Péter Kovács, PhD - Éva Kuruczleki

Statistics I

learning guide

2018

Contents

Preface

In order to understand the news, social and business phenomena and our environment, to interpret the relationships among social and business data correctly we need statistical literacy, reasoning and thinking. This can include knowledge of basic statistical key figures, understanding concepts describing society (e.g. inflation, unemployment, GDP, etc.), basic information about research methods (from the viewpoint of both use and interpretation), basic information about visualization (about both visualization and interpretation) and the knowledge about data sources and the ability to evaluate the used data sources. The huge amounts of data, data sources and visualization tools (for instance Gapminder, OECD, Eurostat, national statistical agencies,) on the internet provide an opportunity to illustrate complex relations with real data relatively easily. At the same time, the misuse of these tools can lead to misinterpretations.

The main goal of the course is to improve your statistical literacy, reasoning and thinking: how can we identify applicability of statistics as a way of solution, the suitable statistical method, interpret data and results in case of a given problem.

This course is two semesters long. The first semester is an introduction to Statistics. We explore basic terms, descriptive statistics (central tendencies, dispersion, other measures), comparison of data (ratios, index numbers) and time series (trend, seasonality, increment of growth, growth rate). We will also focus on the interpretation of the data and results. During the first semester, we perform paper-based exercises and computer based tasks with the use of statistical databases, visualization tools and Excel functions and Excel PIVOT.

Literature for the first 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 containing the key terms, the sources of the materials, suggested learning activities, sample exercises and solutions in each topic; the same topics and similar tasks are discussed during the lectures and seminars with wider explanations.

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

1 Introduction to statistics

Goals

This chapter introduces the basic terms of statistics. The goal of this chapter is to provide the foundations and the framework of statistics for further chapters. This chapter is successful if the Reader

- learns how to distinguish between the levels of measurement,
- becomes able to explain the meaning of descriptive and inferential statistics,
- learns how to identify autonomously the structure of a dataset,
- improves their knowledge on basic terms such as the meaning and types of statistics or the differences between a sample or a population.

Knowledge obtained by reading this chapter: basic terms of statistics, steps of statistical analysis, measurement levels

Skills obtained by reading this chapter:

- statistical communication – basic terminology, making connections between statistical and everyday terms,
- organization – design, plan and carry out analysis following the necessary steps of statistical analyses.

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

This chapter makes the Reader to be autonomous in: differentiation samples from the population, identifying variables and their measurement levels.

Definitions

Statistics: is the science of collecting, organizing, presenting, analyzing, and interpreting numerical data to assist in making more effective decisions.

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

Registers: list of individuals (for instance: economic units, administrative units)

Sample: is a portion, or part, of the population of interest

Descriptive Statistics: Methods of organizing, summarizing, and presenting data in an informative way.

Inferential Statistics: A decision, estimate, prediction, or generalization about a population, based on a sample.

Steps of statistical analysis:

- planning

- data collection
- check and clean the data
- analysis
- presentation, feedbacks

Levels of measurement:

- categorical
    - nominal: Data that is classified into categories and cannot be arranged in any particular order. Example: eye color, gender, religious affiliation.
    - ordinal: involves data arranged in some order, but the differences between data values cannot be determined or are meaningless. Example: During a taste test of 4 soft drinks, Mellow Yellow was ranked number 1, Sprite number 2, Seven-up number 3, and Orange Crush number 4.
- noncategorical, quantitative (metric, scale)
    - interval: similar to the ordinal level, with the additional property that meaningful amounts of differences between data values can be determined. There is no natural zero point. Example: Temperature on the Fahrenheit scale.
    - ratio: the interval level with an inherent zero starting point. Differences and ratios are meaningful for this level of measurement. Example: Monthly income of surgeons, or distance traveled by manufacturer's representatives per month.

Learning activities

In order to learn the basic terms

1. Read Chapter 1 from the book (Page 2-16).
2. Open and explore 1_introduction.ppt.
3. Explore and solve the sample tasks.
4. Check your knowledge: solve the chapter exercises in the book.

Sample tasks

---

1. The bank2.xls file contains employees' data of a bank.
    a. What is the population size?
    b. How many variables are in the data file? What are the measurement levels of the variables?

2. We would like to examine the statistics class.
    a. What is the target population?
    b. What is the population size?
    c. What is 1 unit?

Sample tasks solutions

1. The bank2.xls file contains employees' data of a bank.
    a. What is the population size?
    - N=474

    b. How many variables are in the data file? What are the measurement levels of the variables?
    - 6 variables (But ID is not important from the point of view of properties of individuals, we cannot analyze that in a statistical way.)
    - beginning salary: ratio
    - gender: nominal
    - age group: ordinal
    - current salary: ratio
    - language exam level: ordinal

2. We would like to examine the statistics class.
    a. What is the target population?
    - Those students who are sitting in the room at the moment of examination.

    b. What is the population size?
    - e.g. N=25

    c. What is 1 unit?
    - One student

2 Descriptive statistics

Descriptive statistics is a collection of methods of organizing, summarizing, and presenting data in an informative way. Different methods can be applied in the different measurement levels:

- Nominal level: frequency, relative frequency, distribution (tables, charts), mode
- Ordinal level: frequency, relative frequency, distribution (tables, charts), mode, median
- Quantitative variable (scale level):
    - frequency, relative frequency,
    - distribution (tables, charts), mode
    - measures of central tendencies: mode, median, mean
    - deviation and dispersion
    - measures of the distribution shape (skewness, kurtosis)

2.1 Tables, charts

Goals

This chapter introduces the basic information compressing tools. Learning of this chapter is successful if the Reader is able to do the followings:

- create and interpret basic tables and charts
- use the PIVOT function of Excel
- collect and use data from the website of official statistics.

Knowledge obtained by reading this chapter:

- basic terms of descriptive statistics: frequency, frequency distribution
- simple tables and graphs
- Excel functions, PIVOT

Skills obtained by reading this chapter:

- statistical reasoning – defining elements of statistical situations, describing a population
- statistical communication – organize the date in an easily understandable, visually pleasing way with the help of tables and graphs

Attitudes developed by reading this chapter: openness to data visualization and organization

This chapter makes the Reader to be autonomous in: choosing the proper table or graph to visualize data and to create summary statistics with the help of PIVOT tables

Definitions

Class frequency: The number of observations in each class

A Frequency distribution is a grouping of data into mutually exclusive categories showing the number of observations in each class.

A relative frequency distribution shows the percent of observations in each class.

Tools for publishing statistical data: tables and charts

Format requirements for tables:

- title
- units, titles of rows and columns
- sum
- data source
- notices
- order of categories

Types of charts:

- Scatter
- Line
- Bar
- Pie
- Pictogram
- Cartogram

Learning activities

In order to learn how to create and interpret tables and charts

1. Read Chapter 2 from the book (Page 22-52).
2. Open and explore 2_1_tables_and_charts.ppt.
3. Explore Excel Pivot char function with Easy Excel: http://www.Excel-easy.com/data-analysis/pivot-tables.html
4. Explore and solve the sample tasks.
5. Check your knowledge: solve the chapter exercises in the book.

Sample tasks

---

1. The bank2.xls file contains employees' data of a bank. Solve problems below with Excel **PIVOT tables.**

a. How can we describe the employees by gender? Describe it by table and chart too.
b. How can we describe the employees by language exam level?
c. How can we describe the employees under 40 by gender?
d. How can we describe the employees by gender and language exam in the same time?
e. Describe men and women separately according to language exam level distribution. Compare data.
f. What is the ratio of man on the different language exam levels?

2. Explore statistical databases: HCSO, EUROSTAT, OECD

a. What is the number of unemployment in Hungary in 2014 and 2015? What about the unemployment rate?
b. Consider the methodology.

c. Compare the Hungarian data with the European average.

Sample tasks solutions

---

1. The bank2.xls file contains employees' data of a bank. Solve problems below with Excel **PIVOT tables.**

a. How can we describe the employees by gender?

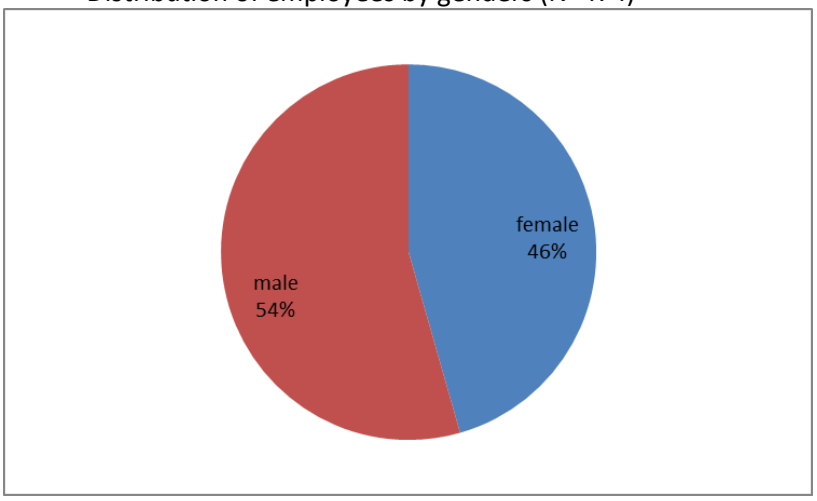
Number of employees by genders

Gender	Number of employees
female	216
male	258
Total	**474**

Source: bank.xls

There are 474 employees in the bank, where 258 persons are male and 216 persons are female.

Distribution of employees by genders (N=474)



Source: bank.xls

46% of the employees are female and 54% of the employees are male.

b. How can we describe the employees by language exam level?

Number of employees by language exam level

Language exam level	Number of employees
No	53
A	196
B	195
C	30
Total	**474**

Source: bank.xls

The number of employees with A level language exam is 196 persons.

c. How can we describe the employees under 40 by gender?

Distribution of employees under 40 by gender

Gender	Distribution, %
female	39.94
male	60.06
Total	**100.00**

Source: bank.xls

60% of the employees who are under 40 are male.

d. How can we describe the employees by gender and language exam in the same time?

Number of employees by gender and language exam, person

Gender	A	B	C	No	Total
female	128	58		30	216
male	68	137	30	23	258
Total	**196**	**195**	**30**	**53**	**474**

Source: bank.xls

The total number of employees is 474.
The total number of men is 258.
The total number of people who have language exam level B is 195.
The total number of female who have language exam level A is 128.

Distribution of employees by gender and language exam, %

Gender	A	B	C	No	Total
female	27.00	12.24	0.00	6.33	45.57
male	14.35	28.90	6.33	4.85	54.43
Total	**41.35**	**41.14**	**6.33**	**11.18**	**100.00**

Source: bank.xls

54 percent of the employees are male.
41 percent of the employees have language exam level B.
27 percent of the employees are female with language exam level A.

e. Describe man and women separately according to language exam level distribution. Compare data.

Distribution of language exam level and by gender, %

Gender	A	B	C	No	Total
female	59.26	26.85	0.00	13.89	100.00
male	26.36	53.10	11.63	8.91	100.00
Total	**41.35**	**41.14**	**6.33**	**11.18**	**100.00**

Source: bank.xls

11 percent of employees have no language exam.
53 percent of the male have language exam level B.
27 percent of the female have language exam level B.
59 percent of the female have language exam level A.

Compare values: by calculating difference or ratio

e.g. 59% and 27%

- 59/27=2.2

- If we consider females, the probability of that a woman has a language exam level A is 2.2 times higher than a female has a language exam level B.

e.g. 53% and 27%

- 53/27=2
- The probability that a male has a language exam level B is 2 times higher that than probability that a female has a language exam level B.
- The chance that we can found a person with language exam level B is two times higher among man than among women.

f. What is the ratio of man on the different language exam level?

Distribution of man and woman level by language exam level, %

Gender	A	B	C	No	Total
female	65.31	29.74	0.00	56.60	45.57
male	34.69	70.26	100.00	43.40	54.43
Total	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

Source: bank.xls

The ratio of men among language exam level A is 34.69%. All of the respondents are men within those who have language exam level C.

2. Explore statistical databases: HCSO, EUROSTAT, OECD

a. What is the number of unemployment in Hungary in 2014 and 2015? What about the unemployment rate?

- Data in Hungarian Central Statistical office can be found in the following link: http://www.ksh.hu/?lang=en
- Go to DATA → TABLES (STADAT), than choose a topic (e.g. Society→ Labour Market) and look for the table which contains the data what is searched for

b. Consider the methodology.

Methodology can be found in each table on the upper left corner (by clicking the link 'Methodology').

c. Compare the Hungarian data with the European average.

International comparisons can be done e.g. with data

- available in HCSO in the topic 'International statistics'
- available in Eurostat http://ec.europa.eu/eurostat
- available in OECD http://stats.oecd.org/

2.2 Measures of central tendencies

Goals

This chapter introduces the basic measures of central tendencies. Learning of this chapter is successful if the Reader:

- can explain the characteristics, uses, advantages, and disadvantages of each measure of central tendencies
- can compute and interpret the mode, the median and the mean of ungrouped and grouped data.

Knowledge obtained by reading this chapter: measures of central tendency – mean, mode, median.

Skills obtained by reading this chapter:

- statistical communication – describing a population with the help of measures of central tendency,
- logical skills – identifying which mean formula is needed in certain situations (i.e. differentiating between arithmetic and harmonic mean formulas or if a weighted formula is needed).

Attitudes developed by reading this chapter: confidence in the application of the measures of central tendency.

This chapter makes the Reader to be autonomous in: applying the measures of central tendency at population data outside of the context of this learning guide.

Definitions

Mode (Mo): is the value of the observation that appears most frequently

Median (Me): is the midpoint of the values after they have been ordered from the smallest to the largest.

- If N (number of cases) is odd: the middle element in the ranked data
- If N (number of cases) is even: the mean of the two middle elements in the ranked data

Mean ($\overline{x}$): is obtained by dividing the sum of all values by the number of values in the data set.

Learning activities

In order to learn the concept and the measurement of central tendencies, definition, calculation and interpretation of dispersion measures (Mo, Me, $\overline{x}$)

1. Read Chapter 3 from the book (Page 65-86).
2. Open and explore 2_2_central_tendencies.ppt.
3. Explore and solve the sample tasks.
4. Check your knowledge: solve the chapter exercises in the book.

Sample tasks

1. There are several pieces of data for 9 employees of a company in the given table:

Gender	Monthly gross salary (thousand HUF)
male	100
male	140
male	120
male	120
female	80
female	90
female	85
female	100
female	105

A) Calculate and interpret the sum of the monthly gross salaries grouped by gender.
B) Calculate and interpret the mean of the monthly gross salaries grouped by gender. Calculate the mean of the monthly gross salaries for all employees using different weighted and unweighted formulas.
C) Calculate and interpret the mode of the monthly gross salaries.
D) Calculate and interpret the median of the monthly gross salaries.

2. The bank.xls file contains employees' data of a bank.

- Calculate with the help of Excel functions the
  - the sum,
  - the mean,
  - the median,
  - the mode of the current salaries.

3. In a company, the blue collar workers' average monthly salary is 120 HUF, and the white collar workers' average monthly salary is 200 thousand HUF. The ratio of white collar workers is higher than the ratio of blue collar workers by 30 percentage points. Calculate the workers' average monthly salary.

4. There was a research about the habits of internet users. It has turned out that the average time a person spent on the internet within the group of primary education level respondents is 20 minutes. The average time a person spent on the internet within the group of secondary education level respondents is 40 minutes; and the average time a person spent on the internet within the group of tertiary education level respondents is 70 minutes. It is also known that the sum of the time spent on the internet within the group of primary education level respondents is 300 minutes. The sum of the time spent on the internet within the group of secondary education level respondents is 1000 minutes; while the sum of the time spent on the internet within the tertiary group of education level respondents is 2800 minutes. **How much is the average time spent on the internet within the respondent groups?**

5. There is a summary about the flight times of pilots in a low-cost airline:

Flight time, hour	Number of pilots, person
2	30
4	10
8	40
12	20
Total	100

A) How many hours did the pilots fly together?
B) What is the mean of flight times?
C) Calculate and interpret the median.
D) Calculate and interpret the mode.

6. There is a summary about distances swam in a swimming pool:

Distance swam, meter	Number of swimmers, person
1-500	20
501-1000	45
1001-2000	60
2001-3000	12

A) What is the average distance swam?
B) Estimate and interpret the mode.
C) Estimate and interpret the median.

Sample tasks solutions

---

1. There are several data for 9 employees of a company in the given table:

Gender	Monthly gross salary (thousand HUF)
male	100
male	140
male	120
male	120
female	80
female	90
female	85
female	100
female	105

A) Calculate and interpret the sum of the monthly gross salaries grouped by gender.

Organize data and results of tasks A and B in a table.

Gender	Number of employees, person	Sum of monthly gross salaries, thousand HUF	Ratio of sum of monthly gross salaries, %	Ratio of people, %	Mean of monthly gross salaries, thousand HUF
Male	4	480	52	44	120
Female	5	460	48	56	92
Total	**9**	**940**	**100**	**100**	**104.4**

$$\mathrm{Sm} = 100 + 140 + 120 + 120 = 480 \quad \text{thousand HUF}$$
$$\mathrm{Sf} = 80 + 90 + 85 + 100 + 105 = 460 \quad \text{thousand HUF}$$
$$\mathrm{S} = \sum \mathrm{S_j} = 480 + 460 = 940 \quad \text{thousand HUF}$$

The sum of the monthly gross salaries is 480 thousand HUF for males.
The sum of the monthly gross salaries is 460 thousand HUF for females.
The sum of the monthly gross salaries is 940 thousand HUF.

Other interpretations e.g.:
52%: Males earn 52% of the total salaries.
44% of the employees are male.

B) Calculate and interpret the mean of the monthly gross salaries grouped by gender. Calculate the mean of the monthly gross salaries for all employees using different weighted and unweighted formulas.

$$\bar{\mathrm{x}}_{\mathrm{f}} = \frac{480}{4} = 120 \quad \text{thousand HUF}$$
$$\bar{\mathrm{x}}_{\mathrm{n}} = \frac{460}{5} = 92 \quad \text{thousand HUF}$$
$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{100+140+120+120+80+90+85+100+105}{9} = 104.44 \quad \textit{thousand HUF}$$
$$\bar{x} = \frac{\sum_{j=1}^{k} f_j \bar{x}_j}{\sum_{i=1}^{k} f_j} = \frac{4 \cdot 120 + 5 \cdot 92}{9} = 104.44 \quad \textit{thousand HUF}$$
$$\bar{x} = \frac{\sum_{j=1}^{k} S_j}{\sum_{j=1}^{k} \frac{S_j}{\bar{x}_j}} = \frac{940}{\frac{480}{120} + \frac{460}{92}} = 104.44 \quad \textit{thousand HUF}$$

The mean of the monthly gross salaries is 120 thousand HUF for males.
The mean of the monthly gross salaries is 92 thousand HUF for women.
The mean of the monthly gross salaries is 104.44 thousand HUF.

C) Calculate and interpret the mode of the monthly gross salaries.
Mo_1=100, Mo_2=120

The most frequent salaries are the 100.000 Ft and the 120.000 Ft.

D) Calculate and interpret the median of the monthly gross salaries.
-Rank cases: 80, 85, 90, 100, 100, 105, 120, 120, 140
-Me=100 thousand HUF

Half of the monthly gross salaries are less or equal to 100 thousand Ft. (The half of the employees earns maximum 100 thousand FT.)

2. The bank.xls file contains employees' data of a bank.

- Calculate with the help of Excel functions the
  - the sum,
  - the mean,
  - the median,
  - the mode of the current salaries.

Solutions:
- sum: use the SUM() function
- mean: use the AVERAGE() function
- median: use the MEDIAN() function
- mode: use the MODE() function

Descriptive statistics about the current salaries

Sum, USD	6525950
Mean, USD	13 768
Median, USD	11550
Mode, USD	12300

Source: bank.xls

3. In a company, the blue collar workers' average monthly salary is 120 HUF, and the white collar workers' average monthly salary is 200 thousand HUF. The ratio of white collar workers is higher than the ratio of blue collar workers by 30 percentage points. Calculate the workers' average monthly salary.

We don't know the number of employees, but the whole company is considered 100%
100%=white collar (W) +blue collar workers (B)
B=W-30

$$100 = W + (W - 30)$$

$$W = 65 \quad (\text{ratio of white collar workers})$$

Ratio of blue collar workers: 100-65=35%

$$\bar{x} = \frac{65 \cdot 200 + 35 \cdot 120}{100} = 172\,\text{HUF}$$

The workers' average monthly salary is 172 HUF.

4. There was a research about the habits of internet users. It has turned out that the average time a person spent on the internet within the group of primary education level respondents is 20 minutes. The average time a person spent on the internet within the group of secondary education level respondents is 40 minutes; and the average time a person spent on the internet within the group of tertiary education level respondents is 70 minutes. It is also known that the sum of the time spent on the internet within the group of primary education level respondents is 300 minutes. The sum of the time spent on the internet within the group of secondary education level respondents is 1000 minutes; while the sum of the time spent on the internet within the tertiary group of education level respondents is 2800 minutes. **How much is the average time spent on the internet within the respondent groups?**

$$avg\ time = \frac{sum\ of\ time}{number\ of\ respondents} \quad \rightarrow \quad number\ of\ respondents = \frac{sum\ of\ time}{avg\ time}$$

Grand mean:

$$\text{Avg time} = \frac{\sum \text{sum of time}}{\sum \text{number of respondents}} = \frac{\sum \text{sum of time}}{\sum \frac{\text{sum of time}}{\text{avg time}}}$$

$$\bar{x} = \frac{\sum_{j=1}^{k} S_j}{\sum_{j=1}^{k} \frac{S_j}{\bar{x}_j}} = \frac{300+1000+2800}{\frac{300}{20}+\frac{1000}{40}+\frac{2800}{70}} = \frac{4100}{80} = 51.25 \text{ min/ person}$$

The average time spent by using internet is 51.25 minutes/person within the respondent group.

5. There is a summary about the flight times of pilots in a low-cost airline:

Flight time, hour (x_i)	Number of pilots, person (f_i)
2	30
4	10
8	40
12	20
Total	100

A) How many hours did the pilots fly together?
S=2*30+4*10+8*40+12*20=660 hours

The pilots flew 660 hours together.

B) What is the mean of flight times?

$$\bar{x} = \frac{660}{100} = 6.6\ hours/\ person$$

The average flight time is 6.6 hours/person.

C) Calculate and interpret the median.
Me= (50.case+ 51.case) / 2 = (8+8) /2 = 8 hours

Half of the pilots flew at least 8 hours.

D) Calculate and interpret the mode.
Mo=8 hours

The most frequent flight time is 8 hours.

6. There is a summary about the distances swam in a swimming pool:

Swam distance, meter	Number of swimmers, person (f_i)	Cumulative frequencies ($f'_{i)}$	Length of class ($h_{i)}$	Data density ($f_i/h_{i)}$	Class mark (midpoint) x_i
1-500	20	20	500	0.04	250
501-1000	45	65	500	0.09	750
1001-2000	60	125	1000	0.06	1500
2001-3000	12	137	1000	0.012	2500

A) What is the average distance swam?

$$\bar{x} = \frac{20 \cdot 250 + 45 \cdot 750 + 60 \cdot 1500 + 12 \cdot 2500}{137} = 1158.76\,\text{m}$$

The average distance swam is 11.76 m.

B-C) Estimate and interpret the mode and the median.
Median

- The median will be in the class where the following is true at first: $\frac{N}{2} \le f_i'$ → 68,5 ≤f_i' → median is in the 3rd class
- Me=1500 m
- The half of the swimmers swam maximum 1500 m.

Mode

- Based on data density, because the lengths of each class are not the same.
- $\frac{f_i}{h_i}$ is highest in the 2nd class
- Mo=750 m
- The most frequent distance swam is 750 m.

2.3 Dispersion

Goals

This chapter introduces the basic measures of dispersion. Learning of this chapter is successful if the Reader is able to do the followings:

- explain the characteristics, uses, advantages and disadvantages of each measure of dispersion
- compute and interpret the range, the variance, the standard deviation and the coefficient of variation of ungrouped- and grouped data too.

Knowledge obtained by reading this chapter: measures of dispersion – range, variance, standard deviation, coefficient of variation.

Skills obtained by reading this chapter:

- statistical communication – describing a population with the help of measures of dispersion, interpreting statistical language,
- logical skills – identifying which formula is needed in certain situations (i.e. differentiating between formulas for elementary data and grouped data); making connections between measures of central tendency and dispersion.

Attitudes developed by reading this chapter: confidence in the application of the measures of dispersion.

This chapter makes the Reader to be autonomous in: applying the measures of dispersion at population data outside of the context of this learning guide.

Definitions

- **Dispersion:** expresses the differences between the values and the value's deviation from the central tendencies.
- **Measures for dispersion:**
    - Difference
        - Range
        - Gini 's average absolute difference
    - Deviation
        - Standard deviation
        - Variance
        - Coefficient of variation
- **Standard deviation (σ):** shows in the unit of the examined variable how the individuals deviate on average from the mean.
- **Coefficient of variation (v):** shows in percentage how the individuals deviate on average from the mean.

Learning activities

In order to learn the concept and the measurement of dispersion, definition, calculation and interpretation of dispersion measures (σ, v)

1. Read Chapter 4 from the book (Page 100-116).
2. Open and explore 2_3_dispersion.ppt.
3. Explore and solve the sample tasks.
4. Check your knowledge: solve the chapter exercises in the book.

Sample tasks

---

1. There are several data for 9 employees of a company in the given table:

Gender	Monthly gross salary, thousand HUF
male	100
male	140
male	120
male	120
female	80
female	90
female	85
female	100
female	105

0) Calculate and interpret the mean. (On paper and in Excel too.)
A) Calculate and interpret the standard deviation. (On paper and in Excel too.)
B) Calculate and interpret the coefficient of variation. (On paper and in Excel too.)
C) Calculate and interpret the standard deviation in each group.
D) Calculate and interpret the variation of coefficient in each group.

2. There is a summary about the flight times of pilots in a low-cost airline:

Flight time, hour	Number of pilots, person
2	30
4	10
8	40
12	20
Total	100

A) Calculate and interpret the standard deviation. Try to use the Excel.
B) Calculate and interpret the coefficient of variation. Try to use the Excel.

3. There is a summary about the distances swam in a swimming pool:

Distance swam, meter	Number of swimmers, person
1-500	20
501-1000	45
1001-2000	60
2001-3000	12

A) Calculate and interpret the standard deviation.
B) Calculate and interpret the coefficient of variation.

4. The bank.xls file contains employees' data of a bank. Calculate with the help of Excel functions the

- the mean,
- standard deviation,
- coefficient of variation.

Sample tasks solutions

---

1. There are several data for 9 employees of a company in the given table:

Gender	Monthly gross salary (thousand HUF)
male	100
male	140
male	120
male	120
female	80
female	90
female	85
female	100
female	105

0) Calculate and interpret the mean. (On paper and in Excel too.)

$$\bar{x} = \frac{100+140+120+120+80+90+85+100+105}{9} = 104.4 \; thousand \; HUF$$

The average salary is 104,4 thousand HUF for all employees.

(Excel solution: use the AVERAGE function)

A) Calculate and interpret the standard deviation. (On paper and in Excel too.)

$$\sigma = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_N-\bar{x})^2}{N}} = \sqrt{\frac{(100-104.4)^2+(140-104.4)^2+\ldots+(105-104.4)^2}{9}} = 18.17 \; thousand \; HUF$$

or

$$\sigma = \sqrt{\frac{100^2+140^2+\ldots+105^2}{9} - 104.4^2} = 18.17 \; thousand \; HUF$$

The monthly gross salaries deviate on average by 18.17 thousand HUF from the mean of the employee's monthly gross salary.

(Excel solution: use the STDEVP function)

B) Calculate and interpret the coefficient of variation. (On paper and in Excel too.)

$$v = \frac{\sigma}{\bar{x}} = \frac{18.17}{104.4} = 0.174 \rightarrow 17.4\%$$

The monthly gross salaries deviate on average by 17.4 % from the mean of the employee's monthly gross salary.

(Excel solution: use mathematical operands)

C) Calculate and interpret the standard deviation in each group.

$$\bar{x}_{males} = \frac{100+140+120+120}{4} = 120\ thousand\ HUF$$

$$\sigma_{males} = \sqrt{\frac{(100-120)^2+(140-120)^2+(120-120)^2+(120-120)^2}{4}} = 14.14\ thousand\ HUF$$

The male's monthly gross salaries deviate on average by 14.14 thousand HUF from the mean of the male's monthly gross salary.

(Excel solution: use functions by selecting data for males)

$$\bar{x}_{females} = \frac{80+90+85+100+105}{5} = 92\ thousand\ HUF$$

$$\sigma_{females} = \sqrt{\frac{(80-92)^2+(90-92)^2+(85-92)^2+(100-92)^2+(105-92)^2}{5}} = 9.27\ thousand\ HUF$$

The female's monthly gross salaries deviate on average by 9.27 thousand HUF from the mean of the female's monthly gross salary.

(Excel solution: use functions by selecting data for females)

D) Calculate and interpret the variation of coefficient in each group.

$$v_{males} = \frac{\sigma_{males}}{\bar{x}_{males}} = \frac{14.14}{120} = 0.1179 \rightarrow 11.79\%$$

The male's monthly gross salaries deviate on average by 11.79% from the mean of the male's monthly gross salary.

$$v_{frmales} = \frac{\sigma_{females}}{\bar{x}_{females}} = \frac{9.27}{92} = 0.1008 \rightarrow 10.08\%$$

The female's monthly gross salaries deviate on average by 10.08% from the mean of the female's monthly gross salary.

2. There is a summary about the flight times of pilots in a low-cost airline:

Flight time, hour (x_i)	Number of pilots, person (f_i)
2	30
4	10
8	40
12	20
Total	100

A) Calculate and interpret the standard deviation.

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{30 \cdot 2 + 10 \cdot 4 + 40 \cdot 8 + 20 \cdot 12}{100} = \frac{660}{100} = 6.6 \text{ hours}$$

$$\sigma = \sqrt{\frac{30(2-6,6)^2 + 10(4-6,6)^2 + 40(8-6,6)^2 + 20(12-6,6)^2}{100}} = 3.69 \text{ hours}$$

or

$$\sigma = \sqrt{\frac{30 \cdot 2^2 + 10 \cdot 4^2 + 40 \cdot 8^2 + 30 \cdot 12^2}{100} - 6.6^2} = 3.69 \text{ hours}$$

The flight times deviate on average by 3.69 hours from the average flight time.

B) Calculate and interpret the coefficient of variation.

$$v = \frac{\sigma}{\bar{x}} = \frac{3,69}{6,6} = 0.56 \rightarrow 56\%$$

The flight times deviate on average by 56% from the average flight time.

3. There is a summary about the distances swam in a swimming pool:

Swam distance, meter	Number of swimmers, person (f_i)	(x_i)
1-500	20	250
501-1000	45	750
1001-2000	60	1500
2001-3000	12	2500

A) Calculate and interpret the standard deviation.

$$\bar{x} = \frac{20 \cdot 250 + 45 \cdot 750 + 60 \cdot 1500 + 12 \cdot 2500}{137} = 1158.76 \text{ meter}$$

$$\sigma = \sqrt{\frac{20(250-1158.76)^2 + 45(750-1158.76)^2 + 60(1500-1158.76)^2 + 12(2500-1158.76)^2}{137}} = 619.69$$

or

$$\sigma = \sqrt{\frac{20 \cdot 250^2 + 45 \cdot 750^2 + 60 \cdot 1500^2 + 12 \cdot 2500^2}{137} - 1158.76^2} = 619.69$$

The distances swam deviate on average by 619.69 meter from the average distance swam.

B) Calculate and interpret the coefficient of variation.

$$v = \frac{\sigma}{\bar{x}} = \frac{619.69}{1158.76} = 0.5348 \rightarrow 53.48\%$$

The distances swam deviate on average by 53.48% from the average distance swam.

4. The bank.xls file contains employees' data of a bank. Calculate with the help of Excel functions the

- the mean,
- standard deviation,
- coefficient of variation.

Solutions:

- mean: use the AVERAGE() function
- standard deviation: use the STDEVP() function
- Coefficient of variation: use mathematical operands

Descriptive statistics about the current salaries

Mean, USD	13 767.83
Standard deviation, USD	6 823.06
Coefficient of variation, %	49.56

Source: bank.xls

2.4 Other Descriptive measures: Concentration, skewness

Goals

This chapter introduces the basic measures of concentration and skewness. Learning of this chapter is successful if the Reader is able to do the followings:

- explain the characteristics, uses, advantages, and disadvantages of each measure of concentration and skewness
- compute and interpret the HI, the HI*, the P and the F measures.

Knowledge obtained by reading this chapter:

- measures of concentration,
- measures of skewness.

Skills obtained by reading this chapter:

- statistical communication – describing the distribution of values within a population with the help of measures of concentration and skewness,
- analytical skills – creating comprehensive descriptive statistical analysis of any grouped and ungrouped population data both paper-based and with the help of Excel functions.

Attitudes developed by reading this chapter: confidence in the application of the measures of central tendency.

This chapter makes the Reader to be autonomous in: combining the methods learned in the previous chapters to create descriptive analyses.

Definitions

Concentration: There is concentration in a population if a few individuals have a large part from the sum of values.

Measures for concentration:

- Lorenz-curve
- Gini's Concentration index
- Herfindahl-index: higher HI values indicate higher level of concentration
- Normalized Herfindahl-index:
    - Higher HI* values indicate higher level of concentration
    - In several industries, HI* values above 0.18 indicate high level of concentration, HI* values below 0.1 indicate low level of concentration.
- Concentration rate (CRn, CR3, CR5, CR10)

Quartiles: Three summary measures that divide a ranked data set into four equal parts.

Boxplot: A boxplot is a graphical summary of data that is based on a five-number (x_{min}, Q1, Q2, Q3, x_{max}) summary.

Skewness: is the measurement of the lack of symmetry of the distribution.

Measures for skewness:

- Pearson (P)
- Fischer (F)

Learning activities

In order to learn the concept and the measurement of dispersion, definition, calculation and interpretation of dispersion measures (HI, HI*, F, P)

1. Read Chapter 4 from the book (Page 117-127).
2. Open and explore 2_4_other_descriptive_measures.ppt.
3. Explore and solve the sample tasks.
4. Check your knowledge: solve the chapter exercises in the book.

Sample tasks

---

1. The revenues of 10 big companies in an industry are known below:

Company name	Weekly revenue (thousand dollar)	Zi
Smallest	5	
Smaller	10	
Small	15	
Lower-medium size	20	
Medium size	25	
Upper-medium size	30	
Successful	45	
More successful	100	
2nd best	250	
Biggest	500	

Fill the empty cells in the table. Describe the concentration of the given industry with the help of Herfindahl-index (on paper and in Excel too).

2. In a company, 11 employees' salaries are known:

Monthly gross salaries (thousand HUF)
110
115
120
125
130

135
170
180
200
250
280

a. Examine the skewness of the monthly gross salaries with the help of P and F index (paper). (Q_1=120, Q_3=200)
b. Create a boxplot based on the salaries (paper).
c. Examine the skewness of the monthly gross salaries with the help of an Excel function.

Sample tasks solutions

---

1. The revenues of 10 big companies in an industry are known below:

Company name	**Weekly revenue (thousand dollar)**	**Zi**
Smallest	5	0.005
Smaller	10	0.010
Small	15	0.015
Lower-medium size	20	0.020
Medium size	25	0.025
Upper-medium size	30	0.030
Successful	45	0.045
More successful	100	0.100
2nd best	250	0.250
Biggest	500	0.500
total	1000	1

Fill the empty cells in the table. Describe the concentration of the given industry with the help of Herfindahl-index (on paper and in Excel too).

$$HI = \sum Z_i^2 = 0.005^2 + 0.01^2 + 0.015^2 + \ldots + 0.25^2 = 0.33$$

$$HI^* = \frac{HI - \frac{1}{N}}{1 - \frac{1}{N}} = \frac{0.33 - \frac{1}{10}}{1 - \frac{1}{10}} = 0.252$$

There is a high level of concentration.

Excel solution:

- Z_i: use mathematical operands
- HI: use the SUMSQ function
- HI*: use mathematical operands

2. In a company, 11 employees' salaries are known:

Monthly gross salaries (HUF)
110
115
120
125
130
135
170
180
200
250
280

a. **Examine the skewness of the monthly gross salaries with the help of P index.** (Q1=120, Q3=200)

$$\bar{x} = \frac{110 + 115 + \ldots + 280}{11} = 165$$

$$\text{Me} = 135$$

$$\sigma = \sqrt{\frac{(110-165)^2 + (115-165)^2 + \ldots + (280-165)^2}{11}} = 54.94$$

$$P = 3 \cdot \frac{\bar{x} - Me}{\sigma} = 3 \cdot \frac{165-135}{54.94} = 1.64$$

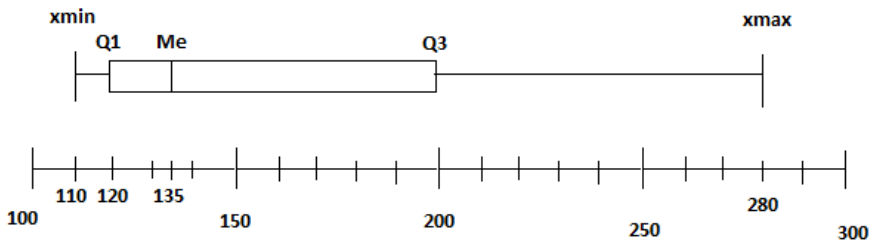
Most of the monthly gross salaries are below the mean.

$$F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)} = \frac{(200-135) - (135-120)}{(200-135) + (135-120)} = \frac{65-15}{65+15} = \frac{50}{80} = 0.625$$

The distribution of monthly gross salaries is skewed to the right.

b. Create a boxplot based on the salaries.

Boxplot based on the salaries



Source: task 2

c. Examine the skewness of the monthly gross salaries with the help of an Excel function.

Excel solution: use the SKEW function

Review Section (Topic 1-2)

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 measurement level of gender (male, female) is nominal.		
Skewness is the measurement of the lack of symmetry of a distribution.		
HI*=0.5 means that the concentration is low.		

2. Find and circle the correct answer from the list.
Median is

a) the most frequent value
b) the midpoint of the values after they have been ordered from the smallest to the large
c) shows how the individuals deviate on average from the mean
d) can be interpreted in nominal measurement level

If we consider the education variable (primary, secondary, tertiary school degrees)

a) that is a variable with ordinal measurement level
b) we cannot define median
c) we can calculate a mean
d) we can calculate standard deviation

3. There was a survey about the employees in a company. It is known that the average age within the employees in the marketing department is 28 years. The average age within the employees in the finance department is 32 years; and the average age within the employees in the production department is 35 years. It is also known that 16 employees work in the marketing department, 12 employees work in the finance department and 47 employees work in the production department. **Calculate** the average age within the company. **Interpret** the result.

4. There was a survey about the weekly statistics learning time. It is known that the average learning time within the BSc students is 6 hours, the average learning time within the MSc students is 4 hours and the average learning time within the PhD students is 2 hours. It is known also that the sum of learning time of BSc students is 120 hours, the sum of learning time of MSc students is 60 hours and the sum of learning time of PhD students is 14 hours. **Calculate** the average learning time among all of the students. **Interpret** the result.

5. The distances (km) completed in a running workout are known in the case of 10 runners:

4, 4, 5, 7, 7, 8, 4, 4, 10, 7

a) **Calculate and interpret** the mode.
b) **Calculate and interpret** the median.

6. Five employees work in a working group, their monthly gross salaries are below (thousand HUF):

140, 170, 200, 280, 300

It is also known that the average monthly gross salary is 218 thousand HUF.

Calculate and interpret the standard deviation.

7. There is a summary about prices of products in a case of a company:

Product price, thousand HUF	**Number of sold products (pieces)**
3.5	26
4.5	45
6.4	50
Total	121

Calculate and interpret the coefficient of variation.

8. The revenues of 4 soap manufacturer companies in an industry are known below:

Soap manufacturer company	**Weekly revenue (thousand dollar)**
Rose	25
Natural	50
Creamy	100
Cleaning	300

Describe the concentration of the soap manufacturer companies with the help of the **normalized Herfindahl-index**.

9. The daily working time (hours) is known in the case of eleven employees:

4, 4, 5, 5, 5, 6, 8, 9, 9, 11, 12

It is known also that Q1=5 hours, Me=6 hours, Q3=9 hours, $\bar{x} = 7.09$ hours, $\sigma = 2.71$ hours

a) **Calculate** the P and F measures. **Interpret** the results.
b) **Create a boxplot** based on the daily working hours.

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

Statement	TRUE	FALSE
The measurement level of gender (male, female) is nominal.	X	
Skewness is the measurement of the lack of symmetry of a distribution.	X	
HI*=0.5 means that the concentration is low.		X

2. Find and circle the correct answer from the list.
Median is

e) the most frequent value
f) the midpoint of the values after they have been ordered from the smallest to the large
g) shows how the individuals deviate on average from the mean
h) can be interpreted in nominal measurement level

If we consider the education variable (primary, secondary, tertiary school degrees)

e) that is a variable with ordinal measurement level
f) we cannot define median
g) we can calculate a mean
h) we can calculate standard deviation

3. There was a survey about the employees in a company. It is known that the average age within the employees in the marketing department is 28 years. The average age within the employees in the finance department is 32 years; and the average age within the employees in the production department is 35 years. It is also known that 16 employees work in the marketing department, 12 employees work in the finance department and 47 employees work in the production department. **Calculate** the average age among in the company. **Interpret** the result.

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{16 \cdot 28 + 12 \cdot 32 + 47 \cdot 35}{16 + 12 + 47} = 33.03 \text{ years}$$

The average age is 33.03 years in the company.

4. There was a survey about the weekly statistics learning time. It is known that the average learning time within the BSc students is 6 hours, the average learning time within the MSc students is 4 hours and the average learning time within the PhD students is 2 hours. It is known also that the sum of learning time of BSc students is 120 hours, the sum of learning time of MSc students is 60 hours and the sum of learning time of PhD students is 14 hours. **Calculate** the average learning time among all the students. **Interpret** the result.

$$\bar{x} = \frac{\sum_{j=1}^{k} S_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} S_j}{\sum_{j=1}^{k} \frac{S_j}{\bar{x}_j}} = \frac{120+60+14}{\frac{120}{6}+\frac{60}{4}+\frac{14}{2}} = 4.62 \text{ hours}$$

The average learning time among all of the students is 4.62 hours.

5. The distances (km) completed in a running workout are known in the case of 10 runners:

4, 4, 5, 7, 7, 8, 4, 4, 10, 7

c) **Calculate and interpret** the mode.
Mo=4 km
The most frequent distance is 4 km.

d) **Calculate and interpret** the median.
Ranked cases: 4, 4, 4, 4, 5, 7, 7, 7, 8, 10

$$Me = \frac{5^{th}\ case + 6^{th}\ case}{2} = \frac{5+7}{2} = 6 \text{ km}$$

Half of the completed distances are maximum 6 km.

6. Five employees work in a working group, their monthly gross salaries are below (thousand HUF):

140, 170, 200, 280, 300

It is also known that the average monthly gross salary is 218 thousand HUF.

Calculate and interpret the standard deviation.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{140+170+200+280+300}{5} = 218 \text{ thousand HUF}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}} = \sqrt{\frac{(140-218)^2 + (170-218)^2 + (200-218)^2 + (280-218)^2 + (300-218)^2}{5}} = 62.10$$

thousand HUF

The monthly gross salaries deviate on average by 62.10 thousand HUF from the average monthly gross salary.

7. There is a summary about prices of products in a case of a company:

Product price, thousand HUF	**Number of sold products (pieces)**
3.5	26
4.5	45
6.4	50

Calculate and interpret the coefficient of variation.

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{26 \cdot 3.5 + 45 \cdot 4.5 + 50 \cdot 6.4}{26 + 45 + 50} = 5.07 \text{ thousand HUF}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{26 \cdot (3.5 - 5.07)^2 + 26 \cdot 4(3.5 - 5.07)^2 + 50 \cdot (6.4 - 5.07)^2}{26 + 45 + 50}} = 1.18$$

thousand HUF

$$v = \frac{\sigma}{\bar{x}} = \frac{1.18}{5.07} = 0.2318 \rightarrow 23.18\%$$

The prices of products deviate on average by 23.18% from the mean of product prices.

8.

The revenues of 4 soap manufacturer companies in an industry are known below:

Soap manufacturer company	Weekly revenue (thousand dollar)	Z_i
Rose	25	$\frac{25}{475} = 0.05$
Natural	50	$\frac{50}{475} = 0.11$
Creamy	100	$\frac{100}{475} = 0.21$
Cleaning	300	$\frac{300}{475} = 0.63$
Total	475	1.00

Describe the concentration of the soap manufacturer companies with the help of the **normalized Herfindahl-index**.

$$HI = \sum_{i=1}^{N} Z_i^2 = 0.05^2 + 0.11^2 + 0.21^2 + 0.63^2 = 0.46$$

$$HI^* = \frac{HI - \frac{1}{N}}{1 - \frac{1}{N}} = \frac{0.46 - \frac{1}{4}}{1 - \frac{1}{4}} = 0.28$$

There is a high level of concentration.

9. The daily working time (hours) is known in the case of eleven employees:

4, 4, 5, 5, 5, 6, 8, 9, 9, 11, 12

It is known also that Q1=5 hours, Me=6 hours, Q3=9 hours, $\bar{x} = 7.09$ hours, $\sigma = 2.71$ hours

c) **Calculate** the P and F measures. **Interpret** the results.

$$P = 3 \cdot \frac{\bar{x} - Me}{\sigma} = 3 \cdot \frac{7.09 - 6}{2.71} = 1.21$$

Most of the daily working times are below the mean.

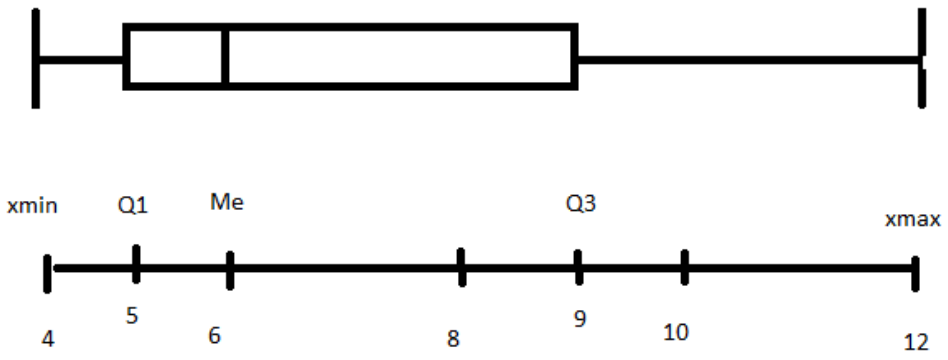
OR

The distribution of daily working times is skewed to the right.

$$F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)} = \frac{(9-6) - (6-5)}{(9-6) + (6-5)} = \frac{2}{4} = 0.5$$

The distribution of daily working times is skewed to the right.

d) **Create a boxplot** based on the daily working hours.



Excel exercises – Seminar part 1

The loan.xls file contains customer data of a bank. Describe the customers based on the aspects below. **Calculate** the results with the help of **Excel (functions, pivot)**, then copy the calculated results and tables into this document. **Interpret the results**.

Aspects:

1. Describe the customers' distribution by status (table and bar chart). Interpret one value from the table.
2. Describe the distribution of customers by status in each education category (table). Interpret one value from the table.

3. Create 5 categories from the debt rate variable. Describe the distribution of customers by debt rate categories (table). Interpret one value from the table.
4. Create descriptive statistics about the years at current workplace. Describe the minimum, maximum, median, mode, mean, standard deviation, coefficient of variation and the skewness of the years at current workplace. Interpret the calculated values.
5. Create a table which contains the mean and the standard deviation of the household income by status. Interpret all the values from a row of the table.

Excel solutions

Watch practice_seminar_part1_excel_solution.wmv

3 Comparison of data

A, B: two data. A can be compared to B

- by making difference: *A-B*
- by making a ratio: A/B

Applications

1. Comparing a group and the total population by
    - relative frequencies (discussed in descriptive statistics)
    - relative sum of values (discussed in descriptive statistics)
2. Comparing quantitative variables measured by values in a temporal or spatial analysis: examination of index numbers.
3. Comparing of changes of a variable in time: analysis of time series.

3.1 Index numbers

Goals

This chapter introduces the index numbers (price-, quantity- and value indices). Learning of this chapter is successful if the Reader is able to do the followings:

- understand the importance and application fields of index numbers
- compute and interpret the simple and aggregate indices.

Knowledge obtained by reading this chapter: knowledge of price, quantity and value indices.

Skills obtained by reading this chapter:

- analytical skills – the ability to compute, understand and interpret data related to economics that have further implications in other subjects as well.

Attitudes developed by reading this chapter: confidence in the application of weighted arithmetic and harmonic mean formulas in a different context other than the measures of central tendency.

This chapter makes the Reader autonomous in: analyzing the changes in values over time observed from different aspects.

Definitions

Simple index: shows the relative change in price, quantity or value of a given product in the current period compared to the base period.

Price index (simple): the average price change for the products of a product.

Quantity index (simple): the average price change for the products of a product.

Value index (simple): expresses the relative change of a phenomenon measured in value of a product.

Aggregate index: shows the relative change in price, quantity or value of a given basket of goods (product group) in the current period compared to the base period.

Price index (aggregate): the average price change for the products of a product group.

Quantity index (aggregate): the average price change for the products of a product group.

Value index (aggregate): expresses the relative change of a phenomenon measured in value of a product group.

Learning activities

In order to learn the concept, calculation and interpretation of index numbers

1. Read Chapter 18 from the book (Page 656-682).
2. Open the 3_1_Index_numbers.ppt.
3. Explore and solve the sample tasks.
4. Check your knowledge: solve the chapter exercises in the book.

Sample tasks

---

1. The given sales data are known in a case of three products:

Product	**2010. January**		**2011. January**	
	Unit price, Ft/kg	quantity, kg	Unit price, Ft/kg	quantity, kg
Marking				
A	53	110	65	96
B	81	175	96	176
C	159	23	176	34

a) Mark the columns. Calculate the values of sales for each product.
b) Calculate and interpret the simple price indices.
c) Calculate and interpret the simple quantity indices.
d) Calculate and interpret the simple value indices.
e) Calculate the aggregate values in both periods.
f) Calculate and interpret the value index for the product group.
g) Calculate and interpret the price indices for the product group.
h) Calculate and interpret the quantity indices for the product group.

2. In the case of a grocery the given turnover data are known:

Product	**Turnover in 2001 (million HUF)**	**Change of the turnover in 2001 compared to 2000 (%)**
Newspaper	30	110
Lottery	50	135
Sweetness	20	121

Calculate the change of the turnover for the grocery in 2001 compared to 2000.

3. There is company which produces three products. The given data are known for the sales:

Product	**Unit prices in 2002 compared to 2001, %**	**Sales in 2001, million HUF**	**Sales in 2002 compared to 2001, %**
A	110	330	130
B	100	430	110
C	120	440	105

a) Calculate the total relative change of sales (value index). Work on paper and in Excel also.
b) Calculate the effects behind the total relative change of revenues (price index, quantity index). Work on paper and in Excel also.
c) Create a coherent interpretation about the calculated results.

4. A farmer has the following revenue data:

Product	**Distribution of revenue in 2003 (%)**	**Price change, % (2004/2003)**	**Quantity change, % (2004/2003)**
Potato	36	+23	-10
Lettuce	40	+15	+10
Carrot	24	+20	-18

a) Calculate the relative revenue change of the farmer.
b) Calculate the effects behind the total relative change of revenues (price index, quantity index).
c) Create a coherent interpretation about the calculated results.

5. In a company, the given data are known for three products:

Products	**Production value (thousand Ft)**		**Production value in 2005 (2001=100.00%)**
	2005		
	at the price of 2001	at current price	
A product	600	500	91.63
B product	500	600	138.00
C product	800	900	112.50
Total	**1900**	**2000**	

a) Calculate the total relative change of production value (value index). Work on paper and in Excel also.
b) Calculate the effects behind the total relative change of revenues (price index, quantity index). Work on paper and in Excel also.

c) Create a coherent interpretation about the calculated results.

Sample tasks solutions

---

1. The given sales data are known in a case of three products:

Product	2010. January			2011. January			Simple quantity-index	Simple price-index	Simple value-index
	Unit price, Ft/kg	Quan-tity, kg	Value	Unit price, Ft/kg	Quan-tity, kg	Value			
Marking	***p0***	***q0***	***v0= q0p0***	***p1***	***q1***	***v1=q1p1***	***iq=q1/q0***	***ip=p1/p0***	***iv=v1/v0***
A	53	110	5830	65	96	6240	0.87	1.23	1.070
B	81	175	14175	96	176	16896	1.01	1.19	1.192
C	159	23	3657	176	34	5984	1.48	1.11	1.636

a) Mark the columns. Calculate the values of sales for each product.

b) Calculate and interpret the simple price indices.
1,23 → The price of the A product increased by 23% from 2010 January to 2011 January.

c) Calculate and interpret the simple quantity indices.
0,87 → The sold quantity of the A product decreased by 13% from 2010 January to 2011 January.

d) Calculate and interpret the simple value indices.
1,07 → The sales value of the A product increased by 7% from 2010 January to 2011 January.

e) Calculate the aggregate values in both periods.

$$\sum_{i=1}^{N} v_0 = \sum_{i=1}^{N} p_0 q_0 = 53 \cdot 110 + 81 \cdot 175 + 159 \cdot 23 = 23662$$

$$\sum_{i=1}^{N} v_1 = \sum_{i=1}^{N} p_1 q_1 = 65 \cdot 96 + 96 \cdot 176 + 176 \cdot 34 = 29120$$

$$\sum_{i=1}^{N} p_0 q_1 = 96 \cdot 53 + 176 \cdot 81 + 34 \cdot 176 = 24750$$

$$\sum_{i=1}^{N} p_1 q_0 = 110 \cdot 65 + 175 \cdot 96 + 23 \cdot 176 = 27998$$

f) Calculate and interpret the value index for the product group.

$$I_v = \frac{\sum v_1}{\sum v_0} = \frac{29120}{23662} = 1.2307$$

$$I_v = \frac{\sum_{i=1}^{n} v_0 i_v}{\sum_{i=1}^{n} v_0} = \frac{5830 \cdot 1{,}070 + 14175 \cdot 1{,}192 + 3657 \cdot 1{,}636}{23662} = 1.2307$$

$$I_v = \frac{\sum_{i=1}^{n} v_1}{\sum_{i=1}^{n} \frac{v_1}{i_v}} = \frac{29120}{\frac{6240}{1{,}070} + \frac{16896}{1{,}192} + \frac{5984}{1{,}636}} = 1.2307$$

The sales value of the product group increased by 23.07% from 2010 January to 2011 January.
OR:
The sales value of the products increased on average by 23.07% from 2010 January to 2011 January.

g) Calculate and interpret the price indices for the product group.

$$I_p^0 = \frac{\sum_{i=1}^{n} q_0 p_1}{\sum_{i=1}^{n} q_0 p_0} = \frac{110 \cdot 65 + 175 \cdot 96 + 23 \cdot 176}{23662} = \frac{27998}{23662} = 1.1832$$

$$I_p^1 = \frac{\sum_{i=1}^{n} q_1 p_1}{\sum_{i=1}^{n} q_1 p_0} = \frac{29120}{96 \cdot 53 + 176 \cdot 81 + 34 \cdot 176} = \frac{29120}{24750} = 1.1766$$

The prices of the products increased on average 18.32% from 2010 January to 2011 January.
OR:
Due to the price changes the sales value of the product group increased by 18.32 % from 2010 January to 2011 January.

h) Calculate and interpret the quantity indices for the product group.

$$I_q^0 = \frac{\sum_{i=1}^{n} q_1 p_0}{\sum_{i=1}^{n} q_0 p_0} = \frac{24750}{23662} = 1.046$$

$$I_q^1 = \frac{\sum_{i=1}^{n} q_1 p_1}{\sum_{i=1}^{n} q_0 p_1} = \frac{29120}{27998} = 1.0401$$

The sold quantities of the products increased on average 4.01% from 2010 January to 2011 January.
OR:
Due to the sold quantity changes the sales value of the product group increased by 4.01% from 2010 January to 2011 January.

2. In the case of a grocery the given turnover data are known:

Product	Turnover in 2001 (million HUF) (v1)	Change of the turnover in 2001 compared to 2000 (%) (iv)
Newspaper	30	110
Lottery	50	135
Sweetness	20	121

Calculate the change of the turnover for the grocery in 2001 compared to 2000.

$$I_v = \frac{\sum_{i=1}^{n} v_1}{\sum_{i=1}^{n} \frac{v_1}{i_v}} = \frac{100}{\frac{30}{1.1} + \frac{50}{1.35} + \frac{20}{1.21}} = 1.237$$

The turnover for the grocery increased by 23.7 percent to 2001 from 2000.

3. A company produces three products. The given data are known for the sales:

Product	Unit prices in 2002 compared to 2001, % (ip)	Sales in 2001, million HUF (v0)	Sales in 2002 compared to 2001, % (iv)
A	110	330	130
B	100	430	110
C	120	440	105

a) Calculate the total relative change of sales (value index). Work on paper and in Excel also.

$$I_v = \frac{\sum_{i=1}^{n} v_0 i_v}{\sum_{i=1}^{n} v_0} = \frac{330 \cdot 1.3 + 430 \cdot 1.1 + 440 \cdot 1.05}{1200} = 1.137$$

Excel solution: use the SUMPRODUCT function and mathematical operands

b) Calculate the effects behind the total relative change of revenues (price index, quantity index). Work on paper and in Excel also.

$$I_p^0 = \frac{\sum v_0 i_p}{\sum v_0} = \frac{330 \cdot 1.1 + 430 \cdot 1 + 440 \cdot 1.2}{1200} = 1.101$$

$$I_q^1 = \frac{I_v}{I_p^0} = \frac{1.137}{1.101} = 1.033$$

Excel solution: use the SUMPRODUCT function and mathematical operands

c) Create a coherent interpretation about the calculated results.

The prices of the products increased on average by 10.1 percent and the quantities of the products increased on average by 3.3 percent from 2001 to 2002. Consequently, the sales values of the products increased on average by 13.7 percent from 2001 to 2002.

4. A farmer has the following revenue data:

Product	Distribution of revenue in 2003 (%) (v0)	Price change, % (2004/2003) (ip)	Quantity change, % (2004/2003) (iq)	iv=ip*iq
Potato	36	+23	-10	1.107
Lettuce	40	+15	+10	1.265
Carrot	24	+20	-18	0.984

a) Calculate the relative revenue change of the farmer.

$$I_v = \frac{\sum_{i=1}^{n} v_0 i_v}{\sum_{i=1}^{n} v_0} = \frac{0.36 \cdot 1.107 + 0.4 \cdot 1.265 + 0.24 \cdot 0.984}{1} = 1.141$$

b) Calculate the effects behind the total relative change of revenues (price index, quantity index).

$$I_p^0 = \frac{\sum v_0 i_p}{\sum v_0} = \frac{0.36 \cdot 1.23 + 0.4 \cdot 1.15 + 0.24 \cdot 1.2}{1} = 1.191$$

$$I_q^1 = \frac{I_v}{I_p^0} = \frac{1.141}{1{,}191} = 0.958$$

c) Create a coherent interpretation about the calculated results.

The farmer's revenue increased by 14.1 percent from 2003 to 2004. This is caused by two factors: the prices and the sold quantities changed too.

Due to the price changes, the farmer's revenue increased by 19.1 percent from 2003 to 2004.

Due to the quantity changes the farmer's revenue decreased by 4.2 percent from 2003 to 2004.

5. In a company, the given data are known for three products:

Products	Production value (thousand Ft)		Production value in 2005 (2001=100.00%) (iv)
	2005		
	at the price of 2001 (q1p0)	at current price (q1p1=v1)	
A product	600	500	91.63
B product	500	600	138.00
C product	800	900	112.50
Total	1900	2000	

a) Calculate the total relative change of production value (value index). Work on paper and in Excel also.

$$I_v = \frac{\sum v_1}{\sum \frac{v_1}{i_v}} = \frac{2000}{\frac{500}{0.9163} + \frac{600}{1.38} + \frac{900}{1.125}} = 1.1233$$

Excel solution: use the SUMPRODUCT function and mathematical operands

b) Calculate the effects behind the total relative change of revenues (price index, quantity index). Work on paper and in Excel also.

$$I_p^1 = \frac{\sum q_1 p_1}{\sum q_1 p0} = \frac{2000}{1900} = 1.0526$$

$$I_q^0 = \frac{I_v}{I_p^1} = \frac{1.1233}{1.052} = 1.067$$

Excel solution: use the SUMPRODUCT function and mathematical operands!

c) Create a coherent interpretation about the calculated results.
The prices of the products were higher on average by 5.26 percent, and the quantities of the products were higher on average by 6.7 percent in 2005 compared to 2001. Consequently, the revenues of the products were higher on average by 12.33 percent in 2005 compared to 2001.

3.2 Time series

Goals

This chapter introduces time series analysis by using basic measures (increment of growth, growth rate) and models (decomposition model). Learning of this chapter is successful if the Reader is able to do the followings:

- calculate and interpret basic measures of time series analysis (increment of growth, growth rate);
- determine, compute and interpret linear and nonlinear trend equations;
- determine and interpret a set of seasonal differences and indices;
- use trend equations and seasonal differences and indices to forecast future time periods.

Knowledge obtained by reading this chapter:

- increment of growth, growth rate;
- time series analysis: linear and nonlinear trend, decomposition models.

Skills obtained by reading this chapter:

- statistical reasoning – applying statistical processes appropriate to the development path of time series data;
- communication – producing, understanding and interpreting time series analysis models.

Attitudes developed by reading this chapter: raising general curiosity for economic and social progress and the application of time series analysis outside of the context of this course.

This chapter makes the Reader autonomous in: creating forecast with the help of time series analysis techniques.

Definitions

Increment of growth: shows the average absolute change of the data per time unit in the given period.

Growth rate: shows the average relative change of the data per time unit in the given period.

Interpolation: forecasting within the period.

Extrapolation: forecasting outside the period.

Components of time series: trend, seasonal component, cyclical component, error term.

Additive decomposition model: assumes that time series data consist of the sum of the components of the time series.

Multiplicative decomposition model: assumes that time series data consist of the product of the components of the time series.

Trend: the long run direction of the time series.

b0 parameter of the linear trend (intercept of the linear trend function): shows the estimation of data when t=0.

b1 parameter of the linear trend (slope of the linear trend function): shows the average change of data from time unit to time unit.

Seasonal component: pattern in a time series within a year. These patterns tend to repeat themselves from year to year.

Learning activities

In order to learn the concept, calculation and interpretation of index numbers

1. Read Chapter 19 from the book (Page 690-714).
2. Open and explore 3_2_time_series.ppt.
3. Explore and solve the sample tasks.
4. Check your knowledge: solve the chapter exercises in the book.

Sample tasks

---

1. Examine the export of a company between 2000 and 2005.

Year	Export (t)
2000	200
2001	210
2002	218
2003	232
2004	240
2005	250

A) Calculate the yearly average export.
B) Calculate the changes
   a. by base ratios (compare to 2000),
   b. by link ratios.
C) Calculate **year by year**
   a. the relative changes,
   b. the absolute changes.
D) Calculate during the given period **the total**
   a. the relative changes,
   b. the absolute changes.
E) Calculate the **yearly average**
   a. relative change (growth rate),
   b. absolute change (increment of growth).

2. Examine the revenues of an enterprise.

Year	Revenue (m USD)	Revenue		Changes to the previous year		Revenue
		1999=100.0 %	Previous year =100.0 %	(m USD)	%	2000=100.0 %
1999						
2000		107.5				
2001		116.1				
2002			98.0			
2003		130.9				
2004	450.0					
2005			106.0			

Calculate the missing values if we know that in 2004 the revenue was 1.5 time higher than in 1999.

3. How many years are needed to decrease the debt of a company by 50%, if the debt of the company is decreasing yearly by 3% on average?

4. Examine the production of a company.

Year	Production, tons
2006	20
2007	22
2008	25
2009	28
2010	30
2011	34

A) Draw the time series. What kind of time series components can we see?
B) Set up a linear trend
   a. on paper
   b. with Excel chart
   c. with Excel functions
C) Interpret the trend parameters.
D) Estimate the value of 2015.

5. Download the date of the R&D expenditures between 1990 and 2014 from HCSO.
   a. Draw the time series. What kind of time series components can we see?
   b. Set up an exponential trend. Interpret the trend parameters. Estimate the R&D expenditure of 2016.
   c. Try to find another estimation.

6. Examine the number of passengers at Budapest Airport.

Quarters		Number of passengers (thousands)
2004.	I.	547
2004.	II.	829
2004.	III.	1254
2004.	IV.	920
2005.	I.	907
2005	II.	1238

2005	III.	1712
2005	IV.	1217
2006.	I.	1266
2006.	II.	1774
2006.	III.	2267
2006.	IV.	1543
2007.	I.	1485
2007.	II.	2084
2007.	III.	2780
2007.	IV.	1902

A) Draw the time series. Identify the components of the series.
B) Fit a linear trend. According to the linear trend **estimate the number of the passengers during the period**. Interpret the trend parameters. Draw the time series and the trend on the same chart. Estimate the number of the passengers of 2015. IV. quarter.
C) **Compute and interpret the seasonal differences**.
D) According to the linear trend and the seasonal variation **estimate the number of the passengers during the period**. Draw the time series, the trend and the estimation on the same chart. Estimate the number of the passengers of 2015. IV. quarter.
E) **Estimate the deseasonalized number of the passengers during the period**.

Sample tasks solutions

---

1. Examine the export of a company between 2000 and 2005.

Year	Export (t)	Export (2000=100%)	Export (Previous year=100%)	Changes to the previous year (%)	Changes to the previous year (t)
2000	200	100	-	-	-
2001	210	105	105	+5.0	+10
2002	218	109	103.8	+3.8	+8
2003	232	116	106.4	+6.4	+14
2004	240	120	103.4	+3.4	+8
2005	250	125	104.2	+4.2	+10

A) Calculate the yearly average export.

$$\bar{y} = \frac{200+210+218+232+240+250}{6} = 225 \quad t$$

B) Calculate the changes
   a. by base ratios (compare to 2000),

   for example:

$$b_{2001} = \frac{210}{200} = 1.05 \rightarrow 105\%$$

$$b_{2002} = \frac{218}{200} = 1.09 \rightarrow 109\%$$

$$b_{2003} = \frac{232}{200} = 1.16 \rightarrow 116\%$$ (See also table above.)

The export in 2003 was higher by 16 percent than in 2000.

b. by link ratios.

for example:

$$l_{2001} = \frac{210}{200} = 1.05 \rightarrow 105\%$$

$$l_{2002} = \frac{218}{210} = 1.038 \rightarrow 103.8\%$$

$$l_{2003} = \frac{232}{218} = 1.064 \rightarrow 106.4\%$$ (See also table above.)

The export in 2003 was higher by 6.4 percent than in 2002.

C) Calculate **year by year**

a. the relative changes,
b. the absolute changes.

See table above.

D) Calculate during the given period **the total**

a. the relative changes,

$$\frac{y_n}{y_1} = \frac{250}{200} = 1.25 \rightarrow 125\% \rightarrow +25\%$$

b. the absolute changes.

$$y_n - y_1 = 250 - 200 = +50\ t$$

E) Calculate the **yearly average**

a. relative change (growth rate),

$$\bar{l} = \sqrt[5]{1.25} = 1.05 \rightarrow 105\% \rightarrow +5\%$$

The yearly average relative change of the export was 5 percent in the given period.

b. absolute change (increment of growth).

The yearly average absolute change of the export was 10 tons in the given period.

2. Examine the revenues of an enterprise.

Year	Revenue (m USD)	Revenue		Changes to the previous year		Revenue
		1999=100.0 %	Previous year =100.0 %	(m USD)	%	2000=100.0 %
1999	300	100.0	-	-	-	300/322.5=93.02
2000	322.5	107.5	107.5	+22.5	+7.5	322.5/322.5=100.0
2001	348.3	116.1	108.0	+25.8	+8.0	348.3/322.5=108.0
2002	341.3	113.8	98.0	-7.0	-2.0	105.8
2003	392.7	130.9	115.1	+51.4	+15.1	121.8
2004	450.0	150	114.6	+57.3	+14.6	139.5
2005	477	159.0	106.0	+27.0	+6.0	147.9

Calculate the missing values if we know that in 2004 the revenue was 1.5 times higher than in 1999.

Possible steps for calculating revenues:

$$\frac{rev_{2004}}{rev_{1999}} = 1.5 \rightarrow \frac{450}{rev_{1999}} = 1.5 \rightarrow rev_{1999} = 300$$

$$\frac{rev_{2000}}{rev_{1999}} = 1.075 \rightarrow \frac{rev_{2000}}{300} = 1.075 \rightarrow rev_{2000} = 322.5$$

$$\frac{rev_{2001}}{rev_{1999}} = 1.161 \rightarrow \frac{rev_{2001}}{300} = 1.161 \rightarrow rev_{2001} = 348.1$$

$$\frac{rev_{2002}}{rev_{2001}} = 0.98 \rightarrow \frac{rev_{2000}}{348.3} = 0.98 \rightarrow rev_{2002} = 341.3$$

$$\frac{rev_{2003}}{rev_{1999}} = 1.309 \rightarrow \frac{rev_{2003}}{300} = 1.309 \rightarrow rev_{2003} = 392.7$$

$$\frac{rev_{2005}}{rev_{2004}} = 1.06 \rightarrow \frac{rev_{2003}}{450} = 1.06 \rightarrow rev_{2005} = 477$$

3. How many years are needed to decrease the debt of a company by 50% if the debt of the company is decreasing yearly by 3% on average.

$$\bar{l} = \sqrt[N-1]{\frac{y_N}{y_1}}$$

$$\bar{l} = 0.97$$

$$\frac{y_N}{y_1} = 0.5$$

$$0.97 = \sqrt[N-1]{0.5}$$

$$0.97 = \sqrt[N-1]{0.5}$$

$$0.97^{N-1} = 0.5$$

$$\ln 0.97^{N-1} = \ln 0.5$$

$$(N-1) \cdot \ln 97 = \ln 0.5$$

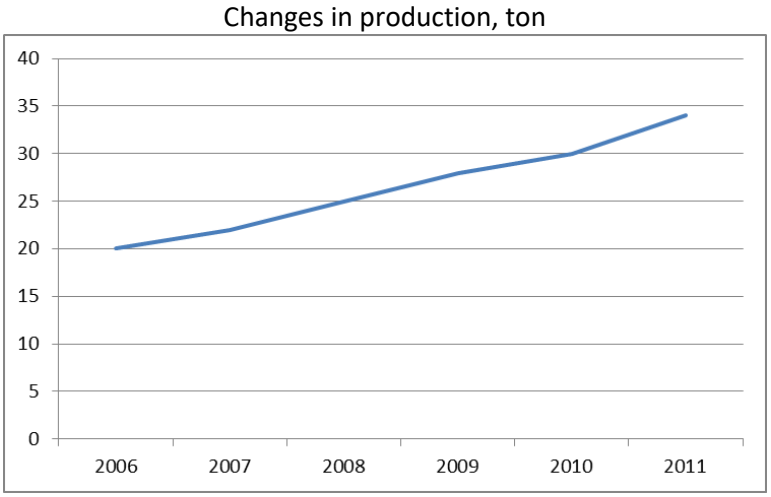
$$N - 1 = \frac{\ln 0.5}{\ln 0.97} = 22.8 \approx 23 years$$

It is possible to reach the target after 23 years or reach the target in the 24th year.

4. Examine the production of a company.

Year	Production, tons
2006	20
2007	22
2008	25
2009	28
2010	30
2011	34

A) Draw the time series. What kind of time series components can we see?



Source: task4

We can identify the trend, but there is no seasonal component. In the case of trend, a linear trend can be applied.

B) Set up a linear trend
   a. on paper

Year	Production, t	t	t*y	t^2
2006	20	1	20	1
2007	22	2	44	4
2008	25	3	75	9
2009	28	4	112	16
2010	30	5	150	25
2011	34	6	204	36
total	159	21	605	91

$$\bar{y} = \frac{20+22+25+28+30+34}{6} = \frac{159}{6} = 26.5$$

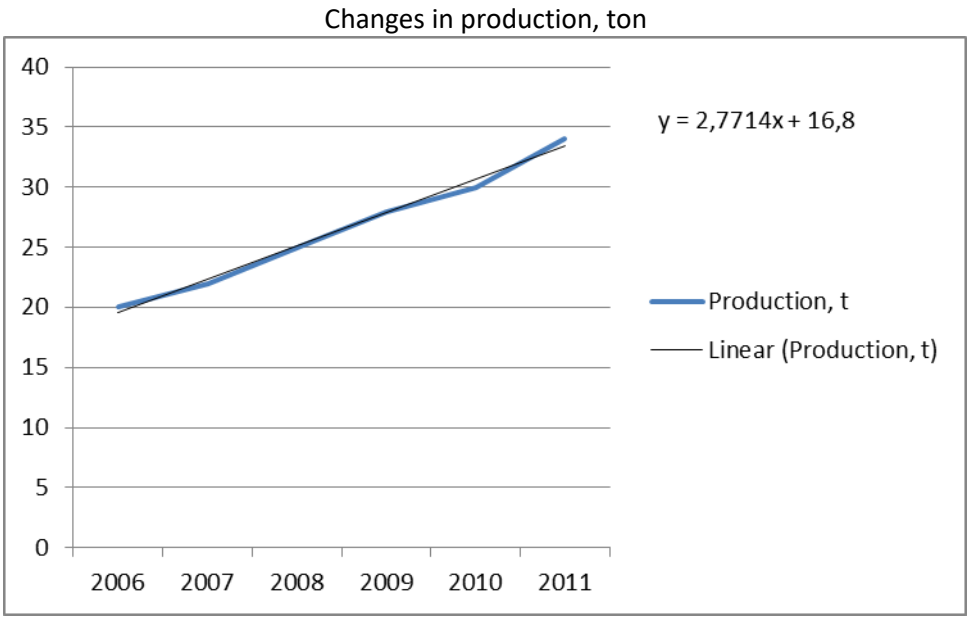
$$\bar{t} = \frac{1+2+3+4+5+6}{6} = \frac{21.5}{6} = 3.5$$

$$n = 6 \quad \sum ty = 605 \quad \sum t^2 = 91$$

$$b_1 = \frac{\frac{\sum ty}{n} - \bar{t} \cdot \bar{y}}{\frac{\sum t^2}{n} - \bar{t}^2} = \frac{\frac{605}{6} - 3.5 \cdot 26.5}{\frac{91}{6} - 3.5^2} = 2.77 \text{ ton}$$

$$b_0 = \bar{y} - b_1 \bar{t} = 26.5 - 2.77 \cdot 3.5 = 16.8 \text{ ton}$$

b. with Excel chart



Source: task 4

Excel solution:
- Create a line chart
- Right click on the line, use the 'Add trendline' option
- Select the 'Display Equation on chart' option

c. with Excel functions

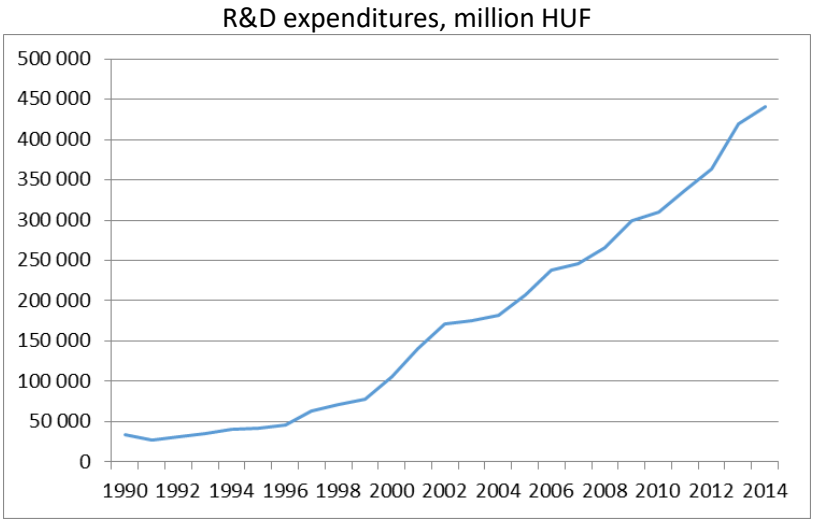
Excel solution:
- use INTERCEPT function for calculating b_0
- use SLOPE function for calculating b_1

C) Interpret the trend parameters.

The estimated production in 2005 was 16.8 tons.
During this period the estimated production increased on average by 2.77 tons yearly.

D) Estimate the value of 2015.
$TREND_{2015}=16.8+2.77*10=44.5$ tons

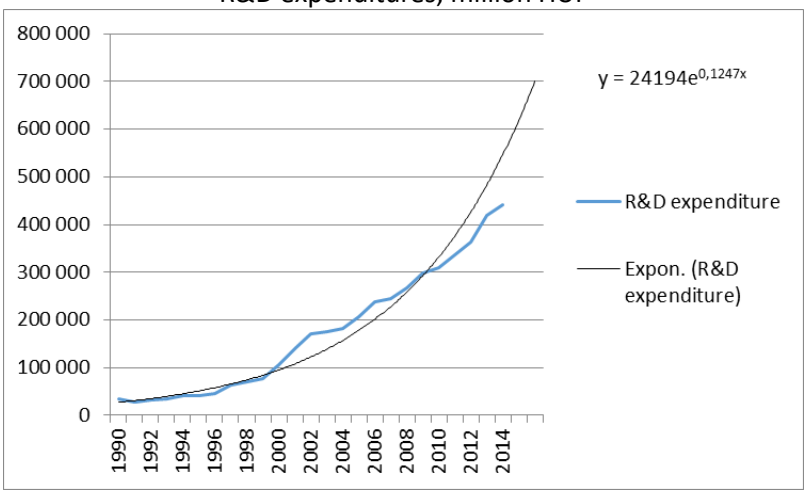
5. Download the date of the R&D expenditures between 1990 and 2014 from HCSO.
   a. Draw the time series. What kind of time series components can we see?



Source: task5

We can identify the trend, but there is no seasonal component. In the case of trend, an exponential trend can be applied.

b. Set up an exponential trend. Interpret the trend parameters. Estimate the R&D expenditure of 2016.



Source: task5

Excel solution:

- Create a line chart
- Right click on the line, use the 'Add trendline' option
  - Select the 'Exponential' trend
  - Write '2' in the Forecast Forward field (If we make a forecast until 2016, 2 more years are needed from the given period.)
  - Select the 'Display Equation on chart' option

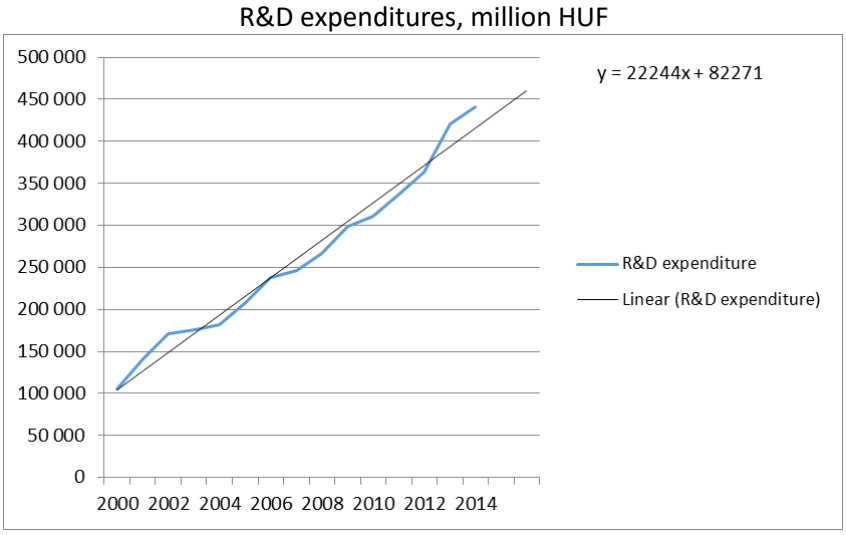
The Excel shows the exponential trend in the $\mathrm{TREND} = b_0 \cdot e^{\ln b_1 t}$ format. We cannot interpret this form, we have to calculate the value of b_1.

$b_1 = e^{\ln b_1} = e^{0,1247} = 1.133$ (It is possible to use the EXP function.)

The final form of the exponential trend: $TREND = 24194 \cdot 1.133^t$

The estimated R&D expenditure in 1989 was 24 194 million HUF.
During this period, the estimated R&D expenditure increased on average by 13.3 percent yearly.

c. Try to find an other estimation.



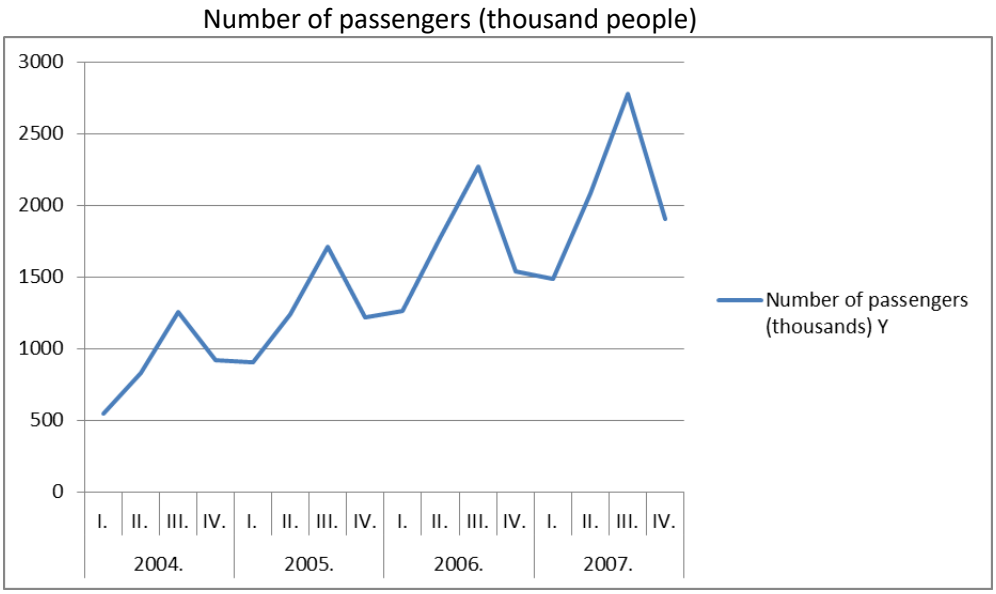
Source: task5

If we consider the period from 2000 to 2014, a linear trend seems to fit better than an exponential trend.

6. Examine the number of passengers at Budapest Airport.

Quarters		Number of passengers (thousands)
2004.	I.	547
2004.	II.	829
2004.	III.	1254
2004.	IV.	920
2005.	I.	907
2005	II.	1238
2005	III.	1712
2005	IV.	1217
2006.	I.	1266
2006.	II.	1774
2006.	III.	2267
2006.	IV.	1543
2007.	I.	1485
2007.	II.	2084
2007.	III.	2780
2007.	IV.	1902

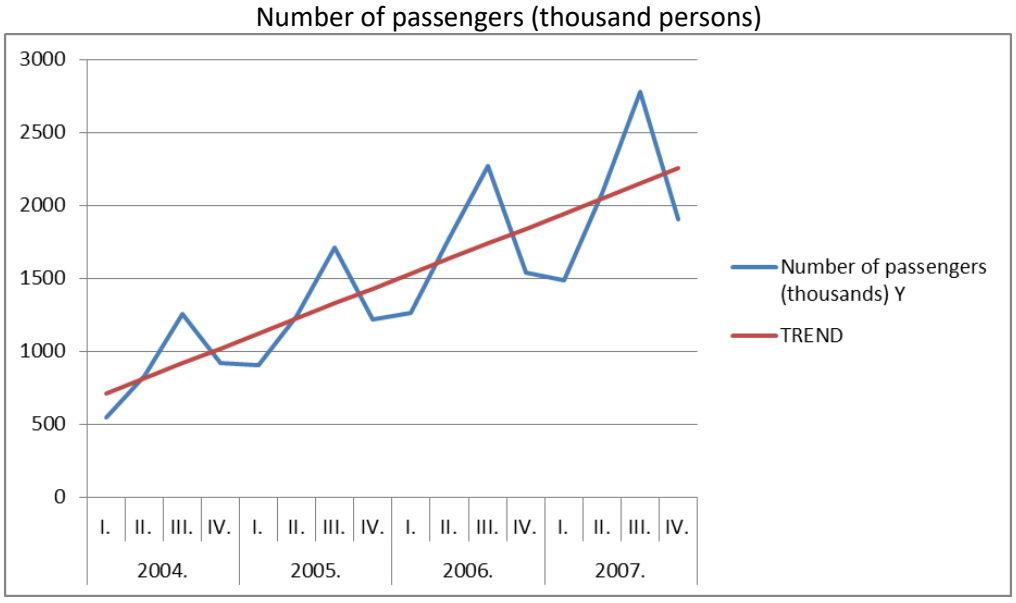
A) Draw the time series. Identify the components of the series



Source: task6

We can identify the trend and the seasonal component also. In the case of trend, a linear trend can be applied.

B) Fit a linear trend. According to the linear trend **estimate the number of the passengers during the period**. Interpret the trend parameters. Draw the time series and the trend on the same chart. Estimate the number of the passengers of 2015. IV. quarter.



Source: task6

- Excel solution:
  - Use functions (INTERCEPT, SLOPE) for calculating b_0 and b_1
  - Calculate the trend with mathematical operands (see numeric results at the end of this task)
- TREND=611.13+102.55*t

- Interpretation:
    - The estimated number of passengers was 611.13 thousand persons in the IV. quarter of 2003.
    - During the period the estimated number of passengers increased quarterly on average by 102.55 thousand persons.
- Forecast on paper: 2015/IV →t=48 → 611.13+102.55*48

C) **Compute and interpret the seasonal differences**.

Seasonal differences

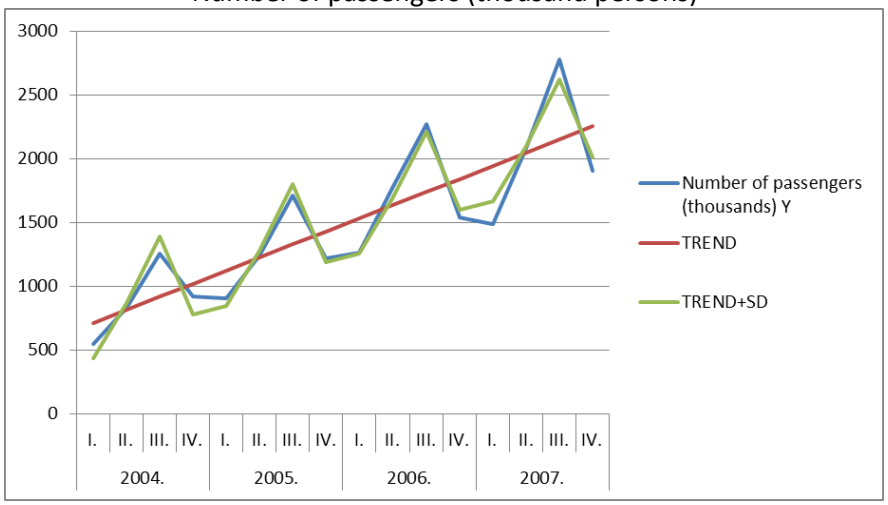
Quarters	**Seasonal differences (thousand persons)**
1	-277.74
2	49.71
3	469.16
4	-241.14
Total	0

Source: task6

- Excel solution:
    - Calculate the Y-TREND values
    - Calculate the quarterly means from the Y-TREND values → raw seasonal differences
    - Check whether the sum of raw seasonal differences = 0. If raw seasonal differences =0 → raw seasonal differences are corrected seasonal differences (If not, a correction is needed.)
- Interpretation:
    - s1: In the first quarters of the period, the observed number of passengers was lower on average by 277.74 thousand persons than the trend/estimation.
    - s2: In the second quarters of the period, the observed number of passengers was higher on average by 49.71 thousand persons than the trend/estimation.
    - s3: In the second quarters of the period, the observed number of passengers was higher on average by 49.71 thousand persons than the trend/estimation.
    - s4: : In the fourth quarters of the period, the observed number of passengers was lower on average by 241.14 thousand persons than the trend/estimation.

D) According to the linear trend and the seasonal variation **estimate the number of the passengers during the period**. Draw the time series, the trend and the estimation on the same chart. Estimate the number of the passangers of 2015. IV. quarter.

Number of passengers (thousand persons)

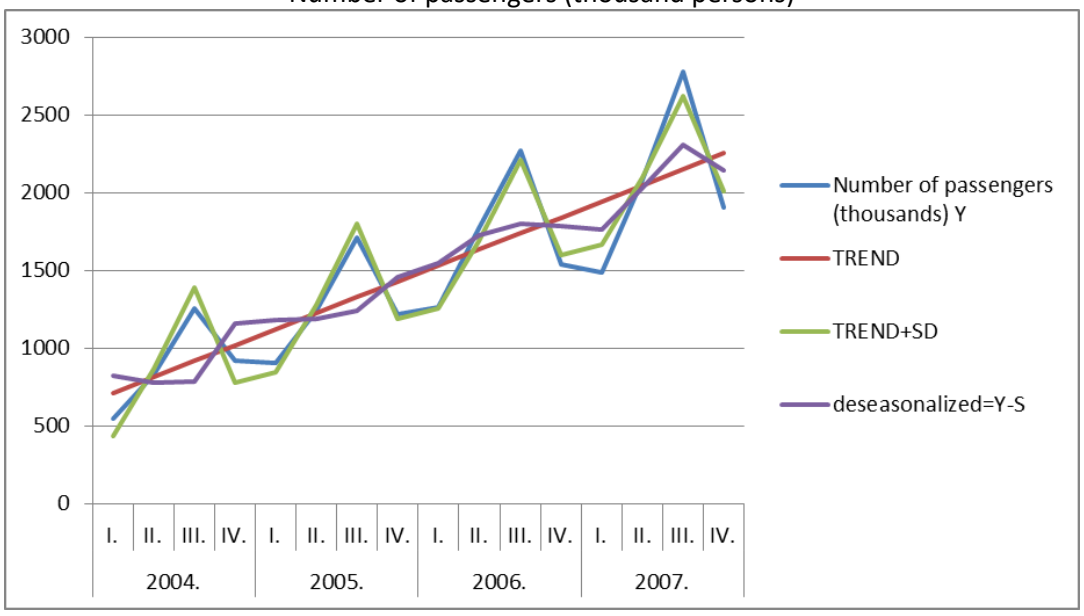


Source: task6

- Excel solution:
  - Calculate the estimation by mathematical operands (see numeric results at the end of this task)
- Forecast on paper: Y=TREND+SD=(611.13+102.55*48)+(-241.14)

E) **Estimate the deseasonalized number of the passengers during the period**.

Number of passengers (thousand persons)



Source: task6

- Excel solution:
  - Calculate the estimation by mathematical operands (see numeric results at the end of this task)

Numeric results of task 6

Quarters		Number of passengers (thousands) Y	t	TREND	Y-TREND	Seasonal differences	TREND+SD	Deseasonalized= Y-S
2004.	I.	547	1	713.68	-166.68	-277.74	435.94	824.74
	II.	829	2	816.23	12.77	49.71	865.94	779.29
	III.	1254	3	918.78	335.22	469.16	1387.94	784.84
	IV.	920	4	1021.33	-101.33	-241.14	780.19	1161.14
2005.	I.	907	5	1123.88	-216.88	-277.74	846.15	1184.74
	II.	1238	6	1226.43	11.57	49.71	1276.15	1188.29
	III.	1712	7	1328.99	383.01	469.16	1798.15	1242.84
	IV.	1217	8	1431.54	-214.54	-241.14	1190.40	1458.14
2006.	I.	1266	9	1534.09	-268.09	-277.74	1256.35	1543.74
	II.	1774	10	1636.64	137.36	49.71	1686.35	1724.29
	III.	2267	11	1739.19	527.81	469.16	2208.35	1797.84
	IV.	1543	12	1841.74	-298.74	-241.14	1600.60	1784.14
2007.	I.	1485	13	1944.29	-459.29	-277.74	1666.56	1762.74
	II.	2084	14	2046.85	37.15	49.71	2096.56	2034.29
	III.	2780	15	2149.40	630.60	469.16	2618.56	2310.84
	IV.	1902	16	2251.95	-349.95	-241.14	2010.81	2143.14

Source: task6

Review Section (Topic 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
The b1 parameter of the linear trend shows the slope of the linear function.		
In a case of a time series which shows quarterly data, 12 seasonal differences can be calculated.		
An aggregate price index shows the relative change in quantity of a given product group in the current period compared to the base period.		

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

i_q

i) is a simple price index
j) is a simple quantity index
k) is an aggregate quantity index
l) is an aggregate price index

In a case of a time series which shows yearly data, the growth rate

i) shows the yearly average relative change of the data in the given period
j) shows the yearly average absolute change of the data in the given period
k) shows the yearly relative change of the data in the given period
l) shows the yearly absolute change of the data in the given period

3. In a case of a football stadium, some data are known:

Ticket	Unit prices in 2013 (2012=100.0%)	Revenue in 2013, million EUR	Revenue in 2013 (2012=100.0%)
full price	105.1	200	102.4
student	102.5	120	104.1

a) Calculate the total relative change of revenues (value index).
b) Calculate the effects behind the total relative change of revenues (price index, quantity index).
c) Create a coherent interpretation about the calculated results.

4. In a case of a cinema, some data are known:

Ticket	Revenue in 2013 (2012=100.0%)	Revenue in 2012, million EUR	Number of sold quantities in 2013 (2012=100.0%)
premier	105.1	200	102.5
regular	102.5	300	98

a) Calculate the total relative change of revenues (value index).
b) Calculate the effects behind the total relative change of revenues (price index, quantity index).
c) Create a coherent interpretation about the calculated results.

5. The number of visitors is known in a case of a museum:

Year	Number of visitors (thousand persons)	Number of visitors (2007=100.0%)	Number of visitors (previous year=100.0%)
2007	303		
2008	310		
2009	315		
2010	322		
2011	330		
2012	335		
2013	340		

a) **Fill the empty cells** in the table. **Interpret** the calculated values in 2009.
b) **Calculate and interpret** the increment of growth.
c) **Calculate and interpret** the growth rate.

6. The foreign trade (export) of a company is known below:

Year	Export, thousand USD	t	t*y	t^2
2003	20			
2004	25			
2005	34			
2006	46			
2007	51			
2008	58			
Total				

a) **Fill the empty cells** in the table.
b) **Set up** a linear trend (t=1,2,…, n).
c) **Interpret** the parameters of the linear trend.
d) **Estimate** the export for 2015 based on the linear trend.

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 b1 parameter of the linear trend shows the slope of the linear function.	X	
In a case of a time series which shows quarterly data, 12 seasonal differences can be calculated.		X
An aggregate price index shows the relative change in quantity of a given product group in the current period compared to the base period.		X

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

i_q

m) is a simple price index
n) <u>is a simple quantity index</u>
o) is an aggregate quantity index
p) is an aggregate price index

In a case of a time series which shows yearly data, the growth rate

m) <u>shows the yearly average relative change of the data in the given period</u>
n) shows the yearly average absolute change of the data in the given period
o) shows the yearly relative change of the data in the given period
p) shows the yearly absolute change of the data in the given period

3. In a case of a football stadium, some data are known:

Ticket	**Unit prices in 2013 (2012=100.0%)** **ip**	**Revenue in 2013, million EUR** **v1**	**Revenue in 2013 (2012=100.0%)** **iv**	**ip coefficient form**	**iv coefficient form**
full price	105.1	200	102.4	1.051	1.024
student	102.5	120	104.1	1.025	1.041

d) Calculate the total relative change of revenues (value index).

$$I_v = \frac{\sum v_1}{\sum \frac{v_1}{i_v}} = \frac{200+120}{\frac{200}{1.024}+\frac{120}{1.041}} = 1.030 \rightarrow 103.0\% \rightarrow +3.0\%$$

e) Calculate the effects behind the total relative change of revenues (price index, quantity index).

$$I_p^1 = \frac{\sum v_1}{\sum \frac{v_1}{i_p}} = \frac{200+120}{\frac{200}{1.051}+\frac{120}{1.025}} = 1.041 \rightarrow 104.1\% \rightarrow +4.1\%$$

$$I_q^0 = \frac{I_v}{I_p^1} = \frac{1.030}{1.041} = 0.990 \rightarrow 99.0\% \rightarrow -1.0\%$$

f) Create a coherent interpretation about the calculated results.

The prices of the products increased on average by 4.1 percent but the quantities of the products decreased on average by 1.0 percent from 2012 to 2013. Consequently, the revenues of the products increased on average by 3.0 percent from 2012 to 2013.

4. In a case of a cinema, some data are known:

Ticket	Revenue in 2013 (2012=100.0%) iv	Revenue in 2012, million EUR v0	Number of sold quantities in 2013 (2012=100.0%) iq	iv coefficient form	iq coefficient form
premier	105.1	200	102.5	1.051	1.025
regular	102.5	300	98	1.025	0.98

d) Calculate the total relative change of revenues (value index).

$$I_v = \frac{\sum v_0 \cdot i_v}{\sum v_0} = \frac{200 \cdot 1.051 + 300 \cdot 1.025}{200 + 300} = 1.035 \rightarrow 103.5\% \rightarrow +3.5\%$$

e) Calculate the effects behind the total relative change of revenues (price index, quantity index).

$$I_q^0 = \frac{\sum v_0 \cdot i_q}{\sum v_0} = \frac{200 \cdot 1.025 + 300 \cdot 0.98}{200 + 300} = 0.998 \rightarrow 99.8\% \rightarrow -0.2\%$$

$$I_p^1 = \frac{I_v}{I_q^0} = \frac{1.035}{0.998} = 1.037 \rightarrow 103.7\% \rightarrow +3.7\%$$

f) Create a coherent interpretation about the calculated results.

The quantities of the products decreased on average by 0.2 percent but the prices of the products increased on average by 3.7 percent from 2012 to 2013. Consequently, the revenues of the products increased on average by 3.5 percent from 2012 to 2013.

5. The number of visitors is known in a case of a museum:

Year	Number of visitors (thousand persons)	Number of visitors (2007=100.0%)	Number of visitors (previous year=100.0%)
2007	303	100.0	-
2008	310	102.3	102.3
2009	315	104.0	101.6
2010	322	106.3	102.2
2011	330	108.9	102.5
2012	335	110.6	101.5
2013	340	112.2	101.5

d) **Fill the empty cells** in the table. **Interpret** the calculated values in 2009.

The number of visitors was higher by 4.0 percent in 2009 than in 2007.
The number of visitors was higher by 1.6 percent in 2009 than in 2008.

e) **Calculate and interpret** the increment of growth.

$$\bar{d} = \frac{340 - 303}{6} = 6.17 \text{ thousand persons}$$

The yearly average absolute change of the number of visitors was 6,17 thousand persons in the given period.

f) **Calculate and interpret** the growth rate.

$$\bar{l} = \sqrt[6]{\frac{340}{303}} = 1.019 \rightarrow 101.9\% \rightarrow +1.9\%$$

The yearly average relative change of the number of visitors was 1,9 percent in the given period.

6. The foreign trade (export) of a company is known below:

Year	Export, thousand USD	t	t*y	t^2
2003	20	1	20	1
2004	25	2	50	4
2005	34	3	102	9
2006	46	4	184	16
2007	51	5	255	25
2008	58	6	348	36
Total	234	21	959	91

e) Fill the empty cells in the table.

f) Set up a linear trend (t=1,2,…, n).

$$\bar{y} = \frac{234}{6} = 39 \quad \bar{t} = \frac{21}{6} = 3.5$$

$$n = 6 \quad \sum ty = 959 \quad \sum t^2 = 91$$

$$b_1 = \frac{\frac{\sum ty}{n} - \bar{t} \cdot \bar{y}}{\frac{\sum t^2}{n} - \bar{t}^2} = \frac{\frac{959}{6} - 3.5 \cdot 39}{\frac{91}{6} - 3.5^2} = 8 \text{ thousand USD}$$

$$b_0 = \bar{y} - b_1 \cdot \bar{t} = 39 - 8 \cdot 3.5 = 11 \text{ thousand USD}$$

TREND=11+8*t

g) Interpret the parameters of the linear trend.

b_0: The estimated value of export in 2002 was 11 thousand USD.
b_1: During the given period, the estimated value of export increased on average by 8 thousand USD yearly.

h) **Estimate** the export for 2015 based on the linear trend.

2015 → t=13
TREND $_{2015}$=11+8*13=115 thousand USD

Excel exercises – Seminar part 2

Open practice_seminar_part2.xls and solve the tasks.

Excel solutions

You can check your results if you open practice_seminar_part2_solution.xls and watch practice_seminar_part2_solution.wmv

4 Sample exam

Paper-based exercises

1. **Answer the following questions. (20 points)**

What can we examine with the help of a Lorenz-curve?	
What is the mode?	
What is the equation of a multiplicative time series model?	
Which index can measure the inflation?	

2. **Three types of balls (basketball, volleyball, football) are sold in a sport shop. The following table is known about the sold balls:**

Price of ball (EUR)	Number of sold balls (pieces)
10	50
17	30
23	20
Total	100

It is also known that the average price of balls is 14.7 EUR.
Calculate and interpret the coefficient of variation of the price of the sold balls. **(15 points)**

3. In a case of a winery, some data are known:

Year	Wine production, million litres
2012	120
2013	113
2014	108
2015	101
2016	95
2017	90

a) Fit a linear trend. Interpret the trend parameters. (15 points)
b) According to the trend give an estimation for 2022. (5 points)

4. In a case of a company, some data are known:

Product	Revenue in 2013, %	Prices in 2014 (2013=100.0%)	Revenue in 2014 (2013=100.0%)
A	60	95.1	105.4
B	40	112.5	106.1

a) Calculate the total relative change of revenues (value index)! (5 points)
b) Calculate the effects behind the total relative change of revenues (price index, quantity index)! (10 points)
c) Create a coherent interpretation about the calculated results! (5 points)

5. There was a survey about time spent weekly with transportation. It is known that the weekly average time spent with transportation by students is 5.5 hours/person, by adults is 7 hours/person on average and by seniors is 3 hours/person on average weekly. It is also known that the sum of the weekly time spent with transportation of students is 165 hours, of adults is 315 hours and of seniors is 105 hours. **Calculate** the weekly average time spent with transportation among all of the respondents! **Interpret** the result! **(15 points)**

6. There is a summary about the annual net revenues of 5 shoe manufacturer companies:

Shoe manufacturer company	Annual net revenue (million HUF)
A	100
B	200
C	250
D	450
E	600
Total	**1600**

Describe the concentration of the shoe manufacturer companies with the help of the **normalized Herfindahl-index**! **Interpret the result! (10 points)**

Solutions

1. **Answer the following questions. (20 points)**

What can we examine with the help of a Lorenz-curve?	*Concentration in the population, e.g. distribution of wealth*
What is the mode?	*The mode is the value of the observation that appears most frequently*
What is the equation of a multiplicative time series model?	*Y=T*S*C*E*
Which index can measure the inflation?	*price index*

2. **Three types of balls (basketball, volleyball, football) are sold in a sport shop. The following table is known about the sold balls:**

Price of ball (EUR)	Number of sold balls (pieces)
10	50
17	30
23	20
Total	100

It is also known that the average price of balls is 14.7 EUR.
Calculate and interpret the coefficient of variation of the price of the sold balls. **(15 points)**

$$\sigma = \sqrt{\frac{50 * (10 - 14.7)^2 + 30 * (17 - 14.7)^2 + 20 * (23 - 14.7)^2}{100}} = 5.139\ EUR$$

$$v = \frac{5.139}{14.7} = 0.349 \sim 34.9\%$$

The prices of the balls deviate on average by 34.9% from the average price.

3. **In a case of a winery, some data are known:**

Year	**Wine production, million litres**	**t**	**t^2**	**t*y**
2012	120	1	1	120
2013	113	2	4	226
2014	108	3	9	324
2015	101	4	16	404
2016	95	5	25	475
2017	90	6	36	540
Total	627=Σy	21=Σt	91=Σt^2	2089=Σt*y

n=6
$\bar{y} = 104.5$
$\bar{t} = 3.5$

c) Fit a linear trend. Interpret the trend parameters. (15 points)

$$b_1 = \frac{6 * 2089 - 21 * 627}{6 * 91 - 21^2} = -6.02857$$

$$b_0 = 104.5 - (-6.02857) * 3.5 = 125.6$$

TREND=125.6-6.03*t

b_0: The estimated value of wine production in 2011 was 125.6 liters.
b_1: In the examined period (*between 2012 and 2017*) the estimated value of wine production decreased by 6.03 liters on average annually.

d) According to the trend give an estimation for 2022. (5 points)

t_{2022}=11 TREND=125.6-6.03*11=59.27 liters

4. In a case of a company, some data are known:

Product	Revenue in 2013, % v_0	Prices in 2014 (2013=100.0%) i_p	Revenue in 2014 (2013=100.0%) i_v
A	60	95.1	105.4
B	40	112.5	106.1

d) Calculate the total relative change of revenues (value index)! (5 points)

$$\mathrm{Iv} = \frac{60 * 0.951 + 40 * 1.125}{100} = 1.0206 \sim +2.06\%$$

e) **Calculate the effects behind the total relative change of revenues (price index, quantity index)! (10 points)**

$$\mathrm{I}_\mathrm{p}^0 = \frac{60 * 1.054 + 40 * 1.061}{100} = 1.0568 \sim +5.68\%$$

$$I_q^1 = \frac{1.0206}{1.0568} = 0.9657 \sim -3.43\%$$

f) Create a coherent interpretation about the calculated results! (5 points)

The company's revenue increased by 2.06 percent from 2013 to 2014. This is caused by two factors: a change in the prices and the quantities sold. Due to the price changes, the company's revenue increased by 5.68 percent, and due to the quantity changes the ticket seller's revenue decreased by 3.43 percent from 2013 to 2014.

OR:

The prices of the products increased on average by 5.68 percent, and the quantities of the products decreased on average by 3,43 percent from 2013 to 2013. Consequently, the revenues of the products increased on average by 2.06 percent from 2013 to 2014.

5. There was a survey about time spent weekly with transportation. It is known that the weekly average time spent with transportation by students is 5,5 hours/person, by adults is 7 hours/person on average and by seniors is 3 hours/person on average weekly. It is also known that the sum of the weekly time spent with transportation of students is 165 hours, of adults is 315 hours and of seniors is 105 hours. **Calculate** the weekly average time spent with transportation among all of the respondents! **Interpret** the result! **(15 points)**

	Weekly average time spent with transportation, hours $\bar{x}_j$	Total time spent weekly with transportation, hours S_j

Students	5.5	165
Adults	7.0	315
Senior	3.0	105
Total		585

$$\bar{x} = \frac{585}{\frac{165}{5.5} + \frac{315}{7} + \frac{105}{3}} = 5.318\ hours$$

People surveyed spent on average 5.318 hours with transportation during the week.

6. **There is a summary about the annual net revenues of 5 shoe manufacturer companies:**

Shoe manufacturer company	**Annual net revenue (million HUF)**	**Z_i**
A	100	0.0625
B	200	0.125
C	250	0.15625
D	450	0.28125
E	600	0.375
Total	**1600**	**1**

n=5

Describe the concentration of the shoe manufacturer companies with the help of the **normalized Herfindahl-index**! **Interpret the result! (10 points)**

$$HI = 0.0625^2 + 0.125^2 + 0.15625^2 + 0.28125^2 + 0.375^2 = 0.263672$$

$$HI^* = \frac{0.263672 - \frac{1}{5}}{1 - \frac{1}{5}} = 0.07959$$

The concentration of the annual net revenues among the shoe manufacturer companies is low (because HI* < 0.1).

5 Excel functions used during seminars

This chapter introduces the built-in Excel functions used during seminars with a short explanation of the function and their usage.

Name	**Description**
=SUM(array)	Calculates the sum of an array of numbers.
=AVERAGE(array)	Calculates the mean of an array of numbers. ATTENTION: this function only works when using individual data, for frequency distribution tables and/or classes, see formulas used in Chapter 2.2 *Measures of central tendency*.
=MEDIAN(array)	Calculates the median of an array of numbers. ATTENTION: this function only works when using individual data, for frequency

	distribution tables and/or classes, see formulas used in Chapter 2.2 *Measures of central tendency*.
=MODE(array)	Calculates the mode of an array of numbers. ATTENTION: this function only works when using individual data, for frequency distribution tables and/or classes, see formulas used in Chapter 2.2 *Measures of central tendency*.
=MIN(array)	Shows the minimum of an array of numbers.
=MAX(array)	Shows the maximum of an array of numbers.
=SKEW.P(array)	Calculates the asymmetry measure of an array of numbers.
=VARP(array)	Calculates the variance of an array of numbers. ATTENTION: this function only works when using individual data, for frequency distribution tables and/or classes, see formulas used in Chapter 2.3 *Dispersion*.
=STDEVP(array)	Calculates the standard deviation of an array of numbers. ATTENTION: this function only works when using individual data, for frequency distribution tables and/or classes, see formulas used in Chapter 2.3 *Dispersion.*
=SUMPRODUCT(array1;array2)	Calculates the sum of the products of two arrays of numbers ($\Sigma a_i * b_i$ for any a and b for all i=1, 2 … n)
=SUMQS(array)	Calculates the sum of the squares of an array of numbers (Σx_i^2 for all i=1, 2 … n)
=LINEST(y array; t array)	This function is used to calculate b_0 and b_1 parameters of a linear trend equation. ATTENTION: to use this function, select two cells next to each other and after selecting the necessary arrays, press Ctrl+Shift+Enter simultaneously to run the function.
=SLOPE(y array; t array)	This function is used to calculate the b_1 parameter of a linear trend equation.
=INTERCEPT(y array, t array)	This function is used to calculate the b0 parameter of a linear trend equation.
=TREND(y array, t array for available time periods; t array for forecast)	This function is used to fit a linear trend at a given examined period of data and to create a forecast based on the linear trend. ATTENTION: to use this function, select the whole column in which you want to fit the trend and create forecast, and after selecting the necessary arrays, press Ctrl+Shift+Enter simultaneously to run the function.
=LOGEST(y array; t array)	This function is used to calculate b_0 and b_1 parameters of an exponential trend equation. ATTENTION: to use this function, select two cells next to each other and after selecting the necessary arrays, press Ctrl+Shift+Enter simultaneously to run the function.