristics).

**Attitudes developed** **by reading this chapter**: confidence in the application of different hypothesis testing methods.

**This chapter makes the Reader to be autonomous in**: differentiating between one and two-sample tests.

Definitions

**Hypothesis testing** is a procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

**Hypothesis** is a statement about the value of a population parameter developed for the purpose of testing.

**Steps of Hypothesis testing:**

* State null and alternate hypotheses
* Select a statistical test
* Compute the value of test statistics
* Setup decision rules
* Make a decision
	+ do not reject nullhypothesis
	+ reject null- and accept alternate hypothesis

**Null Hypothesis H0:** A statement about the value of a population parameter.

**Alternative Hypothesis H1:** A statement that is accepted if the sample data provide evidence that the null hypothesis is false.

**Test statistic:** A value, determined from sample information, used to determine whether or not to reject the null hypothesis.

**Critical value:** The dividing point between the region where the null hypothesis is rejected and the region where it is not rejected.

**Level of Significance (α):** The probability of rejecting the null hypothesis when it is actually true.

**Type I Error:** Rejecting the null hypothesis when it is actually true.

**Type II Error:** Accepting the null hypothesis when it is actually false.

Learning activities

In order to learn how to apply hypothesis testing

1. Read Chapter 10.1-10.5 from the book (Page332-340)!
2. Open and explore 3\_0\_hypothesistesting.ppt!

# One sample tests

Learning activities

In order to learn how to apply hypothesis testing

1. Read Chapter 10.6-10.10 from the book (Page 340-370)!
2. Open and explore 3\_1\_Onesampletests.ppt!
3. Explore and solve the sample tasks!
4. Check your knowledge: solve the chapter exercises in the book!

Sample tasks

**1.** A study found that the mean stopping distance for a school bus traveling 50 miles per hour is 264 feet. The transportation director for the Orlando City Board of Education wants to compare his fleet of buses with national statistics. For a sample of ten buses the mean stopping distance was 270 feet and the standard deviation was 15 feet. Should the director conclude that the stopping distance is more for the Orlando buses? Use the 0.10 significance level.

**2.** The Exchange Bank wishes to determine the mean balance on the mortgages it holds. A sample of 36 mortgages showed the mean balance to be $86,000 with a sample standard deviation of $12,000. Would it be reasonable to conclude that the population mean is less than $90,000? Use the 0.05 significance level.

**3.** The Appliance Center reports on its TV Commercials that a more than 70 percent of their customers have purchased an appliance from them before. The president of the company hired a marketing research firm to independently validate this claim. In a sample of 200 recent buyers, 160 reported that they had, in fact, purchased an appliance from the Appliance Center before. At the 0.01 significance level is the claim of the commercial correct?

**4**. The survey2.sav file contains data about a course evaluation. Is the mean of the preparedness of the lecturer is 5? Answer the question at 5% significance level! Interpret the results!

Sample tasks solutions

**1.** A study found that the mean stopping distance for a school bus traveling 50 miles per hour is 264 feet. The transportation director for the Orlando City Board of Education wants to compare his fleet of buses with national statistics. For a sample of ten buses the mean stopping distance was 270 feet and the standard deviation was 15 feet. Should the director conclude that the stopping distance is more for the Orlando buses? Use the 0.10 significance level.







 Decision rule: retain H0, if t is within 

We accept/retain the nullhipothesis at 5% significance level therefore the average stopping distance is not significantly more for the Orlando buses.

**2.** The Exchange Bank wishes to determine the mean balance on the mortgages it holds. A sample of 36 mortgages showed the mean balance to be $86,000 with a sample standard deviation of $12,000. Would it be reasonable to conclude that the population mean is less than $90,000? Use the 0.05 significance level.







 Decision rule: 

We reject the nullhipothesis at 5% significance level therefore the population mean is less than 90000$.

**3.** The Appliance Center reports on its TV Commercials that a more than 70 percent of their customers have purchased an appliance from them before. The president of the company hired a marketing research firm to independently validate this claim. In a sample of 200 recent buyers, 160 reported that they had, in fact, purchased an appliance from the Appliance Center before. At the 0.01 significance level is the claim of the commercial correct?







 Decision rule: retain H0, if z is within 

We reject the nullhipothesis at 5% significance level therefore more than 70 percent of customers have purchased in that company before, so the claim of the commercial is correct.

**4.** The survey2.sav file contains data about a course evaluation. Is the mean of the preparedness of the lecturer is 5? Answer the question at 5% significance level! Interpret the results!

|  |
| --- |
| **One-Sample Statistics** |
|  | N | Mean | Std. Deviation | Std. Error Mean |
| Preparedness of lecturer | 83 | 4,94 | ,239 | ,026 |

|  |
| --- |
| **One-Sample Test** |
|  | Test Value = 5 |
| t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference |
| Lower | Upper |
| Preparedness of lecturer | -2,293 | 82 | ,024 | -,060 | -,11 | -,01 |

The nullhypothesis of the test is that the mean of preparedness of

lecturer is 5. One-sample t-test can be applied for answering this

question.

At a 5% significance level, we reject the nullhypothesis (sig<0,05) therefore the mean of seminar grades is not 5. With 95% probability, the mean of seminar grades is lower than 5, of which difference is between 0.01 and 0.11 units.

# Two samples tests

Learning activities

In order to learn how to apply hypothesis testing

1. Read Chapter 11 from the book (Page 370-409)!
2. Open and explore 3\_2\_twosamplestests.ppt!
3. Explore and solve the sample tasks!
4. Check your knowledge: solve the chapter exercises in the book!

Sample tasks

**1.** The Anderson’s Super Dollar had two grocery stores in Erie, Pennsylvania. The mean time customers wait in the checkout line at the Byrne Road store is 3.7 minutes with a standard deviation of 0.8 minutes, for a sample of 40 customers. The mean waiting time for the I-90 store is 3.5 minutes with a standard deviation of 0.7 minutes for a sample of 45 customers. At the 0.05 significance level can we conclude there is a difference in the waiting time for the two stores?

**2.** A sample of 200 Lion Store charge customers 50 years old or older showed that 20 did not pay their entire balance at the end of the month. A sample of 300 customers under 30 showed that 50 did not pay their entire balance at the end of the month. At the 0.02 significance level can we conclude that the same percent of the younger customers didn’t pay their entire balance at the end of the month as that of the older customers?

**3.** The mean high temperature for 12 days in July in Detroit, Michigan was 88 degrees with a standard deviation of 4 degrees. The mean high temperature in Hilton Head, South Carolina for 8 July days was 91 degrees with a standard deviation of 3 degrees. At the 0.05 significance level, can we conclude that there is no difference in the average temperatures?

**4.** An egg farmer wanted to determine if increasing the time the lights were on in his hen house would increase egg production. For a sample of eight chickens he determined their

production before and after increasing the amount of time the lights were on.

The data are reported below. At the 0.01 significance level, has

there been an increase in production?

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hen | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Before | 10 | 8 | 5 | 2 | 3 | 7 | 3 | 3 |
| After | 7 | 5 | 6 | 8 | 8 | 8 | 10 | 2 |

**5.** We examine the stress level (1….5) of the students before and after the statistics exam.

|  |  |
| --- | --- |
| **Before** | **After** |
| 5 | 4 |
| 5 | 1 |
| 4 | 3 |
| 4 | 2 |
| 3 | 2 |
| 3 | 1 |
| 2 | 2 |
| 3 | 3 |
| 3 | 2 |
| 4 | 2 |

Can we conclude the equality of the averages? Solve the task on paper and by SPSS as well!

**6.** In an entrance exam we examine the reaction time (sec) of two candidates. We assume the normality of the reaction time. Can we conclude the equality of the averages? Solve the task on paper and by SPSS as well!

See the next page for the data structure you need to import to SPSS

|  |  |
| --- | --- |
| Candidate | Reaction |
| 1 | 0,68 |
| 1 | 0,72 |
| 1 | 0,66 |
| 1 | 0,75 |
| 1 | 0,73 |
| 1 | 0,7 |
| 1 | 0,76 |
| 1 | 0,69 |
| 1 | 0,78 |
| 2 | 0,81 |
| 2 | 0,84 |
| 2 | 0,77 |
| 2 | 0,85 |
| 2 | 0,84 |
| 2 | 0,86 |
| 2 | 0,82 |
| 2 | 0,83 |

Sample tasks solutions

**1.** The Anderson’s Super Dollar had two grocery stores in Erie, Pennsylvania. The mean time customers wait in the checkout line at the Byrne Road store is 3.7 minutes with a standard deviation of 0.8 minutes, for a sample of 40 customers. The mean waiting time for the I-90 store is 3.5 minutes with a standard deviation of 0.7 minutes for a sample of 45 customers. At the 0.05 significance level can we conclude there is a difference in the waiting time for the two stores?

(On paper: we assume the equality of variances)

 

$$H\_{0}: μ\_{1}=μ\_{2}$$

$$H\_{1}: μ\_{1}\ne μ\_{2}$$





 Decision rule: retain H0, if t is within 

We retain the nullhipothesis at 5% significance level therefore there is no difference in the waiting time for the two stores.

**2.** A sample of 200 Lion Store charge customers 50 years old or older showed that 20 did not pay their entire balance at the end of the month. A sample of 300 customers under 30 showed that 50 did not pay their entire balance at the end of the month. At the 0.02 significance level can we conclude that the same percent of the younger customers didn’t pay their entire balance at the end of the month as that of the older customers?

 







 Decision rule: retain H0, if z is within 

We retain the null hypothesis at 2 % significance level therefore the same percent of the younger customers not pay their entire balance at the end of the month.

**3.** The mean high temperature for 12 days in July in Detroit, Michigan was 88 degrees with a standard deviation of 4 degrees. The mean high temperature in Hilton Head, South Carolina for 8 July days was 91 degrees with a standard deviation of 3 degrees. At the 0.05 significance level, can we conclude that there is no difference in the average temperatures?

 

$$H\_{0}: μ\_{1}=μ\_{2}$$

$$H\_{1}: μ\_{1}\ne μ\_{2}$$





 Decision rule: retain H0, if t is within 

We retain the nullhipothesis at 5% significance level therefore there no difference in the average temperatures.

**4**. An egg farmer wanted to determine if increasing the time the lights were on in his hen house would increase egg production. For a sample of eight chickens he determined their production before and after increasing the amount of time the lights were on. The data are reported below. At the 0.01 significance level, has there been an increase in production?

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hen | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Before | 10 | 8 | 5 | 2 | 3 | 7 | 3 | 3 |
| After | 7 | 5 | 6 | 8 | 8 | 8 | 10 | 2 |



(Difference: after – before)



 Decision rule: retain H0, if t is within 

At a 5% significance level we retain the H0, therefore there has not been an increase in production.

**5.** We examine the stress level (1….5) of the students before and after the statistics exam.

|  |  |
| --- | --- |
| **Before** | **After** |
| 5 | 4 |
| 5 | 1 |
| 4 | 3 |
| 4 | 2 |
| 3 | 2 |
| 3 | 1 |
| 2 | 2 |
| 3 | 3 |
| 3 | 2 |
| 4 | 2 |

Can we conclude the equality of the averages? Solve the task on paper and by SPSS as well!

(Let’s use 5% significance level.)

On paper:



Difference: before - after



 Decision rule: retain H0, if t is within 

At a 5% significance level we reject the H0, therefore the means of stress levels before and after the test are different.

SPSS:



The nullhypothesis of the test is that the means of stress level before and after the test is the same.

We can apply paired samples test for answering this question.

At 5% significance level, we reject the nullhypothesis (based on the sig=0.004<0.05 value), therefore the means of stress levels before and after the test are different.

**6.** In an entrance exam we examine the reaction time (sec) of two candidates. We assume the normality of the reaction time. Can we conclude the equality of the averages? Solve the task on paper and by SPSS as well!

On paper:

H0: µ1= µ2

H1: µ1≠ µ2





 Decision rule: retain H0, if t is within 

At 5% significance level, we reject the H0, therefore the means are different.

SPSS:





The nullhypothesis of the test is that the means of reaction times of the two candidates are the same. We want to compare the means of two groups, so independent sample t-test can be applied to answering this question.

We can assume the equality of the variances (sig=0.224>0.05 value shows the equalities of variances), so we should use the first row to make our final decision. In the first row, based on the sig(2-tailed)<0.05 value, so we reject the nullhypothesis at 5% significance level. The means of reaction times of the two candidates are not the same, the first candidate is faster (based on the sample means).

# Review Section (Topic 1-3)

Paper based exercises

**1**. Decide about the following statements whether they are TRUE or FALSE! Put an “X” sign in the correct column!

|  |  |  |
| --- | --- | --- |
| **Statement** | **TRUE** | **FALSE** |
| A confidence interval estimation is concerning about the sample |  |  |
| If you want to test whether a statement is true of false in the population, hypothesis testing can be used |  |  |
| If you do not know the population standard deviation the task cannot be solved, because there is no test for that situation |  |  |

**2**. Find and circle the correct answer from the list!

If you want to compare the average salaries of men and women (in the case when you do not know anything about the population standard deviations)

1. two independent samples t-test can be applied
2. paired t-test can be applied
3. one sample t-test can be applied
4. one sample z-test can be applied

The result of a 95% confidence interval estimation about the mean of age is the following: (24.2;25;8) years. The interpretation:

1. With 97.5% probability, the mean of age is between 24.2 and 25.8 years
2. With 95% probability, the age is between 24.2 and 25.8 years
3. With 95% probability, the mean of age is between 24.2 and 25.8 years
4. With 95% probability, the mean of age is between 24.2 and 25.8 percent

**3.**

There was a survey in a University about the sporting habits of students. With a help of an online questionnaire, a sample with 150 elements was collected. Based on this sample, the proportion of those who were regularly doing sports is 70 percent.

1. Develop a 90% interval for the proportion of those who were regularly

doing sports!

1. Interpret the result!

**4.**

According to a TV commercial the price of X washing powder is lower than 2000 HUF/piece in most retail shops.

There was a survey about unit prices of the X washing powder in several retail shops. 160 washing powders were examined in different places. Based on this sample with 160 elements, the mean of washing powder prices is 1900 HUF/piece, with a standard deviation of 110 HUF/piece.

Can we assume at 5% significance level that the mean of washing powder prices is lower than 2000 HUF/piece?

**5.** Economist students were asked about expected starter salaries in a survey. The results from the sample in the following table:

|  |  |  |
| --- | --- | --- |
| Gender | Number of sample | Expected starting salary, thousand HUF/month |
| mean | standard deviation |
| Male | 200 | 240 | 12 |
| Female | 300 | 230 | 11 |

(The sample standard deviations can be considered to be equal.)

Can we assume at 5% significance level that the average expected starting salary of male and female respondents are the same?

Paper based Solutions

**1**. Decide about the following statements whether they are TRUE or FALSE! Put an “X” sign in the correct column!

|  |  |  |
| --- | --- | --- |
| **Statement** | **TRUE** | **FALSE** |
| A confidence interval estimation is concerning about the sample |  | X |
| If you want to test whether a statement is true of false in the population, hypothesis testing can be used | X |  |
| If you do not know the population standard deviation the task cannot be solved, because there is no test for that situation |  | X |

**2**. Find and circle the correct answer from the list!

If you want to compare the salaries of men and women (in the case when you do not know anything about the population standard deviations)

1. two independent samples t-test can be applied
2. paired t-test can be applied
3. one sample t-test can be applied
4. one sample z-test can be applied

The result of a 95% confidence interval estimation about the mean of age is the following: (24.2;25;8) years. The interpretation:

1. With 97.5% probability, the mean of age is between 24.2 and 25.8 years
2. With 95% probability, the age is between 24.2 and 25.8 years
3. With 95% probability, the mean of age is between 24.2 and 25.8 years
4. With 95% probability, the mean of age is between 24.2 and 25.8 percent

**3.** There was a survey in a University about the sporting habits of students. With a help of an online questionnaire, a sample with 150 elements was collected. Based on this sample, the proportion of those who were regularly doing sports is 70 percent.

1. Develop a 90% interval for the proportion of those who were regularly doing sports!



Condition: n\*p;n\*q>10 → 105; 45>10

 (63.8;76.2)%





1. Interpret the result!

With 90% probability, the proportion of those who regularly doing sports is between 63.8 and 76.2 percent.

**4.** According to a TV commercial the price of X washing powder is lower than 2000 HUF/piece in most retail shops.

There was a survey about unit prices of the X washing powder in several retail shops. 160 washing powders were examined in different places. Based on this sample with 160 elements, the mean of washing powder prices is 1900 HUF/piece, with a standard deviation of 110 HUF/piece.

Can we assume at 5% significance level that the mean of washing powder prices is lower than 2000 HUF/piece?







 Decision rule: retain H0, if t is within 

We reject the nullhypothesis at 5% significance level, so the mean of

washing powder prices is lower than 2000 HUF/piece.

**5.** Economist students were asked about expected starting salaries in a survey. The results from the sample in the following table:

|  |  |  |
| --- | --- | --- |
| Gender | Number of sample | Expected starting salary, thousand HUF/month |
| mean | standard deviation |
| Male | 200 | 240 | 12 |
| Female | 300 | 230 | 11 |

(The sample standard deviations can be considered to be equal.)

Can we assume at 5% significance level that the average expected starting salary of male and female respondents are the same?

 







 Decision rule: retain H0, if t is within 

We reject the nullhypothesis at 5% significance level, so the average expected starting salary of male and female respondents are not the same.

SPSS – Seminar part 1

The employee.sav file contains a random sample of a banks’ employees. Solve these problems with SPSS.

1. Describe the current salary with frequency table, mode, mean, median, standard. deviation. Interpret it!
2. Modify the type of the gender from string to numeric! Prepare a frequency table!
3. Can we assume that the average starting salary is equal to $20000? (=0,05)
4. Is there a significant difference between the average starting and current salary? (=0,05)
5. Is there a significant difference between the male’s average current salary and the female’s average current salary?

SPSS solutions

Check the results by watching the spss\_test1.avi video. The interpretations can be found here.

The employee.sav file contains a random sample of a banks’ employees. Solve these problems with SPSS.

1. Describe the current salary with frequency table, mode, mean, median, standard. deviation. Interpret it!

|  |
| --- |
| **Statistics** |
| Current Salary  |
| N | Valid | 474 |
| Missing | 0 |
| Mean | $34,419.57 |
| Median | $28,875.00 |
| Mode | $30,750 |
| Std. Deviation | $17,075.661 |

The mean of current salaries is 34419.57 $. The most frequent current salary is 30750 $. Half of the salaries are maximum 28875 $. The current salaries deviate on average by 17075.661 $ from the mean.

1. Modify the type of the gender from string to numeric! Prepare a frequency table!

|  |
| --- |
| **Gender** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Female | 216 | 45,6 | 45,6 | 45,6 |
| Male | 258 | 54,4 | 54,4 | 100,0 |
| Total | 474 | 100,0 | 100,0 |  |

45.6 percent of the employees are female and 54.4 percent of the employees are male.

1. Can we assume that the average starting salary is equal to $20000? (=0,05)

|  |
| --- |
| **One-Sample Statistics** |
|  | N | Mean | Std. Deviation | Std. Error Mean |
| Starting Salary | 474 | $17,016.09 | $7,870.638 | $361.510 |

|  |
| --- |
| **One-Sample Test** |
|  | Test Value = 20000 |
| t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference |
| Lower | Upper |
| Starting Salary | -8,254 | 473 | ,000 | -$2,983.914 | -$3,694.28 | -$2,273.55 |

The nullhypothesis of the test is that the mean of starting salaries is 20000$.

One-sample t-test can be applied for answering this question.

At a 5% significance level, we reject the nullhypothesis (sig<0.05) therefore the mean of starting salaries is not 20000$.

The mean of starting salaries is lower than 20000$.

1. Is there a significance difference between the average beginning and current salary? (=0,05)

|  |
| --- |
| **Paired Samples Statistics** |
|  | Mean | N | Std. Deviation | Std. Error Mean |
| Pair 1 | Current Salary | $34,419.57 | 474 | $17,075.661 | $784.311 |
| Starting Salary | $17,016.09 | 474 | $7,870.638 | $361.510 |

|  |
| --- |
| **Paired Samples Test** |
|  | Paired Differences | t | df | Sig. (2-tailed) |
| Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference |
| Lower | Upper |
| Pair 1 | Current Salary - Starting Salary | $17,403.481 | $10,814.620 | $496.732 | $16,427.407 | $18,379.555 | 35,036 | 473 | ,000 |

The nullhypothesis of the test is that the mean of starting salaries and the mean of current salaries are the same.

Paired samples T-test can be applied for answering this question.

At 5% significance level, the nullhypothesis is rejected (sig<0.5), so the mean of starting salaries and the mean of current salaries are not the same.

The mean of current salaries is higher than the mean of starting

salaries.

1. Is there a significant difference between the male’s average current salary and the female’s average current salary?

|  |
| --- |
| **Group Statistics** |
|  | Gender | N | Mean | Std. Deviation | Std. Error Mean |
| Current Salary | Female | 216 | $26,031.92 | $7,558.021 | $514.258 |
| Male | 258 | $41,441.78 | $19,499.214 | $1,213.968 |



The nullhypothesis of the test is that the male’s average current salary and the female’s average current salary are the same.

We want to compare the means of two groups, so independent sample t-test can be applied to answering this question.

We cannot assume the equality of the variances (sig<0.05 value shows that variances are not the same), so we should use the second row to make our final decision.

In the second row, based on the sig(2-tailed)<0.05 value, we reject the nullhypothesis at 5% significance level. The male’s and female’s average current salaries are not the same.

Males have higher average current salary than females (based on the sample means).

# Relationships, causal models

Goals

**This chapter introduces the types of relationships between variables**. Learning of this chapter is successful if the Reader is able to do the followings:

* choose and apply the appropriate test statistics to a given problem
* interpret the results
* apply SPSS for hypothesis testing and interpret the results.

**Knowledge obtained by reading this chapter:** types of relationships, relationship-testing both paper- and SPSS based.

**Skills obtained by reading this chapter:**

* statistical reasoning – testing the significance and strength of relationship between two variables depending on their measurement levels;
* statistical communication – conducting relationship analyses and providing complex interpretation for the results of the analyses;
* logical skills – identifying relationship types

**Attitudes developed** **by reading this chapter**: confidence in the application of different relationship analyses.

**This chapter makes the Reader autonomous in**: differentiating between variables’ measurement levels and relationship types.

Definitions

**Types of relationships between variables**

* Deterministic
* Independent
* Stochastic

**Types of stochastic relationships**:

* Association (Between 2 categorical variables)
* Mixed relationship (Between 1 categorical and 1 metric variable)
* Rank correlation (Between the ranks of 2 metric variables)
* Correlation (Among 2 or more metric variables)

Learning activities

1. Open and explore 4\_0\_typesofrelationships. ppt!

# Crosstabs analysis

Goals

This chapter introduces crosstabs analysis. Learning of this chapter is successful if the Reader is able to do the followings:

* examine a relationship between two categorical variables
* apply Chi-square test on paper and by SPSS and interpret the results.

Definitions

**Crosstabs analysis:** a method which can be used for measuring relationship between categorical variables

**Expected frequencies**: hypothetic frequency of a class if the two variables are independent

**Observed frequencies**: frequency of a class in the sample

Learning activities

In order to learn the concept, calculation and interpretation in the topic of crosstabs analysis

1. Read Chapter 17.7 from the book (Page 666-679)!
2. Open and explore 4\_1\_Crosstabs analysis.ppt!
3. Explore and solve the sample tasks!
4. Check your knowledge: solve the chapter exercises in the book!

Sample tasks

**1.** In a public opinion research, the opinions about an economic TV program were examined. The following table is known about this research:

|  |  |  |
| --- | --- | --- |
| Profession | Opinion about the TV program | Total |
| Good | Fair enough | Bad |
| EconomistLawyerOther diploma | 100100100 | 200 60 60 | 100 40 40 | 400200200 |
| Total | 300 | 320 | 180 | 800 |

1. Calculate the expected frequencies in the case of independence!
2. Examine the relationship between the profession and the opinion about the TV program (α=0.05)?
3. Calculate and interpret the coefficients measuring the strength of the relationship (if it makes sense calculating them)!

**2.** Answer the following questions based on the bank.sav file!

1. Create a combination (contingency) table which contains the employees’ profession and gender!
2. Compute the employees’ distribution by gender!
3. Compute the expected frequencies in the case of independence!
4. Is there any relationship between the profession and gender at 5% significance level? If there is, calculate a coefficient measuring the strength of the relationship!

Sample tasks solutions

**1.** In a public opinion research, the opinions about an economic TV program were examined. The following table is known about this research:

|  |  |  |
| --- | --- | --- |
| Profession | Opinion about the TV program | Total |
| Good | Fair enough | Bad |
| EconomistLawyerOther diploma | 100100100 | 200 60 60 | 100 40 40 | 400200200 |
| Total | 300 | 320 | 180 | 800 |

1. Calculate the expected frequencies in the case of independence!



 

  

  

1. Examine the relationship between the profession and the opinion about the TV program (α=0.05)?

H0: There is no significant relationship between the profession and the opinion about the TV program/ the profession and the opinion about the TV program are independent

H1: There is a significant relationship between the profession and the opinion about the TV program/ the profession and the opinion about the TV program are dependent from each other

Condition for application: fij\*>5





We reject the H0 at 5% significance level, so there is a significant relationship between the profession and the opinion about TV program. .

1. Calculate and interpret the coefficients measuring the strength of the relationship (if there is a sense of calculating them)!





There is a weak relationship between the profession and the

opinion about TV program.

**2.** Answer the following questions based on the bank.sav file!

1. Create a combination (contingency) table which contains the employees’ profession and gender!

|  |
| --- |
| **Profession \* Gender Crosstabulation** |
| Count  |
|  | Gender | Total |
| male | female |
| Profession | employee with primary school qualification | 27 | 0 | 27 |
| employee with secondary school qualification | 157 | 206 | 363 |
| employee with tertiary school qualification | 64 | 9 | 73 |
| head of divison with secondary school qualification | 4 | 1 | 5 |
| head of division with tertiary school qualification | 6 | 0 | 6 |
| Total | 258 | 216 | 474 |

1. Compute the employees’ distribution by gender!

|  |
| --- |
| **Profession \* Gender Crosstabulation** |
|  | Gender | Total |
| male | female |
| Profession | employee with primary school qualification | Count | 27 | 0 | 27 |
| % within Gender | 10,5% | 0,0% | 5,7% |
| employee with secondary school qualification | Count | 157 | 206 | 363 |
| % within Gender | 60,9% | 95,4% | 76,6% |
| employee with tertiary school qualification | Count | 64 | 9 | 73 |
| % within Gender | 24,8% | 4,2% | 15,4% |
| head of divison with secondary school qualification | Count | 4 | 1 | 5 |
| % within Gender | 1,6% | 0,5% | 1,1% |
| head of division with tertiary school qualification | Count | 6 | 0 | 6 |
| % within Gender | 2,3% | 0,0% | 1,3% |
| Total | Count | 258 | 216 | 474 |
| % within Gender | 100,0% | 100,0% | 100,0% |

1. Compute the expected frequencies in the case of independence!

|  |
| --- |
| **Profession \* Gender Crosstabulation** |
|  | Gender | Total |
| male | female |
| Profession | employee with primary school qualification | Count | 27 | 0 | 27 |
| Expected Count | 14,7 | 12,3 | 27,0 |
| % within Gender | 10,5% | 0,0% | 5,7% |
| employee with secondary school qualification | Count | 157 | 206 | 363 |
| Expected Count | 197,6 | 165,4 | 363,0 |
| % within Gender | 60,9% | 95,4% | 76,6% |
| employee with tertiary school qualification | Count | 64 | 9 | 73 |
| Expected Count | 39,7 | 33,3 | 73,0 |
| % within Gender | 24,8% | 4,2% | 15,4% |
| head of divison with secondary school qualification | Count | 4 | 1 | 5 |
| Expected Count | 2,7 | 2,3 | 5,0 |
| % within Gender | 1,6% | 0,5% | 1,1% |
| head of division with tertiary school qualification | Count | 6 | 0 | 6 |
| Expected Count | 3,3 | 2,7 | 6,0 |
| % within Gender | 2,3% | 0,0% | 1,3% |
| Total | Count | 258 | 216 | 474 |
| Expected Count | 258,0 | 216,0 | 474,0 |
| % within Gender | 100,0% | 100,0% | 100,0% |

1. Is there any relationship between the profession and gender at 5% significance level? If there is, calculate a coefficient measuring the strength of the relationship!

|  |
| --- |
| **Chi-Square Tests** |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 79,757a | 4 | ,000 |
| Likelihood Ratio | 97,263 | 4 | ,000 |
| Linear-by-Linear Association | 12,786 | 1 | ,000 |
| N of Valid Cases | 474 |  |  |
| a. 4 cells (40,0%) have expected count less than 5. The minimum expected count is 2,28. |

Based on the note below the table categories are needed to be recoded. After recoding:

|  |
| --- |
| **Profession \* Gender Crosstabulation** |
|  | Gender | Total |
| male | female |
| Profession | employee with primary school qualification | Count | 27 | 0 | 27 |
| Expected Count | 14,7 | 12,3 | 27,0 |
| % within Gender | 10,5% | 0,0% | 5,7% |
| employee with secondary school qualification | Count | 157 | 206 | 363 |
| Expected Count | 197,6 | 165,4 | 363,0 |
| % within Gender | 60,9% | 95,4% | 76,6% |
| employee with tertiary school qualification | Count | 64 | 9 | 73 |
| Expected Count | 39,7 | 33,3 | 73,0 |
| % within Gender | 24,8% | 4,2% | 15,4% |
| head of divison | Count | 10 | 1 | 11 |
| Expected Count | 6,0 | 5,0 | 11,0 |
| % within Gender | 3,9% | 0,5% | 2,3% |
| Total | Count | 258 | 216 | 474 |
| Expected Count | 258,0 | 216,0 | 474,0 |
| % within Gender | 100,0% | 100,0% | 100,0% |

|  |
| --- |
| **Chi-Square Tests** |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 79,318a | 3 | ,000 |
| Likelihood Ratio | 95,565 | 3 | ,000 |
| Linear-by-Linear Association | 11,972 | 1 | ,001 |
| N of Valid Cases | 474 |  |  |
| a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,01. |

|  |
| --- |
| **Symmetric Measures** |
|  | Value | Approximate Significance |
| Nominal by Nominal | Phi | ,409 | ,000 |
| Cramer's V | ,409 | ,000 |
| N of Valid Cases | 474 |  |

The nullhypothesis of the test is that there is no significant relationship between the profession and gender.

We examine a relationship between two categorical variables, therefore crosstabs analysis can be applied for answering this question.

Based on the note below the table, the application condition was not met in more than 20% of the cells, so categories were recoded. After recoding, the application condition is met.

At a 5% significance level, we reject the nullhypothesis (Pearson Chi Square sig<0.05), so there is a significant relationship between the profession and gender. The relationship between the examined variables has a medium strength (C=0.409).

The ratio of employees with tertiary education qualification is 24.8% among male respondents. The ratio of employees with secondary education qualification is 95.4% among female respondents. Within males, the employees with tertiary education qualification is typical, and within female the employees with secondary education is typical.

# Analysis of variance

Goals

This chapter introduces the analysis of variance (Oneway ANOVA). Learning of this chapter is successful if the Reader is able to do the followings:

* examine a relationship between a categorical (independent) variable and a metric (dependent) variable
* create the ANOVA table, apply F test and calculate the variance coefficient on paper and by SPSS and interpret the results.

Definitions

**Dependent variable** is the variable being predicted or estimated

**Independent variable** provides the basis for estimation. It is the predictor variable

**ANOVA** (ANalysis Of VAriance):a method which can be used for comparing means for more than two independent samples = examining relationships between categorical and metric variables. The method is also used for testing a regression models’ fit or testing multiple correlation coefficient.

**Variance coefficient**: shows the proportion of variance in metric variable explained by categorical variable

Learning activities

In order to learn the concept, calculation and interpretation in the topic of ANOVA

1. Read Chapter 12 from the book (Page 410-460)!
2. Open and explore 4\_2\_anova.ppt!
3. Explore and solve the sample tasks!
4. Check your knowledge: solve the chapter exercises in the book!

Sample tasks

**1.** 100 students were asked about the income an „average” entrepreneur has. The data of the sample are in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| Training programme | Number of respondents (person) | Mean of assumed monthly income (thousand HUF) | Standard deviation of assumed monthly income (thousand HUF) |
| Business Administration and Management | 60 | 150 | 12,1 |
| Commerce and marketing | 30 | 100 | 11,8 |
| Finance and accounting | 10 | 130 | 12,0 |

It is also known that the assumed monthly income follows normal distribution in each training programme group, and the variances of the assumed monthly incomes can be considered equal.

1. Is there any relationship between the training programme and the assumed monthly income (α=0,05)
2. If it makes sense, calculate the H and H2 measures!

**2.** A company has four types of machines and it sets out to examine the productivity of its machines. Based on a sample from production data, the following data are known:

|  |  |  |
| --- | --- | --- |
| Type of machine | Number of observed productivity data | Mean of productivity (pieces/hour) |
| A | 30 | 29,9 |
| B | 32 | 29,75 |
| C | 30 | 29,23 |
| D | 31 | 29,97 |
| Total | 123 | 29,72 |

It is also known that SStotal=115,04. The productivity is not skewed to the right strongly in each machine type group, and the variances of productivities can be considered equal.

1. Is there any relationship between the type of machine and productivity (α=0,05)?
2. If it makes sense, calculate the H and H2 measures!

**3.** Examine based on the productivity.sav file if there is any relationship between the types

of machines and productivity (α=0,05)! Prepare a complex analysis

(conditions for application, if it is necessary: reason for rejecting H0-Post

Hoc test, H, H2)!

**4.** Examine based on MA.sav file if there is any relationship between the social class and the number of assets (α=0,05)! Prepare a complex analysis (conditions for application, if it is necessary: reason for rejecting H0-Post Hoc test, H, H2)!

Sample tasks solutions

**1.** 100 students were asked about the income an „average” entrepreneur has. The data of the sample are in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| Tranining programme | Number of respondents (person) | Mean of assumed monthly income (thousand HUF) | Standard deviation of assumed monthly income (thousand HUF) |
| Business Administration and Management | 60 | 150 | 12,1 |
| Commerce and marketing | 30 | 100 | 11,8 |
| Finance and accounting | 10 | 130 | 12,0 |

It is also known, that the assumed monthly income follows normal distribution is each training programme group, and the variances of the assumed monthly incomes can be considered equal.

1. Is there any relationship between the training programme and the assumed monthly income (=0,05)?

H0: there is no relationship between the training programme and the assumed monthly income











At 5% we reject the nullhypothesis, so there is a significant relationship between the training programme and the assumed monthly salary.

ANOVA Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source of variance | Sum of Squares | df | Mean square | F |
| treatment  | 50100 | 2 | 25050 | 173,91  |
| error | 13972.15 | 97 | 144,04 |
| Total | 64072.15 | - | - | - |

*SStotal=SSerror+SStreatment=50100+13972.15=64072.15*

b) H and H2





78.2% of variance of the monthly assumed income is explained by the tranining programme. The remaining 21.8% is explained by other factors.

There is a strong relationship between the training programme and the assumed monthly income.

**2.** A company has four types of machines and it sets out to examine the productivity of its machines. Based on a sample from production data, the following data are known:

|  |  |  |
| --- | --- | --- |
| Type of machine | Number of observed productivity data | Mean of productivity (pieces/hour) |
| A | 30 | 29.9 |
| B | 32 | 29.75 |
| C | 30 | 29.23 |
| D | 31 | 29.97 |
| Total | 123 | 29.72 |

It is also known that SStotal=115.04. The productivity is not skewed to the right strongly in each machine type group, and the variances of assumed monthly incomes can be considered equal.

1. Is there any relationship between the type of machine and productivity (α=0,05)?

H0: there is no relationship between the type of machine and productivity.









At 5% we reject the nullhypothesis, so there is a significant relationship between type of machines and the productivity.

1. If it makes sense, calculate the H and H2 measures!





9% of variance of the productivity is explained by the types of machines. The remaining 91% is explained by other factors.

There is a weak relationship between the training programme and the assumed monthly income.

**3.** Examine based on productivity.sav file if there is any relationship between the types of machines and the productivity (α=0,05)! Prepare a complex analysis (conditions for application, if it is necessary: reason for rejecting H0-Post Hoc test, H, H2)!

The nullhypothesis of the test is that there is no significant relationship between the productivity and the types of machines/ means of productivities are equal in the types of machines.

We examine a relationship between a categorical and a metric variable, therefore ANOVA can be applied for answering this question.

|  |
| --- |
| **Descriptives** |
| productivity (pieces/hour)  |
|  | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | Minimum | Maximum |
| Lower Bound | Upper Bound |
| A | 30 | 29,9000 | ,92289 | ,16850 | 29,5554 | 30,2446 | 28,00 | 32,00 |
| B | 32 | 29,7500 | ,91581 | ,16189 | 29,4198 | 30,0802 | 28,00 | 32,00 |
| C | 30 | 29,2333 | 1,00630 | ,18372 | 28,8576 | 29,6091 | 28,00 | 31,00 |
| D | 31 | 29,9677 | ,91228 | ,16385 | 29,6331 | 30,3024 | 28,00 | 32,00 |
| Total | 123 | 29,7154 | ,97106 | ,08756 | 29,5421 | 29,8888 | 28,00 | 32,00 |

|  |
| --- |
| **Test of Homogeneity of Variances** |
| productivity (pieces/hour)  |
| Levene Statistic | df1 | df2 | Sig. |
| ,963 | 3 | 119 | ,413 |

The variance homogeneity is assumed based on the Levene-test sig>0.05 value, so the ANOVA table is considered for answering the main question.

|  |
| --- |
| **ANOVA** |
| productivity (pieces/hour)  |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 10,006 | 3 | 3,335 | 3,779 | ,012 |
| Within Groups | 105,034 | 119 | ,883 |  |  |
| Total | 115,041 | 122 |  |  |  |

At 5% significance level, we reject the H0 (sig<0.05), so there is a significant relationship between the productivity and the types of machines/ means of productivities are not equal in the types of machines.

|  |
| --- |
| **Multiple Comparisons** |
| Dependent Variable: productivity (pieces/hour)  |
| Tukey HSD  |
| (I) machine type | (J) machine type | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval |
| Lower Bound | Upper Bound |
| A | B | ,15000 | ,23876 | ,923 | -,4721 | ,7721 |
| C | ,66667\* | ,24258 | ,034 | ,0346 | 1,2987 |
| D | -,06774 | ,24061 | ,992 | -,6947 | ,5592 |
| B | A | -,15000 | ,23876 | ,923 | -,7721 | ,4721 |
| C | ,51667 | ,23876 | ,139 | -,1055 | 1,1388 |
| D | -,21774 | ,23676 | ,794 | -,8347 | ,3992 |
| C | A | -,66667\* | ,24258 | ,034 | -1,2987 | -,0346 |
| B | -,51667 | ,23876 | ,139 | -1,1388 | ,1055 |
| D | -,73441\* | ,24061 | ,015 | -1,3614 | -,1074 |
| D | A | ,06774 | ,24061 | ,992 | -,5592 | ,6947 |
| B | ,21774 | ,23676 | ,794 | -,3992 | ,8347 |
| C | ,73441\* | ,24061 | ,015 | ,1074 | 1,3614 |
| \*. The mean difference is significant at the 0.05 level. |



If we consider the pairwise comparisons of each group means, the average productivity of machine C is lower than the average productivity of machine A and D.

|  |
| --- |
| **Measures of Association** |
|  | Eta | Eta Squared |
| productivity (pieces/hour) \* machine type | ,295 | ,087 |

There is a weak relationship between the training programme and the assumed monthly income. 9% of variance of the productivity is explained by the types of machines. The rest 91% is explained by other factors.

**4.** Examine based on the MA.sav file if there is any relationship between the social class and the number of assets (α=0.05)! Prepare a complex analysis (conditions for application, if it is necessary: reason for rejecting H0, H, H2)!

The nullhypothesis of the test is that there is no significant relationship between the social class and the number of assets / means of number of assets are equal in the social classes.

We examine a relationship between a categorical and a metric variable, therefore ANOVA can be applied for answering this question.

|  |
| --- |
| **Descriptives** |
| number of assets, pieces  |
|  | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | Minimum | Maximum |
| Lower Bound | Upper Bound |
| lower class | 220 | 4,4955 | 2,04180 | ,13766 | 4,2242 | 4,7668 | ,00 | 9,00 |
| lower-middle class | 817 | 4,6707 | 1,96988 | ,06892 | 4,5355 | 4,8060 | ,00 | 10,00 |
| middle class | 1469 | 4,9408 | 2,06004 | ,05375 | 4,8353 | 5,0462 | ,00 | 10,00 |
| upper class | 137 | 5,5182 | 2,25930 | ,19303 | 5,1365 | 5,9000 | 1,00 | 9,00 |
| Total | 2643 | 4,8502 | 2,05255 | ,03993 | 4,7719 | 4,9285 | ,00 | 10,00 |

|  |
| --- |
| **Test of Homogeneity of Variances** |
| number of assets, pieces  |
| Levene Statistic | df1 | df2 | Sig. |
| 2,612 | 3 | 2639 | ,050 |

(Note: This sig value (0.05) is a rounded value. You can check the exact value (0.049) by double clicking in SPSS output, and then double clicking on the value again.)

The variance homogeneity cannot be assumed based on the Levene-test sig=0.049<0.05 value, so the Welch test should be considered for answering the main question.

|  |
| --- |
| **Robust Tests of Equality of Means** |
| number of assets, pieces  |
|  | Statistica | df1 | df2 | Sig. |
| Welch | 9,358 | 3 | 443,409 | ,000 |
| a. Asymptotically F distributed. |

At 5% significance level, we reject the H0 (sig<0.05), so there is a significant relationship between the social class and the number of assets / the means of assets are not the same in the social classes.

|  |
| --- |
| **Multiple Comparisons** |
| Dependent Variable: number of assets, pieces  |
| Tamhane  |
| (I) social class | (J) social class | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval |
| Lower Bound | Upper Bound |
| lower class | lower-middle class | -,17529 | ,15395 | ,830 | -,5827 | ,2321 |
| middle class | -,44532\* | ,14778 | ,017 | -,8368 | -,0538 |
| upper class | -1,02279\* | ,23708 | ,000 | -1,6512 | -,3944 |
| lower-middle class | lower class | ,17529 | ,15395 | ,830 | -,2321 | ,5827 |
| middle class | -,27003\* | ,08740 | ,012 | -,5002 | -,0398 |
| upper class | -,84750\* | ,20496 | ,000 | -1,3930 | -,3020 |
| middle class | lower class | ,44532\* | ,14778 | ,017 | ,0538 | ,8368 |
| lower-middle class | ,27003\* | ,08740 | ,012 | ,0398 | ,5002 |
| upper class | -,57747\* | ,20037 | ,027 | -1,1113 | -,0436 |
| upper class | lower class | 1,02279\* | ,23708 | ,000 | ,3944 | 1,6512 |
| lower-middle class | ,84750\* | ,20496 | ,000 | ,3020 | 1,3930 |
| middle class | ,57747\* | ,20037 | ,027 | ,0436 | 1,1113 |
| \*. The mean difference is significant at the 0.05 level. |



If we consider the pairwise comparisons of each group means, the means of number of assets are different in each social class except of the lower- and lower-middle class. Higher social classes show higher means of number of assets.

|  |
| --- |
| **Measures of Association** |
|  | Eta | Eta Squared |
| number of assets, pieces \* social class | ,107 | ,011 |

There is a weak relationship between the social class and the number of assets (H=0,107). 1.1 percent of the variance of the number of assets is explained by the social class, the rest 98.9 percent can be explained by other factors.

# Correlation and regression analysis

Goals

This chapter introduces correlation and regression analysis. Learning of this chapter is successful if the Reader is able to do the followings:

* examine a relationship between a metric variables and build regression models
* calculate the coefficient of correlation and build regression models on paper and by SPSS and interpret the results.

Definitions

**Dependent variable** is the variable being predicted or estimated

**Independent variable** provides the basis for estimation. It is the predictor variable

**Correlation analysis**: a method which can be used for measuring the relationship between metric variables

**Regression analysis**: in a regression analysis we use the independent variable (x) to estimate the dependent variable (y)

**Coefficient of Correlation** is a measure of the strength of the relationship between two variables

**b0**: is the y-intercept of a linear regression function. It is the estimated y value when x=0

**b1**: is the slope of a linear regression function. If x increases by one unit, the estimated y increases or decreases on average by b1 unit.

**Coefficient of determination**: is the proportion of the total variation in the dependent variable (y) that is explained by the independent variable (x)

**Standard error of the estimates** (in a regression model): shows the average difference between the observed and the estimated values of y

Learning activities

In order to learn the concept, calculation and interpretation in the topic of ANOVA

1. Read Chapter 13 from the book (Page 461-511)!
2. Open and explore 4\_3\_correlationandregression.ppt!
3. Explore and solve the sample tasks!
4. Check your knowledge: solve the chapter exercises in the book!

Sample tasks

**1.** In a case of 20 employees the years spent in the education system and the monthly salary data are known:

|  |  |
| --- | --- |
| **Years spent in theeducation system (year)** | **Monthly salary (thousand HUF)** |
| 19 | 167,5 |
| 12 | 125 |
| 12 | 93,7 |
| 16 | 172,2 |
| 19 | 188,6 |
| 12 | 59,9 |
| 15 | 101,5 |
| 12 | 120 |
| 12 | 81,38 |
| 12 | 75,4 |
| 17 | 82,3 |
| 15 | 117 |
| 16 | 160,02 |
| 15 | 73,15 |
| 12 | 83,7 |
| 12 | 88,8 |
| 16 | 185 |
| 8 | 52,1 |
| 11 | 78 |
| 17 | 121 |

1. Describe the relationships between the years spent in the education system and the monthly salary with the help of coefficient of correlation!
2. Is the relationship between the years spent in the education system and the monthly salary significant (α=0,05)?
3. Solve task a) and task b) by SPSS!

**2.** In the case of 20 companies, the value of fixed assets and the productivity were observed. The observed data can be found in the following table:

|  |  |
| --- | --- |
| Value of fixed assets, million HUF/person, x | Productivity, pieces/person, y |
| 1,1 | 1,4 |
| 1,8 | 2,4 |
| 1,3 | 2,1 |
| 2,2 | 3,7 |
| 2,5 | 3,9 |
| 2,4 | 3,3 |
| 1,6 | 2,0 |
| 1,6 | 2,8 |
| 2,8 | 4,4 |
| 2,1 | 3,2 |
| 1,9 | 3,4 |
| 2,2 | 3,1 |
| 1,9 | 2,8 |
| 1,5 | 2,5 |
| 1,3 | 1,6 |
| 3,0 | 4,4 |
| 2,3 | 3,5 |
| 2,7 | 3,8 |
| 1,8 | 3,0 |
| 1,7 | 2,5 |

$$\left(\sum\_{}^{}x=39,7;\sum\_{}^{}y=59,8; \sum\_{}^{}xy=126,71; \sum\_{}^{}x^{2}=84,07; \sum\_{}^{}y^{2}=192,48; \sum\_{}^{}e^{2}=1,502;\right)$$

SST=13,678

1. Create a scatter plot in SPSS! What can we assume based on the scatter plot?
2. Fit a  linear regression curve to the data! Interpret the regression parameters!
3. Estimate the productivity for a company, where the value of assets is 2 million HUF/person!
4. Calculate and interpret the standard error of the estimate!
5. Calculate the coefficient of determination! Interpret the result!
6. Solve the B,D,E tasks by SPSS!

Sample tasks solutions

1. In a case of 20 employees the years spent in the education system and the monthly salary data are known.

1. Describe the relationships between the years spent in the education system and the monthly salary with the help of coefficient of correlation!















There is a strong relationship with a positive direction between years spent in the education system and the monthly salary.

1. Is the relationship between the years spent in the education system and the monthly salary significant?







At a 5% significance level, we reject the nullhypothesis, so there is a significant relationship between the years spent in education and the monthly salary.

1. Solve task a) and task b) by SPSS!

|  |
| --- |
| **Correlations** |
|  | Years spent in education system (year) | Monthly salary (thousand HUF) |
| Years spent in education system (year) | Pearson Correlation | 1 | ,730\*\* |
| Sig. (2-tailed) |  | ,000 |
| N | 20 | 20 |
| Monthly salary (thousand HUF) | Pearson Correlation | ,730\*\* | 1 |
| Sig. (2-tailed) | ,000 |  |
| N | 20 | 20 |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). |

At a 5% significance level, there is a significant (sig<0.05), strong relationship with a positive direction (rxy=0.730) between the years spent in the education system and the monthly salary.

**2.** In the case of 20 companies, the value of fixed assets and the productivity were observed.

$$\left(\sum\_{}^{}x=39.7;\sum\_{}^{}y=59.8; \sum\_{}^{}xy=126.71; \sum\_{}^{}x^{2}=84.07; \sum\_{}^{}y^{2}=192.48; \sum\_{}^{}e^{2}=1.502;\right)$$

SST=13,678

1. Create a scatter plot in SPSS? What can we assume based on the scatter plot?



A linear, strong, and positive relationship can be assumed based on the scatter plot.

1. Fit a  linear regression curve to the data! Interpret the regression parameters!









If the value of assets is 0 million HUF/person, the estimated productivity is -0.028 pieces/person.

If the value of assets increases by 1 million HUF/person, the estimated productivity increases on average by 1.52 pieces/person.

1. **Estimate the productivity for a company, where the value of assets is 2 million HUF/person!**

x=2

 million HUF/person

1. **Calculate and interpret the standard error of the estimate!**

 pieces/person



The average difference between the observed and estimated values of productivity is 0.29 pieces/person (9.7 %).

1. **Calculate the coefficient of determination! Interpret the result!**



89% of the total variation of productivity is explained by (variation) the value of assets. The rest 11% is explained by other factors.

1. **Solve the B,D,E tasks by SPSS!**

|  |
| --- |
| **Model Summary** |
| R | R Square | Adjusted R Square | Std. Error of the Estimate |
| ,943 | ,890 | ,884 | ,289 |
| The independent variable is value\_assets\_millionHUF\_person\_x. |

89% of the total variation of productivity is explained by (variation) the value of assets. The rest 11% is explained by other factors.

The average difference between the observed and estimated values of productivity is 0.29 pieces/person (9.7 %).

|  |
| --- |
| **Model Summary and Parameter Estimates** |
| Dependent Variable: productivity\_pieces\_person\_y  |
| Equation | Model Summary | Parameter Estimates |
| R Square | F | df1 | df2 | Sig. | Constant | b1 |
| Linear | ,890 | 145,903 | 1 | 18 | ,000 | -,028 | 1,521 |
| The independent variable is value\_assets\_millionHUF\_person\_x. |



If the value of assets is 0 million HUF/person, the estimated

productivity is -0.028 pieces/person.

If the value of assets increases by 1 million HUF/person, the

estimated productivity increases on average by

1.52 pieces/person.

# Review Section (Topic 4)

Paper based exercises

**1**. Decide about the following statements whether they are TRUE or FALSE! Put an “X” sign in the correct column!

|  |  |  |
| --- | --- | --- |
| **Statement** | **TRUE** | **FALSE** |
| The values of the H measure can be between -1 and +1. |  |  |
| The values of the coefficient of correlation can be between -1 and +1. |  |  |
| The “b0” shows the estimated value of variable “y” if x=0. |  |  |

**2**. Find and circle the correct answer from the list!

If we examine the relationship between the starting salary (thousand USD) and the current salary (thousand USD)

1. we discuss the relationship between two categorical variables
2. correlation analysis can be applied
3. crosstabs analysis can be applied
4. ANOVA can be applied

In a  regression model the standard error of the estimates shows

1. the average difference between the observed and estimated values of the “y” variable.
2. the average difference between the observed and estimated values of the “x” variable.
3. the proportion of the total variation in the dependent variable (y) that is explained by the variation in the independent variable (x)
4. the proportion of the total variation in the independent variable (x) that is explained by the variation in the dependent variable (y)

**3.** Sporting habits were examined in a frame of a survey based on a sample with 400 elements. The following data are known about the sample:

|  |  |  |
| --- | --- | --- |
| Gender | Sporting habits | Total |
| Doing sports regularly  | Not doing sports regularly  |
| Male | 80 | 120 | 200 |
| Female | 70 | 130 | 200 |
| Total  | 150 | 250 | 400 |

At 5% significance level, is there a significant relationship between gender and sporting habits?

**4.** The following data are known about a sample based on a questionnaire:

|  |  |  |
| --- | --- | --- |
| Qualification | Number of respondents (person) | Daily average internet usage time (minutes) |
| Primary education qualification | 20 | 10 |
| Secondary education qualification | 40 | 40 |
| Tertiary education qualification | 20 | 100 |
| Total | 80 | 47.5 |

It is also known that the daily internet usage time follows normal distribution in each qualification group, and variances of daily internet usage time can be considered equal; and SST=200000

At 5% significance level, is there any relationship between qualification and daily internet usage time?

**5.** In the case of 25 companies, the number of employees (person, x) and the number of manufactured products (thousand pieces/week, y) are known. The following results are also known:



Calculate and interpret the linear coefficient of correlation!

**6.** A shoe manufacturer company produces different pairs of shoes. In the case of 25 pairs of shoes, it was examined how long they remain in good condition so the durability is known (month, x). The price of a pair of shoes (thousand HUF/pair of shoes, y) is also known, and the following results were calculated based on the data:



1. Fit a  linear regression curve to the data (Calculate b1 and b0, write the model equation)! Interpret the regression parameters!
2. Estimate the price of a pair of shoes with a durability of 24 months!
3. Calculate and interpret the standard error of the estimate!
4. Calculate the coefficient of determination! Interpret the result!

Paper based solutions

**1**. Decide about the following statements whether they are TRUE or FALSE! Put an “X” sign in the correct column!

|  |  |  |
| --- | --- | --- |
| **Statement** | **TRUE** | **FALSE** |
| The values of the H measure can be between -1 and +1. |  | X |
| The values of the coefficient of correlation can be between -1 and +1. | X |  |
| The “b0” shows the estimated value of variable “y” if x=0. | X |  |

**2**. Find and circle the correct answer from the list!

If we examine the relationship between the starting salary (thousand USD) and the current salary (thousand USD)

1. we discuss the relationship between two categorical variables
2. correlation analysis can be applied
3. crosstabs analysis can be applied
4. ANOVA can be applied

In a  regression model the standard error of the estimates shows

1. the average difference between the observed and estimated values of the “y” variable.
2. the average difference between the observed and estimated values of the “x” variable.
3. the proportion of the total variation in the dependent variable (y) that is explained by the variation in the independent variable (x)
4. the proportion of the total variation in the independent variable (x) that is explained by the variation in the dependent variable (y)

**3.** Sporting habits were examined in a frame of a survey based on a sample with 400 elements. The following data are known about the sample:

|  |  |  |
| --- | --- | --- |
| Gender | Sporting habits | Total |
| Doing sports regularly  | Not doing sports regularly  |
| Male | 80 | 120 | 200 |
| Female | 70 | 130 | 200 |
| Total  | 150 | 250 | 400 |

At 5% significance level, is there a significant relationship between

gender and sporting habits?

H0: There is no relationship between gender and sporting habits.

H1: There is a relationship between gender and sporting habits.

 

 





We retain the H0 at 5% significance level, so there is no significant relationship between gender and sporting habits.

**4.** The following data are known about a sample based on a questionnaire:

|  |  |  |
| --- | --- | --- |
| Qualification | Number of respondents (person) | Daily average internet usage time (minutes) |
| Primary education qualification | 20 | 10 |
| Secondary education qualification | 40 | 40 |
| Tertiary education qualification | 20 | 100 |
| Total | 80 | 47.5 |

It is also known, that the daily internet usage time follows normal distribution in each qualification group, and variances of daily internet usage time can be considered equal; and SST=200000

At 5% significance level, is there any relationship between qualification and daily internet usage time?

H0: There is no relationship between qualification and daily internet usage

time.

H1: There is a relationship between qualification and daily

internet usage time.









At 5% we reject the nullhypothesis, so there is a significant relationship between qualification and daily internet usage time.

**5.** In the case of 25 companies, the number of employees (person, x) and the number of manufactured products (thousand pieces/week, y) are known. The following results are also known:



Calculate and interpret the linear coefficient of correlation!





There is a strong relationship with a positive direction between the number of employees and the number of manufactured products.

**6.** A shoe manufacturer company produces different pairs of shoes. In the case of 25 pairs of shoes, it was examined how long they remain in good condition so the durability is known (month, x). The price of a pair of shoes (thousand HUF/pair of shoes, y) is also known, and the following results were calculated based on the data:



1. Fit a  linear regression curve to the data (Calculate b1 and b0, write the model equation)! Interpret the regression parameters!







If the durability is 0 month, the estimated price of the pair of shoes is 5.591 thousand HUF/pair of shoes.

If the durability increases by 1 month, the estimated price of a pair of shoes increases on average by 0.446 thousand HUF/pair of shoes.

1. Estimate the price of a pair of shoes with a durability of 24 months!

x=24

 thousand HUF/pair of shoes

1. Calculate and interpret the standard error of the estimate!

 thousand HUF/pair of shoes

The average difference between the observed and estimated values of the price of a pair of shoes is 1.73 thousand HUF/pair of shoes.

1. Calculate the coefficient of determination! Interpret the result!



85.2% of the variation of the price of a pair of shoes is explained by the durability. The rest 14.8% is explained by other factors.

SPSS exercises – Seminar part 2

**1.** The finance.sav database contains data from a questionnaire which measured the financial knowledge of tertiary school students. In the topics, higher points mean higher financial knowledge.

1. Is there any relationship between the participation in financial education and the type of school at 5% significance level?
2. Is there any relationship between the general economic knowledge and the type of school at 5% significance level?

**2**. In a company, there was a survey among the employees about the years spent in education (year, x) and about the monthly salary (HUF, y).

1. Fit a linear regression model to the data! Interpret the regression parameters!
2. Interpret the standard error of the estimate and the coefficient of determination!

|  |  |
| --- | --- |
| Years spent in education (year, x) | Monthly salary (HUF, y) |
| 19 | 167500 |
| 12 | 92260 |
| 12 | 93700 |
| 16 | 172200 |
| 19 | 188600 |
| 12 | 59900 |
| 15 | 101500 |
| 12 | 96000 |
| 12 | 81380 |
| 12 | 75400 |
| 17 | 82300 |
| 15 | 117000 |
| 16 | 160020 |
| 15 | 73150 |
| 12 | 83700 |
| 12 | 88800 |
| 16 | 185000 |
| 8 | 52100 |
| 11 | 78000 |
| 17 | 121000 |
| 12 | 77900 |
| 9 | 77900 |
| 16 | 192000 |
| 12 | 86300 |
| 13 | 89000 |
| 10 | 63500 |
| 8 | 68500 |
| 17 | 146700 |
| 13 | 97600 |
| 15 | 101250 |

SPSS solutions

You can check your results if you open and watch the practice\_seminar\_test\_part2.wmv and the interpretations can be found here.

**1**. The finance.sav database contains data from a questionnaire which measured the financial knowledge of tertiary school students. In the topics, higher points mean higher financial knowledge.

1. Is there any relationship between the participation in financial education and the type of school at 5% significance level?

|  |
| --- |
| **Participation in financial education \* Type of school Crosstabulation** |
|  | Type of school | Total |
| grammar school | vocational school | vocational technical school |
| Participation in financial education | yes | Count | 190 | 711 | 126 | 1027 |
| % within Type of school | 21,3% | 31,2% | 30,5% | 28,6% |
| no | Count | 700 | 1571 | 287 | 2558 |
| % within Type of school | 78,7% | 68,8% | 69,5% | 71,4% |
| Total | Count | 890 | 2282 | 413 | 3585 |
| % within Type of school | 100,0% | 100,0% | 100,0% | 100,0% |

|  |
| --- |
| **Chi-Square Tests** |
|  | Value | df | Asymp. Sig. (2-sided) |
| Pearson Chi-Square | 30,928a | 2 | ,000 |
| Likelihood Ratio | 32,205 | 2 | ,000 |
| Linear-by-Linear Association | 20,824 | 1 | ,000 |
| N of Valid Cases | 3585 |  |  |
| a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 118,31. |

|  |
| --- |
| **Symmetric Measures** |
|  | Value | Approx. Sig. |
| Nominal by Nominal | Phi | ,093 | ,000 |
| Cramer's V | ,093 | ,000 |
| N of Valid Cases | 3585 |  |

We examine a relationship between two categorical variables, therefore crosstabs analysis can be applied for answering this question.

The nullhypothesis of the test is that there is no significant relationship between the participation in financial education and the type of school.

Based on the note below the table, the application condition of the test is met.

At a 5% significance level, we reject the nullhypothesis (Pearson Chi Square sig<0.05), so there is a significant relationship between the participation in financial education and the type of school.

The relationship between the examined variables is weak (C=0.093).

The ratio of those who have participated in financial education is higher among vocational school (31.2%) than in the whole sample (28.6%). The ratio of those who have not participated in financial education is higher among grammar school (78.7%) than in the whole sample (71.4%).

1. Is there any relationship between the general economic knowledge and the type of school at 5% significance level?

|  |
| --- |
| **Descriptives** |
| general economic knowledge, points  |
|  | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | Minimum | Maximum |
| Lower Bound | Upper Bound |
| grammar school | 890 | 1,2798 | ,88947 | ,02982 | 1,2213 | 1,3383 | ,00 | 3,00 |
| vocational school | 2282 | 1,3129 | ,94630 | ,01981 | 1,2740 | 1,3517 | ,00 | 3,00 |
| vocational technical school | 413 | ,9613 | ,86656 | ,04264 | ,8774 | 1,0451 | ,00 | 3,00 |
| Total | 3585 | 1,2642 | ,92986 | ,01553 | 1,2337 | 1,2946 | ,00 | 3,00 |

|  |
| --- |
| **Test of Homogeneity of Variances** |
| general economic knowledge, points  |
| Levene Statistic | df1 | df2 | Sig. |
| 18,469 | 2 | 3582 | ,000 |

|  |
| --- |
| **Robust Tests of Equality of Means** |
| general economic knowledge, points  |
|  | Statistica | df1 | df2 | Sig. |
| Welch | 28,302 | 2 | 1062,186 | ,000 |
| a. Asymptotically F distributed. |

|  |
| --- |
| **Multiple Comparisons** |
| Dependent Variable: general economic knowledge, points  |
| Tamhane  |
| (I) Type of school | (J) Type of school | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval |
| Lower Bound | Upper Bound |
| grammar school | vocational school | -,03311 | ,03580 | ,732 | -,1187 | ,0524 |
| vocational technical school | ,31852\* | ,05203 | ,000 | ,1940 | ,4430 |
| vocational school | grammar school | ,03311 | ,03580 | ,732 | -,0524 | ,1187 |
| vocational technical school | ,35162\* | ,04702 | ,000 | ,2390 | ,4642 |
| vocational technical school | grammar school | -,31852\* | ,05203 | ,000 | -,4430 | -,1940 |
| vocational school | -,35162\* | ,04702 | ,000 | -,4642 | -,2390 |
| \*. The mean difference is significant at the 0.05 level. |

|  |
| --- |
| **Measures of Association** |
|  | Eta | Eta Squared |
| general economic knowledge, points \* Type of school | ,119 | ,014 |

We examine a relationship between a categorical and a metric variable, therefore ANOVA can be applied for answering this question.

The nullhypothesis of the test is that there is no significant relationship between the general economic knowledge and the type of school.

The variance homogeneity cannot be assumed based on the Levene-test sig<0.05 value, so the Welch test should be considered for answering the main question.

At 5% significance level, we reject the H0 (sig<0.05), so there is a significant relationship between the general economic knowledge and the type of school.

If we consider the pairwise comparisons of each group means, the mean of general economic knowledge points in vocational technical school is significantly lower than the mean of general economic points in grammar school or in vocational school.

There is a weak relationship between the general economic knowledge and the type of school (H=0.119). 1.4% of the variance in general economic knowledge is explained by the type of school (H2=0.014). The rest 98.6% is explained by other factors.

**2.** In a company, there was a survey among the employees about the years spent in education (year, x) and about the monthly salary (HUF, y).

1. Fit a linear regression model to the data! Interpret the regression parameters!

|  |
| --- |
| **Coefficients** |
|  | Unstandardized Coefficients | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| years | 10921,262 | 1641,214 | ,783 | 6,654 | ,000 |
| (Constant) | -41765,041 | 22664,205 |  | -1,843 | ,076 |

$$\hat{y}=-41765+10921∙x$$

If the years spent in education is 0 years, the estimated monthly

salary is -41765.041 HUF. (no sense)

If the years spent in education increases by 1 year, the estimated monthly salary increases on average by 10921.262 HUF.

1. Interpret the standard error of the estimate and the coefficient of determination!

|  |
| --- |
| **Model Summary** |
| R | R Square | Adjusted R Square | Std. Error of the Estimate |
| ,783 | ,613 | ,599 | 26130,885 |
| The independent variable is years. |

The average difference between the observed and the estimated values of the monthly salary is 26130.885 HUF.

61.3% of the variation of the monthly salary is explained by the years spent in education. The remaining 28.7% is explained by other factors.

This teaching material has been compiled at the University of Szeged, and is supported by the European Union. Project identity number: EFOP-3.4.3-16-2016-00014