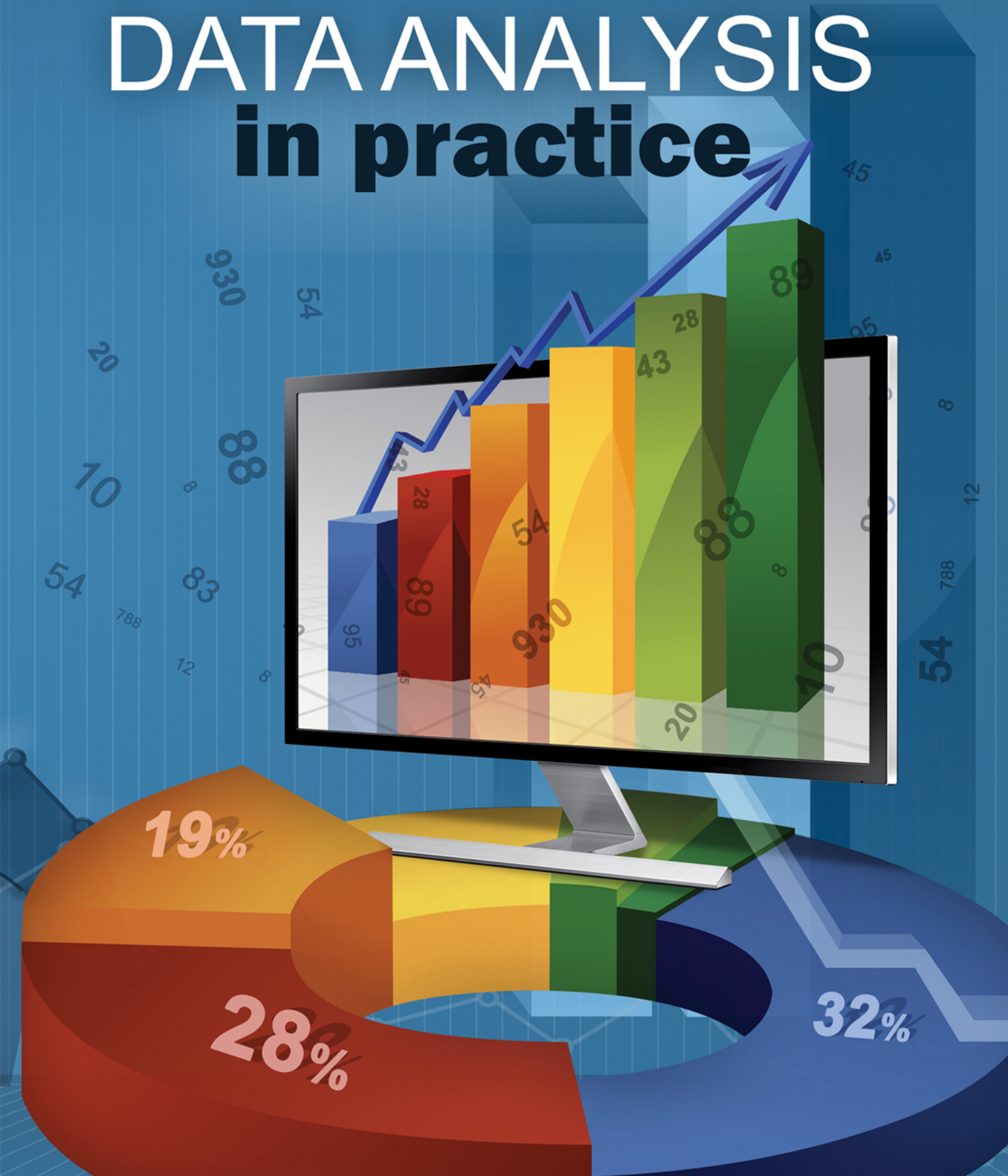




UNIVERSITY OF PÉCS  
FACULTY OF HEALTH SCIENCES

Pongrác Ács

# DATA ANALYSIS in practice



# **DATA ANALYSIS IN PRACTICE**

**Pécs, 2015.**



PÉCSI TUDOMÁNYEGYETEM  
UNIVERSITY OF PÉCS



# DATA ANALYSIS IN PRACTICE

Author & editor: Dr. Pongrác Ács

Authors:

Dr. Pongrác Ács

Dr. András Oláh

Dr. Annamária Karamánné Pakai

László Bence Rapos

Reviewed by:

Dr. Csilla Szabó

Dr. László Balogh

Translated by:

Kata Füge

Technical editor: Gábor Varga

Publisher: University of Pécs, Faculty of Health Sciences

This publication was prepared with the funding of the Social Renewal Operational Programme-4.1.2.

E-13/1/KONV-2013-0012.

*Pécs, 2015.*

**ISBN 978-963-642-371-1**

The book was prepared within the Social Renewal Operational Programme-4.1.2. E-13/1/KONV-2013-0012 tender.



HUNGARIAN  
GOVERNMENT

European Union  
European Social  
Fund



INVESTING IN YOUR FUTURE

# CONTENT

---

FOREWORD .....	6
<b>1. A THEORETICAL OVERVIEW OF SCIENTIFIC RESEARCH (ANNAMÁRIA KARAMÁNNÉ PAKAI, ANDRÁS OLÁH).....</b>	<b>9</b>
1.1. INTRODUCTION.....	9
1.2. SOURCES OF KNOWLEDGE IN EVERYDAY UNDERSTANDING AND SCIENTIFIC RESEARCH.....	9
1.3. PROCESS OF SCIENTIFIC RESEARCH .....	11
1.4. CHOOSING THE RESEARCH TOPIC, DEFINING THE SCIENTIFIC PROBLEM .....	11
1.5. THE MAIN OBJECTIVES OF SCIENTIFIC RESEARCH .....	14
1.6. BASIC REQUIREMENTS OF FORMULATING A HYPOTHESIS.....	15
1.7. PLANNING A SCIENTIFIC RESEARCH AND THE STEPS OF PREPARATION.....	16
1.8. CONCEPTUALIZATION AND OPERATIONALIZATION.....	23
1.9. TYPES OF RESEARCH .....	26
1.10. ETHICAL QUESTIONS OF SCIENTIFIC RESEARCH .....	29
1.10.1. <i>Research including human participants</i> .....	29
1.10.2. <i>Regulations in Hungary</i> .....	31
<b>2. LITERATURE REVIEW IN PRACTICE. USING THE MOST POPULAR DATABASES FOR LITERATURE REVIEW (ANNAMÁRIA KARAMÁNNÉ PAKAI, ANDRÁS OLÁH) .....</b>	<b>33</b>
2.1. EXPLORING AND GATHERING RELEVANT SCIENTIFIC LITERATURE .....	33
2.2. SOURCES OF A LITERATURE REVIEW.....	34
2.3. USE OF INTERNET DURING LITERATURE REVIEW .....	35
2.4. ELECTRONIC LIBRARIES .....	36
2.5. ELECTRONIC BIBLIOGRAPHIES.....	41
2.6. SPECIALIZED ONLINE DATABASES .....	45
2.6.1. <i>EBSCOhost</i> .....	46
2.6.2. <i>MATARKA (Hungarian Periodicals Table of Contents Database)</i> .....	52
2.6.3. <i>MOKKA (Hungarian National Common Catalogue)</i> .....	55
2.6.4. <i>OVID</i> .....	57
2.6.5. <i>ScienceDirect</i> .....	62
2.6.6. <i>Scopus</i> .....	64
2.6.7. <i>SpringerLink</i> .....	68
2.7. SPECIALIZED DATABASES.....	70
2.7.1. <i>MEDLINE</i> .....	70
2.7.2. <i>PubMed</i> .....	71
2.7.3. <i>SPORTDiscus</i> .....	73
2.7.4. <i>Digital Library of the University of Physical Education</i> .....	73
2.8. ELECTRONIC BOOKS AND JOURNALS .....	75
2.8.1. <i>Open Access journals</i> .....	75
2.8.2. <i>Directory of Open Access Journals</i> .....	76
2.9. ONLINE SEARCH ENGINES.....	77

2.10. QUOTATIONS AND INTERTEXTUAL REFERENCES.....	77
<b>3. BASIC STATISTICAL CONCEPTS, TYPES OF VARIABLES AND CRITERION VARIABLES (PONGRÁC ÁCS).....</b>	<b>85</b>
3.1. DEFINITION OF STATISTICS .....	85
3.2. STATISTICAL DATA.....	86
3.3. TYPES OF VARIABLES AND SCALES .....	87
<b>4. EDITING ONLINE QUESTIONNAIRES IN PRACTICE (LÁSZLÓ BENEC RAPOSA).....</b>	<b>90</b>
4.1. INTRODUCTION.....	90
4.2. PREREQUISITES OF THE QUESTIONNAIRE, BASIC PRINCIPLES OF PROVIDING GENERAL INFORMATION AND PREPARATION.....	91
4.3. QUESTIONNAIRE-EDITING IN PRACTICE USING GOOGLE DOCS .....	100
4.4. SUMMARY OF RESPONSES.....	108
4.5. QUESTIONNAIRE EDITING – A PRACTICAL EXAMPLE .....	111
<b>5. THE SPSS USER INTERFACE, IMPORTING AND EXPORTING DATA, BECOMING FAMILIAR WITH MENUS (PONGRÁC ÁCS) .....</b>	<b>122</b>
5.1. THE SPSS USER INTERFACE .....	122
5.2. IMPORTING DATA .....	127
5.3. BECOMING FAMILIAR WITH MENUS .....	132
<b>6. DATA CLEANING, SIMPLE ANALYSES WITH PRIMARY DATA, DATA MANIPULATION (PONGRÁC ÁCS)141</b>	
<b>7. DESCRIPTIVE STATISTICS, TABLES AND GRAPHS (PONGRÁC ÁCS, LÁSZLÓ BENEC RAPOSA) .....</b>	<b>162</b>
7.1. THEORETICAL BACKGROUND, DESCRIPTIVE STATISTICAL INDICATORS .....	162
7.2. STATISTICAL TABLES .....	175
7.3. STATISTICAL GRAPHS, DIAGRAMS .....	183
<b>8. ASSOCIATION AND CORRELATION ANALYSIS (PONGRÁC ÁCS) .....</b>	<b>203</b>
8.1. A THEORETICAL BACKGROUND FOR STATISTICAL RELATIONSHIPS .....	203
8.2. ASSOCIATION AND CROSSTAB ANALYSIS.....	204
8.3. CORRESPONDENCE ANALYSIS .....	221
8.4. MIXED ASSOCIATION.....	224
8.5. CORRELATION ANALYSIS .....	227
<b>9. INFERENCE STATISTICS (PONGRÁC ÁCS).....</b>	<b>233</b>
9.1. INTRODUCTION, THEORETICAL BACKGROUND.....	233
9.2. STATISTICAL ESTIMATION .....	235
9.3. DIFFERENCE TESTS.....	241
9.4. HYPOTHESIS TESTING (PARAMETRIC AND NON-PARAMETRIC TEST IN PRACTICE).....	242
<b>10. AN INTRODUCTION TO REGRESSION ANALYSIS (PONGRÁC ÁCS ) .....</b>	<b>265</b>

10.1. TWO-VARIABLE LINEAR REGRESSION .....	265
10.2. MULTIPLE LINEAR REGRESSION.....	269
<b>11. LITERATURE .....</b>	<b>274</b>
<b>12. APPENDIX.....</b>	<b>280</b>
<b>13. SUPPLEMENT (TABLES).....</b>	<b>282</b>

## Foreword

Our main objective writing this electronic coursebook was to provide a solid basis for data analysis methods applied when writing a thesis. The authors are aware of the fact that the knowledge on data analysis methodology have been changing continually and it has become so vast that it is almost impossible to summarize the subject in detail and to incorporate its elements as a user. Our goal was not this – we did not strive to present all of this knowledge for the reader. Instead, we attempted to provide a practical and useful methodology guide for thesis-writing, an obligatory task at all levels of Hungarian higher education. Teachers involved in tertiary education are responsible for the supervision of research work of bachelor-, masters- and PhD students, they teach research methodology courses, and thus the experience they have gained clearly reflects which methodologies are the most preferred and useful ones for testing the most frequently proposed hypotheses. Students are required to demonstrate a basic knowledge on scientific research issues while doing literature reviews, preparing for tests and presentations, writing their theses or participating in university research activities.

It is important to point out that this electronic coursebook exploring data analysis shall neither be considered as an IT coursebook, nor as a statistics coursebook. Although most of the content focuses on analyses using statistical methodology, our main goal was to emphasize practical approaches as well as some of the fundamental theoretical explanations. For the statistical topics presented here our intention was to be easily understandable for the reader while we hoped to keep professional integrity as well. We believe that this book in itself will be enough to show the highlighted methods, and thus there is no need to use another statistical book to understand these concepts. However, we do suggest further reading material for those who would like to study these topics more extensively.

This volume has been compiled with a specific concept in mind. The first chapter (*Theoretical overview of scientific research*) provides basic information on various types of scientific research, their steps and specific research objectives. We believe it is important to keep the practical approach in focus here as well, therefore we present a research proposal for a thesis in detail, hopefully proving useful for those who will need to prepare one in the future.

The second chapter describes in detail the literature that forms the basis of scientific research with specific reference to principles applied for theses (*Literature review in practice. The most popular databases in literature review*). Chapter 2 therefore attempts to show the most often used electronic databases through which the readers themselves will be able to find important literature in their own field of research.

The third chapter (*Basic statistical concepts, types of variables and criterion variables*) briefly describes the most essential concepts and notions used in data analysis.

In the fourth chapter (*Editing online questionnaires in practice*) we will try to describe online surveys, a method used widely during the compilation of a primary database. After introducing some basic information and discussing the major issues with a specific application chosen for this purpose, an example will be offered as to how to design such questionnaires and analyse the collected data. The material of this example will be used in the databases of further chapters.

From the fifth chapter on – similarly to several other statistical coursebooks – we will present the SPSS software, by detailing the methods most generally used for data analysis. All databases used in this coursebook can be accessed from the website [www.etk.pte.hu](http://www.etk.pte.hu). We aimed to keep the concept of the book throughout the chapters describing the SPSS environment as well. Accordingly, in chapter five (*The SPSS user interface, importing and exporting data, becoming familiar with menus*) we will describe the software and its basic features.

Chapter 6 (*Data cleaning, simple analyses with primary data, data manipulation*) focuses on describing the basic methods for “data manipulation” using specific examples.

The reader may find a theoretical and practical guide for descriptive statistics in Chapter 7 (*Descriptive statistics, tables and graphs*), where we present the statistical tables and figures that are the most often used when presenting data.

Chapter 8 (*Association and correlation analysis*) tries to explain the basic connection and correlation between statistical variables.

Chapter 9 (*Inferential statistics*) describes methods that are the most often used during research data analysis (statistical estimation and testing hypotheses). Obviously, a certain amount of theoretical knowledge is required for this topic discussed at the beginning of the chapter.

The last, tenth chapter (*Introduction to regression analysis*) offers a short introduction into the issue of regression analysis, using short and practical examples.

As an advisory comment I would like to draw the reader’s attention to the fact that even after reading this coursebook several times he may not be able to solve a research problem for the first time; as prior practice on the computer will be inevitable. Our experience as researchers show that the number of hours spent with practice on the computer strongly correlates with the success of problem-solving.



It is also a fact that the research findings published in the field of health and medicine– and also sport sciences – are an infinite source of information for researchers and anyone else interested. This huge storehouse of information lends itself to critical analysis and rigorous scrutiny and provides us with the opportunity to publish new results.

I would like to express my gratitude to my co-authors who, focusing on my concept, have contributed to this volume with precise but also hands-on chapters, maintaining the high quality of this electronic material. I would like to thank our colleagues who encouraged and inspired us to write this book. Also, I am grateful to all those teachers and mentors who helped us during our first steps as researchers and have supported us ever since. I would like to say special thanks to Alexandra Makai, who – as a PhD-student – completed the tiring job of proofreading and corrections.

I am also grateful for our proof-reader to correct the material with the required criticism and professionalism, and thus providing invaluable assistance in finalising the material.

I offer this work for the memory of Professor József Pintér, who had drawn my attention to the challenging situation of scientific research methodology and, more specifically, research conducted in sport sciences.

I would like to invite my colleagues and students alike to help upgrading and actualising this work with their comments and suggestions, providing me the opportunity to correct any mistakes that might have been left in the material (for which only the author is responsible).

Pécs, 22 September 2014

Pongrác Ács  
author

# **1. A THEORETICAL OVERVIEW OF SCIENTIFIC RESEARCH (Annamária Karamánné Pakai, András Oláh)**

## **1.1. Introduction**

Nowadays a basic expectation towards professionals working in the scientific field is to conduct evidence-based research that has its grounds in research methodology and biostatistics and use the findings of national and international research activities. This knowledge does not only contribute to the successful completion of the researchers' own work but also accommodate the understanding and critical assessment of scientific publications and lectures given at scientific conferences and professional meetings (Lampek – Kívés 2012, Pakai – Kívés 2013).

## **1.2. Sources of knowledge in everyday understanding and scientific research**

*Scientific research* is a planned human work in a set time and place based on pre-set hypotheses in pursuit of solving a problem. The notion neatly illustrates that science is a systematic effort in which rational thinking, evidence-based statements and logical reasoning are required.

Human beings fundamentally strive to get to know and understand the world around them. We distinguish between two ways of understanding reality. *Empirical reality* is everything that a person observes for himself or herself during his or her whole life. An example for this could be a healthcare professional, who works with patients and during his or her daily work understands their personality, behaviour, and reaction to a particular disease and who can recognize how to make quick decisions in life-threatening situations. Personal experiences and observations can be used as reference and a way to justify the validity, reliability and objectivity of certain information. The role of coaches in sports is quite similar, as they use their past experience of trainings and tournaments in their decision-making processes. Consensual reality refers to the process where a person accepts and agrees to the knowledge that is transferred by others through his own system of criteria. An example for this is a patient who wants to recover will follow the advice of the physiotherapist during therapy .

To satisfy one's curiosity, it is possible to complete various observations or smaller experiments in everyday life, but mistakes will naturally occur along the way. . Generally it can be stated that everyday observations are usually random, cursory and only partially conscious. Most students are unable to recall what kind of skirt the teacher was wearing, or

whether she was wearing a skirt at all. The fundamental reason for this is what we call *loose observation*, which might lead to the problem of *over-generalisation*, when conclusions drawn from a single incident will be generalised and is regarded as evidence. . Such situations may occur most often due to lack of time, for example when general conclusions are drawn, for instance, when asking only a handful of students about the conditions of the dormitory. Over-generalizing may lead to the next type of mistake referred to as *selective perception*. If a certain correspondence has been established between certain things in everyday life, which is also supported by a theory, then people may tend to focus only on specific events or situations that are proof of the given correspondence, not recognizing several other considerations. This factor has a great role in the development of racial or ethnic stereotypes. *Imaginative expansions and illogical thinking* may also occur during everyday observations, especially in cases when a person finds no relevant explanation for certain phenomena incomprehensible to him or her. For example, this is the case when someone is preparing for a bad mark after several good ones.

Another typical case is when previous information affects the perception of the current situation, thus *understanding is biased*. For example, if a student fails an exam, he might suppose that the teacher asked difficult questions deliberately. There are several situations in a person's life that are hard to understand or explain, so people tend to believe that certain things or phenomena are impossible to understand or mystical. Certainly, *mystification* is a typical phenomenon when students bring their mascots and lucky charms to exams. Apart from mascots, sportsmen often like to wear a certain piece of clothing for enhancing the chances of luck (e.g. Hungarian goalkeeper Gábor Király's grey trousers).

Considering the above-mentioned flaws, all scientific professionals are required to have the ability to select; we must use targeted, planned and adequate methods in getting to know reality; and make statements that are logically and empirically established. As a result, scientific examinations will display fewer perception deficiencies. After all when conducting scientific research, we consciously observe things, which in turn will reduce the number of mistakes as well. Objectification of conscious examination can be completed by using tools helping measurement. To avoid over-generalization, scientists have to make sure they use a larger number of samples for their examinations, and also they should attempt to repeat their examinations. Selectivity can be avoided during scientific work if the scientist prepares and follows a conscious, general plan that has a specific objective of the research and draws conclusions based on valid results. If the scientist finds contradicting results, he should continue the examination. Scientists should not only be conscious and careful during data

collection but also in their reasoning. Publicity in science also provides a good opportunity for supervision of the same work by a different group of researchers. It is a scientific principle that everything can be understood, and thus there is no place for mystification. Furthermore, no research project shall be closed early, as all questions should be considered open at the end of the day. It is a common phenomenon nowadays, that older results are being proven wrong by new findings.

Summarizing the above-mentioned ideas we may state that cognition is realised through everyday situations and life actions, and data derived from them are random. In contrast, science is a conscious, planned and systematic activity, during which the researcher is aware of the risks of mistakes, but he does his utmost to prevent them (Babbie 2004, Bíróné et al 2011, Gőcze 2011).

### **1.3. Process of scientific research**

First of all, the researcher defines the problem, sets an objective and draws up the hypothesis, after which he collects data in a planned manner, following the rules of conducting research. Afterwards, using the chosen statistical software, data is being processed, analysed, and finally results are published to inform others, keeping the rules of scientific research referring to content and frame (Lampek – Kívés 2012). The steps of scientific research are depicted in Figure 1/1.

### **1.4. Choosing the research topic, defining the scientific problem**

The first step of a research process is to choose a topic to be examined and define the problem. This is influenced by the researcher's interest, motivation, his or her understanding of the relevant literature, and also the specific aim of the task (for example the preparation of a thesis). The source of the specific problem may be rooted in everyday practice (for instance when a yet unsolved problem is in the focus of interest); it may derive from a theory of the chosen discipline or from the relevant literature. The choice of topic for students in Hungarian higher education may also be influenced by questions that arose at lectures or practice work, in previous theses of older students, but the topic lists provided by the institution for the Scientific Student Group (TDK) is also typically relied on..

However, there are a few questions to be clarified prior to the research work. The problem to be examined should be considered from several aspects to confirm whether there is indeed a need for the completion of the given examination. The answer to this question may be

provided by consulting both national and international scientific publications. It is very important that the literature explored with the appropriate search techniques should be reviewed appropriately.

Further information on the topic will be provided in the chapter entitled "*Literature review in practice. The most popular databases for literature review*".

Reviewing the relevant literature is useful for researchers for several reasons: it does not only help precisely define the problem but understanding new result may also inspire professionals for further research.

It is also important to make sure that the chosen topic really needs to be observed. Reviewing the literature may also help the researcher see how much scientific interest there is for the given field. If it is a field with a large number of investigations conducted, it is advised to find a less discovered subfield, or follow the directions for further research mentioned in the literature.

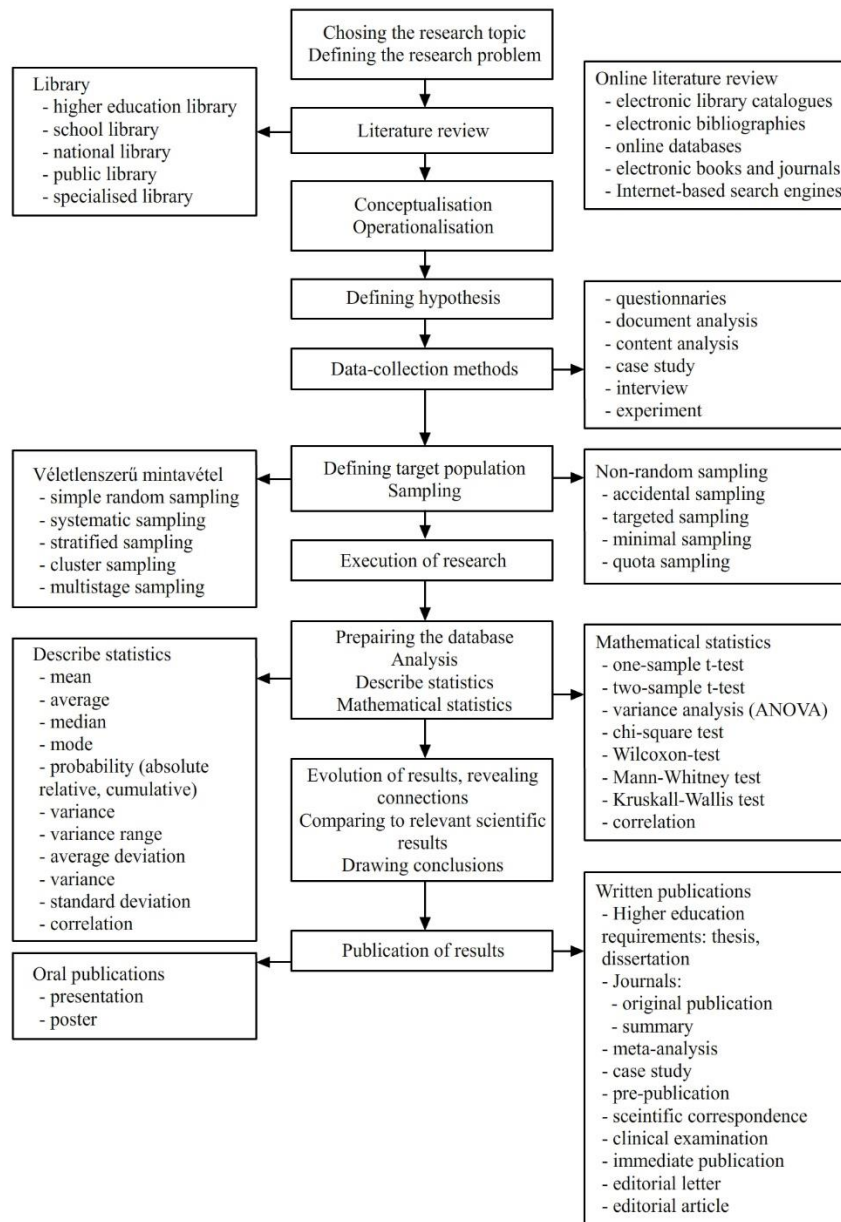
It should also be considered whether the researcher possesses the appropriate professional knowledge to cover the topic, and whether he has the abilities required for planning the research project, reviewing the literature, doing the statistical analysis of the data and drawing conclusions. Nowadays a good command of English is also a basic requirement for researchers as most of the literature is written in this language.

Senior researchers should also consider how practical the results are likely to be, and whether these result will contribute to the efficiency of relevant practice.

Deadlines should be clarified prior to the final choice and decision. It should be thought about whether there is enough time available for completion of the research and the deadline should be precisely given.

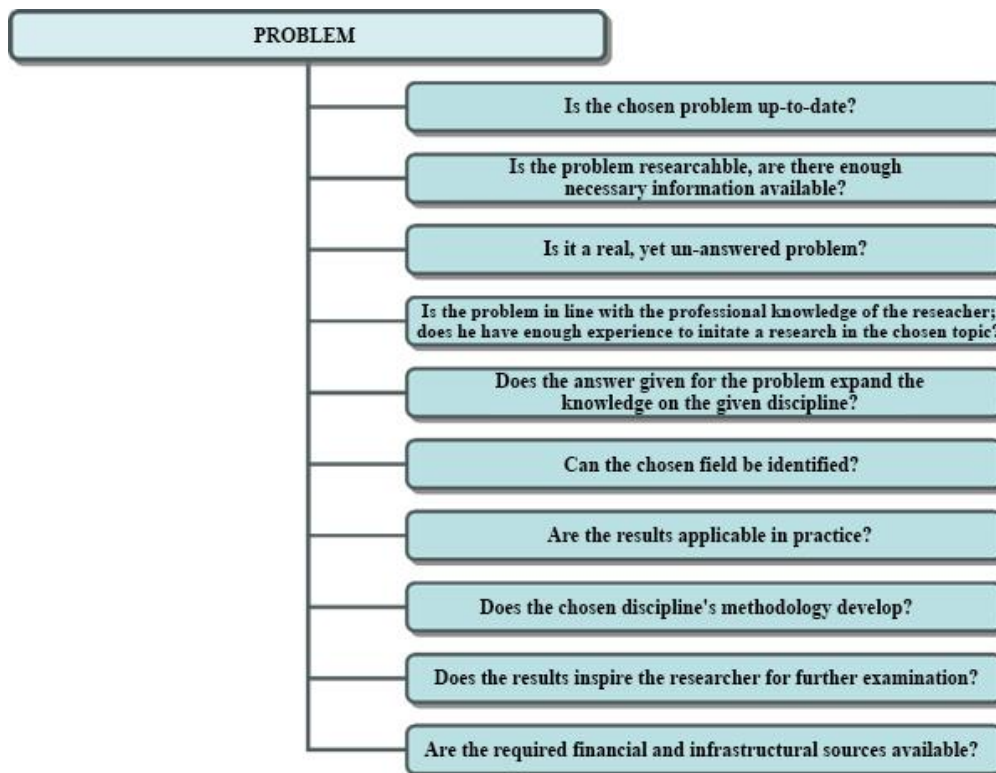
If the questions indicated on the figure1/2 have been attempted to be answered by the researcher but the relevant literature produced no evidence and left national issues open as well, then the research questions must be rephrased and reconducted.

An experienced researcher may derive information from literature research to identify what the fields are in focus of the profession and which topics are given priority – all of which may influence future publications.



**Figure 1/1. Steps of the research process**  
 Source: Lampek – Kivés 2012, Betlehem 2010,  
 edited by the authors

However, despite the above advice, difficulties may occur when defining a given problem. In such cases it is a wise idea to examine the chosen topic, continue reviewing the literature or consult an expert of the field to define the problem more precisely (Papp 2013, Cserné Adermann 1999, Falus 2004, Ács 2009).



**Figure 1/2. Examining how topical the research problem is**  
 Source: Edited by the authors based on Falus 2004

### 1.5. The main objectives of scientific research

When the researcher aims to describe a phenomenon or examine certain principles, he or she may specify several objectives that can be divided into three groups. *Uncovering research* may be carried out when only the key objective is to provide orientation in the given field, and the researcher does not know the topic very well. Here, the main task of the examination may be to decide whether it is worth doing a more in-depth research in the field. Uncovering research may also be used when the effectiveness of a method – to be used in practice later – needs to be tested. Such examinations are not very large scale and, accordingly, their costs are low, too. Their result should not be used to jump at hasty conclusions, although they may prove useful in different way: in inductive research they help deciding whether it is worth continuing research in the field, they may help paving the way to finding the exact definition of the research problem, and they may be applied to define hypotheses and identify the most appropriate data-collection method(s). Uncovering research most often uses observations, interviews and case studies as a tool. A good example for this is the examination carried out in the Kútvölgyi Clinical Block of Semmelweis University in Budapest, Hungary focusing on the medication distribution process that nurses follow. The authors used the method of direct observation to recognize the incidence of mistakes in medication connected to the process of

the medication distribution in the hospital and to identify types of mistakes in practice (Lám et al 2011).

The aim of *descriptive research* is to provide a precise description of a phenomenon, event, situation or a selected group. Such works may serve as a good basis for further research conducted with an explanatory objective. An example of descriptive research can be found in the publication of Imre Boncz et al (2013), in which they analyse the participation data of the 2008-2009 national breast screening programme organized by the National Health Insurance Fund of Hungary (OEP). The research defined the proportion of 45-65 year-old women who participated in the 4<sup>th</sup> phase of the programme (2008-2009), either for screening or for diagnostic imaging of the breast.

*Explanatory research activities* are carried out with the aim of revealing cause and effect relationships. This type of research does not only use descriptive statistics to analyse the examined event, phenomenon or group, but also aims to explain certain features between which a certain relationship is assumed. For example, the Hungarostudy 2002 programme lead by Mária Kopp was a health examination representative for the Hungarian adult population based on age, gender and size of residence. Each chapter of the monograph based on this work reviews the most important data explaining cause and effect relationships in the given field (Kopp – Kovács 2006).

The above-mentioned three types of research do not differ sharply from one another, but the emphasis also varies in different phases of the research. In practice, a more extensive research programme may include all three types.

## **1.6. Basic requirements of formulating a hypothesis**

A hypothesis is the preliminary assumption of the expected results of the research, which should be either accepted or rejected during the examination using appropriate data-collecting and statistical methods. The formulation of hypotheses may be inductive or deductive.

*Inductive hypothesis formulation* assumes the existence of principles examined in practice, while *deductive methods* start with a theoretical statement as the basis of the hypothesis, thus these methods start from a general theory and move on to specific cases.

A Basic set of requirements regarding the formulation of a hypothesis:

- It should be set prior to the research, to serve as a guideline.
- It should be framed as a simple, compact statement (declarative sentence) based on the information already possessed by the researcher.



- It should indicate the relationship between dependent and independent variables.
- It should use unequivocal, clearly formulated notions.
- The appropriate data-collection methods should be available for the verification of the formulated hypothesis.
- The hypothesis should be unambiguously accepted or rejected by the chosen data-collection method and applied statistical tests.
- Relevant and verified hypotheses should be kept.
- All in all, hypotheses should provide an answer for the problem that is being examined during the research project.

The number of hypotheses that should be formulated in a given research is unequivocally defined by the objectives and the conditions of the examination. For example, the ideal number of hypotheses in a thesis is maximum 5.

The following example shows the required compounds of formulating a hypothesis. A basic task during this formulation is to define the relationship between the dependent and independent variables. In the cause and effect relationship, the *dependent variable* is the effect while the *independent variable* is the cause. The core question of the research is to describe in what ways the independent variable defines the dependent variable. The following example shows the relationship of the dependent and independent variables in the formulated hypothesis:

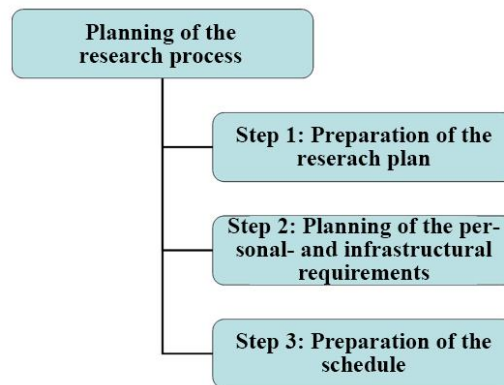
- Wound healing disorientation occurs less frequently during the application of Triclosan suture than during the application of traditional sutures (Pakai et al 2013).

In this example, the dependent variable is the wound healing disorientation, which is influenced by the independent variable, i.e. the type of suture (Triclosan or traditional).

### **1.7. Planning a scientific research and the steps of preparation**

There are several questions that arise in the first phase of a research work and during the formulation of the hypothesis, such as the following: What would the researcher like to examine: the population or a sample? Which is the most appropriate data collection method to be used (a questionnaire, a document analysis, an experiment, etc.)? Which variables may be used to quantify the formulated hypotheses?; Which statistical methods should be used to test the hypotheses?; What may be the limitations of the research?. All of these questions encourage the researcher to precisely plan the process of the research and to assemble a

general, exhaustive research plan. The preparation of the research plan focuses on three basic domains (Figure 1/3.).



**Figure 1/3. Planning the research process**

Source: Edited by the authors based on Hornyacsek 2013

The *first domain of planning* is the definition of the basic items of the research, which must be included in any research plan (Table 1/1):

- The definition of the problem and the objective of the research. Based on the relevant literature, it should be described in 5-10 sentences what the research sets out to examine, why it is worth examining the chosen topic and what practical outcomes are expected.
- The formulation of one or more hypotheses, the number of which depends on the objectives and conditions of the chosen examination. To complete a quality research it is advised to formulate no more than five hypotheses.
- Type of the research. The research plan should only list whether the research is going to be quantitative or qualitative, prospective or retrospective, longitudinal or cross-sectional, etc.
- A short description of the target population.
- Criteria of inclusion and exclusion. Features that will ensure the homogeneity of the sample should be considered already in the planning phase.
- It should be decided in the phase of selecting a sample whether it is going to be random or non-random.
- Sample size of the examined group and the control group (if any).
- Exact location of the examination.
- Planned time of the examination, defined in an interval.

- Choice of data-collection methods and tools (questionnaire, document analysis, content analysis, case study, interview, experiment).
- Description of the examined (dependent and independent) variables.
- Description of the chosen statistical analysis.
- Short summary of the expected results.

It is advised to write an overview of the relevant literature during the preparation of the research plan. Apart from the bibliographic data of the publication, there are a few other features to be considered: a short description of the background issues of the article; the definition of the type, location and time of the research; the method of selecting the sample and a short description of the method; the list of data collection methods and a short description of the most important and relevant statistical tests (Table 1/1).

The *second phase of planning* involves the consideration of personal and infrastructural requirements. It is a basic individual prerequisite that the researcher has the appropriate professional knowledge and expertise in the selected topic; he or she must be familiar with both the international and the national literature on the subject and should also have the ability to generalize and select information and maintain an effective and persistent workflow. The leader of the research project should also consider who to include in the research team. In case of a large scale research it might be necessary to include colleagues to help with data collection and recording. It is also advised to invite colleagues who are expert in statistical analysis and publication of the results. Some topics might require inviting professionals of other fields as well. Infrastructural provisions, tools and certain machines should also be available to complete the research. The appropriate financial resources should be available – the source of this might be the research fund of the given institution, a tender or the support of foundations or other organizations.

The *third phase of planning* is the preparation of a schedule that defines the deadlines of the certain stages during the research. Table 1/3 above summarizes the schedule of a thesis. A proper schedule supports the preparation of a successful thesis, helps meeting deadlines, highlights parallel tasks and challenges, and offers opportunities for corrections (Hornyacsek 2013).

**Table 1/1. Research plan of a thesis (sample)**

Title/topic of research: Health examination of a multinational company				
<p><b>Definition of the problem:</b> the incidence of tumours and cardiovascular diseases had grown by the end of the 20<sup>th</sup> century; stress at the workplaces gradually increases and thus the incidence of connected diseases has also become also higher. It should be made clear for all workers that the conditions provided at one's workplace has a great impact on employees' general health conditions, thus the main aim here is to increase interest in healthy lifestyle and health promotion programmes.</p> <p>Work flows with significant risks and dangers imply higher level of stress. In such workplaces the organization must be on constant alert to react quickly. Constant pressure, the growing level of adrenaline, irregular breath and constant irritation of the muscles may also be harmful to health. Employers, employees and altogether the whole society may pay a high price for the negligence of these factors. Regular physical activity supports a healthier lifestyle, well-being, and better accomplishment in work. It also helps prevent problems with the heart, weight control and blood pressure, helps stress relief, provides a general well-being and, on the long run, supports a healthier and more active aging.</p>				
<p><b>Aims:</b> the aim of the current work is to examine the health of employees working in the production line and in offices in multinational companies in terms of internal medicine and locomotor diseases. The results of the examination may serve as a basis for the organization of further dissemination- and health promotion programmes.</p>				
<p><b>Hypotheses:</b></p> <p>We presume that more than 25% of the workers in production have a higher rate of pulse and blood pressure than normal; 20% of these workers have a higher level of cholesterol than normal; and 5% has a higher rate of glycaemia rate than normal.</p> <p>We presume that due to the unbalanced workload, the pain in the neck, back and-lumbar regions and the positivity of muscle power and stretching is present among more than 25% of workers in the production line in the past one year.</p> <p>We presume that the rate of diagnosed chronic diseases and the number of days in sick-leave is significantly higher among workers on the conveyor belt than those working in other parts of production or in administrative positions.</p>				
<b>Type of research:</b> cross-sectional, quantitative				
<b>Target group of the research:</b> Production and office workers at a multinational company (N=120 people)				
<b>Inclusion criteria:</b> employee of the company agrees to participation		<b>Exclusion criteria:</b> is not an employee of the company does not agree to participation		
In-case sample choice is random: -----		In-case sample choice is non-random: quota sampling based on gender, age and number of years in work		
<b>Sample size of the examined group and the control group (if there is one):</b> 100 people working in the production 20 people working in administration				
<b>Location of the research:</b> for-profit multinational company operating in the Western-Transdanubian region of Hungary				
<b>Time of the examination:</b> July 2012 – January 2013				
<b>Methods and tools for data collection</b> (description of question groups, indication of sources):				
<b>Standardized questionnaire:</b> Referring to health status	<b>Self-edited questionnaire:</b> Demographic data	<b>Structured interview:</b> -	<b>Document analysis:</b> institution's medical database, chronic diseases, number of days spent in sick-leave	<b>Other:</b> Internal medicine: blood pressure, glycaemia, heart rate, oxygen-saturation, level of cholesterol, , abdominal circumference Physical: magnitude of movement in case of major joints, examination of muscular power and stretch measurement:

<b>Indication of source:</b> self-fill questionnaire of the National Residential Health Examination 2003	
<b>Dependent variables:</b> heart rate, blood pressure, glycaemia, cholesterol, neck- and back pain, muscle power and stretch, back and lumbar pain, lifestyle of low physical activity, addictions, level of stress, health status	<b>Independent variables:</b> working in production work position time-span of getting to work
<b>Applied statistical tests:</b> descriptive statistics: absolute and relative frequencies, mean, standard deviation, median, mode mathematical statistics: correlation, one- or two-sample t-tests, one way analysis of variance (ANOVA), Chi-square test Results are considered significant if $p < 0.05$ .	
<b>Applied statistical software:</b> PASW Statistics 18.0	
<b>Expected results:</b> We expect to uncover yet undiagnosed internal medicinal and locomotor diseases, so their treatment can be administered. We also wish to uncover workplace stress, and the informational and health promotion programmes based on our research may decrease the occurrence of these problems and is expected to increase workers' health consciousness. Furthermore, a growth of general joint protection among workers is expected.	

Source: Bajsz 2013

**Table 1/2. Review of a relevant scientific paper (sample)**

<p><b>Data of the publication</b> /reference (author, year of publication year of publication, title, name of journal, year, number of publication, page number): Móczár Cs., Borda F., Faragó K., Borgulya G., Brauniczer F., Vörös V. (2007) Egészséges életmód hatása túlsúlyos és elhízott betegeken, Orvosi Hetilap 148(2):65–69.</p>
<p><b>Problem and aim:</b> Understanding principles of healthy lifestyle and incorporating them into everyday practices is a core principle in the treatment of obese and overweight patients. The aim of the prevention programme within the Direct Patient Care Model Experiment of Kecskemét, Hungary was to decrease cardiovascular risk factors of overweight and obese patients and prevent cardiovascular diseases.</p>
<p><b>Type of research:</b> quantitative, prospective</p>
<p><b>Target group, inclusion and exclusion criteria, sampling:</b> 2489 overweight or obese adults were screened and included in the programme, based on their BMI measure. Those patients were participating who attended their GPs and had BMI higher than 25 kg/m<sup>2</sup>. Inclusion of patients and recognition of the presence of exclusion criteria (exclusion of secondary reasons for weight-gain) was carried out on the basis of their GP's examination and the available medical documentation of the patients. Inclusion was accepted only after the description of the prevention programme and signature of consent.</p>
<p><b>Location and time of examination:</b> Kecskemét and its agglomeration 1 April 2001 to 31 March 2002; Follow up took two years, and the examination ended in 31 March 2004.</p>
<p><b>Data collection method:</b> A data sheet was completed after inclusion to the programme. It recorded data on the following medical features: identification number, age, gender, type of occupation, frequency of physical activity, smoking habits (yes-no categories), alcohol consumption, eating habits. Source document of the examination included the patient's medical records, – the source of the BMI –, abdominal circumference, blood pressure and heart rate (in aestivation and loading tests after 15 squats), the results of laboratory tests such as fasting serum cholesterol, HDL-cholesterol, triglyceride, serum glucose.</p>
<p><b>Statistical analysis:</b> Linear lines were fit to patients' BMI-time graphs using linear regression. Changes in the BMI were detected by two multivariate statistical methods: a) decision tree method and b) covariance analysis. General trends of the BMI were analysed by the Wilcoxon-test.</p>
<p><b>Results, conclusions and recommendations:</b> Small but significant decrease was recognized in the BMI (average decrease: 0.5605; p &lt; 0.001), and in the level of cholesterol (average decrease: : 0.23; p &lt; 0.001). Statistically significant decrease was recognized in other parameters of metabolisms: fasting blood glucose 0.15 mmol/l (p &lt; 0.001), and 0.19 mmol/l (p &lt; 0.03), triglyceride: 0.18 (p &lt; 0.001), and 0.08 mmol/l (not significant), no significant difference was detected in case of HDL-cholesterol. Blood pressure in aestivation was decreased by 5.9 Hgmm (p &lt; 0.001) by the end of the first year and another 0.11 Hgmm (not significant) by the end of the second year. Based on these results the prevention programme can be considered successful, and altogether the cardiovascular risks of patients decreased over the two years.</p>

Source: edited by the authors based on Móczár et al 2007

**Table 1/3. Schedule of thesis preparation**

Research step	Time frame	Exact time
Choosing the topic, contacting the supervisor, consulting the supervisor, writing a literature review, preparing the research model, preparing the research plan	3 and a half months	<b>1 September – 15 December 2014</b>
Submission of research plan	1 week	<b>10-15 September 2014</b>
Institution's declaration on acceptance of the research topic	1 month	<b>16 December 2014 – 30 January 2015</b>
Continuing literature review	2 months	<b>February – March 2015</b>
Specification and finalization of data collection method applied	1 month	<b>April 2015</b>
Data collection, data recording	4 months	<b>May – August 2015</b>
Statistical analysis of data	2 months	<b>September – October 2015</b>
Writing the literature review of the thesis	2 weeks	<b>January 2016</b>
Writing the results and the conclusion section of the thesis	1 month	<b>February – March 2016</b>
Supervisor's opinion and correction	1 week	<b>March 2016</b>
Filing documentation	2 days	<b>end of March 2016</b>
Printing the thesis and copying it on a disk	3 days	<b>beginning of April 2016</b>
Deadline of thesis submission		<b>beginning of April 2016</b>
Preparation for the defence of the thesis, assembling the presentation	2 weeks	<b>June 2016</b>
Defence of thesis		<b>second week of June 2016</b>

Source: edited by the authors based on Majoros 1997

## 1.8. Conceptualization and operationalization

There are various terms used when phrasing the problem, addressing the research questions and formulating the hypotheses. The source of such terms may either be the relevant literature or other scientific results. Some of these terms are clear for everyone; there is no need for a specific explanation, for example the term “suture” used during surgery. On the other hand there are complex terms as well, which may mean different things to different people such as stress, health, quality of life, depression, burnout or fear, where a specific description or definition is essential to offer.

**Conceptualization** is the process in which the researcher unequivocally defines the abstract notions and variables applied. It is also important to clarify dimensions and the appropriate indicators. Dimensions are specific aspects or perspectives of a given term that may be further divided into main and sub-dimensions. A fitting example for this could be the work of Gyöngyvér Salavecz et al, a validation of the short form of the Hungarian version of the Effort–Reward–Imbalance Questionnaire. The examination focuses on measuring stress at work places. According to Synergist’s workplace stress model, if the work effort and its reward are not proportional, then the stress it causes may lead to the decrease of health.

A worker receives a reward for his efforts in his work based on the principle of reciprocity in accordance with the society’s rules. If this is not the case, and the worker’s reap small rewards, then the disequilibrium may lead to physical (e.g. cardiovascular-) or mental (e.g. burnout or depression) diseases (Siegrist 1997). This model does not only consider sources of stress in the workplace, but also looks at the individual features (Salavecz et al 2006). The three main dimensions of stress at workplace are (1) efforts at workplace, (2) rewards at workplace and (3) overtasking. These three are already complex dimensions, but they are further separated into subdimensions. *Efforts at workplace* include time confusion occurring during work, interruptions, distracting factors and growing efforts. Further subdimensions of *reward* are financial rewards, recognition of employers and colleagues, safety and promotional possibilities. The dimension of *overtasking* may be further separated into the subdimensions of ability to be extracted from work or work overload. Finally we reach the indicators that can be unequivocally measured: for example, ability to be extracted from work may be measured with the help of questions, scales or indices, such as “When arriving home it is easy to relax and I do not think about problems at work”. Indicators are regarded as measures that signal the presence or absence of the given notion for a researcher. Indicators should be interchangeable, they should be able to substitute one another, and the same notion should be measurable with the help of indicators with different relationships. After precisely



defining the abstract notions, the next step is to determine the specific measurement techniques and steps. *Operationalization* is the process of defining the steps through which the abstract notions defined during research will become empirically examinable, i.e. all those questions are being formulated that will be asked during the research by various data collection tools (survey, interview, document analysis).

Apart from considering the applied data collection method, the magnitude of measurement and measurement size should also be dealt with during operationalization. The magnitude of measurement refers to the range of examined values relevant for the research. For example, when examining the financial situation of a poor social class, there is no use to define the limit of food consumption in 200,000 HUF (approx. 630 EUR). An exact definition of *measurement levels* is inevitable, as both data collection and the definition of the statistical methods are scale-dependent. Levels of measurement are further discussed in *Chapter 4*. It is well depicted by the process of conceptualization and operationalization how various subjects, phenomena, and attributes and variables of certain features are defined (Table 1/4).

For example, attributes of level of education as a variable may be: primary school, high school diploma, higher education degree; attributes of hair colour as a variable may be black, brown, blond or red; or the variable of glycaemia may show different values in case of various patients (Pakai – Kívés 2013, Falus 2004, Héra – Ligeti 2006).

**Table 1/4. Variables and their attributes of a BSc-level student studying at the Faculty of Health Sciences, University of Pécs**

variable	attributes
age	19 years 20 years 21 years 22 years
gender	male female
location of campus	Pécs Kaposvár Szombathely Zalaegerszeg
specialization	nurse health visitor physiotherapist paramedical dietician midwife public health inspector health promotion recreation management diagnostic imaging analyst medical diagnostic laboratory analyst
year	1 <sup>st</sup> year 2 <sup>nd</sup> year 3 <sup>rd</sup> year 4 <sup>th</sup> year
programme	full-time correspondent

Source: edited by the authors

There are three basic requirements for measurement: validity, reliability and objectivity. **Validity** reflects how the chosen data collection tool actually measures what the researcher set out to measure. It may have various types:

- Content validity: all aspects of the notion examined are considered.
- Construct validity: the applied measurement tool meets the scientific expectations.
- Current validity: shows how much correspondence there is between the new measurement tool applied in the current study and measurement tools applied in previous studies.
- Predictive validity: shows how much correspondence there is between the results of the current study and the result of a possible future study on the same topic.

*We talk about reliability* when repeated measurements with the same tools bring the same or similar results as the original measurement. *Objectivity* shows how capable the researcher is to complete the examination objectively and independently or, in other words, the results of the measurement must not depend on any other factor only on the factor being observed (e.g. the performance of a student) (Falus 2004).

## 1.9. Types of research

Another question that should be clarified at the beginning of all scientific research is what the most effective research methodology is to reach the aim of the examination. The decision might be influenced by personal, infrastructural or financial aspects based on the differences between various examinations. There are several ways to differentiate between various types of research –the following paragraph will look at the most important concepts.

We may distinguish between basic, applied, and developmental research activities depending on the level of examination. *Basic research* is an experimental, observational, systematic or theoretical work based on empirical observation that focuses on understanding certain phenomena or principles of the world. Its chief aim is to gather new information and knowledge to support a more detailed theory. Pure basic research does not set out to have any specific practical usage, while the main idea of targeted basic research is to form a basis for finding the answer to future problems. Basic research is most often carried out by higher education institutions, universities, or research institutions. Here is one example from the field of pharmaceutical studies: researchers of Northumbria University examined the effect of dietary supplements with certain content of fish oils. Altogether 22 young adults (average age=21.96 years) took part in the double-blind, placebo-controlled examination. The first group of participants received fish oil rich in docosahexaenoic acid, the second group of participants received fish oil rich in eicosapentaenoic acid (EPA), and the third group received olive oil. Results show that regular intake of fish oil with various amount of omega3 did not influence mental functions significantly, although it did decrease mental tiredness and increase reaction time (Jackson et al 2012). *Applied research activities* focus on answering problems of practice using the findings of basic research. Such examinations, for example, include market research, where the service-provider uses the results to plan a layout or change a business strategy. Another example is a research activity conducted by pharmaceutical companies with substances or pharmaceutical products. The objective of developmental research is to expand and evolve various theories, methods, tools and products. A good example for this kind of research is the work involving three various departments of the

Budapest University of Technology and Economics supported by the Economic Competitiveness Operational Programme of the European Union, in which a device that monitors health status at home operable without a special knowledge was developed. A working group from the Faculty of Health Sciences (Zalaegerszeg campus) of the University of Pécs, cooperated in the programme by maintaining information flow and monitoring the machine's development (Jobbágy et al 2008, Karamánné et al 2006).

Naturally, time also plays an important role during research. An examination may take place at a specific time or also over a longer period of time. In case of *cross-sectional studies*, data collection on the examined sample is usually made only once, at a specific time, without any repetition. This way, this specific type of examination focuses on the individual's state, opinion or status of a specific time, although questions may arise that refer to the past (a few months or years back), but the reliability of data in such a case is questionable. A cross-sectional study conducted by Réka Vajda et al examined the knowledge on cervical cancer screening among women with daughters aged 9-14 (Vajda et al 2014). *Longitudinal research*, however, takes a longer period of time. The examination is repeated with the same sample after a specific amount of time has elapsed. An example for this is the research that examined changes in the lifestyle of healthcare workers during a period between 1986 and 2006, collecting data every four years. At the beginning of the research, participants had normal weight and did not suffer from any type of chronic disease. Results showed an average 1.5 kg increase in weight and an altogether 7.6 kg weight-gain within the span of 20 years (Mozaffarian et al 2011). Longitudinal studies may be divided into three types: *trend research* examines changes taking place in the same sample. *Cohort studies* look at smaller samples and detect changes of the sample through time. *Panel research* examines the same sample several times. Longitudinal studies offer larger quantity and more reliable information on the sample than cross-sectional studies, but they are a lot more expensive to carry out. When it comes to time, we should also mention prospective and retrospective research. In *prospective research* the tasks of measurement and observation are completed during the examination, therefore data is basically collected during the examination. In retrospective research, the analysis focuses on events and data collected prior to the date of examination. It is important to remember that experiments fall in the category of prospective research, while observational studies may belong to both of the above-mentioned types. Cohort studies must be differentiated from prospective studies – they may also be retrospective in nature (for example when examining archived documents to study the population of a given city). Case-

control studies, however, are always retrospective (for example when two groups are considered: one with a certain disease and one without) (Reiczigel 2005).

Basically there are three strategies available for a researcher in the planning phase, based on the type of the research question. Choosing the right strategy is of key importance as it will define each step in the researcher's forthcoming work. The aim of *qualitative studies*, most typically used for quality analysis and seldom producing numerical results, is to provide deeper, more detailed knowledge to highlight underlining relationships and to help better understanding a problem. This type of study usually seeks answers for questions such as "Why?" and "How?", that arose during the formulation of the problem. Data collection takes place in a small, non-representative sample that has specific limits. The researcher builds a confidential relationship with the respondents during the examination, and he may be able to collect vital information that may not be available during a quantitative research. The disadvantage of qualitative studies is that they are not applicable for collecting objective, numerical data, they have low reliability and a high risk of subjectivity. The most often used methods of qualitative studies are observations, experiments, in-depth interviews, delphi studies, focus groups studies and case studies.

Another group of research is *quantitative studies*, used for gathering information that has a measurable quantity. These types of studies provide information required by the study on the sample using structured surveys, structured interviews or structured observations. The large amount of reliable and numerical data collected may be analysed with various descriptive and mathematical statistics. However, despite the substantial differences, *both qualitative and quantitative research* may be applied during a complex and comprehensive study, such as in case of the work of Szabó et al (2009) carried out in the Addictology Department of the Psychiatric Ward of the Hospital of Zala County, dealing with patients with a drinking problem. In the research programme, participants were asked to produce a handwritten CV. An analysis was carried out by content analysis software referred to as ATLAS ti. 5.00, focusing on the frequency of words with a social content. First of all, all words were extracted from the CVs by the researchers with the help of this particular software, then words referring to human relationships were highlighted and organized into categories based on their meaning. Finally the frequency of the interconnected social words was examined.

## **1.10. Ethical questions of scientific research**

Researchers throughout their scientific work have a great responsibility; not only do they have to possess the appropriate professional knowledge and experience but they also have to meet ethical expectations.

The leader of the research programme is responsible for all ethical issues. The individual rights of every participant must be respected and anonymity must be ensured. The research must provide genuine, detailed information about the objectives and expected duration of the research, its advantages and risks – all this may serve as a basis for a confidential relationship between the respondent and the researcher. An examination must be subject to the consent of the participant, which may be withdrawn any time. Risks that the participants should run must be minimized, and providing safe circumstances is also a fundamental requirement. Finally, the time that the participant spends on the programme must also be appreciated.

When having to meet ethical expectations, scientific researchers must carry out their work in accordance with the basic rules of their profession, making sure they establish and reach high quality standards, and they do their work accurately. Data protection, the usage of sources, and ethical rules of research results should be carefully dealt with to avoid ethical misdemeanour. It is unethical to manipulate or falsify results, or omit incriminating data. Since a researcher uses results of other scientists in his or her work, it is very important to apply precise references to keep science clear and avoid the violation of copyright (Falus 2004, Cserné 1999, Majoros 2004, Dempsey 1999).

### **1.10.1. Research including human participants**

Human beings have been participating in scientific research for a long time, although it had not always been a beneficiary experience for them and often it happened contrary to their consent. This concept is proven by one of Herodotus' notes as well, written in the 7<sup>th</sup> century based on the stories of Hephaestus' priests. Pharaoh Psammetichus ordered two new-borns to be taken by a shepherd to raise them among his flock and do not ever speak to the babies. This way the pharaoh wanted to see what language the children will start speaking, as he supposed that particular language to be the primal language of humankind.

Centuries of history have shown that in most cases research with human subjects involved slaves, prisoners of war, or other people who were not considered equal. The basic rights of human subjects involved in research during World War II were often violated as well. Such experiments included Mengele's twin studies in a laboratory based in the Auschwitz concentration camp. Nowadays, however, there are strict regulations that must be applied

when conducting both international and national experiments involving human subjects. For example, the basic requirements for experiments involving human subjects are set by the 10 principles of the Nuremberg Code, securing the dignity and safety of participants (Table 1/5.). These rules also highlight that participants who are involved in the research must be informed and asked for their consent to complete research prior to the examination (Kerpel-Fronius 2008, Kovács 2007).

The World Medical Association accepted the Helsinki Declaration in 1964, providing guidance for professionals in medical research. The Ethical Principles of Medical Research consisting of 35 items is revised and edited every 6-8 years. The declaration has three main parts: (1) the introduction; (2) the basic rules applicable for all medical research; and (3) the supplementary principles for biomedical research carried out during medical care.

The various committees of the Council of Europe and the Steering Committee on Bioethics (CDBI) have a significant role in the ethical regulation of medical and biomedical research activities. The Oviedo Convention on Human Rights and Biomedicine took place in 1997, and was also signed by Hungary. The obligatory regulation required by the CDBI to comply with is the following: human cloning is forbidden, organ trade is forbidden, biomedical research is strictly regulated, and also genetic examinations with a medical purpose are strictly regulated. Furthermore, there are other recommendations published covering xenotransplantation, the protection of rights of patients with mental disorders, and examinations involving human-biological substances.

The various directives of the European Union on human examinations, pharmaceutical experiments and animal testing should also be considered (Fésűs 2014).

**Table 1/5. The Nuremberg Code**

- The voluntary consent of the human subject is absolutely essential.
- The experiment should be such as to yield fruitful results for the good of society, unprocurable by other methods or means of study, and not random and unnecessary in nature.
- The experiment should be so designed and based on the results of animal experimentation and a knowledge of the natural history of the disease or other problem under study, that the anticipated results will justify the performance of the experiment.
- The experiment should be so conducted as to avoid all unnecessary physical and mental suffering and injury.
- No experiment should be conducted, where there is an *apriori* reason to believe that death or disabling injury will occur; except, perhaps, in those experiments where the experimental physicians also serve as subjects.
- The degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved by the experiment.
- Proper preparations should be made and adequate facilities provided to protect the experimental subject against even remote possibilities of injury, disability, or death.
- The experiment should be conducted only by scientifically qualified persons. The highest degree of skill and care should be required through all stages of the experiment of those who conduct or engage in the experiment.
- During the course of the experiment, the human subject should be at liberty to bring the experiment to an end, if he has reached the physical or mental state, where continuation of the experiment seemed to him to be impossible.
- During the course of the experiment, the scientist in charge must be prepared to terminate the experiment at any stage, if he has probable cause to believe, in the exercise of the good faith, superior skill and careful judgment required of him, that a continuation of the experiment is likely to result in injury, disability, or death to the experimental subject.

Source: edited by the authors based on <http://www.hhs.gov/ohrp/archive/nurcode.html>

### **1.10.2. Regulations in Hungary**

A scientific permission is required prior to the research. In Hungary, an application for such a permission must be submitted to the appropriate research ethics committee.

Based on Decree 23/2002 (V.9.) of the Minister of Health on medical research involving human subjects (revised by Decree 31/2009. (X.20.) of the Minister of Health) the following committees are entitled to issue professional ethical permissions:



- Scientific and Research Ethics Committee of the Medical Research Council (ETT TUKEB)
- Regional Research Ethics Committee (REKEB)
- Institutional Research Ethics Committee (IKEB)

Prior to submitting a research permission it is recommended to consult the relevant Hungarian legislation in force (Supplement Nr. 1), that would secure validation of basic human rights during research involving human subjects. It is important to be informed about the requirements to be met when submitting such a request. Information on the procedures can be found on the webpage of the Medical Research Council at [www.ett.hu](http://www.ett.hu) (available in Hungarian). It should also be highlighted that a research plan must be submitted by the professional leader of the programme concerning both invasive and non-invasive examinations (Alexin – Lelovics 2009).

## **2. LITERATURE REVIEW IN PRACTICE. USING THE MOST POPULAR DATABASES FOR LITERATURE REVIEW (Annamária Karamánné Pakai, András Oláh)**

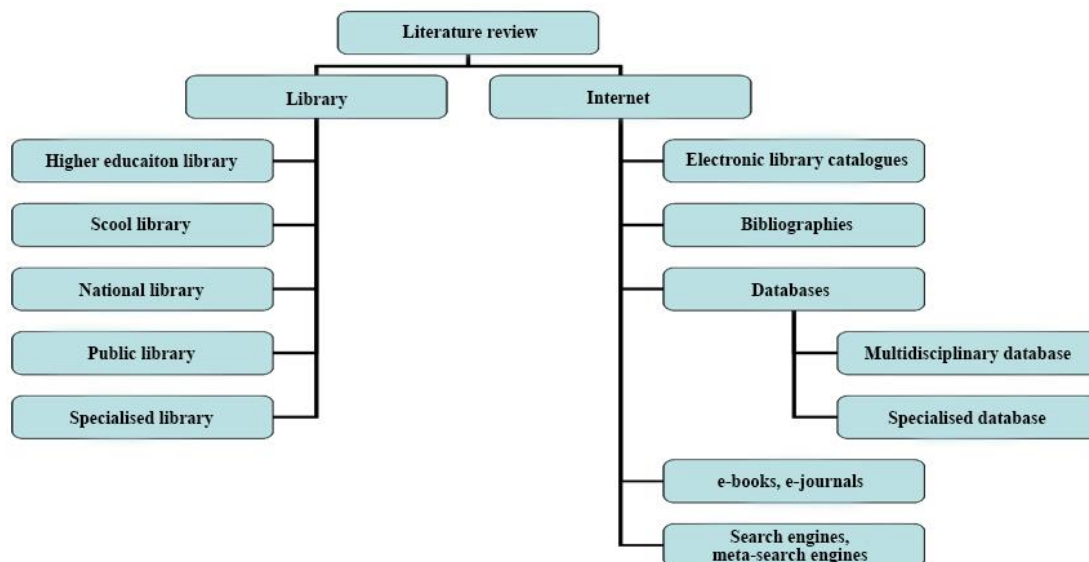
### **2.1. Exploring and gathering relevant scientific literature**

After selecting and identifying the research problem, it is inevitable to explore and review the relevant (i.e. important, determining materials of the field) scientific literature written on the subject both in a national and international context. Nowadays there is a wide range of libraries, specialized libraries and electronic databases available for researchers to do this job, most of which provide full-text access to most publications (Figure 2/1.).

Up-to-date, evidence-based scientific work can only be carried out if the researcher is familiar with the chosen topic. When looking at previous works on the subject, the researcher must comprehend what sort of international and national studies were published in the chosen field. An appropriate knowledge of a topic can only be accumulated if the researcher is familiar with the authors, types of research, problem-focues, target groups, samples and data collection methods (survey, document analysis, experiment, etc.) most often applied in the given field to support hypotheses; and if he is familiar with the new issues recommended to explore on the basis of previous results; and what sort of limitations did previous research activities may have had.

Reviewing the relevant literature helps precisely define the research question(s), understand the scientific concepts, and also assemble data collection method and tools, in the analysis of results, and in the preparation of the scientific publication based on the results.

Not only is literature review is not only a time-consuming task, but it also requires a thorough knowledge on research methodology and statistics, as the researcher must be able to assess the quality of the given publication (at least from a few aspects) and the reliability of the information and results.



**Figure 2/1. Starting points of a literature review**

Source: edited by the authors

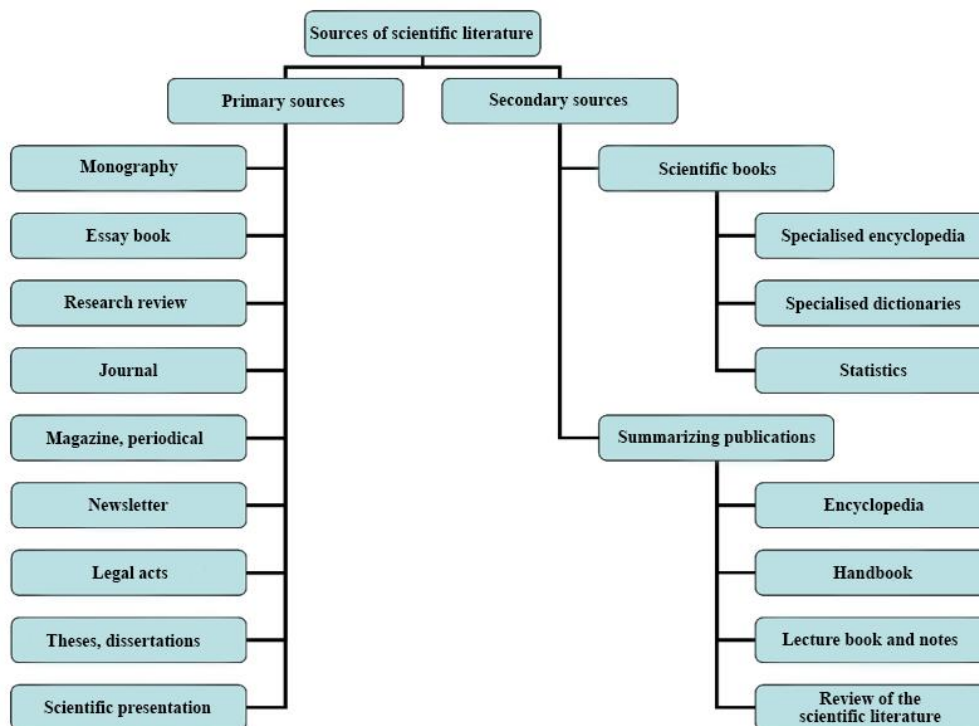
## 2.2. Sources of a literature review

Some of the relevant sources required for a scientific publication is available in institutional or specialized libraries. Primary and secondary sources may be used during the review (Figure 2/2), but the importance of grey literature should also be highlighted.

**Primary sources** are original works covering a specific topic of a field and describe the research results of the author directly, in the form of a journal article, dissertation or conference presentation.

**Secondary sources** are summaries based on the primary sources of a given field that provide an exhaustive picture of the topic. They do not only describe the primary source used, but may also analyse, argue, complement or even prove it.

**“Grey literature”** consists of works which are used by a relatively small circle of professionals, while they may be quite useful for the researcher since works of higher education institutions or governmental or other organizations belong to this category, and these are often difficult to access since they are not available in shops.



**Figure 2/2. Sources of scientific literature**  
 Source: edited by the authors based on Oláh 2008

### 2.3. Use of internet during literature review

Online-based literature reviews may have various starting points, and these will be discussed in the forthcoming paragraphs. They may include:

- Catalogues of electronic libraries
- Electronically available bibliographies
- On-line databases
- Electronic books or journals
- Online search engines

Supplementary services may be:

- newsletter, mailing lists and websites of professional organizations or associations, such as:
  - newsletter of the Hungarian Higher Education Sport Association  
 (available at: [www.mefs.hu/hirlevel](http://www.mefs.hu/hirlevel)),
- the website of the Hungarian Sport Sciences Association  
 (available at: [www.sporttudomany.hu](http://www.sporttudomany.hu)),

- the website of the Hungarian Table Tennis Association  
(available at: [www.moatsz.hu](http://www.moatsz.hu))

Before starting the literature review it is important to collect the most relevant Hungarian and English (or other foreign language) keywords and concepts, and to define synonyms and the various aspects of our search (such as age, gender, language or time). The most important factor of a successful database-search is a correct definition of search terms.

## **2.4. Electronic libraries**

Library catalogues offer the possibility to search documents available in the specific library by various features. Nowadays most libraries have already switched from the traditional card-based catalogues to a technologically more modern electronic catalogue system. The latter offers the major advantage that data of the library's stock is available not only from the given library but also from external PC's, and thus it may be decided whether the literature we would like to access is available in the given library or not, and if yes, what status it has (available to borrow, or accessible only in the reading room).

### **Online Public Access Catalogue (OPAC)**

OPAC is a free, online, public library database accessible by anyone regardless of location or time. It offers a wide range of possibilities to search; works are distributed by author (editor or other writer), title or title fragments, keyword (referring to the content of the document), topic, publisher, ISBN and ISSN number (individual identification number of book or journal), UDC (Universal Decimal Classification, not used at the University of Pécs), location (library, institutions, hospitals – with the help of the librarian), warehouse location (call number (with the help of the librarian), year of publication (the same number should be entered to both queries, this is a filtering condition), type of document (book, periodical, file, visual file, etc.; this is also a filtering condition), and language (a filtering condition as well).

We used the electronic catalogue of Central Library of University of Pécs as a source of the examples in the following paragraphs. It is available at: [www.lib.pte.hu](http://www.lib.pte.hu)

### **Search mode**

In most cases there are three search modes offered by online catalogues: 1) simple search; 2) advanced search and 3) browse. *Simple searches* may be carried out by author, title, keyword, publisher, or identification number (ISBN, ISSN). The record/page display can be set between

4-100 search results (Figure 2/3). An *incomplete search* may be applied when the exact content of the notion of interest is unknown or we wish to enter only a fragment of it. This may be carried out by omitting the beginning or the end of the word, but may also be used as a substitution in the middle of the word. The most often used symbols for this purpose are the dollar sign (\$), the hashtag sign (#), the asterix (\*), the percentage sign (%), or the question mark (?), but substitution of certain letters may also be carried out by the application of an underscore (\_).

Example: we are going to receive all records referring to health (e.g. healthy, health care, health impairment, health sociology, etc.) if we apply the incomplete search option typing “health%”.

**Advanced search** engines provide the opportunity to combine multiple simple searches. In this case the so called Boole-operators (or logic operators) are used to connect various searches in the database. There are three operators in use to generate connections: **and**; **or** and **not**; which generate several connections during various searches (Figure 2/4).

AND (logical AND connection) provides records that include all search terms

Example: **sport and health**

If one of the search words are **sport** and the other one is **health**, then the search for “**sport and health**” will provide records that include both of these words, regardless of the order of the words.

OR (logical OR connection) provides records that include at least one of the search terms

Example: **sport or health**

Resulting records will include both the term *sport* and the term *health*, and also those ones where both terms occurs (this common scope can be defined by the use of the ‘*and*’ operator). The order of the words does not matter in this case either, thus *sport or health = health or sport*.

NOT (logical NOT or exclusion) will provide records that include the term indicated first but will not include the words indicated afterwards

Example: **sport not health**

Resulting records will include the term *sport* but any record will be deleted that also includes the word *health*. In this case the order of the search terms is really important: there will be very different results for the search query *sport not health* than for the query *health not sport*.

When searching, the search term should be inserted into the appropriate place, and the appropriate type should be chosen from the drop-down list. The location of documents, the year of publication, the language and type of document (i.e. book, journal, article, film, etc.) may also be set. The search process can be launched by clicking the *SEARCH* button.

**Browsing** offers the opportunity to search standardized classification data of a specific database (authors, titles, key words) (Figure 2/5). This is a useful option when the researcher is not sure about the spelling of the searched term or the name of the author, or the specific form of the term used in the given database. In this case, symbols of incomplete search may only be used at the end of the word. Browsing is a kind of help provided by the database to find the appropriate search term. This way not only the list of authors' names is searchable, but also specific names (e.g. "George Bush"), as well as any other data (title, keywords, etc.) containing the word "bush".

### **Display of records**

A short list of records appears after we launched a search based on the provided search terms, which contains the most essential information to identify the document, such as the name of the author, the title of the work, year of publication and the type of the document. There are several options to display the resulting records: they may be listed by the number of copies available, length (short or long), year of publication, alphabetical order based on the author's name or title, in ascending or descending order. Clicking *DETAILS* will provide further information about the document's placement and availability (Figure 2/6 and 2/7).

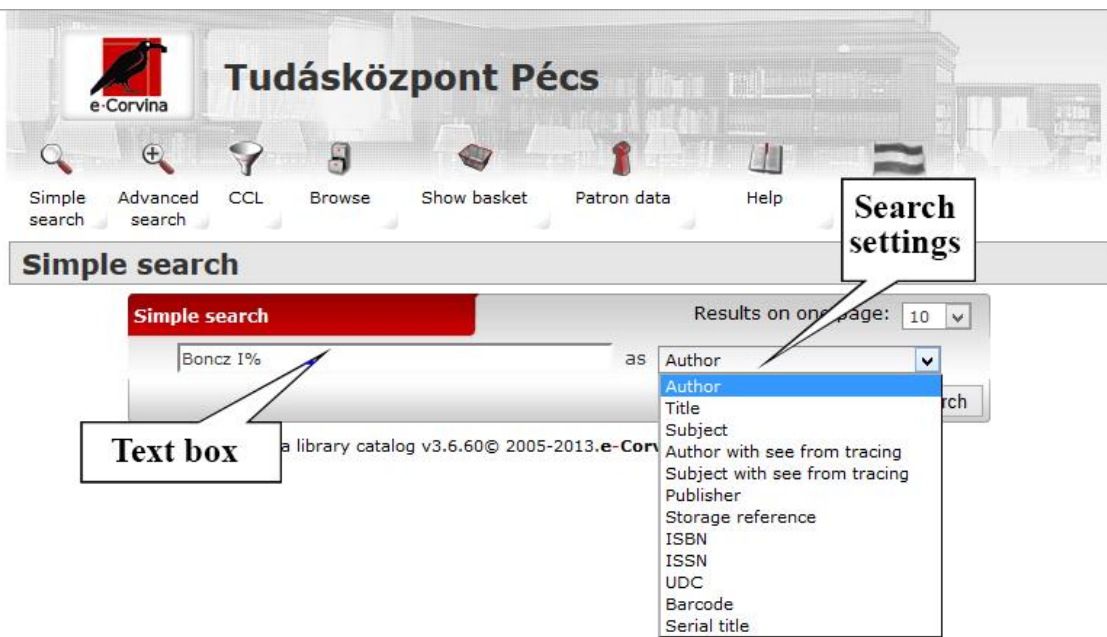


Figure 2/3. Simple search in the library catalogue

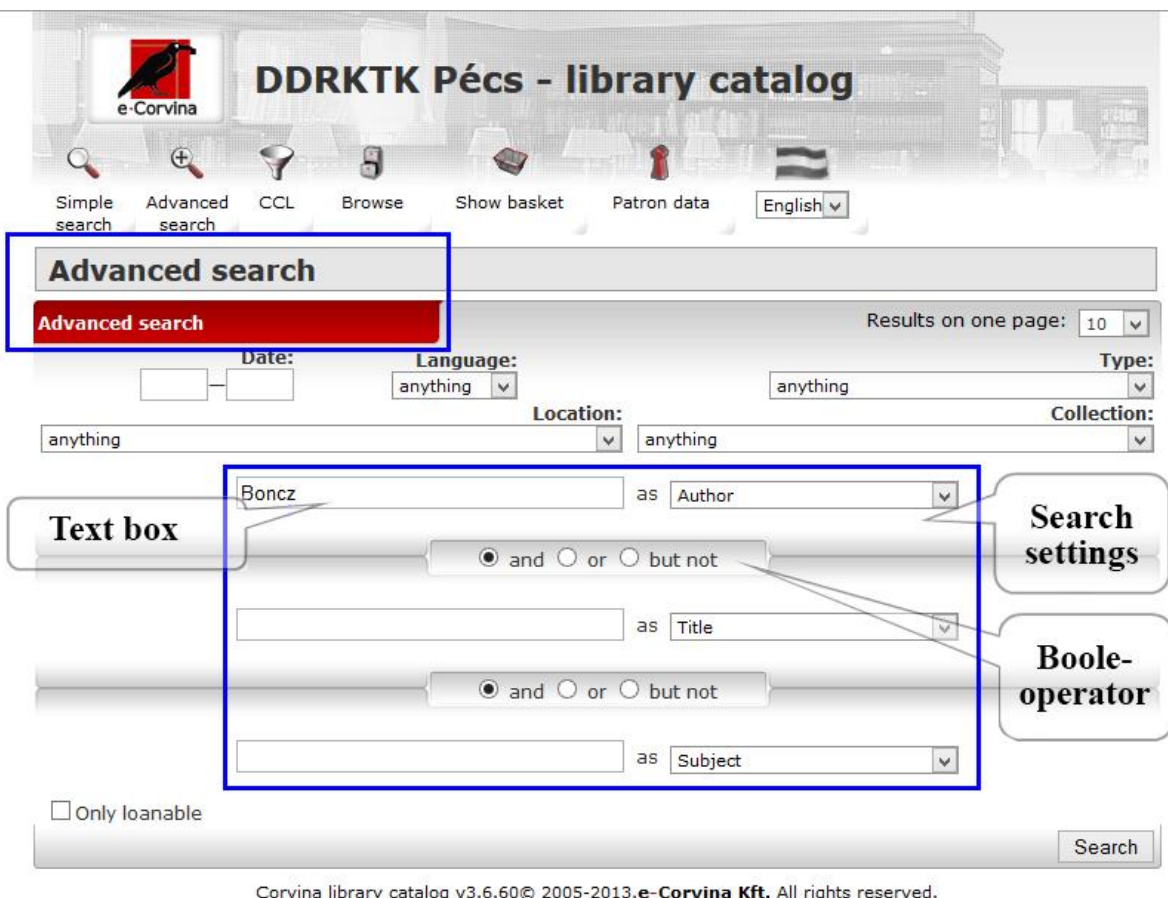


Figure 2/4. Advanced search in the library catalogue



**DDRKTK Pécs - library catalog**

Simple search | Advanced search | CCL | Browse | Show basket | Search history | Patron data | English

**Result list**

Total 38 results. Order: Author ascendent | Detail: Short | #/page: 10

#	Basket	Author	Title	Date	Type	Long
1.	<input type="checkbox"/>	Boncz Dániel	Ilyen a Boncz! : történetek Boncz Gézaról /	2007	Book	<a href="#">Details</a>
2.	<input type="checkbox"/>	Boncz Ferenc	A vallás körüli felségjogok /	1894	Book	<a href="#">Details</a>
3.	<input type="checkbox"/>	Boncz Ferenc	A lelkészi congrua /	1888	Book	<a href="#">Details</a>
4.	<input type="checkbox"/>	Boncz Ferenc	A katolikus főpapi hagyatékok körüli eljárás és erre vonatkozó főbb rendeletek /	1878	Book	<a href="#">Details</a>
5.	<input type="checkbox"/>	Boncz Ferenc	A magyar közigazgatási törvénytudomány kézikönyve : A törvényhozás legújabb állása szerint /	1876	Book	<a href="#">Details</a>
6.	<input type="checkbox"/>	Boncz Ferenc	A magyar közigazgatási törvénytudomány kézikönyve : A törvényhozás legújabb állása szerint /	1876	Book	<a href="#">Details</a>
7.	<input type="checkbox"/>	Boncz Ferenc	Magyar államjog /	1877	Book	<a href="#">Details</a>
8.	<input type="checkbox"/>	Boncz Géza	Az őrlület határa /	[1999]	Book	<a href="#">Details</a>
9.	<input type="checkbox"/>	Boncz Imre PTE ETK	Kutatásmódszertani és egészségügyi statisztikai alapismeretek : jegyzet valamennyi szak számára /	2012	Book	<a href="#">Details</a>
10.	<input type="checkbox"/>	Boncz Imre PTE ETK	Kutatásmódszertani és egészségügyi statisztikai alapismeretek : jegyzet valamennyi szak számára /	2004	Book	<a href="#">Details</a>

Download 1 2 3 4

Corvina library catalog v3.6.60© 2005-2013.e-Corvina Kft. All rights reserved.

Figure 2/5. Browsing in the library catalogue

**DDRKTK Pécs - library catalog**

Simple search | Advanced search | CCL | Browse | Show basket | Search history | Patron data | English

**Result list**

Total 38 results. Order: Author ascendent | Detail: Short

#	Basket	Author	Title	Date	Type	Long
1.	<input type="checkbox"/>	Boncz Dániel	Ilyen a Boncz! : történetek Boncz Gézaról /	2007	Book	<a href="#">Details</a>
2.	<input type="checkbox"/>	Boncz Ferenc	A vallás körüli felségjogok /	1894	Book	<a href="#">Details</a>
3.	<input type="checkbox"/>	Boncz Ferenc	A lelkészi congrua /	1888	Book	<a href="#">Details</a>
4.	<input type="checkbox"/>	Boncz Ferenc	A katolikus főpapi hagyatékok körüli eljárás és erre vonatkozó főbb rendeletek /	1878	Book	<a href="#">Details</a>
5.	<input type="checkbox"/>	Boncz Ferenc	A magyar közigazgatási törvénytudomány kézikönyve : A törvényhozás legújabb állása szerint /	1876	Book	<a href="#">Details</a>
6.	<input type="checkbox"/>	Boncz Ferenc	A magyar közigazgatási törvénytudomány kézikönyve : A törvényhozás legújabb állása szerint /	1876	Book	<a href="#">Details</a>
7.	<input type="checkbox"/>	Boncz Ferenc	Magyar államjog /	1877	Book	<a href="#">Details</a>
8.	<input type="checkbox"/>	Boncz Géza	Az őrlület határa /	[1999]	Book	<a href="#">Details</a>
9.	<input type="checkbox"/>	Boncz Imre PTE ETK	Kutatásmódszertani és egészségügyi statisztikai alapismeretek : jegyzet valamennyi szak számára /	2012	Book	<a href="#">Details</a>
10.	<input type="checkbox"/>	Boncz Imre PTE ETK	Kutatásmódszertani és egészségügyi statisztikai alapismeretek : jegyzet valamennyi szak számára /	2004	Book	<a href="#">Details</a>

Download 1 2 3 4

Corvina library catalog v3.6.60© 2005-2013.e-Corvina Kft. All rights reserved.

Figure 2/6. Result list

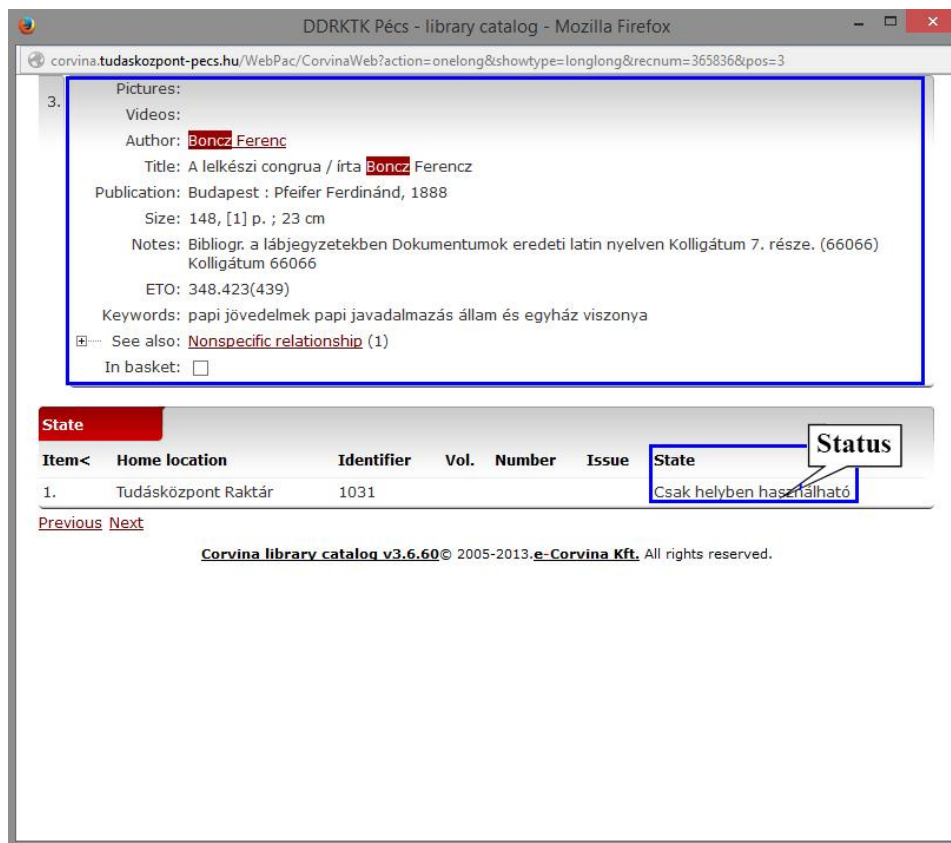


Figure 2/7. Details of the result list

## 2.5. Electronic bibliographies

Apart from library catalogues there are several other, regularly updated electronic databases and up-to-date specialized bibliographies that might help us review the relevant literature in detail. **Bibliographies** are specific lists based on certain features of the referred documents. They may be classified based on the topic, the place of publication, the year of publication, or document type, and most often are arranged alphabetically or by the year of publication. The role of **specialized bibliographies** is to collect and systematically list publications of a given scientific field, such as Bibliographia Medica Hungarica.

### Bibliographia Medica Hungarica (MOB)

Hungary's journal publications and letters from the field of medicine and other related studies are handled by the Health Science Library of the National Institute for Quality- and Organizational Development in Healthcare and Medicines' Directorate General of IT and Health System Analysis. This is an online database open for the public that was previously published in print as well, and since 2006 the data has been published on a CD-ROM four times a year. Each disc contains approximately 1000 new records. Books, journals and other

documents may be searched in this database such as temporary publications, dissertations, and conference books. An incomplete search may be carried out with the use of the percentage sign (%).

### ***Search methods***

The MOB interface – similarly to the library catalogue – provides opportunities for both simple and advance search. A *simple search* can be based on title, author, journal or key words. Researchers are advised to use make searches by title and key words, as the results will be displayed in the alphabetical order of the records' titles. Articles at the beginning of the title are not required to be indicated as it would influence the alphabetical order of the resulting records. (Figure 2/8)

An *advanced search* is carried out by indicating more than one search terms, which have an AND connection to each other (logical AND). The list of the results is similar to that of a simple search.

Browsing is available on the basis of a MOB publication numbers, authors, key words or journal titles.

### ***Display of results***

Results are displayed on the interface in a list containing 20 items and distributed on a maximum of 20 pages, adding up to a maximum of 400 records. Clicking the title of a record will provide a detailed description and a short summary of the document. Web links of certain journals have been indicated since 2011, helping to navigate to their own websites. Furthermore, the detailed description provides a link for the given database record that can be embedded to other sites or texts. Clicking several boxes next to the relevant records will produce an isbd-format file containing the detailed description of all the chosen records (maximum 20) without their short summary. However, choosing from records is only possible from one result page at a time.

Relevant records can be viewed later as well, and the download option is also available (Figures 2/10 and 2/11).

### ***Access:***

The MOB database is available at <http://www.eski.hu>

Kérjük, hogy használat előtt olvassa át a Súgót!

Egyszerű keresés    Simple search

Cím    Boncz I%    Keres

Cím  
 MOB  
 Módszertani  
 Szerzők  
 Tárgyszavak  
 Folyóirat  
 Kulcsszó

title  
 MOB  
 methodological  
 Authors  
 Subject Words  
 journal  
 Keyword

Search

Figure 2/8. Simple search at the MOB interface

Összetett keresés    Advanced search

Szerzők    Boncz I%

Szerzők    Betlehem J%

Szerzők    Oláh A%

Keres    Authors    Advanced search according to authors

Search

- Kistérségi egyenlőtlenségek az otthoni szakápolás vonatkozásában a dél-dunántúli térségben
- A kivonuló mentődolgozók egészségi állapotát befolyásoló főbb tényezők hazánkban
- A munka hatása a kórházi ápolók jóllétére Magyarországon az EU csatlakozáskor
- Az OEP otthoni szakápolási kassza igénybevételének területi egyenlőtlenségei
- Az otthoni szakápolás egészségbiztosítási vonatkozásainak elemzése Magyarországon
- Az otthoni szakápolás igénybevételének területi egyenlőtlenségei a dél-dunántúli térségben
- Tudományos közlések az egészségtudományban
- Önkéntes ápolói nyilvántartás Németországban

Mind kijelölve:    
    
    

Figure 2/9. Advanced search and result list

Kérjük, hogy használat előtt olvassa át a Súgót!

Egyszerű keresés    Simple search

Szerzők    Boncz I%    Keres

<< 1 2 3 4 >>    Results list    Search

1.  10 éves a HBCS! : a Homogén Betegségcsoportok (HBCS) rendszerének tapasztalatai finanszírozói oldalról
2.  A 2002. évi szervezett lakossági emlőszűrés monitorozásának eredményei
3.  A 2007. április 1-i reform hatása a dél-dunántúli egészségügyi intézmények piaci részesedésére
4.  A 2007. április 1-jei egészségügyi reformintézkedések hatása az összes kórházi ágyszámra
5.  60 év alatti combnyaktörötték csavaros osteosynthesiseit követő további ellátások és rizikó tényezők kapcsolata
6.  Az aktív fekvőbeteg szakellátás finanszírozásának visszavezetése a depresszív TVK irányába 2010-2012. között
7.  Az akut myocardialis infarctus betegségterhe Magyarországon, 2003-2005
8.  Az akut stroke előfordulása és betegségterhe hazánkban, OEP-adatok alapján
9.  Attitűdváltozások 18-19 éves fiúk körében a "FÜGE" drogreprevenációs program hatására
10.  Batthyány-Strattmann László : a ferences-herceg-orvos

Figure 2/10. Result list

Részletek

Results

MOB link

A cikk állandó MOB linkje:

<http://mob.gyemsi.hu/detailsperm.jsp?PERMID=88448>

**MOB:** 2005/4

**Szerzők:** Boncz Imre; Hoffer Gábor; Sebestyén Andor; Dózsa Csaba; Ember István

**Tárgyszavak:** EMLŐ DAGANATAI; SZŰRŐVIZSGÁLATOK

**Folyóirat:** Magyar Onkológia - 2005. 49. évf. 2. sz.  
[<http://www.medicalonline.hu/kiadvany.php?paperId=75> ]

Link to the journal

A 2002. évi szervezett lakossági emlőszűrés monitorozásának eredményei / Boncz Imre [et al.]  
Bibliogr.: p. 114-115. - Abstr. hun., eng.  
In: Magyar Onkológia. - ISSN 0025-0244. - 2005. 49. évf. 2. sz., p. 109-115. : ill.

CÉL: Dolgozatunk célja az emlőrákszűrés és az azzal összefüggő egészségügyi ellátás igénybevételének területi, időbeni és populációs jellegzetességeinek elemzése. MÓDSZEREK: Az elemzéshez felhasznált adatok az Országos Egészségbiztosítási Pénztár rutinszerűen gyűjtött finanszírozási adatokat tartalmazó adatbázisából származnak. A kiinduló lakossági kört a 2002. naptári évben a szervezett mammográfiás szűrésen részt vett nők képezték (N=314 395). Az időbeli elemzéseknél kiindulási időpontnak (T0) a "42400 Mammográfiás szűrés" kóddal azonosított mammográfiás szűrés elvégzésének időpontját tekintettük, és vizsgáltuk az ezt követő diagnosztikai (T1 időpont) és

Article abstract

Figure 2/11. Details of the relevant result

## 2.6. Specialized online databases

English-language specialized online databases generally contain scientific journals, relevant books and conference publications from various scientific fields, such as natural sciences, medicine, health sciences, social sciences or engineering. Some databases tend to comprise only of abstracts or bibliographic data of publications, while in many cases full-texts, illustrations, graphs and tables are also available. Resulting records may be downloaded, printed, or sent via e-mail. Generally, most of these databases (as well as special features like the full-text access) are available from the libraries' intranet, but some institutions (e.g. the library of the University of Pécs) offers home access after certain changes in settings. The number of individual subscription for accessing these databases is low due to the high charges, although free access for a certain period of time is offered from time to time. Online

databases can be grouped into multidisciplinary databases, collecting publications from several fields, and specialized databases that offer access to the bibliography of a given field.

## MULTIDISCIPLINARY DATABASES

### 2.6.1. EBSCOhost

One of the most often used online services in the field of health sciences is EBSCOhost, operated by the Boston-based *EBSCO Information Services* of EBSCO Publishing. The company offers bibliographies and electronic journal databases via a standardized search interface. Libraries' access is defined by the license-contract with EBSCO Publishing. It should be kept in mind throughout the literature search that an *embargo* might occur in case of certain journals, meaning that full-texts of the latest articles are available only after a period of time (e.g. 3, 6, 9, 12 months).

#### *Search options*

Before starting the actual literature search in our database (Figure 2/12), it is advised to set the *Languages* used by the researcher (the database is available in Hungarian as well). Most of the literature listed by EBSCOhost is in English, although a limited number of publications in other languages is available as well, which also have their title and bibliographic data in English. The search language of the database is English, with both American and British grammar and spelling available.

The second step is to *select the database*. Currently the University of Pécs has access to the following databases, where there is a Hungarian summary available, to help the researchers select wisely.

- *Academic Search Complete* – a multidisciplinary database primarily for higher education- and research institutions.
- *Business Source Premier* – a research database most often used in the field of business.
- *EconLit* – economics database.
- *Eric* – pedagogy- and education sciences database.
- *MasterFILE Premier* – general database.
- *Health Source-Consumer Edition* – offers consumer health information from the field of medicine, dietetics, paediatrics, sports medicine and general epidemiology.

- *Health Source Nursing/Academic Edition* – medical database that also offers the *Lexi-PAL Drug Guide* covering the patient information materials of various generic medicines.
- *MEDLINE* – medical database.
- *Regional Business News* – full-text database of regional business studies.
- *Library, Information Science & Technology Abstracts* – database of the field of library- and information studies.
- *GreenFILE* – database covering topics of environmental protection, global warming, alternative fuels, etc.
- *Newspaper Source Plus* – a database that synthesizes articles from more than 700 journals and magazines.

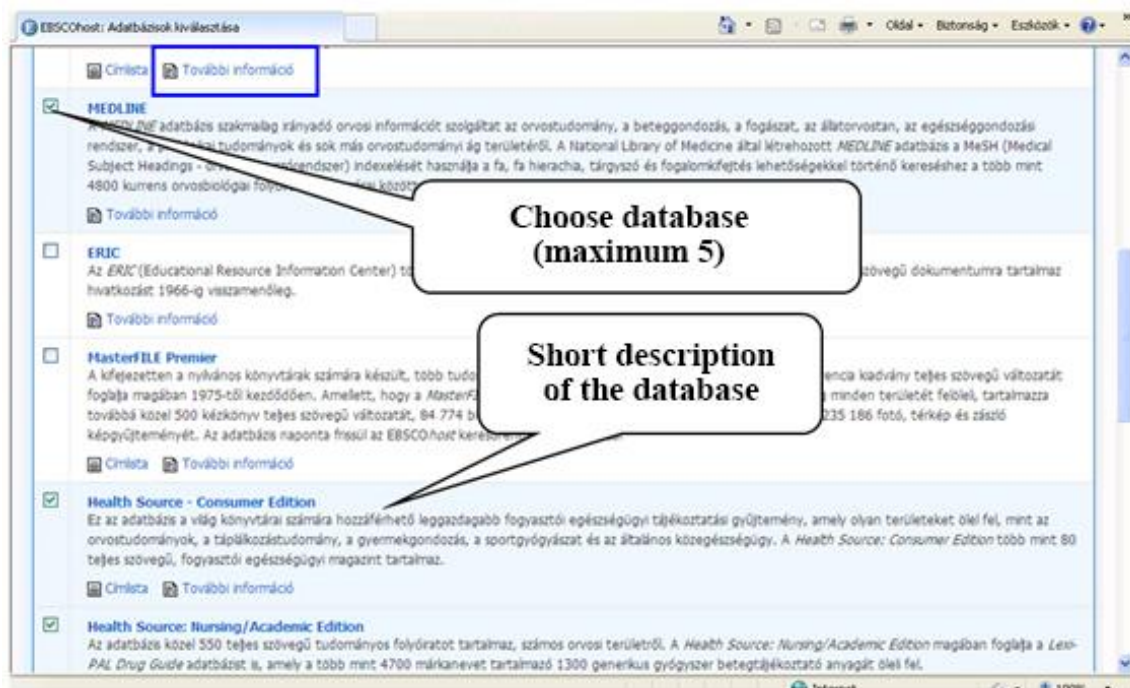
When clicking the titles list a detailed description of the journal-list and publication-data of the specific database will come up on your screen (Figure 2/13).

A *search* can be launched from one or more databases. In case the researcher wishes to review only one database, he should click on its name directly and then click on the word *Continue*. In case of wishing to cover several databases, all of them should be indicated (ticked) before clicking *Continue*.



**Figure 2/12. Accessing the EBSCOhost database**





**Figure 2/13. Browsing in databases**

After accessing the required site the search term should be inserted into the search box, and then the user should click either on *basic search* or on *advanced search*. The latter provides several search boxes where you can search by full text, author's name, title, key words, journal title, etc.

Right under the search box the service providers offer the opportunity to widen or narrow the results. The *expanders* are the following (Figure 2/14):

- *Boolean/Phrase*: Search carried out with the help of Boole-operators.
- *Find all my search terms*: The search automatically uses the AND logic operator, showing only those results that include all the search terms defined by the user simultaneously.
- *Find any of my search terms*: The search automatically uses the OR logic operator, showing those results that include at least one of the search terms defined by the user.
- *Smart text search*: This option allows the insertion of longer texts (maximum of 5,000 characters) into the search boxes, producing a compressed key word-based file through which the search within the abstract of journal articles is carried out. It may be used both with expanders and refiners, and it is also advised to be used together with full-text searches.
- *Apply related terms*: It searches not only the search term but also its synonyms and plural forms.

- *Search within the full text of the articles:* In this case the search applies for the full text of the article. It is a useful feature, but if the search terms are too general then there might be a lot of irrelevant items among the results.

***Narrowing options*** that may be used together or one by one are listed below:

- *Full Text:* Only displays results with full-text access.
- *Scholarly (Peer Reviewed) Journals:* Only displays peer-reviewed scientific articles.
- *Publication:* Only displays articles that were published in the journal requested by the user.
- *References Available:* References cited by the article are also accessible parallel with the research, if their text or bibliographic data is available in one of the EBSCO-databases.
- *Published Date from ... to ...:* a publication date or interval may also be defined during the search.
- *Publication Type:* Narrows results by the type of publication.
- *Number Of Pages:* The search focuses only on articles with a specific page number (length) defined by the user.

### ***Displaying results***

The search box with the given search term is visible at the top of the results page. Below this we can see all *Search Results* that fit the indicated search terms. The system highlights the search term (with bold and italic letters) in the appearing results list.

The *Page Options* button can be found above the search results, and clicking on it will provide a settings option for the format of the results. Thus the results list may be read in a *standard-, brief-, title only- or detailed* format. Also, this is the option where you can set how many results you would like to see on one page (*Results per page*).

Clicking the icon next to the publication's title will reveal the document's most important data: *Authors' names, Source, Publication Date, Publication Type, Subject Terms, Abstract and Database*.

The results page provides the opportunity to further specify one's search – more narrowing can be done using the columns on the right and left. The vertical menu on the left side offers more search options, such as *Full Text, References Available, Scholarly (Peer Reviewed) Journals, Publication Date, Subject, Publication, Company, Age, Geography, and Database*. (Figure 2/16)

The expression *PDF Full Text* next to the results indicates if the full-text version of the article is available. Clicking this option immediately opens the full content of the article, with all the figures, pictures and graphs it contains, too.

There is a new option to connect to the Web of Science database by clicking *Find author's articles in Web of Science*, which might provide useful information on further publications of the authors (Figure 2/15).

Organization is also possible in the results folder, by clicking *Share*, then *Add to folder*. The publications listed accordingly will be saved even after exiting EBSCOhost. The required documents may be saved and printed, and there are options for exportation as well (Figure 2/17).

### Access

The EBSCOhost service can be accessed through the following link: <http://search.epnet.com>

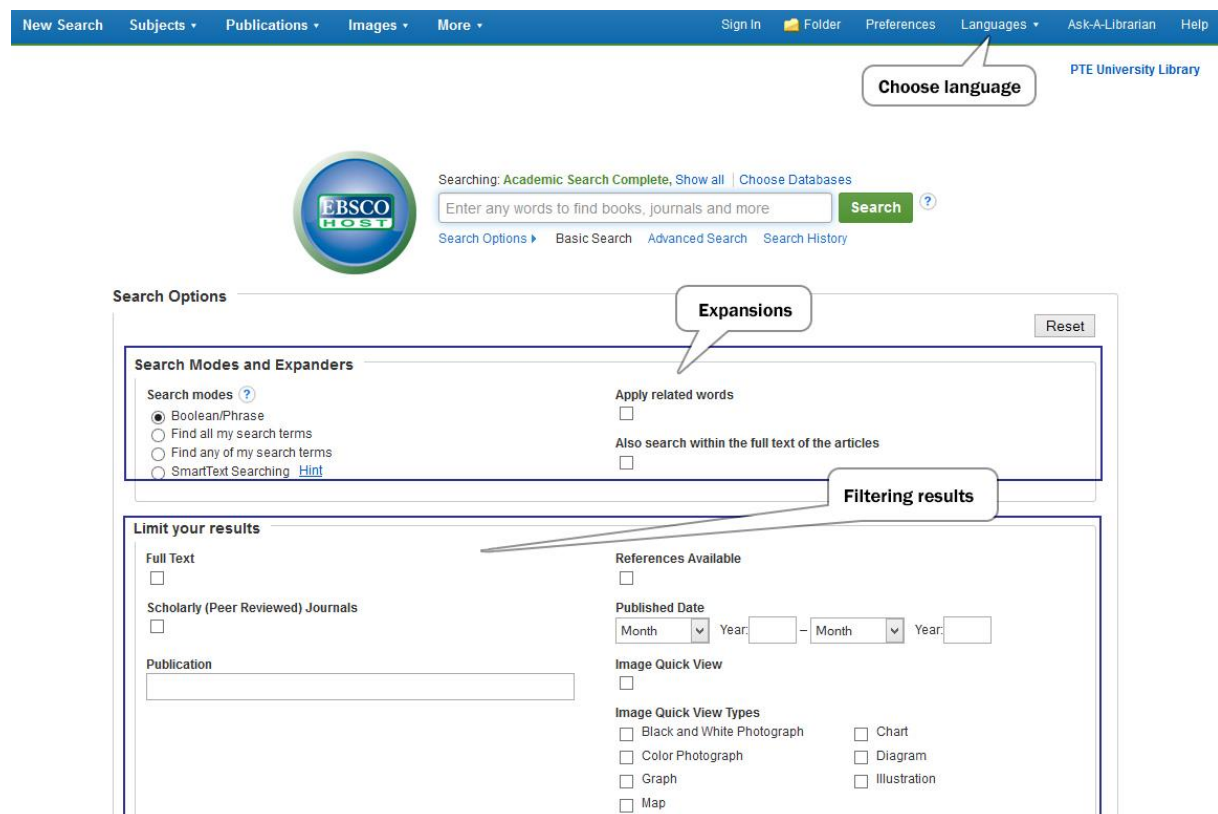


Figure 2/14. Basic search

The screenshot shows the EBSCO database search interface. At the top, there is a navigation bar with options like 'New Search', 'Subjects', 'Publications', 'Images', and 'More'. A search bar contains the term 'stress management' and a 'Search' button. Below the search bar, there are options for 'Basic Search', 'Advanced Search', and 'Search History'. The main content area displays 'Search Results: 1 - 10 of 31,824'. Three search results are visible, each with a title, author information, and a 'Full text' link. Callouts point to the search box, the 'Filtering results' sidebar on the left, and the search results themselves. The sidebar includes sections for 'Refine Results', 'Limit To', and 'Source Types'. The search results include titles like 'SMARTPHONE FOR SELF-MANAGEMENT OF PSYCHOLOGICAL STRESS...' and 'Chronic Stress, Cortisol Dysfunction, and Pain: A Psychoneuroendocrine Rationale for Stress Management in Pain Rehabilitation'.

Figure 2/15. Results list

This screenshot shows the same search results as Figure 2/15, but with a focus on the filtering options. The 'Filtering results' sidebar on the left is expanded, showing various filters such as 'Source Types', 'Subject: Thesaurus Term', 'Subject: Major Heading', 'Subject', 'Publication', 'Company', 'Age', 'Gender', 'Geography', 'NAICS/Industry', and 'Database'. The 'Database' section is checked for 'All Databases'. The main content area shows search results for 'Evaluating the effectiveness of a stress management training on teachers and physicians' stress related outcomes' and 'Stress management in dental students: a systematic review'. Callouts point to the 'Filtering results' sidebar and the search results. The search results include titles, author information, and 'Full text' links. The sidebar also includes a 'Subject: Thesaurus Term' section with a dropdown arrow.

Figure 2/16. Filtering options

The screenshot displays a library search interface. At the top, there are navigation tabs: 'New Search', 'Subjects', 'Publications', 'Images', and 'More'. A search bar contains the text 'stress management'. Below the search bar, there are options for 'Basic Search' and 'Advanced Search'. The main content area shows search results. The first result is a PDF full text document titled 'An Endocrine Hypothesis for the Genesis of Atrial Fibrillation: The Hypothalamic-Pituitary-Adrenal Axis Response to Stress and Glycogen Accumulation in Atrial Tissues.' by Embi, Abraham A.; Scherlag, Benjamin J. The source is 'North American Journal of Medical Sciences', dated 2014. The publication type is 'Academic Journal'. The subjects listed are 'STRESS management; ADRENAL glands; ANATOMY; ATRIAL fibrillation; HYDROCORTISONE; ENDOCRINE diseases; HYPOTHALAMIC-pituitary-adrenal axis; TISSUES -- Wounds & injuries'. The abstract states: 'Background: The underlying role of intracellular glycogen in atrial fibrillation is unknown.' Below the abstract, there are options to 'PDF Full Text (1.9MB)', 'Add to folder', and 'Detailed Record'. The left sidebar contains a 'Refine Results' section with 'Current Search' set to 'stress management'. It also has 'Limit To' options for 'Full Text', 'References Available', and 'Scholarly (Peer Reviewed) Journals'. A 'Publication Date' range is set from 1937 to 2015. The 'Source Types' section is checked for 'All Results' and lists 'Academic Journals (15,670)', 'Journals (13,426)', 'Magazines (4,743)', 'News (2,634)', and 'Trade Publications (1,536)'. The right sidebar shows 'Newsfeeds' and 'Web News' sections with various news items and 'Related Images'.

**Figure 2/17. Details of the result item**

## 2.6.2. MATARKA (Hungarian Periodicals Table of Contents Database)

The Hungarian Periodicals Table of Contents Database (MATARKA) was established in spring of 2002 by the Library, Archives and Museum of the University of Miskolc, Hungary. The freely accessible online service covers the table of content of more than 1500 Hungarian specialized journals published in Hungary as a library consortium, in which the Library of the University of Pécs is also a member of. The coverage of various journals may differ, which means that not all the volumes of a journal are accessible.

### *Search options*

The service offers searches not only in the table of contents of the volumes, but also the by a simple or advanced search in the keywords of the titles.

*Simple search* also provides several options: a maximum of 5 words can be inserted into the search box of the *quick search* option. There is an AND connection between the search terms, providing results only if all terms are found in the specific item, but result may be narrowed down even further. The other option when doing simple search is to fill in four different search boxes: (1) author, (2), co-authors, (3) title – key words and (4) title – fragments. More

than one term may be inserted into these search boxes, connected by an automatic AND connection. However, it is not necessary to fill in all of the above-mentioned search boxes.

Additionally, your search may be narrowed (also simultaneously) by the year of publication, field of research, title of journal, full-text accessibility, or accessibility through the archived articles of the Electronic Periodicals Archive of the National Széchényi Library (EPA). (Figure 2/18)

*Browsing* is also possible by authors and key words listed in the database.

Furthermore, a search may be carried out among journals as well: they are listed in an alphabetical order and a search can be launched by giving the initial letter or the scientific field you wish to explore. A summarizing table of the journals will come up on the screen after clicking on the initial letter. Clicking the on the required journal will display the number of covered volumes, years, and online access link; and the table of content will be displayed after clicking the number of the publication. The following data is available on the journal you search: title data, publisher, first (and last) year of publication, ISSN number, research field, full-text accessibility, website and language.

### ***Display of results***

Resulting records are displayed by their year of publication, starting with the latest documents. Name(s) of author(s), title and other data (title of the journal, year of publication, volume and number, page number) of the publication are also shown (Figure 2/19). The full text is available for approximately 15% of the articles, while some of them may be accessed only for a certain fee.

Other supplementary information is also available on the website: clicking the *help* icon displays the list of previously completed documents and publications along the list of names of the collectors (participating libraries). Opinions and advice about the service may be given by clicking *visitors' book*, while clicking *connect* will provide help concerning any questions or accessibility issues. The menu *statistics* offers statistical data on the database itself, such as number of visitors, and basic documents of the MATARKA Association.

### ***Access:***

The MATARKA database is accessible through the following link: <http://www.matarka.hu>

MATARKA - Hungarian Periodicals Table of Contents Database

Search Partners Statistics Help Guest book Contact

Simple search Advanced search Browse Periodicals Shopping cart Registration Login

**Initial search** (search in title and author together) **Quick search**

**Search** (several fields may be searched together; truncate with % or \*)

**Author:** Ács Pongrác **Search**

**Further author:**

**Author - keywords:** [Help](#)

**Title - keywords:** [Help](#)

**Title - fragment:** [Help](#)

**Limit your search**

from 1800 to 2015

Select a subject

Only full text articles

Only articles from [EPA](#)(Electronic Periodical Archives and Database)

**You can choose more than one journal by holding CTRL and clicking on the journal title**

Select a journal

- 2000 : irodalmi és társadalmi havi lap
- 4D : tájépítészeti és kertművészeti folyóirat
- Abstracta botanica
- Across languages and cultures
- Acta : A Csiki Székely Múzeum és a Székely

**Filtering results**

Sponsored by SZÉCHENYI TERV and NEMZETI KULTURÁLIS ÖRÖKSÉG MINISZTERIUMA and OM and nka

best viewed in 1024x768 with Mozilla and JavaScript enabled

Figure 2/18. Simple search with multiple authors

MATARKA - Hungarian Periodicals Table of Contents Database

Search Partners Statistics Help Guest book Contact

Simple search Advanced search Browse Periodicals Shopping cart Registration Login

8 hit(s) for **ács pongrác**  
8 hit(s) for the combination

Select all Deselect choices

Add to shopping cart You can use the shopping cart to list, save or order article copies.

Page of hits:

- Authors: Melczer Csaba - Melczer László - Szabados Sándor - Ács Pongrác*  
Szívelégtelen betegek életminőségét mérő validált kérdőívek összehasonlító vizsgálata = Comparative examination of validated questionnaires evaluating the quality of life in heart failure patients.  
[\[Teljes szöveg \(PDF\)\]](#)  
*Egészség-Akadémia*, 2012. (3. évf.) 1. sz.
- Authors: Ács Pongrác - Márkus Gábor - Oláh István*  
Út a mecenatúrától a sportszponzoráció felé, avagy a sporttámogatások egy új korszakának kezdetén  
*Marketing & Management*, 2012. (46. évf.) 4. sz. 14-25. p.
- Authors: Stocker Miklós - Ács Pongrác*  
A sportolás növelésével elérhető gazdasági haszon mértéke  
[\[Teljes szöveg \(PDF\)\]](#)  
*Magyar sporttudományi szemle*, 2012. (13. évf.) 3. (51.) sz. 20-26. p.
- Authors: Ács Pongrác - Hécz Roland - Paár Dávid - Stocker Miklós*  
A fitness (m)értéke : A fizikai inaktivitás nemzetgazdasági terhei Magyarországon  
[\[Teljes szöveg \(PDF\)\]](#)  
*Közgazdasági szemle*, 2011. (58. évf.) 7-8. sz. 689-708. p.  
Teljes szöveg (1995/1 számtól): [Elektronikus Periodika Archivum](#)
- Authors: Ács Pongrác*  
A sportolók területi mozgásai, avagy a sportolói vándorlás  
[\[Teljes szöveg \(PDF\)\]](#)

**Figure 2/19. Results list**

### 2.6.3. MOKKA (Hungarian National Common Catalogue)

The Hungarian National Common Catalogue is a bibliographic and location-identifier database offering search options in Hungary's catalogues of national, scientific, higher-educational and county- and city-based libraries.

#### *Search options*

Both a *simple* and an *advanced* search can be carried out by author, title, key words or any type of subject heading. The process can be started by clicking the *Search* button after inserting the search term into the search box (Figure 2/20).



## Display of results

The resulting items show which library hosts the required document and the search can be continued in the database of that specific library, providing information on whether it is available for loan or not (Figure 2/21).

## Access

MOKKA can be accessed through the following website: <http://www.mokka.hu>

The screenshot shows the MOKKA advanced search interface. The search criteria are:

- Field 1: sport, as Author
- Field 2: strenght, as Title
- Field 3: (empty), as Topic

The search is configured with the 'And' operator. The 'Filter your search!' section includes options for Date of publication, Place of publication, Language, and Document type (Monograph, Periodical, Notated music, Cartographic material, Computer file, Musical sound recording, Moving image, projected medium, Manuscript, language material). A tag cloud is visible at the bottom.

Figure 2/20. Advanced search

The screenshot displays the Mokka search interface. At the top, there is a logo for 'mokka Magyar Országos Közös Katalógus' and navigation links for 'MOKKA-ODR catalogue', 'Info Portal', and 'MOKKA Association'. A search bar contains the query 'author = sport' and shows '918 hits. 1-10. shown.'. Below the search bar are navigation buttons for 'Empty Basket', 'All to basket(918)', 'View basket(0)', and 'Back to the search'. The results are displayed in a table with columns: Number, Names, Main title, Year of publication, Document type, and Show. The first two results are by Abád Józsefné and Abádné Hauzer Henriette, both from 1977. The third result is 'The 26th World Table Tennis Championships' from 1961. The fourth result is '30th Stress and Anxiety Research Society Conference' from 2009. On the right side, there are active filters for 'Publish date' and 'Language'. The 'Publish date' filter shows counts for years from 1989 to 1999. The 'Language' filter shows counts for Hungarian (723), German (99), English (94), French (13), Multilingual (6), Italian (4), slv (4), ukr (4), and maa (2).

**Figure 2/21. Results list**

#### 2.6.4. OVID

Access to electronic information for experts of various fields (medicine, social science, engineering and human sciences) is offered by Ovid's classic content provider. It is important to mention that this provider has published a CD-ROM version of the MEDLINE database for the first time. The information service has three pillars here: databases, electronic journals and electronic books. Besides the MEDLINE database only the journals of the Wolters Kluwer publisher are available.

##### *Search options*

An online search can be carried out from the database, books or journals. If the user chooses to *search by database*, then the first step is to choose from the database list. One search offers the choice of maximum five databases. The search process can be started by clicking the *Select Resource(s)* button on the bottom left side of the page.

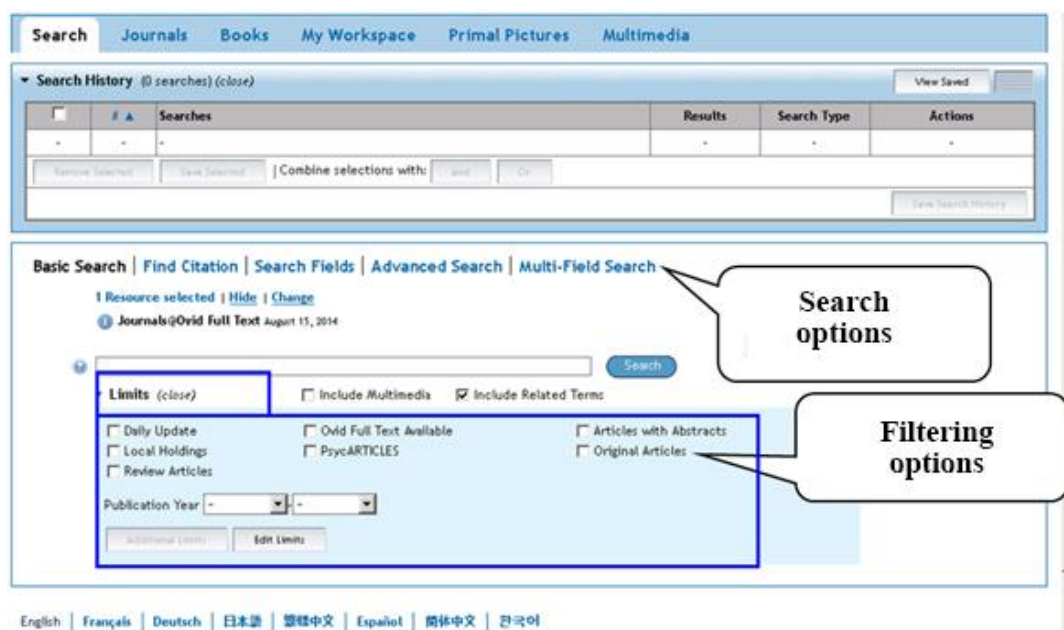
There is also an option to search other databases, by clicking *Change*, which navigates the user back to the list of databases. Ovid offers six types of searches: *Basic Search*, *Find Citation*, *Search Tools*, *Search Fields*, *Advanced Ovid Search*, *Multi-Field Search* (Table 2/1)

**Table 2/1. Search options in Ovid**

Type of search	Description
<b>Basic Search</b>	Search in a natural language, where it is not necessary to know the syntax of the search and the special terminology of the field. The results are ranked by the search terms provided, but relevant results are not significant. The spell check and search functions can be applied for relevant terms.
<b>Find Citation</b>	This type of search is useful to find a <i>specific publication</i> if we know some of the item's bibliographic data, e.g. journal, title, or author of the publication.
<b>Search Tools</b>	They help to define the appropriate <i>subject heading fitting the given key words</i> , and also offers a search option. These options are part of "Advanced Ovid Search".
<b>Search Fields</b>	This option helps searching by various aspects, such as name of institution, DOI-number, ISSN-number, publisher, country, etc.
<b>Advanced Ovid Search</b>	It is a regular advanced search, which offers a higher number of search options, such as by key word, author, title or journal. If there is only one database requested, then Ovid offers a search by subject heading as well.
<b>Multi-Field Search</b>	This option offers a quicker search by combining several search fields, so publications fitting multiple search requirements can be listed as results.

Source: Csajbók 2009

The search term defined by the user should be inserted into the search box during basic search (Figure 2/22). An advanced search may be carried out by *Keywords, Author, Title, or Journal*. List of results may be narrowed by date or document type by clicking *Limits*, providing the resulting *Abstracts, Full Texts or Results*.



**Figure 2/22. Simple search**

**Search** Journals Books My Workspace Primal Pictures Multimedia

▼ Search History (0 searches) (close) View Saved

<input type="checkbox"/>	# ▲	Searches	Results	Search Type	Actions
-	-	-	-	-	-

| Combine selections with:

[Basic Search](#) | [Find Citation](#) | [Search Fields](#) | [Advanced Search](#) | [Multi-Field Search](#)

1 Resource selected | [Hide](#) | [Change](#)

📘 Journals@Ovid Full Text August 15, 2014

Enter keyword or phrase (\* or \$ for truncation)

Keyword  Author  Title  Journal

► Limits (expand)  Include Multimedia

**Figure 2/23. Advanced search**

### ***Display of results***

Resulting data provide information on the following: title, name(s) of author(s), title of the journal, bibliographic data of the publication (year of publication, number of volume, issue and pages); and the abstract is accessible as well. If it is authorized by the provider, the article may be downloaded in a pdf format or may also be saved (Figure 2/24).

*Complete Reference* may be found in the right column of the results page, providing further information on the given record: not only the bibliographic data, but also the institution of the authors, MeSH key words of the article, a short summary, and a link to the article in case of free access (Figure 2/25). Relevant data can be printed, saved or exported in word or pdf format.

An interesting feature is that Ovid opens the access of a different database every month, and will become available upon registration. You can register for the database by clicking *Resource of the Month*, and for journals by clicking *Journal of the Month*.

### **Primal Pictures**

Apart from the search option, an exhaustive, dynamic interactive multimedia is available on the anatomy of the human body through *Primal Pictures*. These 3-D animations help understanding the topic, accompanied by presentations of the biomechanical and surgical processes (Figure 2/26).

### ***Access***

<http://www.ovid.com/site/index.jsp>

Results Tools Options

All  Print Email Export Add to My Projects Keep Selected

Search Information

Clear Selected View: Title Citation Abstract 10 Per Page 1 GO Next

**You searched:**  
 stress management.mp.  
 [mp=title, abstract, full text, caption text]  
 - Search terms used:  
 management  
 stress

**Search Returned:**  
 9869 text results

**Sort By:**

Customize Display

**Filter By**

Add to Search History

+ Selected Only ( 0 )

- Years  
 All Years  
 Current year  
 Past 3 years  
 Past 5 years  
 Specific Year Range

1.  **The Relationship of Chronic and Momentary Work Stress to Cardiac Reactivity in Female Managers: Feasibility of a Smart Phone-Assisted Assessment System.**  
 Lumley, Mark A. PhD; Shi, Weisong PhD; Wiholm, Clairy PhD; Slatcher, Richard B. PhD; Sandmark, Helene PhD; Wang, Shinan MS; Hytter, Anders PhD; Arnetz, Bengt B. MD, PhD  
*Psychosomatic Medicine.*  
 [Original Article: PDF Only]  
 AN: 00006842-900000000-99173.  
 B 13 Raktari jelzet: P 59 (Psychosom Med); USA, New York, ISSN 0033-3174; 1960:22 -  
**Status**  
 Publish Ahead of Print, POST AUTHOR CORRECTIONS, 30 July 2014  
 View Abstract

PDF (341KB) + My Projects

- Ovid Full Text
- Table of Contents
- Abstract Reference
- Complete Reference
- Find Similar
- Find Citing Articles
- Library Holdings
- Request Permissions
- Internet Resources

2.  **Web-Based Parenting Skills Program for Pediatric Traumatic Brain Injury Reduces Psychological Distress Among Lower-Income Parents.**  
 Raj, Stacey P. MA; Antonini, Tanya N. MA; Oberjohn, Karen S. MA; Cassidy, Amy PhD; Makoroff, Kathi L. MD; Wade, Shari L. PhD  
*Journal of Head Trauma Rehabilitation.*  
 [Original Article: PDF Only]  
 AN: 00001199-900000000-99766.  
**Status**

- Table of Contents
- Abstract Reference
- Complete Reference
- Find Similar
- Find Citing Articles
- Library Holdings

Figure 2/24. Results list

1  GO Search Results | Next

1.

**Accession Number** 00006842-900000000-99173.

**Author** Lumley, Mark A. PhD; Shi, Weisong PhD; Wiholm, Clairy PhD; Slatcher, Richard B. PhD; Sandmark, Helene PhD; Wang, Shinan MS; Hytter, Anders PhD; Arnetz, Bengt B. MD, PhD

**Institution** From the Departments of Psychology (M.A.L., R.B.S.), Computer Science (W.S., S.W.), and Family Medicine and Public Health Sciences, Cardiovascular Research Institute and Institute of Environmental Health Sciences (B.B.A.), Wayne State University, Detroit, Michigan; Department of Public Health and Caring Sciences (C.W., B.B.A.), Uppsala University, Uppsala, Sweden; Department of Public Health Sciences (H.S.), Malardalen University, Vasteras, Sweden; and School of Business and Economics (A.H.), Linnaeus University, Vaxjo, Sweden

**Title** **The Relationship of Chronic and Momentary Work Stress to Cardiac Reactivity in Female Managers: Feasibility of a Smart Phone-Assisted Assessment System.[Article]**

**Source** Psychosomatic Medicine.

**Status** Publish Ahead of Print, POST AUTHOR CORRECTIONS, 30 July 2014

**Local Message** B 13 Raktari jelzet: P 59 (Psychosom Med); USA, New York, ISSN 0033-3174; 1960:22 -

**Abstract**  
 Objectives: To evaluate a wireless smart phone-assisted (SPA) system that assesses ongoing heart rate (HR) and HR-triggered participant reports of momentary **stress** when HR is elevated during daily life. This SPA system was used to determine the independent and interactive roles of chronic and momentary work **stress** on HR reactivity among female managers.  
 Methods: A sample of 40 female managers reported their chronic work **stress** and wore the SPA system during a regular workday. They provided multiple reports of their momentary **stress**, both when triggered by increased HR and at random times. Relationships among chronic **stress**, momentary **stress**, and HR were analyzed with hierarchical linear modeling.  
 Results: Both chronic work **stress** (b = 0.08, standard error [SE] = 0.03, p = .003) and momentary work **stress** (b = 1.25, SE = 0.62, p = .052) independently predicted greater HR reactivity, adjusting for baseline

- Ovid Full Text
- Table of Contents
- Abstract Reference
- Find Similar
- Find Citing Articles
- Library Holdings
- Request Permissions
- Internet Resources

Figure 2/25. Information on the chosen record



**Figure 2/26. Primal Pictures**

### 2.6.5. ScienceDirect

Science Direct, a database operated by the scientific publishing group *Elsevier* operating since 1997, offers full-text online access to approximately 2,500 paper-based journals from the field of medicine, natural sciences, social sciences and engineering, and it offers access to journals of other publishers and indexes for 26,000 books as well. For articles published in Elsevier journals before 1995, however, only bibliographic data is available. The service offers search options in bibliographic data, and provides access to databases and full texts simultaneously. An advantage of ScienceDirect is that it offers access to some journal articles before they would appear in print.

#### *Search options*

ScienceDirect offers a really wide set of services: the search may be basic, advanced (*Advanced search*) or expert (*Expert search*), alongside the option of browsing (*Browse*). During basic search, any type of information may be inserted into the search box on the opening page, although these might be further refined by indicating the *author's name*, *journal or book title*, *publication data*, *volume*, *issue* and *page*. The search process can be started by clicking *Submit Quick Search* (Figure 2/27).

The *advanced search* option is based on document types (journal, book, reference works or images). Each type of document has different search forms with different data fields and

search interface. Bool-operators and characters of incomplete searches may be used during work, and words in quotation marks will be searched word-for-word.

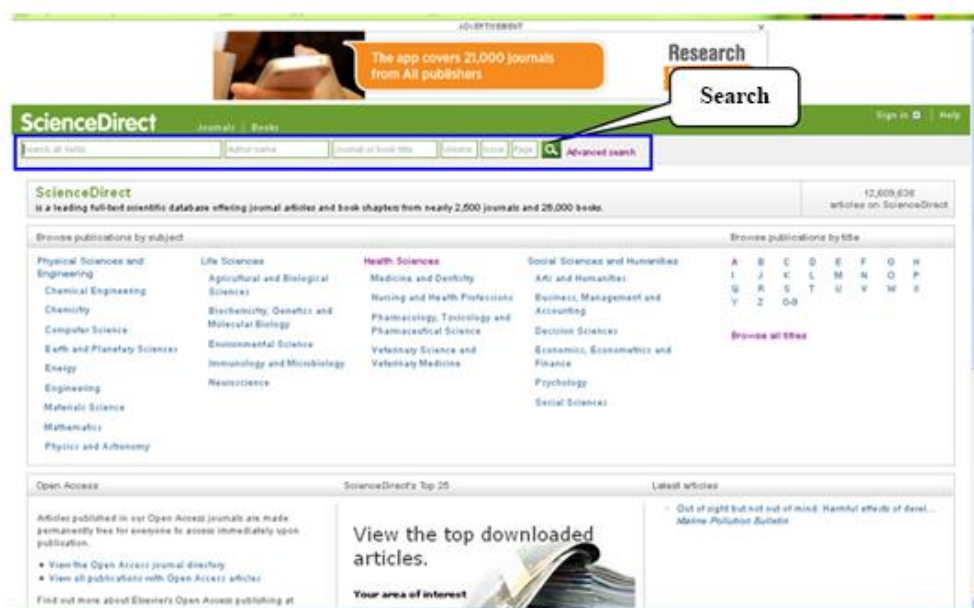
An alphabetical list of all journals and books covered by ScienceDirect can be accessed by clicking *Browse publications by title* on the homepage of the database. Titles can be viewed by clicking the appropriate initial letter, and the list can be further narrowed. The scientific field can be chosen from the list on the left hand side, and by clicking *All publications*, the user can choose from a range of document types such as *All journals*, *Books*, *Book Series*, *Handbooks*, or *Reference Works*. Articles are grouped by access and divided into categories such as those that require *subscription*, those that are *open access*, and the ones that only *contain* some sort of *open access*.

### ***Display of results***

The following bibliographic data are displayed by each resulting record: title and type of publication, name of the journal, publication data and author(s)'s name(s). The box with green lines next to the record indicates whether the full-text of the article is available, and a white box indicates that no subscription is available, so only the abstract is available. Records can be further narrowed down by date and title of the publication, topic, and content type. Full texts may be accessed in a pdf format, and also may be saved and printed (Figure 2/28).

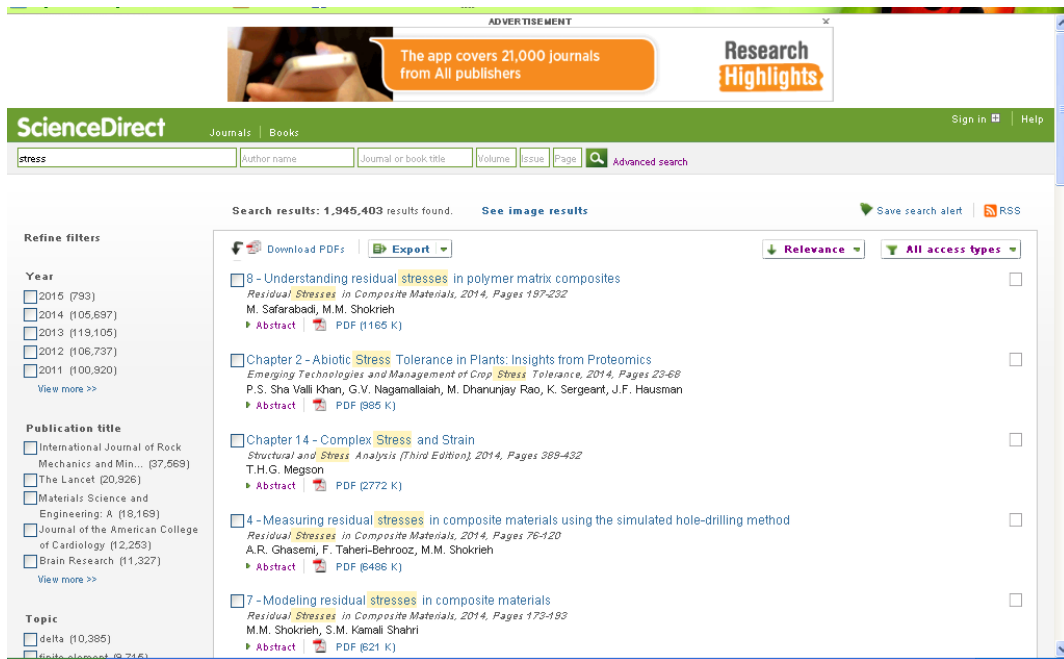
### ***Access:***

ScienceDirect can be accessed via the following link: <http://www.sciencedirect.com>



**Figure 2/27. Search interface at ScienceDirect**





**Figure 2/28. Results list**

### 2.6.6. Scopus

The Scopus multidisciplinary database has been operated by Elsevier publishing since 2004, which is known to be the largest search engine for abstracts and references. It covers documents of the following disciplines: life- and medical sciences, chemistry, physics, mathematics, engineering, social sciences, psychology, economics, biology, agronomy, environmental protection and general sciences. References can be viewed going back to as early as 1996. The daily updated database does not only cover scientific works published in English, but also those in French, German, Spanish and Chinese.

#### *Search options*

Scopus offers a wide range of search options: *document search*, *author search*, *affiliation search* and *advanced search*, which can be carried out on the opening page, all of which consist of search boxes and search interfaces (Figure 2/29).

At the *Document Search* form the most relevant search terms should be inserted into the search box. Search terms should be connected by Boole-operators. Searches may be further detailed by time span, document type and scientific field. When searching for an author, the family name and the first name initials should be inserted into the data box, then the search process can be started by clicking *Search*.

### ***Display of results***

The results page of *Document Search* displays the group of *Document Results* that indicates the number of publications covering the given topic. In case of 100 or more results it is strongly advised to narrow down one's search by focusing on specific aspects (*Refine results*) of *Year*, *Author Name*, *Subject Area*, *Document Type*, *Source Title*, *Keyword*, *Affiliation*, *Country*, *Source Type*, and *Language*. The resulting record contains the title of the publication, name(s) of the author(s), year of publication, other data of the publication, topic and number of citations. The list can be arranged by date, number of citations, order of relevance and alphabetical order of the author's name or the title of publication (Figure 2/30). Citations data can be accessed by clicking *citation overview*. On the next page a table displays the yearly citation number of the given article going back to as early as 1996. This information can also be arranged by date or in an ascending or descending order. The author may discard citations himself by clicking on *Exclude from citation overview*.

When searching by author the resulting list displays the name and name-alterations of the author, as well as the number of publications. Clicking on the name of the chosen author will reveal his summarized publication index indicating the number of publications and citations, the Hirsch-index, the co-authors, the covered topics and the number of references available. Furthermore, there are graphs and summarized tables helping to visualize the number of publications of the author (*View Author Evaluator*, *View citation overview*, *View h-Graph*). The *Affiliation Search* offers an insight into the publication data of a given institution, with the possibility of arranging results by *date*, *citation* and *relevance* (Figures 2/31 and 2/32).

### **Note**

We would like to draw the reader's attention to the fact that the Hirsch-index of the same author listed in Scopus and in Web of Science may differ because each database calculates the summarized publication list of the given author on the basis of on its own data.

### **Access**

Scopus can be accessed via the following link: <http://www.scopus.com>.

**Document search** | Author search | Affiliation search | Advanced search | Browse Sources | Analyze Journals

Search for... *Eg., "heart attack" AND stress* **Article Title, Abstract, Keywords**

**Limit to:**

**Date Range (inclusive)**  
 Published **All years** to **Present**  
 Added to Scopus in the last **7** days

**Document Type**  
**ALL**

**Subject Areas**  
 Life Sciences (> 4,300 titles.)  
 Health Sciences (> 6,800 titles. 100% Medline coverage)  
 Physical Sciences (> 7,200 titles.)  
 Social Sciences & Humanities (> 5,300 titles.)

**Resources**  
 Follow @Scopus on Twitter for updates, news and more  
 Access training videos  
 Learn about alerts and registration

**About Scopus**  
 What is Scopus  
 Content coverage

**Language**  
 日本語に切り替える  
 切换到简体中文

**Customer Service**  
 Help and Contact  
 Live Chat

**About Elsevier**  
 Terms and Conditions  
 Privacy Policy

Copyright © 2014 Elsevier B.V. All rights reserved. Scopus® is a registered trademark of Elsevier B.V. Cookies are set by this site. To decline them or learn more, visit our Cookies page.

Figure 2/29. Search interface

**Scopus** | Register | Login

Search | Alerts | My list | Settings | Contact | Tutorials

Save | Set alert | Set feed

View 5 patent results | Analyze results | Sort on: Date Cited by Relevance

Search

**Filtering**

**Refine**

**Year**  
 2014 (2)  
 2012 (1)  
 2010 (2)  
 2009 (3)  
 2007 (6)

**Author Name**  
 Olah, A. (24)  
 Betelehem, J. (12)  
 Jozsa, R. (11)  
 Háberg, F. (9)  
 Cornelissen, G. (9)

**Subject Area**  
 Medicine (9)  
 Pharmacology, Toxicology and Pharmacetics (8)

**Results list**

<input type="checkbox"/> Economic burden of long-term care of rheumatoid arthritis patients in Hungary	Honvith, Z., Sebestyén, A., Ósterle, A., (.), Bagosi, G., Boncz, I.	2014 The European Journal of Health Economics	0
<input type="checkbox"/> Az otthoni szakápolás egészségbiztosítási vonatkozásainak elemzése Magyarországon [Home nursing care in Hungary]	Cs. Honvith, Z., Sebestyén, A., Molics, B., (.), Bagosi, G., Boncz, I.	2014 Orvosi Hetilap	0
<input type="checkbox"/> Is 'meaningfulness' a general mediating factor? The salutogenic revolution of question-setting in health science and occupational psychology	Varga, K., Toth, Á., Rozsár, J., (.), Betelehem, J., Jeges, S.	2012 European Journal of Mental Health	0
<input type="checkbox"/> A kivonuló mentődolgozók egészségi állapotát befolyásoló főbb tényezők hazánkban [Major contributing factors of self perceived health in Hungarian ambulance personnel]	Betelehem, J., Honvith, A., Gondócs, Z., (.), Boncz, I., Oláh, A.	2010 Orvosi Hetilap	2
<input type="checkbox"/> Prehospital emergency care in Hungary: What can we learn from the past?	Gondócs, Z., Olah, A., Marton-Simora, J., (.), Schaefer, J., Betelehem, J.	2010 Journal of Emergency Medicine 39 (4), pp. 512-518	1 Cited by
<input type="checkbox"/> Stress, geographic disturbance, isolation and circadian	Olah, A., Jozsa, R.	2008 Neurobiology Research	1

Figure 30. Results list

**Scopus** Register | Login

Search | Alerts | My list | Settings Live Chat | Help and Contact | Tutorials

Author last name "Oláh", Author first name "Andras" [Edit](#)

3 author results [About Scopus Author Identifier](#) Sort on: Document Count | Author (A-Z) ...

Show exact matches only  Show documents  View citation overview  Request to merge authors

**Refine**

**Source Title**

- 3rd International Symposium on Wireless Pervasive Computing ISWPC 2008 Proceedings (1)
- American Journal of Public Health (1)
- Annales Des Telecommunications Annals of Telecommunications (1)
- Annals of the New York Academy of Sciences (1)
- Applied Ecology and Environmental Research (1)

**Affiliation**

- Budapesti Corvinus Egyetem (1)
- Budapesti Muszaki ES (1)

<input type="checkbox"/> <b>Oláh, András</b>	1 Oláh, András Oláh, András Oláh, A.	24 Medicine ; Pharmacology, Pecs Tudományegyetem Pecs Toxicology and Pharmaceutics ; Nursing; ...	Pecs	Hungary
<input type="checkbox"/> <b>Oláh, András L.</b>	2 Oláh, András Oláh, András Oláh, A.	11 Computer Science ; Engineering ; Energy, ...	Pamany Peter Katolikus Egyetem	Budapest Hungary
<input type="checkbox"/> <b>Oláh, Andras Bela</b>	3 Oláh, Andras Bela Oláh, A. B.	3 Agricultural and Biological Sciences ; Engineering	Budapesti Corvinus Egyetem	Budapest Hungary

Display  results per page < Page 1 >

**Figure 2/31. Choosing an author from the results list**

**Scopus** Register | Login

Search | Alerts | My list | Settings Live Chat | Help and Contact | Tutorials

**Oláh, András** [About Scopus Author Identifier](#) [View potential author matches](#)

Pecs Tudományegyetem, Faculty of Health Sciences, Pecs, Hungary  
Author ID: 9243195400

Other name formats: Oláh, András  
Oláh, András  
Oláh, András  
[View More](#)

**Documents:** 24 [View Author Evaluator](#)

**Citations:** 132 total citations by 102 documents [View citation overview](#)

**h Index:** 7 The h Index considers Scopus articles published after 1995. [View h-Graph](#)

**Co-authors:** 109

**Subject area:** Medicine , Pharmacology, Toxicology and Pharmaceutics [View More](#)

**24 Documents** | Cited by 102 documents since 1996 | 109 co-authors

24 documents [View in search results format](#) Sort on: Date Cited by ...

[Export all](#) | [Add all to my list](#) | [Set document alert](#) | [Set document feed](#)

Economic burden of long-term care of rheumatoid arthritis patients in Hungary	Horváth, Z., Sebestyén, A., Ósterle, A., (...), Bagosi, G., Boncz, I.	2014	The European Journal of Health Economics	0
---	---	------	--	---

[View at Publisher](#)

**Follow this Author** Receive emails when this author publishes new articles

[Get citation alerts](#)

[Add to ORCID](#)

[Request author detail corrections](#)

Publication range: 2004 - Present
References: 557
<b>Source history:</b>
Canadian Medical Association Journal <a href="#">View documents</a>
Journal of Advanced Nursing <a href="#">View documents</a>

**Figure 2/32. Summary of the author's scientific work**

### **2.6.7. SpringerLink**

An interdisciplinary database called SpringerLink, enjoying immense popularity with researchers, offers full-text access to the books and journals published by Springer Publishing. A proof of the high scientific quality of the service is that it contains journals with high impact factor and those also published by scientific associations. The database focuses on the fields of medicine, life sciences, chemistry, geography, IT, physics and astronomy, engineering, environmental protection, law, economics and social sciences. There are not only English-language records, but also German, Italian, Spanish and Dutch publications, depending on the topic. Bibliographic data (tables of content and abstracts) can be viewed without subscription, and full text publications can be accessed with subscription or via the library intranet of higher education institutions using the EISZ sub-host.

#### ***Search options***

The modern Google-like interface of the service makes searching an easy job. The user should type the research term(s) into the search bar, and he may also choose to use Boole-operators. When using an exact term, it is advised to insert it using quotations marks. The column on the left side of the results offers a narrowing option (*Refine Your Search*) according to *Content Type*, *Article*, *Chapter*, *Reference Work Entry*, *Protocol* and *Book*, and also *Discipline* according to 24 thematic listing, *Sub-discipline*, source document (*Published In*) and *Language* (Figures 2/33 and 2/34). The journals and books also include and index, bibliography and, in some cases, and author's index as well.

#### ***Display of results***

The resulting hits contain the type and title of the document. For books there is also information available about the title and the year of publication and, in case of articles, the name of the author and the title of the journal is also indicated. There are some documents in the results list that are only available for a fee, but full-texts of books and journals published by the Springer publishing group can be downloaded in a pdf format by clicking *Download PDF* (Figure 2/34 and 2/35).

#### **Access:**

The database of SpringerLink can be accessed via the following link: <http://link.springer.com>

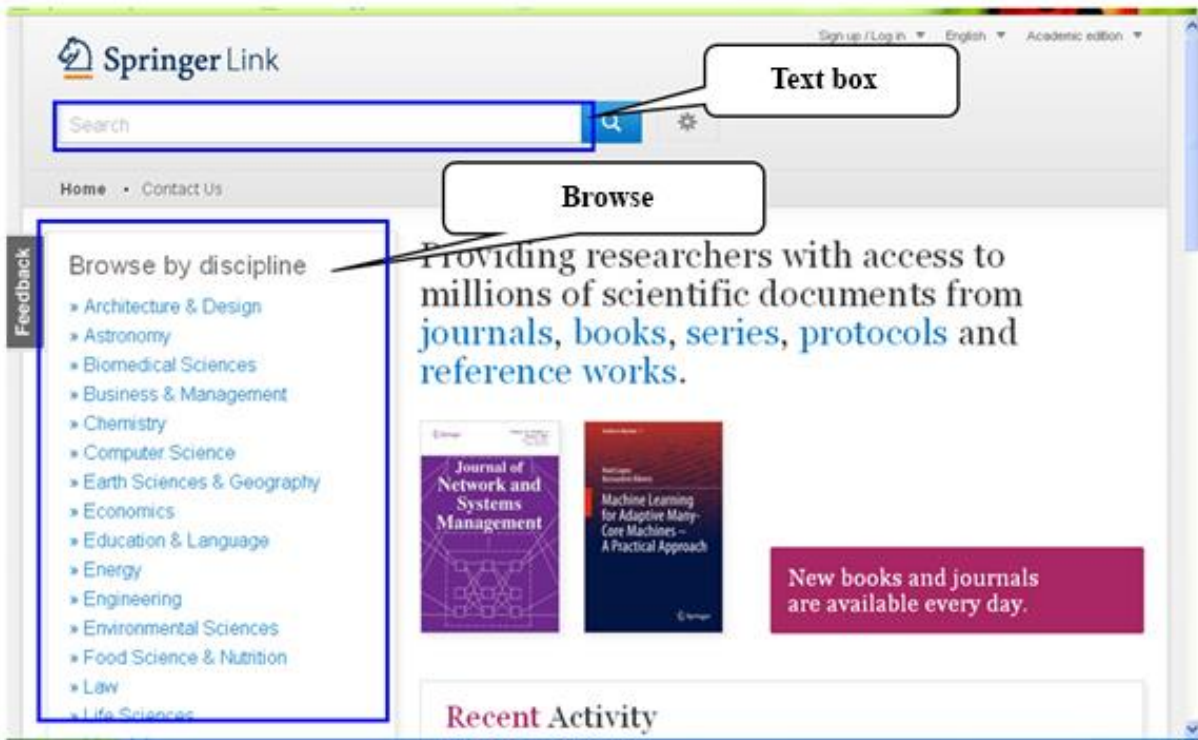


Figure 2/33. Search interface

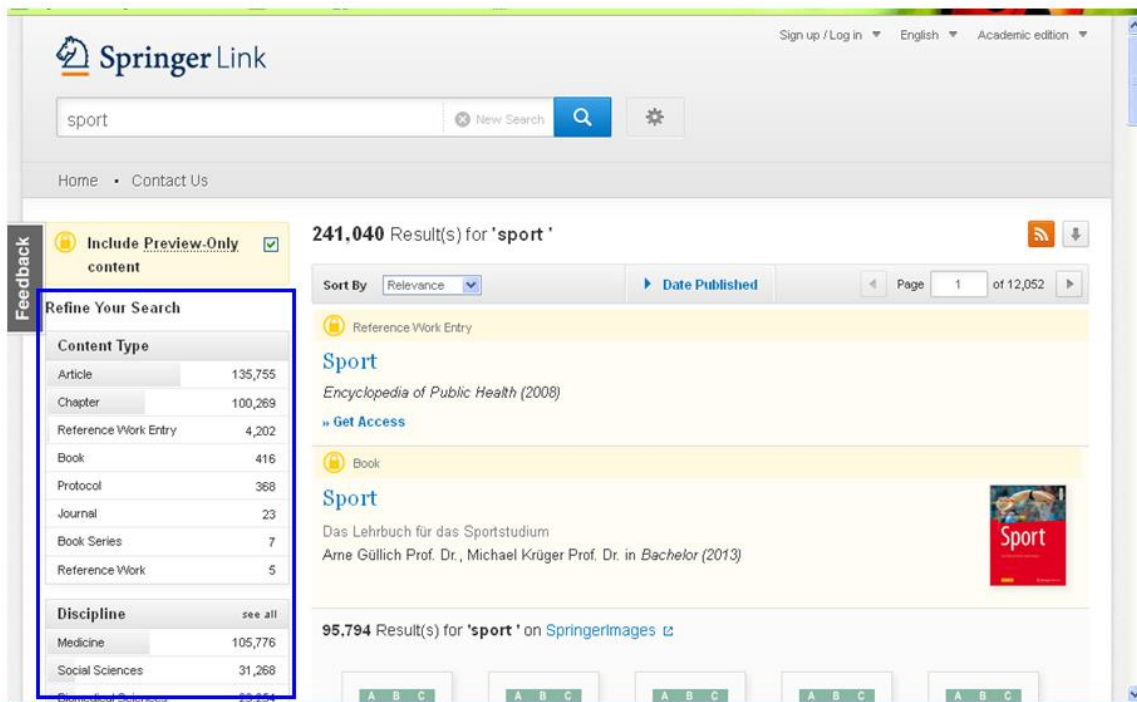


Figure 2/34. Results list

The screenshot shows a search results page for the keyword 'sport'. At the top, it indicates '135,755 Result(s) for 'sport' within Article'. The page is filtered to show 'Article' content only. On the left, there is a 'Refine Your Search' sidebar with the following filters:

- Content Type:** Article (selected)
- Discipline:**
  - Medicine: 71,965
  - Biomedical Sciences: 23,786
  - Life Sciences: 19,103
  - Social Sciences: 12,493
  - Psychology: 10,280
- Subdiscipline:**
  - Orthopedics: 23,731
  - Internal: 21,403
  - Surgery: 10,687
  - Human Physiology: 9,210
  - Rheumatology: 5,758
- Published In:**
  - Knee Surgery, Sports Traumatology, Arthroscopy: 4,238
  - European Journal of Applied Physiology: 3,856

The main results area shows three articles:

- Sport voor de huisartsboekbespreking huisarts sport**  
Dit is een informatief en goed leesbaar boekje voor huisartsen die tijdens het spreekuur niet goed raad weten met problemen en vragen over **sport**. Volgens de auteurs is het boekje bedoeld voor de in **sport** geïntere...  
Sjoerd Hobma in *Huisarts en Wetenschap* (2002)  
[» Get Access](#)
- Der neue Range Rover Sport: Technische Daten**  
ATZextra (2013)  
[» Get Access](#)
- Josef Hackforth an der TU München Ordinarius für Sport, Medien und Kommunikation**  
Michael Schaffrath in *Publizistik* (2001)  
[» Download PDF](#) (215 KB)

Below these, a fourth article is partially visible: **Bericht vom SFMS (Société Française de Médecine du Sport)-**

**Figure 2/35. Refined results list**

## 2.7. Specialized databases

### 2.7.1. MEDLINE

Medline is one of the most well-known, largest, and professionally acclaimed medical bibliographic databases produced by the National Library of Medicine. It covers more than 4,800 scientific journals from the fields of medicine, nursing, dentistry, veterinary medicine, allied health and pre-clinical sciences. The database also has 13 Hungarian journals on its index, and for four of them full texts are available (*Acta Veterinaria Hungarica*, *Hungarian Oncology*, *Hungarian Medical Journal*, and *Pathology Oncology Research*). The database has covered scientific publications since 1965, and currently offers full-text access to the quarter of the indexed journals. The database applies the key word listing of the Medical Subject Heading system (MeSH), which provides the advantage that publications with a similar content but different terminology can be linked. Bibliographic data are provided by the following resources: Index Medicus, International Nursing Index, Index for Dental Literature, PreMedline, AIDSline, BioethicsLine, and HealthSTAR. The service is available free of charge via the intranet of University of Pécs through EBSCOhost of EBSCO Publishing, and it is also available through the Ovid database. The search mechanism is similar to the one described for EBSCOhost.

### **2.7.2. PubMed**

The National Library of Medicine operating in the United States established the most widespread and popular medical bibliographic database, which is accessible free of charge also for the public. The service covers mostly abstracts and bibliographic data dating back to 1950 – full-text articles are available only through the university- or via the clinical library's intranet systems.

#### ***Search options***

After entering the search interface of Pubmed, there are options in English-language for both simple and advanced search. In case of simple search, the one-line search bar can be filled in with the name of the author(s), the title of the publication, or the title of the journal. A *Spell Check Feature* suggesting corrections or alternate spelling options, is also available during search to help experts to minimizing typing mistakes. A database involved by the NCBI (National Centre for Biotechnology Information) can be chosen from the drop-down menu. The searching process can be started by clicking on the *Search* button (Figure 2/36).

An advanced search can be launched by clicking the *Advanced* button under the search bar. Search questions can be inserted in the listed search boxes, by author, book, MeSH key word, journal, etc. Boole-operators that connect search terms can be chosen from the drop-down menu next to the search boxes, and clicking the *Search* button initiates the process.

The *Search history* available under the search bar provides a summarized table on when the search with a specific term was carried out and how many results it produced.

#### ***Display of results***

Data in the results list provide information on the following aspects: title of the article, name(s) of the author(s) and publication data. The abstract of the article can be viewed after clicking the title of the relevant publication. Clicking *Full text links* in the right column offer the total content of the record. Available publications can be printed, saved or sent via e-mail (Figure 2/37).

#### ***Access***

PubMed can be accessed via <http://www.ncbi.nlm.nih.gov/pubmed>



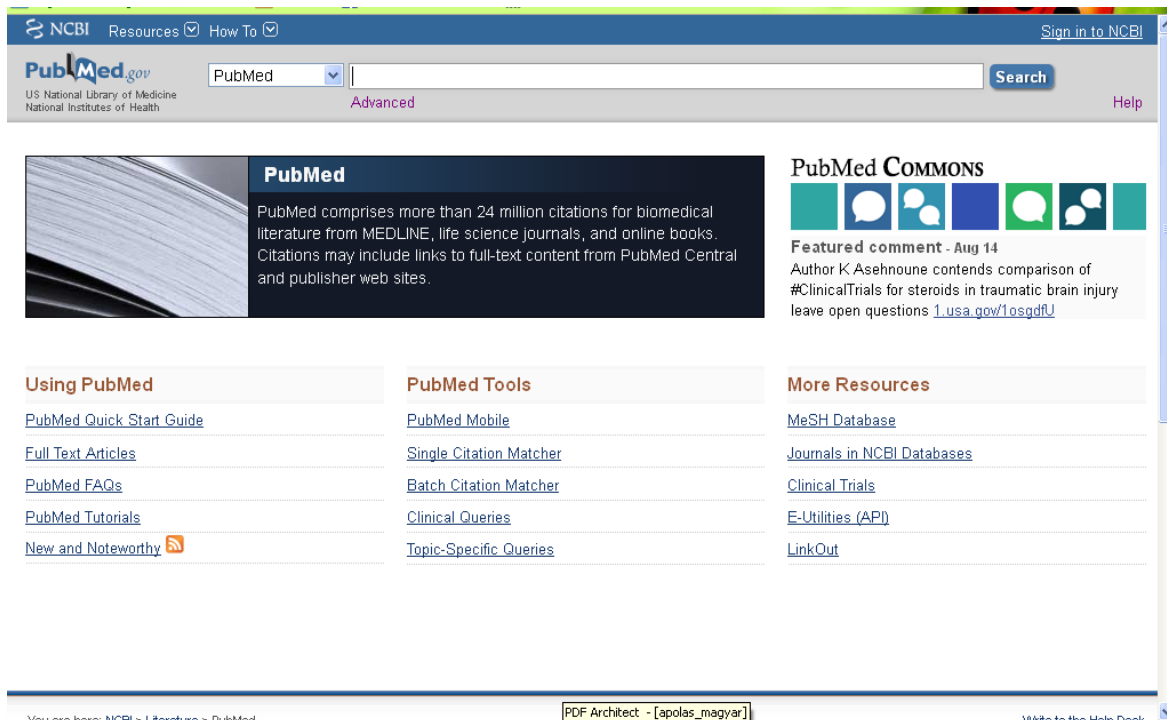


Figure 2/36. Search interface

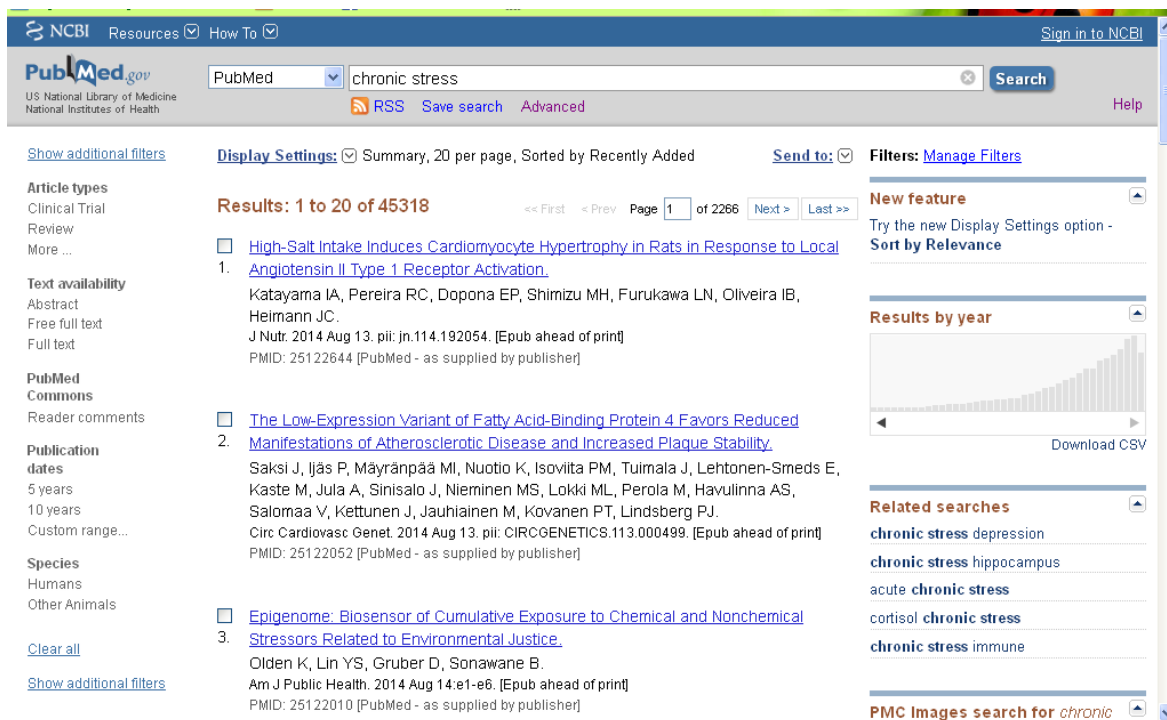


Figure 2/37. Results list

### **2.7.3. SPORTDiscus**

The exhaustive database SPORTDiscus covering scientific literature on sports medicine and rehabilitation can be accessed through the EBSCOhost platform. Most of the indexed journals and documents published in English, French, Spanish and Italian can be accessed in full-text, covering the following fields: biomechanics, sports medicine, exercise, kinesiology, sports, sport psychology, nutrition, health care and therapeutic programmes, physical fitness, physical therapy and rehabilitation. Currently this service is not available from the intranet of the University of Pécs.

#### ***Search options***

The database can be accessed from the search interface of EBSCOhost. The search process is similar to the one described for the EBSCOhost platform, following the choice of language.

### **2.7.4. Digital Library of the University of Physical Education**

The collection of the Digital Library of the University of Physical Education in Hungary focuses on relevant literature published at the field of physical education and sport sciences. To make the system more user-friendly, the indexed literature is organized into collections, where journal articles, books, conference publications, PhD-dissertations, archive sports journals, and some theses are available, alongside sport-related videos and pictures. The Digital Library provides full-text access to journals of the University of Physical Education, such as *Kalokagathia* or the Hungarian Journal of Sports Sciences.

#### ***Search options***

Searching the database can be done in a simple or advanced manner. A simple search starts by inserting search terms into the search box. Incomplete search, Boole-operators, brackets and proximity operators may also be used. An advanced search can be carried out focusing on the title, author, geographic term, metadata and full text. Also, the user may define the time interval, collection or type of media too (picture, text, audio or video) too (Figure 2/38).

#### ***Display of results***

Results may be displayed in the following forms: *Short form, Table view or Full view*.

The first two forms show the title, name(s) of the author(s) and a pdf icon as well. Information on key words, geographic terms, place and time of publication, length and the collection are revealed when clicking on the title of the publication (Figure 2/39).

## Access

The Digital Library of the University of Physical Education can be accessed via

<http://tf.hu/oktatas/konyvtar/tf-digitalis-konyvtar/digitalis-dokumentumok/>

The screenshot shows the search interface of the Semmelweis University Digital Library. At the top, there is a header with the university's logo (TF 1925) and the text "Semmelweis Egyetem Testnevelési és Sporttudományi Kar Digitális könyvtár". A navigation bar includes "Search", "Results", "Previous Searches", "Search Bases", and "My Space". A search bar is prominently displayed with a "Search" button. Below the search bar, there are options for "Simple Search" and "Advanced Search", a "Select collection" dropdown menu (set to "General"), and a "GO" button. A search criteria section includes a text input for "A word or phrase:" and radio buttons for "Contains", "Exact", and "Starts With". Below the search bar, there is a "Collections" section with a grid of categories and their respective item counts: "Témakörök szerint" (3803), "Folyóiratcikkek" (1567), "Könyvek" (11), "PhD-értékelések" (113), "Válogatott TF-szakdolgozatok" (308), "Szakdolgozatok (TF-intranet)" (1315), "Régiségek" (24), "Konferenciakiadványok" (3), "Különnyomatok (TF-intranet)" (1), and "Videók" (3). The footer of the interface shows "© 2007 Ex Libris".

Figure 2/38. Search interface

The screenshot displays the search results list for the query "W-All Words= ács pongrác" in the "General Sáro" collection. The results are sorted by "Ranking" and show records 1 through 18 of 18. The interface includes a header with the university's logo and name, a navigation bar with "Search", "Results", "Previous Searches", "Search Bases", and "My Space", and a search bar containing the query. Below the search bar, there are options for "Brief view", "Table view", and "Full view", and a "Sort by" dropdown menu set to "Ranking". The results list consists of six items, each with a thumbnail image, a title, a subtitle, and the author's name. Item 1: "A sportolás növelésével elérhető gazdasági haszon mértéke" (Economic benefits of increasing p) by Stocker Miklós. Item 2: "Serdülők életmódja és testneveléssel kapcsolatos véleményük A felnőtte világ útján..." by Rétsági Erzsébet. Item 3: "A közilabda szponzorációjának, valamint a sportszakmai siker és a szponzoráció mértéke köz" by Veres Péter. Item 4: "100 év az egyetemi-fiskolai sport szolgáltatásban 1907-2007". Item 5: "Ausztriába migráló magyar labdarúgók motivációs tényezői" by Tóth Zsolt. Item 6: "VIII. Országos Sporttudományi Kongresszus : Program és előadáskivonatok" by n.n.

Figure 2/39. Results list

## **2.8. Electronic books and journals**

Researchers nowadays are provided considerable assistance by online electronic journals, whose number is growing every year. These type of sources offer numerous advantages: the time-span of the publication process is becoming a lot shorter as printing works can be avoided; access through a PC does not depend on location or time, so articles are available anytime and anyplace; the structure of the documents is similar to the printed ones; and the relevant documents can be downloaded, saved or printed as well. The content of some e-journals are downloadable free of charge or after registration, but many require subscription.

### **2.8.1. Open Access journals**

Online *Open Access* and *Full Text* scientific publications and documents that support free information-flow and have numerous positive values have become increasingly important over the past decade. Open access allows both experts and laymen to access relevant scientific literature comfortably and with equal opportunities, regardless of location or time. Another advantage is that it generates a growth in the number of references, as the more people read a given publication, the more chance it has for being used. The consequence is not only the growth in reference-numbers, but also the growth in the popularity of the institution. As access is free for everyone, users won't not have to pay for the articles. Access is also made easy by the process that these articles can be found quickly by using general search engines such as Google, if the appropriate search terms are typed in. Digital documents can be saved, printed or sent via e-mail by the users. Unfortunately, the growing subscription costs result in fewer and fewer journals available for researchers in libraries, which is referred to as a 'journal crisis'. One possible solution for this information shortage is the spread of open access: professionally, it could enhance international and multidisciplinary cooperation, and experts in less wealthy countries could also access relevant, evidence-based scientific results. Quick access also supports the effectiveness of research, making it easier to initiate a scientific debate, and eventually providing quicker and more effective answers to research questions. It is important to remember that copyright laws are to be observed in these cases as well. Nowadays, some licenses offer the opportunity for authors to define the legal frames of the usage of their own publications. Finally, it should be mentioned that the open access to preprints (prior to print publishing) may play an important role in priority publications, which is a hugely important issue, for example, for patenting issues. Open access document are usually provided by digital libraries (Molnár – Németh 2009; Bánhegyi 2003, Bánhegyi 2009).

## 2.8.2. Directory of Open Access Journals

The *Directory of Open Access Journals* collects and provides free access to scientific journals of various fields. Apart from English, journals can be read in Spanish, Portuguese, French, German, Italian, Russian, Turkish, Japanese, Chinese and Malaysian language as well.

Users can perform basic or advanced searches. During basic search, the search term should be inserted into the search box, and then hitting *Search* will start the process.

Advanced search offers options to indicate search terms by *Title, Keywords, Subject, ISSN, DOI, Journal Country, Journal Language, Publisher, Abstract, Author, Year, Journal title, or Journal Alternative Title.*

Connections can be made by using Boole-operators. Results can be arranged according to *Relevance, Date added to DOAJ, Title, Article, or Publication date.* There are also various options to refine results list, by *Publication, Subject, Journal Language, Journal Country, Publisher, Publication charge, Journal License, Date of publication-Articles* and *Journal title-Articles* as well.

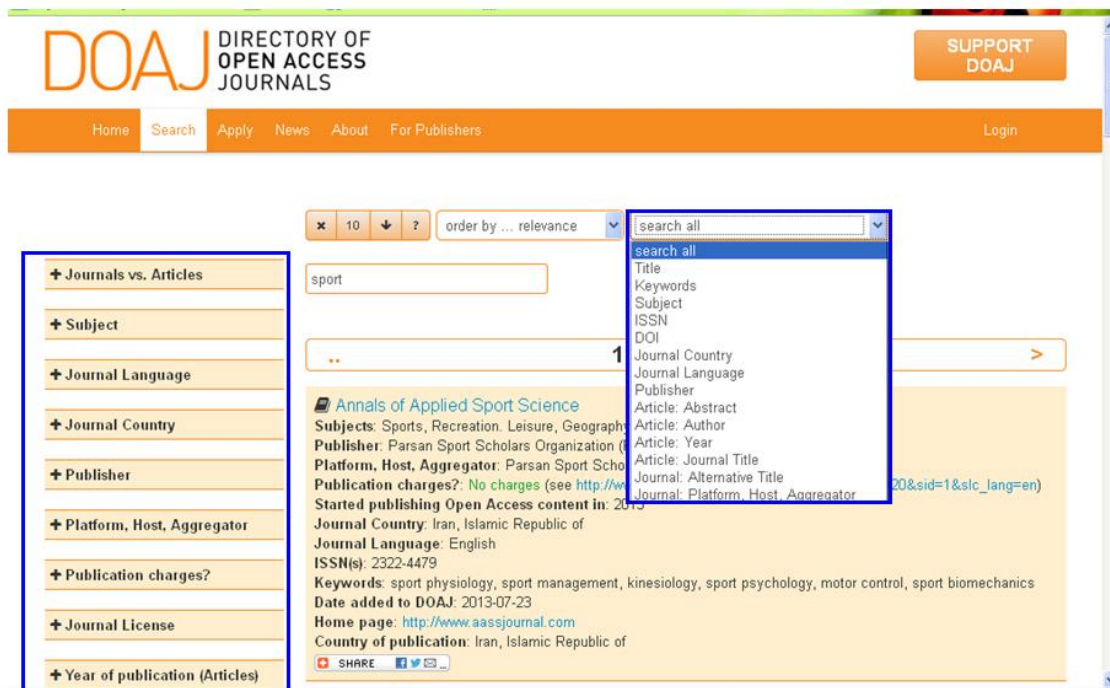
Each item of the result list can be freely accessed, downloaded or printed (Figures 2/41 and 2/42).

### Access

DOAJ can be accessed via <http://www.doaj.org>

The screenshot displays the DOAJ website's search interface. At the top left, the logo reads "DOAJ DIRECTORY OF OPEN ACCESS JOURNALS". A "SUPPORT DOAJ" button is located in the top right. Below the logo is a navigation menu with links for Home, Search, Apply, News, About, For Publishers, and Login. The main content area features a search bar with the text "Search DOAJ" and a search icon. Below the search bar are checkboxes for "journals" and "articles", and a link for "[Advanced Search]". To the right of the search bar, a statistics box shows: "9,958 Journals", "5,844 searchable at Article level", "134 Countries", and "1,705,751 Articles". Below the search bar is a section titled "Directory of Open Access Journals (DOAJ)" with a brief description and a link to the "Application Form". Below that is a "Latest News" section with a link to "Proactive not reactive". On the right side, there are links for "FAQs", "Features", "Open Access Information", "Download metadata", and "New Journals Feed". At the bottom right, there are links for "Our sponsors", "Our members", and "Our publisher members", along with social media icons for Facebook, Twitter, and LinkedIn. A URL is visible at the bottom left: "www.linkedin.com/company/directory-of-open-access-journals-doaj-".

Figure 2/40. Search interface



**Figure 2/41. Search options**

## 2.9. Online search engines

Using general search engines to track various documentation is also popular among researchers. Search engines are useful to browse through documents, pictures and software alike, as the process in this case covers the whole of the Internet. Key word-based searches can be carried out in a basic or advanced way. One of the most popular search engines is Google, *which is* also available in Hungarian (<http://google.com>). Other search engines (with portal services) include Excite (<http://excite.com>), Yahoo (<http://yahoo.com>), Microsoft Internet Start (<http://msn.com>), HotBot (<http://hotbot.com>), and Lycos (<http://lycos.com>). It is important to mention that via the Internet an incredibly large amount of information becomes available, and finding the most relevant and genuine scientific publications may be challenging.

## 2.10. Quotations and intertextual references

Naturally, ideas and relevant works of other authors are also used in the preparation phase of writing a scientific paper. It is a basic requirement to refer to these word-for-word or loosely quoted ideas. With a view to ensuring scientific accuracy and preventing the violation of copyright laws, the referred source must always be indicated precisely. It is regarded as plagiarism to use other authors' words, ideas or scientific results as one's own. References are advised to be used when they support the original statement of the researcher; serves as a

disproof of a certain idea or concept; or contains an important data or theory that is necessary for the understanding of the topic. Therefore, it is acceptable to use the words, ideas or results of other scientists, but a precise reference with accurate bibliographic data on the source is a prerequisite. Another basic principle of giving references is that they should be unambiguous and the source should be reliable. It should be common practice throughout the writing of a thesis to refer to primary sources and not apply secondary sources without checking. A list of books consulted with precise indication of bibliographic data is useful throughout the literature review, as it will come handy during the preparation of the reference list of the thesis and helps finding relevant sources again.

Sources can be indicated in form of a footnote or at the end of the text under *References*.

There are three basic techniques to refer to various sources:

- In case of ***word-for-word quotations*** the referred text must be indicated by quotation marks. The author's name, year of publication and page numbers should be indicated in brackets. If the document has three or more authors, then only the first one should be mentioned, followed by the abbreviation *et al.*
  - **Example:** Professor Selye defined stress the following way: “*stress is a condition with a specific group of symptoms, including all unspecific mutations within a certain biological system.*”... (Selye 1936, 32-45.)
- ***Paraphrasing*** (reference to content) means that the author describes the idea of another author with his own words. In such a case it should be kept in mind that the described line of thought should be the same as the original one. The reference in brackets – here also – should include the name of the original author and the year of publication. However, indicating the page number is not necessary.
- ***Cross-reference*** means that although the original publication was unavailable, it was referred to in a third author's publication. In such cases the brackets should not only include the data of the source author, but it should be followed by the data of the original document.
  - **Example:** Dr. István Karsai highlighted in a speech given at the 8<sup>th</sup> National Sport Sciences Conference that... (Karsai 2012)

Referencing has strict rules for layout, and there are two specific methods.

Numerical listing provides sources according to the references' place in the text. The name-method (Harvard-system) requires indicating sources by the name of the author and year of

publication. The number in front of the alphabetically listed name of the authors shall not be taken into consideration.

### **Preparation of a reference list in a scientific work**

The relevant literature used and referred to during the preparation of the scientific paper should be listed after the text in a summarized manner. In Hungarian there are different terms to describe reference lists and bibliographies. *Reference lists* showcase those relevant documents that were consulted before the preparation of the text, even if there is no direct reference to the source in the text. In contrast, *bibliographies* analyse the basic and necessary literature in connection with the given topic. Their aim is not only to be used for identification, but also for traceability. A record in a bibliography is a basic entry of references or notes used. There are precise rules on how to form bibliographic entries:

- Hungarian standard: MSZ ISO 690 (since 1991),
- International standards, such as APA (American Psychological Association), Harvard British Standard, AMA (American Medical Association), Vancouver/ICMJE.

We would like to draw the reader's attention to the fact that requirements of reference lists may vary by each higher education institution or department, and scientific journal. Description of requirements should always be thoroughly consulted, as there might be specific necessities regarding references and bibliographies, which may differ from the above-mentioned (generally accepted) norms.

### **General principles of bibliographic entries**

- References should be listed alphabetically, each record in a new line.
- Alphabetical order is based on the initial letters of the first author's family name.
  - In case of Hungarian authors, the record should start with the authors family name (without a comma), followed by the initial letter of the first name, and a dot.

#### **Example:**

- Ács P. (2009): A sportolók területi mozgásai, avagy a sportolói vándorlás. Tér és társadalom. 23(3),147.
- The authors should be listed in the order indicated by the original article's cover page.



- Academic degree and abbreviations of other degrees should not be listed among bibliographic data, accordingly they should not be indicated in a reference list or bibliography (e.g.: dr.; prof; PhD)
- When referring to international authors the family name of the author followed by a comma, then the initial letter of the given name and a dot should be indicated

**Example:**

- Baxter, P. – Jack, S. (2008): Qualitative case study methodology: Study design and implementation for novice researchers. *The Qualitative Report*, 13(4), 544-559.
- If the author has more than one given name, each of its initial letters should be indicated, separated by space.

**Example:**

- Adams, J.M.G. – Tyson, S. (2000): The Effectiveness of Physiotherapy to Enable an Elderly Person to Get up from the Floor. *Physiotherapy*, 86(4), 185-189.o.
- The format of referring to international authors should be applied when listing a Hungarian author's work published in another language, thus a comma should be put between the family name and given name of the author.

**Example:**

- Oláh, A., Katona, Gy., Gál, N., Müller, A., Damasdi, M., Boncz, I., Betlehem, J. (2012): The comparison of two minimal invasive surgeries, the tension-free vaginal tape (TVT) and the transobturator tape (TOT) in terms of efficiency and the complications. South Eastern Europe Health Sciences Journal, 2(2), 82-87.

- In the case of edited books, the family name and given name of the editor should be indicated, followed by the (ed.) or (eds.) abbreviation.

**Example:**

- Oláh A. (szerk.) (2012): Az ápolástudomány tankönyve. Budapest, Medicina Kiadó.

- If a document has several authors, the authors' names should be listed using hyphens and spaces. If one author has more than one family name, then there should be no spaces put next to the hyphen.

**Example:**

- Pakai A. – Kívés Zs. (2013): Kutatásról ápolóknak. 2. rész: Mintavétel és adatgyűjtési módszerek az egészségtudományi kutatásokban. Nővér, 26 (3), 20-43.

- Multiple works from the same authors should be listed according to the year of the publication. If there were several documents published in the same year, then they should be listed alphabetically according to the title of the source, denoted by a/b/c letters.
- The number of the given volume in the case of books with multiple volumes should be indicated by digits.
  - **Example:** (Vol. 1.)
- If a book has been published for the first time, the number of the edition should be indicated after the title of the book.
  - **Example:** 2<sup>nd</sup> edition
- References of works published in other languages should follow the rules of the specific language.
- No Hungarian items should be inserted into bibliographic records in other languages. The title should be written in the language of the source, and a Hungarian translation may be given in square brackets.

- If the used work has no indication of authors, then the title and year of the publication should be given in brackets.
- Reference lists also use general abbreviations as shown by Table 2.

**Table 2/2. Summary of general abbreviations used in references and bibliographies**

<b>indicating the edition</b>	
<b>edited</b>	edit.
<b>amplified</b>	amp.
<b>edition</b>	ed.
<b>corrected edition</b>	corr.
<b>indicating the publisher</b>	
<b>Budapest</b>	Bp.
<b>Pub. location not available</b>	Pub. loc. n.a.
<b>Publisher</b>	Pub.
<b>Publisher not available</b>	Pub. n.a.
<b>indicating the journal</b>	
<b>volume</b>	vol.
<b>number</b>	no.
<b>indicating length</b>	
<b>page</b>	p. (from the Latin “pagina”),
<b>page range</b>	pp.

*Formats of a bibliography in case of various types of documents may be the following:*

**Book:**

**Name of the author (Year of publication): Full title. Place of publication: Publisher.**

*Ács P. (2009): Sporttudományi kutatások módszertana. Pécs: Pécsi Tudományegyetem Természettudományi Kar Testnevelés- és Sporttudományi Intézet.*

**Books with more than three authors:**

**The name of the first author et al. (Year of publication): Title. Place of publication: Publisher.**

*Csermely P. et al (1999): Kutatás és közlés a természettudományokban. Budapest: Osiris Kiadó.*

**Book chapter:**

**Name of the author (Year of publication): Title of the book chapter or study. In: Name of the editor(s) (ed(s).): Title of the book. Place of publication: Publisher, Page (from-to)**

*Staunder A. (2007): Stressz és stresszkezelés. In: Kállai J., Varga J., Oláh A. (szerk.): Egészségpszichológia a gyakorlatban. Budapest: Medicina Könyvkiadó Zrt, 153-176. p.*

### **Edited books:**

**Name(s) of the author(s) (ed(s.)) (year): Title of the book. Place of publication: Publisher.**

*Oláh A.(szerk.)(2012): Az ápolástudomány tankönyve. Budapest: Medicina Kiadó.*

### **Book without an author:**

**Title of the book (Year of publication) Place of publication: Publisher.**

*Magyar Statisztikai Évkönyv, 2010 (2011) Budapest: KSH.*

### **Journal:**

**Name(s) of the author(s) (Year of publication): Title of the article. Title of the journal, Number of volume. Number of issue, page number (from-to)**

*Melczer Cs., Melczer L., Szabados S., Ács P.(2012): Szívelégtelen betegek életminőségét mérő validált kérdőívek összehasonlító vizsgálata. In: Egészség-Akadémia, 3.1.54-60. p.*

### **International journal article (similar to references in a Hungarian-language article):**

*Oláh, A., Katona, Gy., Gál, N., Müller, A., Damasdi, M., Boncz, I., Betlehem, J. (2012): The comparison of two minimal invasive surgeries, the tension-free vaginal tape (TVT) and the transobturator tape (TOT) in terms of efficiency and the complications. In: South Eastern Europe Health Sciences Journal, 2.2. pp.82-87.*

### **Reference to a conference presentation:**

**Name(s) of the speaker(s): Title of the speech. Place of the speech. Time of the speech.**

*Pakai A. (2012): A tudományos közlések módszertani alapjai a táplálkozástudomány területén: Előadás. Budapest, „A tét a jövőnk: A táplálkozás és a mozgás összhangjában”.: A Magyar Dietetikusok Országos Szövetsége XIII. Szakmai Konferenciája. 2012.11.17.*

### **Reference to laws:**

**Title of the law (year of publication). Title of the journal, Number of the Volume. Number of the issue, Page number (from-to)**

2010. évi XCII. törvény egyes egészségügyi és szociális tárgyú törvények jogharmonizációs célú módosításáról (2010). In. *Egészségügyi Közlöny* 60. 20. 2986-2989. p.

**Reference to electronic sources and documents:**

**Name(s) of the author (or editor, organization or source) (year of publication): Title of the study or website: subtitle (if available). Name of the website, URL to the website, date of downloading**

*Fidy, J., Makara, G. (2005). Biostatisztika. Retrieved from:*

URL:<http://www.tankonyvtar.hu/hu/tartalom/tkt/biostatisztika-1/ch11.html> {2013.06.05}

**Standard**

**Abbreviations used by the country issuing the standard. Title of the standard. Place of publication: Publisher, Year of publication. Page number.**

*MSZ ISO 690 (1990) Bibliográfiai hivatkozások. Budapest: Magyar Szabványügyi Testület, 22. p.*

### 3. BASIC STATISTICAL CONCEPTS, TYPES OF VARIABLES AND CRITERION VARIABLES (Pongrác Ács)

#### 3.1. Definition of statistics

This chapter aims to clarify the most essential basic concepts for data analysis in practice. Similarly to other scientific fields, statistics also has its own terminology. The following chapters will present the most basic methods of analysing data gathered during data collection, which requires a certain amount of basic statistical knowledge. Nowadays all scientific data-evaluation and analysis are carried out electronically (an admittedly effective way), and thus we are going to explain basic concepts and describe certain analytical methods through the SPSS environment.

*Statistics* is a scientific method of collecting, describing, analysing, evaluating and publishing information on major phenomena and processes. In line with the international literature, we may differentiate between *descriptive statistics*, *inferential statistics* and *statistical decision theory*.

*Descriptive statistics* basically includes methods of collection, analysis and a compact description of numerical data. Its most important subtopics are data collection, data visualization, data grouping and classification, the completion of simple arithmetic operations, and result explanation. This field of statistics applies simpler statistical methods and uses relevant data of the population exquisitely. Excel software is the most often used tool to carry out descriptive statistical analysis, as it is clear and easy to use.

*Inferential statistics* helps forming statements on certain phenomena and processes that are based not solely on direct observations. To put it simply, it helps gathering numerical data that is not measurable directly, but can be obtained through complex mathematical-statistical methods. Inferential statistics is strongly based on mathematical-statistics and probability theory, and it is therefore important to mention that inferences are always based on a certain sample (sample population) in this case. This book will discuss two subfields: estimations and testing hypotheses.

*The statistical decision theory* provides numerical information on the optimal choice between several options, taking random circumstances into consideration as well. Apart from empirical statistical observation and inferences it also provides opportunities for experts to form a subjective opinion. The statistical decision theory combines elements of probability theory and game theory involving the results of statistical observations.

### 3.2. Statistical data

Generally speaking, statistical data are results of some specific measurement (e.g. the number of the population, or a qualitative feature).

*Secondary data* are the ones that are measured and collected by someone else (or come from another source), while *primary data* are the ones that were collected by the researcher himself for the specific purposes of an examination.

We may distinguish between *basic data* that is collected through primary measurements and counting, and *derived data* that is the result of a calculation carried out using more than two basic data. Constantly used statistical derived data is often referred to as an *index-number* (e.g. BMI, population density).

According to their nature, data can be grouped as:

- *ascertainable*, that is *qualitative*, or
- *measurable*, that is *quantitative* data.

The two types differ in terms of data collection methods and the degree of measurability. We may claim that qualitative data can always be transformed into quantitative data (categories, classes, ranks, etc.), while often it does not work vice versa. Mathematical calculations can be easily carried out by quantitative data (e.g. average count, contraction, etc.), while this would be meaningless in case of qualitative data, such as groups or categories.

According to Ozsváth and Ács, based on their values or set of values, data can be:

- alternative or binary (dummy),
- discrete,
- continuous.

There are only two (optional) values available in case of *binary data* (e.g. male/female). “0-1” values (“yes-no”) are quite often used, although such digits might cause some confusion when it comes to division. Some statistical methods use coding with “1”, in these cases this dummy variable should be created.

*Discrete data* have the feature of being “score-like” which means that there is no continuity among values, and the range between values cannot be interpreted. A typical example for this are ranks, numbers of pieces or years, categorizations, etc. Discrete variables have a finite number of criterion variables (e.g. types of diseases occurring in the past year or number of children in a family).

*Continuous data* can be provided with optional precision, and also the range between any two values can be interpreted. “Continuity” is a feature of a measurable set of values, in which a measurement can be carried out with infinite precision (e.g. weight, blood pressure, etc.).

*Items* in statistics are the examination units on which the observation and measurement focuses. Statistical *population* is the sum of items in a statistical observation. Features and characteristics of the items in a population are highly important from the statistical point of view – they are called *variables* or *criteria*. *Criterion variables* are the possible outcomes of the criteria or the variable (Ozsváth-Ács 2011).

### 3.3. Types of variables and scales

Generally, variables can be *quantitative*, *qualitative*, *time-dependent* and *location-dependent*. Some statistical methods, such as analysis of regression, require a different type of definition for a variable. Thus we may distinguish between *dependent* and *independent* variables. Dependent variables are always influenced by independent variables – the two have a cause and effect relationship.

**Independent variables** are those that are expected to have a definite role in the problem we examine. The values (or attributes) of these variables are influenced or chosen by the researcher.

The researcher is interested in the “behaviour” (distribution) of **dependent variables**. He may observe what effect the changing of independent variables has on dependent variables (trying to recognize what kind of connection the two have). For example, the aim is to examine stress and depression in a challenging situation among nurses. The level of stress and depression are going to be the dependent variables that should be measured. The independent variables may be those that might influence these feelings, such as age, gender, or level of challenge – these are going to be the variables to be modified by the researcher. The level of challenge can be changed according to circumstances, while age and gender cannot be changed, only selected as independent items to be taken into account.

Data analysis in practice – especially the one carried out electronically – requires a precise definition of the measurement levels and measurement scales of the variables.



Types of measurement scales:

- nominal scale,
- ordinal scale,
- interval scale,
- ratio scale.

**Nominal (or categorical) scale** is the simplest scale that provides only a small amount of information. Classification and grouping of the items is applicable only to distinguish. The scale can only be interpreted to see whether the examined items are equal or different – but calculations (subtraction or division) with the values of the scale cannot be carried out. Here the coding of the observed items is arbitrary. Consequently, nominal scales include categories and groups, any digits used here can only be regarded as coding, and they often include binary data only (“two categories”, “yes-no”, “same-different”). If there is more than one category given, the value set is also larger, but it would still represent discrete values. It is important to highlight, that the values of the nominal scale, when summed up, cannot be compared or added to one another, as they cannot be organized or averaged either; that is, there is no “smaller or bigger”, “better or worse”, etc. A nominal scale always represents qualitative data, and thus never includes data with continuous distribution.

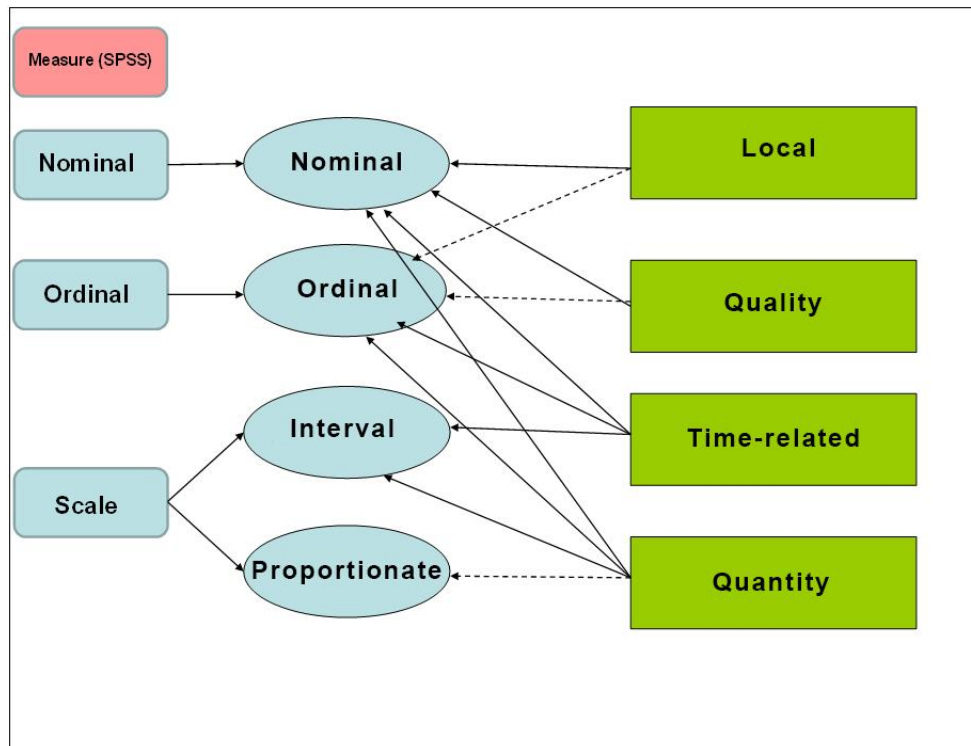
**Ordinal (or ordered categorical) scales** distinguish items and show a specific order. We can put our observed objects and items in order based on the relations corresponding to the order.. The scale defines a ranking difference, but will not show the difference between the positions (e.g. how much more effective is one operation than the other). Ordinal scales always include discrete and fundamentally qualitative data.

**Interval scales** are also known as metric scales, applied to measure the difference between two values, therefore it can describe how much bigger or better one item is than another one. An important feature of interval scales is that it has no real zero point, thus the zero value does not mean the lack of the given feature.

**Proportional scales** offer the largest amount of information and the highest measurement level. The proportion of any two values on the scale can be interpreted. The scale has a real zero point, which means that the zero value unambiguously indicates the lack of the given feature. Any kind of mathematical calculation can be carried out with the scale. Generally, all currently used unit scales can be considered as proportional scales. The precision of the measurement is only a technical question in this case. The only issue regarding its usage is

that units of measure which seem to be similar but have different dimensions may have different numeric systems.

The difference between two metric (interval and proportional) scales can be defined by finding the zero point, as it is arbitrary in case of the interval scale.



**Figure 3/1. The connections between criteria and measurement scales and how they are indicated in the SPSS environment**

Source: edited by the authors based on Pintér – Rappai 2007, p. 31.

It is important to mention, that the most advanced of the above-mentioned scales is the proportional scale, also indicated by the number of possible arithmetic calculations. It can be transformed into any other type of scale. The more advanced a scale is, the larger the number of analyses and comparisons it may be used for. “The definition of scale types is really important, as it defines unambiguously what types of analyses can be carried out – it can cause huge differences if dependent or independent variables are measured on a metric or a non-metric scale.” (Sajtos – Mitev 2007, 25. o.)

## 4. EDITING ONLINE QUESTIONNAIRES IN PRACTICE (László Bence Raposa)

### 4.1. Introduction

Questionnaires are a data collection tool popular in several scientific fields, especially in the field of social sciences. Data and samples of scientific examinations may be provided not only by experimental, analytical and measured data but also by answers provided by respondents in specific questionnaires. The main role of these questionnaires is to collect respondents' answers, mostly in a structured, pre-set way, following a given line of questions. The forms of questionnaires have changed in parallel with the evolution of telecommunication, and nowadays we may distinguish between surveys that are carried out through a postal service, a telephone or the Internet, or those that are completed personally (interviews or in-depth interviews). Surveys to be filled in on a personal basis may be divided into further categories, i.e. those that collect answers on paper (*P2P - paper-and-pencil*) and those using computers (*CAPI - Computer-Aided-Personal-Interview*) (Lógó 2007). Ways and criteria of usage may vary in case of different survey formats, so it is important to choose the one that best serves the aim of the research (Bauer et al 2007).

**Table 4/1. Prerequisites of applicability for various survey types**

	Survey type			
	Personal	Postal	Telephone	Online
<b>Content</b>	Any topic can be referred to	More useful to collect general opinion	Any topic can be referred to	Rich content (pictures, videos, animations)
<b>Cost</b>	High	Low	Low	Very low
<b>Source of error</b>	Relatively few, errors of the interviewer	Relatively high, sample errors	Relatively low, due to providing answers	Self-replies, sampling errors
<b>Time required</b>	1-2 weeks	8-10 weeks	3-4 weeks	1-2 weeks

Source: Bauer et al 2007

Nowadays online questionnaires are gaining popularity against regular, paper-based surveys, as the Internet is becoming the most basic source of information, and generally part of

everyday life. Also, there are several advantages of Internet-based surveys to be listed that prove their growing popularity:

- they are quicker and more effective than paper-based questionnaires, also in terms of data collection, processing and analysis
- they are a lot more cost-efficient (there are no printing and other costs)
- respondents are available in a lot wider scale compared to being contacted personally
- they provide a significantly higher respondent-rate
- responses are often more representative of the reality and are more honest as users experience a higher level of anonymity (e.g. recognition of handwriting is not possible)
- some question types are only applicable with the help of the Internet and telecommunication tools (such as using videos, etc.).

#### **4.2. Prerequisites of the questionnaire, basic principles of providing general information and preparation**

As a prerequisite of defining the questions of the survey, it is important to carry out research in order to become familiar with the topic in detail. The best work can be done if every relevant question is considered.

If the above-mentioned factors have been explored, then the problem can be defined properly, and the questions can be asked and structured accurately and appropriately.

The aim of the examination should be set at the very beginning. It is not worth diverting from it during the process as it may influence the results, however, research opportunities and aspects, identified during the research process, may also be considered. Null hypotheses and hypotheses should be set up that the research aims to support or reject. It is very important to keep in mind that hypotheses are always statements and never questions.

The next step is to formulate the main questions of the survey and fit them into the hypotheses.

The sample size should also be given (estimated number of respondents) as an inappropriate number of responses for certain questions may bias results and lead to false conclusions. The expected results should be predicted prior to defining and editing the exact questions, as it might cause confusion during the analysis if the questions do not cover all the aspects of the examined field. Finally, it is important to mention that if the research aims to complete a more complex statistical analysis, then the survey's scales and their levels and measures

should be homogenized.<sup>1</sup> Prior to editing online questionnaires it is important to consider which basic criteria will ensure that the survey provides the relevant information. The completion of the survey should be only a small effort for the respondent, questions should be easy to answer and should cover information that the respondent can respond to. Furthermore, the questions should motivate respondents and raise their interest, to prevent disinterested answering. Therefore the most appropriate style should be adopted, which makes the least possible mistakes occur..

Also, the topic and aim of the questionnaire should be defined at the beginning, along with its structure and logical order of the items. After defining the topic, aim and structure, questions should be divided into different groups according to subtopics.

Generally, questions should follow the logical path that leads from general, easy-to-answer questions to those that are more topic-specific. This structure is advised to be followed unless there is a special reason to do otherwise.

According to the above-mentioned aspects, questions can be categorized in the following way:

### **1. Questions about the respondent**

While most questionnaires are anonymous, a certain amount of information should still be collected about respondents, such as their gender, age, educational background and some optional socio-demographic or socio-economic aspects that may be useful for the purposes of the research. These answers are important as they provide the basis for our comparisons based on gender, class and age-groups.

### **2. General information**

This group of questions covers the general experiences, perceptions and knowledge of the respondents on the given topic. It might also be useful during the analysis as it is highly important how much the respondent is aware of the basic knowledge of the researched topic and whether there is any bias in his responses.

(Example: What do you think about sports generally? How satisfied are you with your own physical activity? What are your experiences of the local sport facilities?)

---

<sup>1</sup> <http://www.kerdoivem.hu/advice/> (2014-07-16)

### **3. Topic-specific questions**

The questions of this group refer to the exact and crucial knowledge and opinion of the given topic that will form the basis of the analysis. This is the most important section in terms of survey-edition, and thus it requires special attention, for generally this is the longest part of the survey.

(Example: What do you think about the physiological effects of cardio training? How satisfied are you with the ...Sport Facility's services? How often do you carry out cardio training?)

### **4. Problem solving, suggestions**

This is the section where the respondent may provide his opinion and suggestions on a given topic that may be used in the future.

However, it should be kept in mind that this is not an obligatory part of every survey. Rather, it has a sort of marketing feature and is more often a part of organizational development questionnaires.

It is highly useful to add a feedback section at the end of the survey with open questions to evaluate the questions and add opinions and suggestions. The latter may form the basis of further research and this section may also shed some light on hidden mistakes<sup>2</sup>.

(Example: In your opinion, what is the reason why you don't do sports weekly? What other questions would you find relevant in connection with physical activity? In your opinion, how well does this questionnaire evaluate attitudes and sporting habits?)

### **General knowledge of the questions**

The following paragraphs consider any important and practical advice by means of which we can obtain a database that will be easy to analyse. Most of the information is on the method and structure of framing the questions, the most vital issue when conducting surveys.

1. Questions should be short, easy and to the point. A simple style should be followed to ensure that respondents can clearly understand what the questions are about. Complicated and illogical questions and superfluous complexity should be avoided.

---

<sup>2</sup> <http://www.kerdoivem.hu/advice/> (2014-07-16)

2. Too many questions should not be asked on the same page. Although the matrix question-layout helps examining the frequency of more than one factor, it should only be applied for interrelated questions. Whenever possible, these items should be divided into separate questions.
3. It should always be kept in mind to hold the attention of the respondent throughout the whole questionnaire – otherwise the returned questionnaires will be vague and spontaneous, and will have missing answers that won't be appropriate for analysis, and the ratio of missing data will be high.
4. Filters and forwarding items should be used in case of questions reflecting habits and opinions. The lack of this feature will lead to asking several questions that are irrelevant to the respondent, again resulting in inaccurate and missing answers. (Example: Do you do sports? – if the answer is no, then the following questions should not be “What sports do you do?”) Forwarding questions should always be available to support smooth responding in case of negative answers as well.
5. Prior to publishing the questionnaire, it should be made sure that all possible answers are provided for the respondents. If we are not confident in this issue, an open answer of “Other” could be added, where there respondent can decide to add his answer if it is not available from the list.
6. Questions should always be formed precisely and unambiguously.
7. Suggestive and inductive questions that may influence objective responses should be avoided. (Example: What would you choose to eat to preserve your health: healthy fruits or fatty meat? You do not eat unhealthy food, do you?, etc.).
8. When asking about frequency or measure, it is important to always precisely define what the given answer choices represent, as they may often be differently interpreted by various people (e.g. generally, often, occasionally). Unambiguous measurements and time-frames should also be given (e.g. cm, 2-3 times a week, etc.).
9. The options “I don't know” and “It's hard to tell” are not recommended to be given as answers, as these cannot be used for analysis. They might be applied in some specific cases, but only if answering the question might pose real problems to the respondents.
10. Contradictory words should not be used in the questions, as they may confuse the reader. Synonyms are advised to express similar meaning (for example: use applied/not used instead of applied/not applied).
11. A special attention should be paid to sensitive topics (such as income level) and also the possible answers to these questions, even in case of anonymous surveys.

Categories and intervals are useful for these cases (from-to items). Whenever possible, intervals should also include the extreme measures as well.

12. An opportunity to refuse the answer should also be given in case of sensitive or inconvenient issues, indicating “unwillingness to answer”. Lacking this option may result in the refusal of the complete questionnaire, the loss being greater than only one piece of missing data.
13. Maintaining respondents’ attention is not only important in terms of wording, but also in terms of survey length. Too long questionnaires are inappropriate; the application of all questions should be justified. Eye-catching questions should be asked at the very beginning, making a good impression on the respondent. Also, asking open-ended questions should feature in the second part of the survey.
14. A pilot test should be administered to test the questionnaire before publishing it for a wider audience. The grammar of the questionnaire should also be checked by a proof-reader<sup>3</sup>!

### **Types of questions**

The following paragraphs will introduce the most often used items in a questionnaire. First of all, we will describe the two basic types of questions.

#### **Open-ended questions**

In case of open-ended questions respondents use their own words to provide an answer instead of choosing one or more options from a given list. It can be useful when the research examines respondents’ emotions and opinions, and sets out to analyse what factors may play a role in their decision-making.

A statistical analysis of open-ended questions can be challenging, and coding is often impossible due to the diverse terminology used by respondents.

---

<sup>3</sup> <http://www.kerdoivem.hu/advice/> (2014-07-16)

<http://www.kerdoivem.hu/questions/> (2014-07-16)



## **Closed questions**

Closed questions are the opposites of open-ended questions, and here respondents have to choose the one they find most appropriate from a given list of answers, without having the option of forming their own response.

A statistical analysis is easy to complete, answers can be easily coded and treated. There are several types of closed questions: based on the number of answers we may distinguish between alternative (typically true-or-false questions), selective (several options) and scale (quality ranking from 1 to 5) questions.

## **Scale-questions**

Scale-questions are useful to analyse emotions and attitudes, and are typically used to define to what extent they are positive or negative. There are several types of scales, such as ordinal, interval, proportional and nominal scales. The most often used range for answers is between 1 and 5. The symmetry of levels in the scale should be maintained, and the natural attitude (answer 3) should be as far from the negative end as from the positive end. It is also important that if scales with various lengths are used, the number of levels should always be odd. This is the only way to maintain symmetry and to materialize the neutral opinion (answer 3 on a five-level scale).

Generally, practice uses a maximum of 7-level scales, as the more levels a scale has, the smaller difference there is between answers, which means that the it will become more difficult to objectively define the opinion and distinguish between two neighbouring choices. “Likert-scales” measure the level of agreement with a statement, using the same levels, and the researcher’s job is to frame the statements. For this reason Likert-scales are useful for the homogenization of scales (when we have several complex statistical tests) (Lógó 2007).

Classic 7-level „Likert-scale”:

- Strongly agree
- Agree
- Somewhat agree
- Undecided
- Somewhat disagree
- Disagree
- Strongly disagree

### **Applicable questionnaire items**

Items are introduced using the questionnaire editing software Google Docs™, which is one of the most appropriate tools for scientific purposes: it is easy to use and freely accessible for anyone with a Google-account (Gmail). It has a user-friendly interface and uses links that can be shared via e-mail and embedded to websites and social media sites. The resulting database can be downloaded as an Excel worksheet<sup>4</sup>. However, the most important feature is that there are no limitations in terms of number of questionnaires, respondents or shares (Farkas 2012).

#### **„Text”**

Short texts of any information can be inserted (e.g.: contact, e-mail address).

A screenshot of a questionnaire item. On the left, there is a vertical blue bar. To its right, the word "Contact" is displayed in a bold, black font. Below the text is a light gray rectangular text input field with a thin border.

**Figure 4/1. Appearance of questionnaire item “Text”**

#### **“Paragraph Text”**

Also as an opportunity to insert own answers, “Paragraph Text” can be used to insert longer content (e.g. opinion or suggestion).

A screenshot of a questionnaire item. On the left, there is a vertical blue bar. To its right, the text "Opinion or suggestion" is displayed in a bold, black font. Below the text is a wide, light gray rectangular text input field with a thin border.

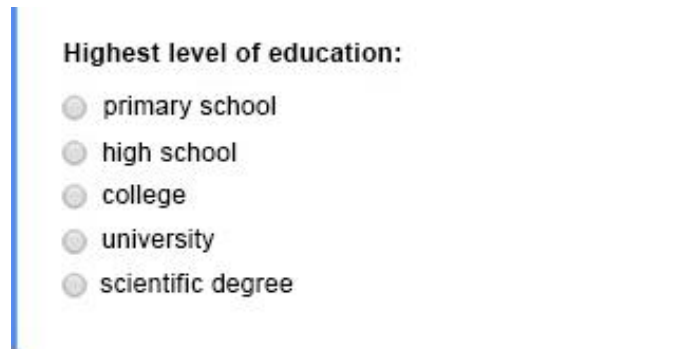
**Figure 4/2. Appearance of questionnaire item “Paragraph Text”**

#### **“Multiple Choice”**

This type of questions is used when only one answer should be chosen – this might have logical or statistical reasons as well. An extra supplementary option is to click “*Add Other*”, providing the opportunity of giving an answer that is not available from the list.

---

<sup>4</sup> **Excel-worksheet:** statistical software of Microsoft Office.



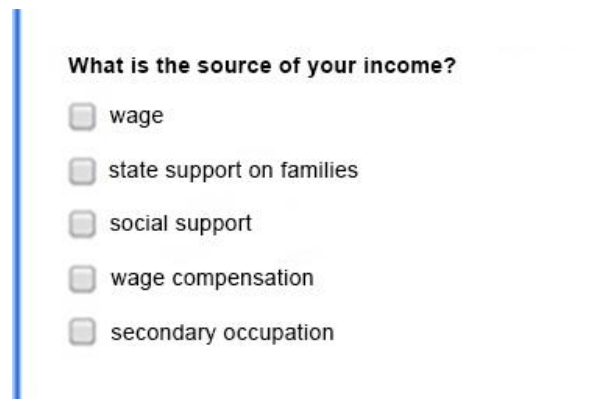
Highest level of education:

- primary school
- high school
- college
- university
- scientific degree

**Figure 4/3. Appearance of questionnaire item “Multiple Choice”**

### “Checkboxes”

This is similar to the “Multiple Choice” option, except that more than one answer can be chosen. The option of “Add Other” is available here as well.



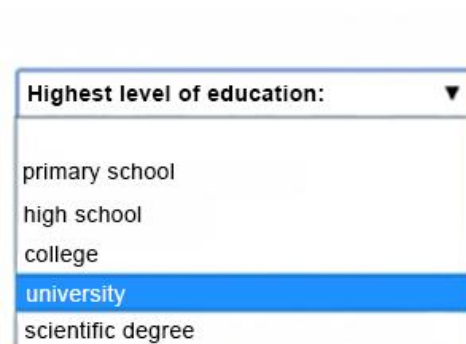
What is the source of your income?

- wage
- state support on families
- social support
- wage compensation
- secondary occupation

**Figure 4/4. Appearance of questionnaire item “Checkboxes”**

### “Choose from a list”

It has the same function as “Multiple Choice”, except that here the required answer can be chosen from a drop-down list.



Highest level of education: ▼

- primary school
- high school
- college
- university**
- scientific degree

**Figure 4/5. Appearance of questionnaire item “Choose from a list”**

### “Scale”

This item offers the opportunity to set the range or the scale of the answers giving a from-to value, measuring from 1 to 5 (useful for questions of frequency as in “all the time/often/never”). This item – when applied to measure agreement – is called a Likert-scale (“strongly disagree – strongly agree”).

How often do you do some kind of sport or physical activity?

1 2 3 4 5

never      always

Figure 4/6. Appearance of questionnaire item “Scale”

### ”Grid”

It is similar to the “Scale” item, except that more than one factor can be grouped for one question based on topic, and they may be weighed according to the requirements of the survey by adding frequency labels. In this case not only a from-to-frequency range is given.

How often do you practice the following sports?

	never-month)	occasionally (once in a	every now and then (at least once in every two weeks)	often (weekly, but maximum 2 times)	always (at least 3 times a week)
football	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
basketball	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
handball	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
volleyball	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
running	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

Figure 4/7. Appearance of questionnaire item “Grid”

## “Date”

This item is used when the exact date of an event is required (e.g. date of birth).

Date of birth

YYYY .mm. dd .

2014.

M	T	W	TH	F	SA	SU
30	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3

Figure 4/8. Appearance of questionnaire item “Date”

## ”Time”

Item “Time” is almost exactly the same as item “Date”, except that here the precise hour of the event can be given.

Start of the training

Hour . Minute . AM./PM.

Figure 4/9. Appearance of the questionnaire item “Time”

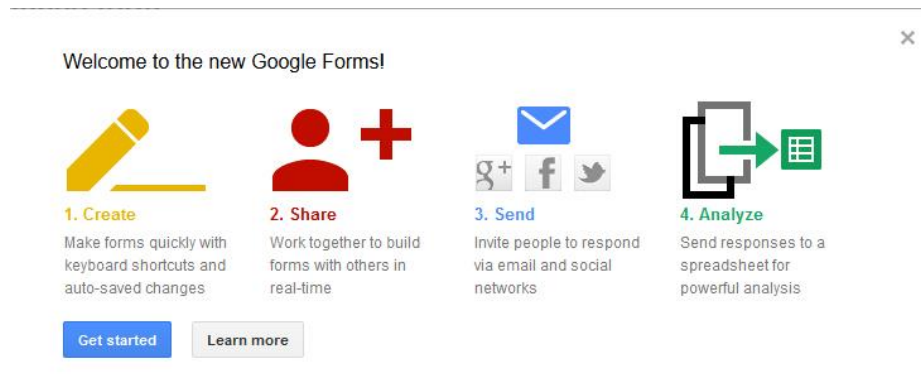
### 4.3. Questionnaire-editing in practice using Google Docs

In the previous chapters there has been mention of the advantages and application of Google Docs™ and Google Form. This chapter provides an introduction to the interface through a specific example.

There are four main steps of extracting the database from a Google Form, and there is also a wizard available for the first few instances. First, the questionnaire file should be set (“Quickly assemble forms by keyboard shortcuts and automatic saves of changes.”). It is important to mention, that more than one person can edit the forms simultaneously; the process is similar to a teamwork carried out by common e-mails or the Google Calendar (“Create forms simultaneously with others.”). When the questionnaire is ready, the link – that can be shared as an URL – should be shared. It can be embedded into any social media site or webpage, and can also be sent via e-mail. The last step is when the returned answers are

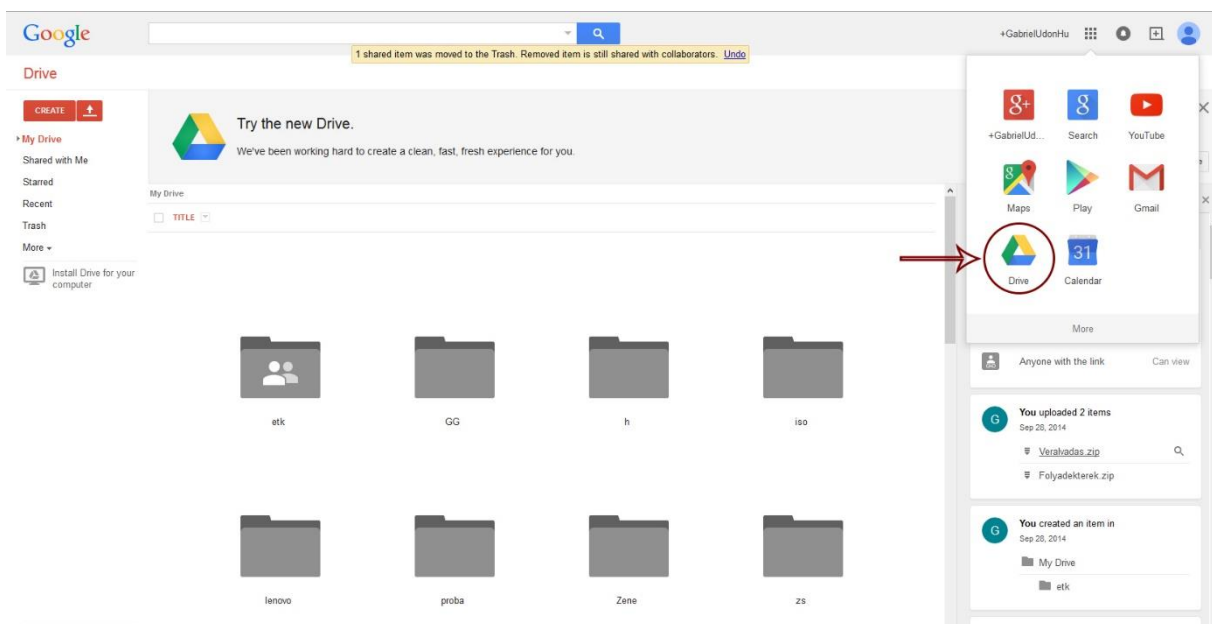
summarizedw by the system – this can be extracted as an Excel file and downloaded and saved (“Carry our effective analysis by exporting answers into worksheets.”).

Although the software comes with a built-in analysis feature (with graphs and figures), it might not be appropriate for use in case of more complex scientific work and in cases where a more complex statistical analysis is required (Farkas 2012).



**Figure 4/10. Visual description of Google Forms**

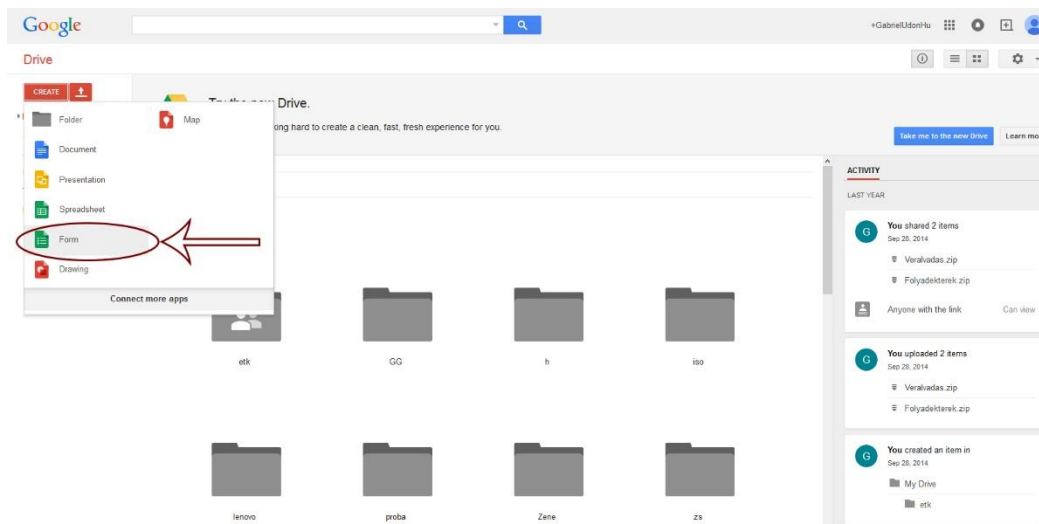
To create a questionnaire, a Google account should be accessed, from which one should access Google Drive, left-clicking the box-like icon in the top-right corner of the Gmail interface, choosing the icon “Drive”.



**Figure 4/11. Path from a Google account to the interface of Google Drive**

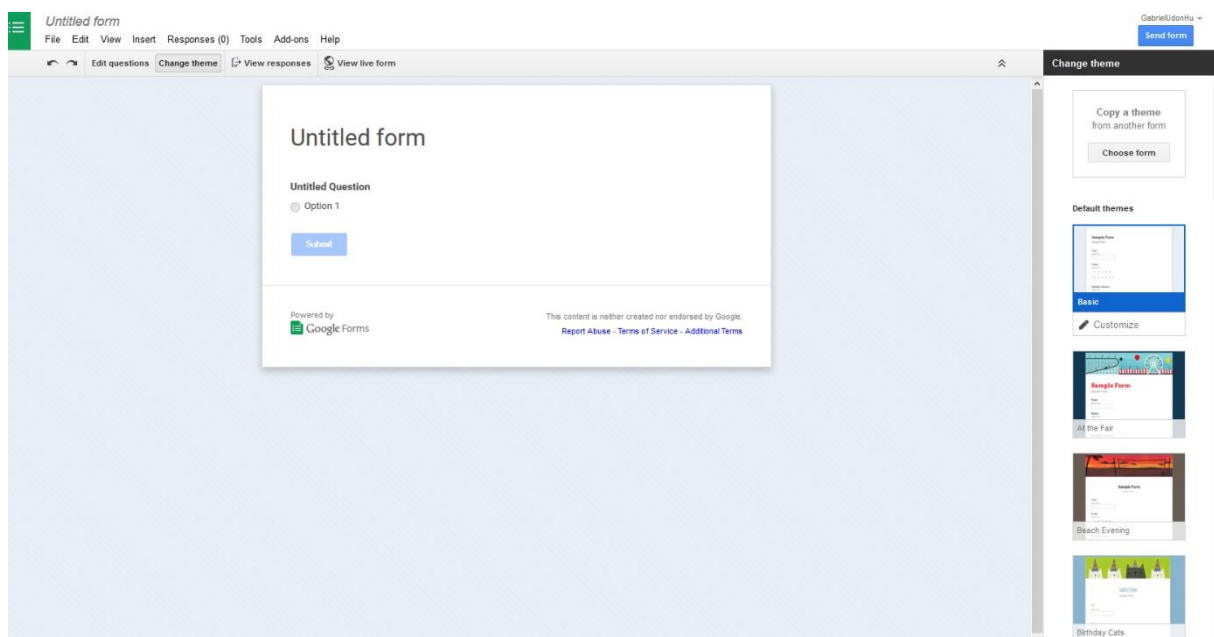
If this process does not work for any reason, there are other methods to try: after accessing the Google account a new tab should be opened in the web browser, where the following link should be copied: <https://drive.google.com/>.

After accessing Drive, a new form should be created by clicking *Create* in the menu list on the left side, then choosing *Form* from the drop-down list.



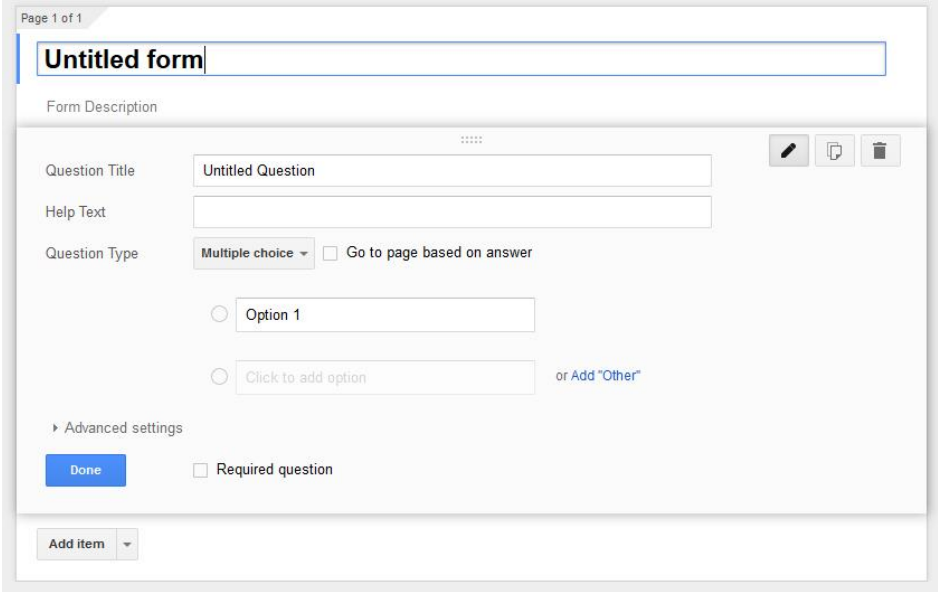
**Figure 4/12. Creating a new form in Drive**

After creating the form, the next thing to set is the title and style of the questionnaire. Style can depend on the creator, although it is probably the *default theme* that fits scientific needs best.



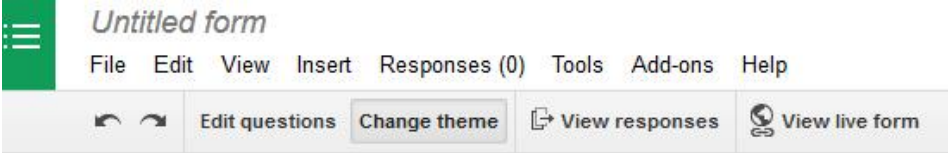
**Figure 4/13. Setting title and style of the form**

After setting title and style we may access the editing interface of the questionnaire, where we can edit the survey through each item. Adding questions is possible by clicking *Add Item*, in the pop-up window. The title of the question should be inserted into the box *Question Title*, and extra information and aid for precise answering can be provided in the *Help Text* section below. Clicking *Question Type* allows setting the required type of the item, as described under “Applicable Questionnaire Items”. The following figure shows the exact placement of the item’s features.



**Figure 4/14. Question-editing interface**

We may proceed from one question to the next according to above-mentioned instructions to assemble the required survey, although there is also a possibility for editing later. The given question can be clicked and removed anytime if the sequence of the questions proves to be inappropriate. Other corrections can be made by clicking the top-right corner of the panel shown above.



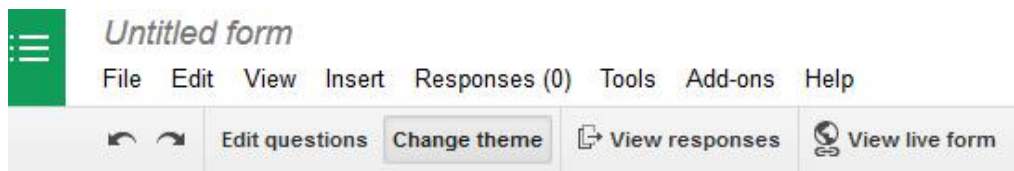
**Figure 4/15. Icons for correction**



The three icons seen above have different functions. The *pencil* icon (first on the left) is useful for editing the text of the questions and correcting grammatical mistakes. The *documents* icon (in the middle) helps to make a copy of the given questions. For example, if the researcher is curious about a similar factor in a given type of question, it should not be rewritten and set again.

The *trash* icon (the third one) is naturally useful for deleting a given question, in case it proves to be superfluous.

When editing is completed, there are some other options at the top corner of the interface. These will be described in the following paragraphs, although they are highly similar to those used by similar software, thus we will only describe the differing ones.



**Figure 4/16. Editing bar of the form**

The *File* menu offers general options, from which the most important are *Download as*, *Embed* and *Send e-mail to co-workers*. By clicking the menu *Download as* we can download all the answers provided for the questionnaire in a .csv format. *Embedding* provides the html-code<sup>5</sup> that can be embedded to websites. *Send e-mail to co-workers* is self-explanatory, and it is worth mentioning that the quick button *Send form* at the right corner of the screen does the same trick.

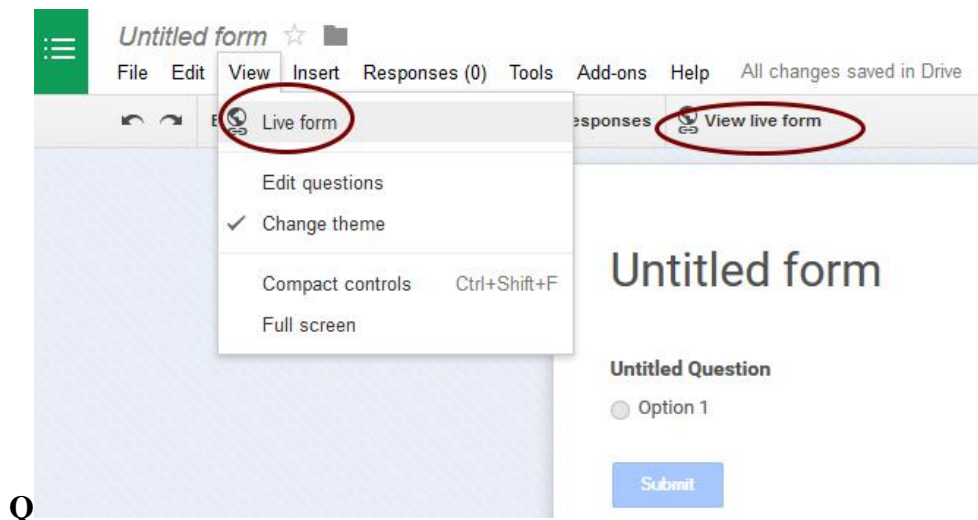
---

<sup>5</sup> **HTML:** HyperText Markup Language

**Figure 4/17. Various functions of the file menu**

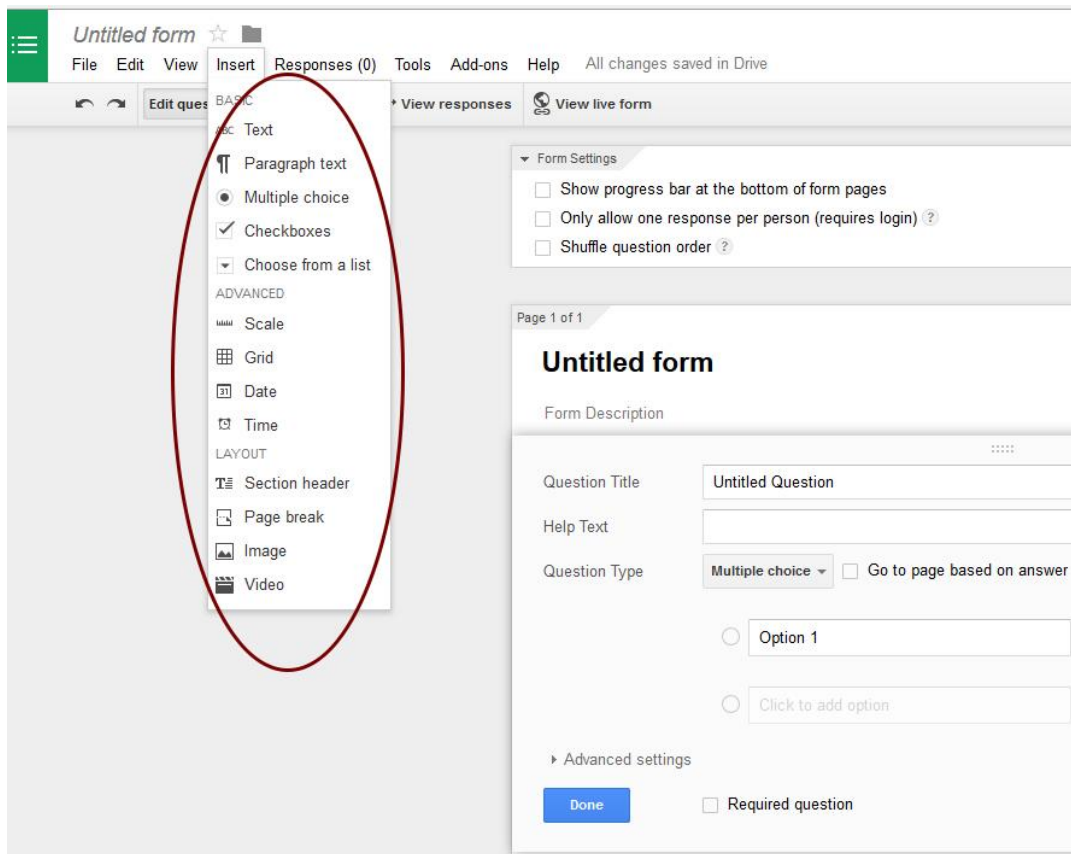
Stepping forward and redoing completed actions can be done by clicking *Edit*.

The *View* menu helps setting how we would like to see the survey – there is also an option of full screen view, but more importantly, choosing *Live view* shows the final layout of the questionnaire. The same can be obtained by clicking the quick button *View live form*.



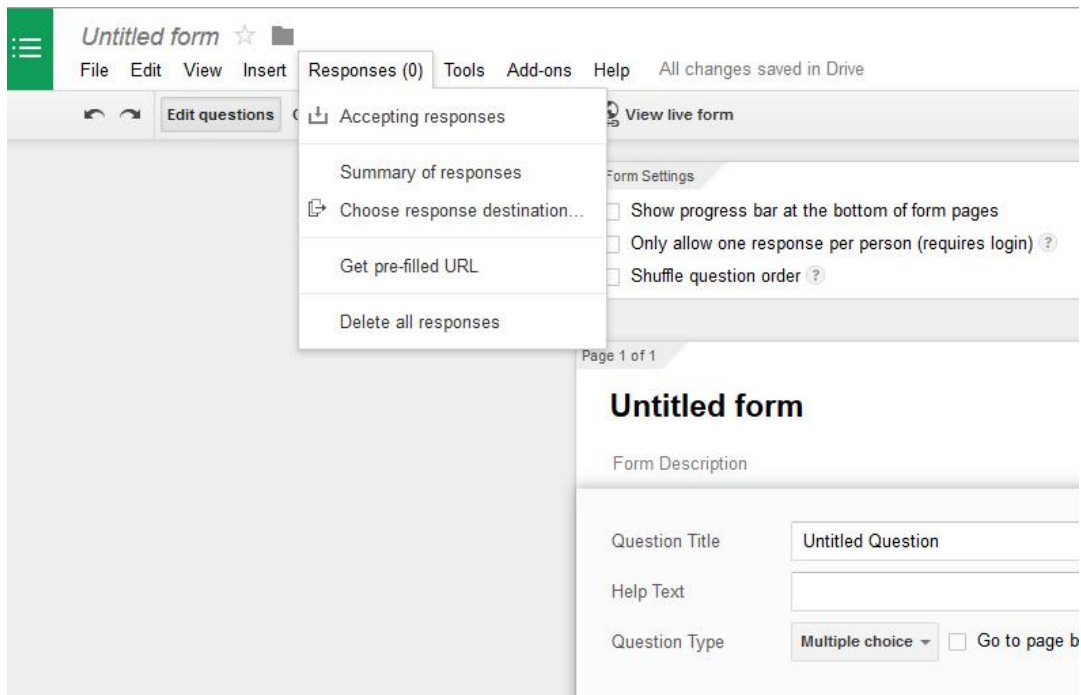
**Figure 4/18. Options of the View menu**

The options of the *Insert* menu helps adding the above-mentioned types of questions and are useful for editing layout, such as *Section Header*. This becomes important when the researcher would like to separate certain groups of questions and he would also like to highlight this separation for its importance in terms of logic or content. The option *Page Break* has the same function similarly to a regular text editing software. Inserting pictures and videos can be done by clicking on *Picture* or *Video*.



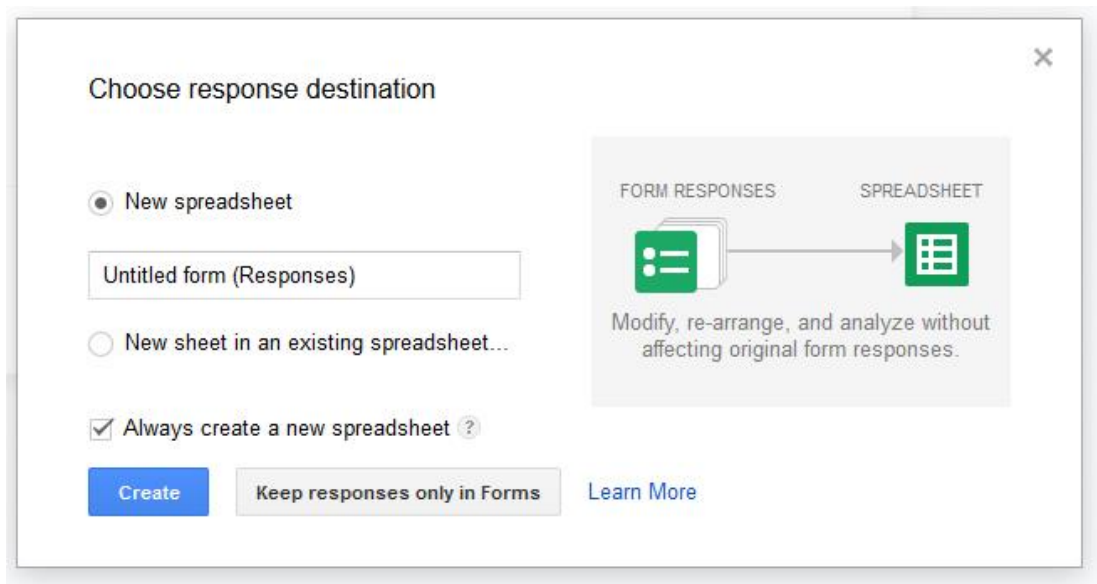
**Figure 4/19. Options of the *Insert* menu**

The researcher can work with the collected answers by clicking *Responses*. *Accepting responses* means that the questionnaire is active, publishable, and the responses can be received. Clicking this menu can also inactivate responses (they will not be available for others until being reset).



**Figure 4/20. Options of the *Responses* menu**

By clicking on *Choose Response Destination* we can define whether the responses should be inserted into a new Excel-workbook, or into a new worksheet of an already existing workbook.



**Figure 4/21. Choosing the destination of responses**

By clicking *Delete All Responses* the researcher may test the same questionnaire on a new sample or group, and all old data will be deleted completely.

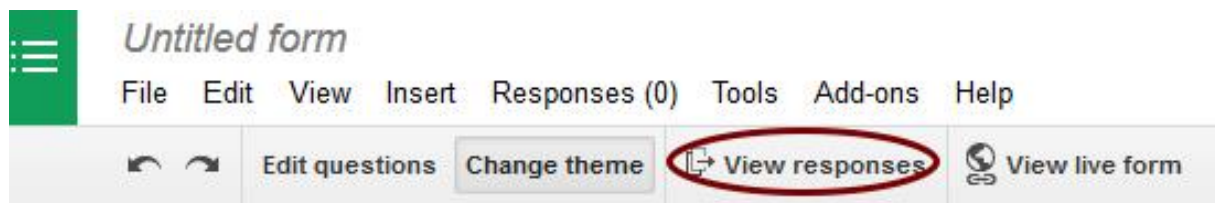
We are going to describe the option *Summary of Responses* in the following section.

Other options of the menu bar are *Tools* and *Help*: the first one provides a script<sup>6</sup> editor and a general editor, while the latter can be useful for reporting problems, describing keyboard shortcuts and accessing important information.

#### 4.4. Summary of responses

As mentioned above, the application can summarize received answers and results. However, we suggest that the researcher should complete filtering, analysis and further statistical tests on the database by himself with the help of the extracted Excel worksheet.

Collected answers and results can be viewed by clicking *See responses* – this way the application automatically provides the table filled with the data and options for further action.



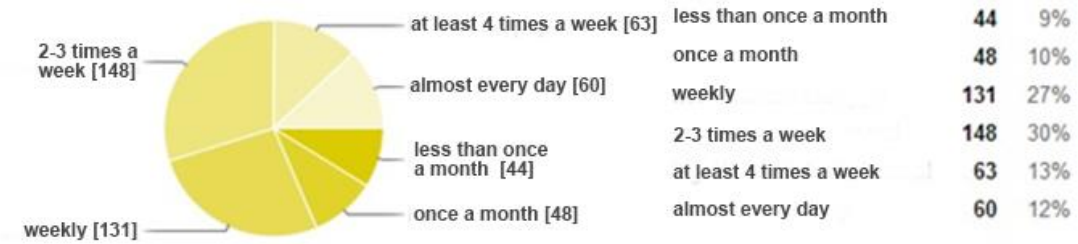
**Figure 4/22. View results**

Before describing the option for further actions, there is one more thing to look at: the *Summary* menu. This creates graphs and diagrams based on the responses of the questionnaire, although the researcher has no option to select or draw focus, so this feature is only useful for providing information but not for statistical analysis.

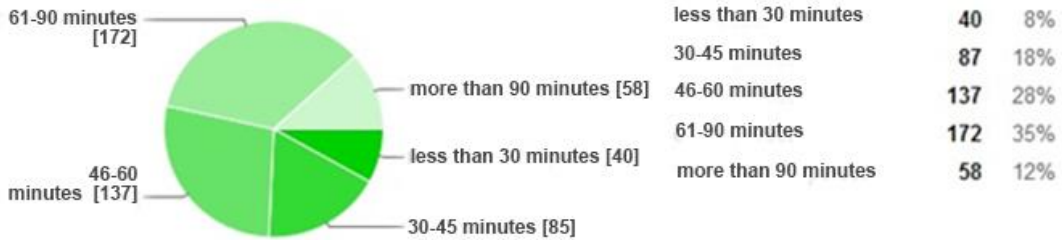
---

<sup>6</sup> **script**: a short programme often used for automatizing certain small tasks

**How often do you do sports since you are attending university (apart from obligatory PE lessons)?**



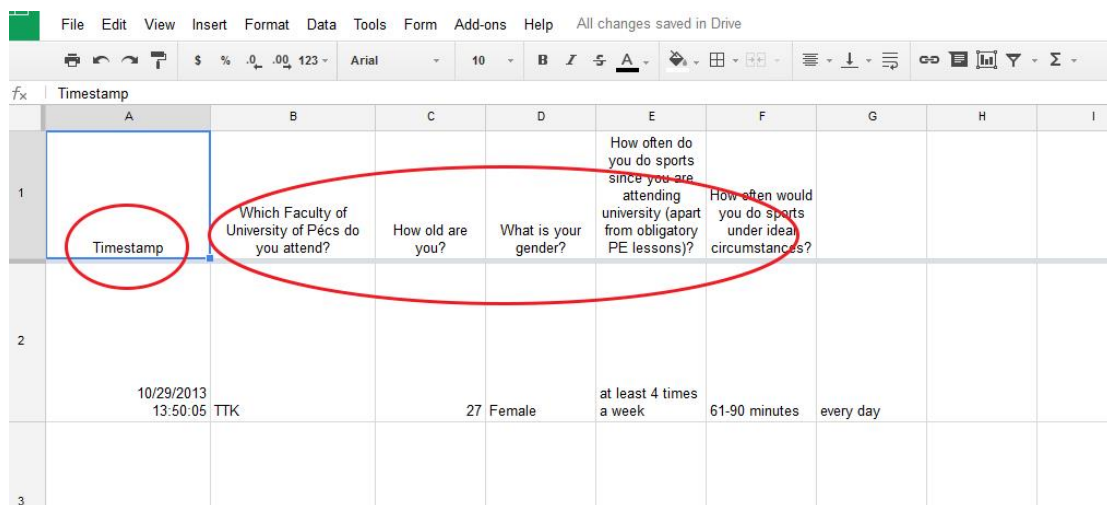
**How much time do you spend on doing sports at once?**



**Figure 4/23. Visual summary of results**

Options for further editing the responses can be found in the Excel-interface. In the same manner as before, we are only going to mention those features here that we find important in terms of synthesizing results and that are different from the features of basic software.

*Timestamp* appears automatically in the top left grey bar, providing information on the date and time of response. The top cells of the rest of the columns contain the question for which the results are given. Right-clicking these column titles and then *Sort* will list data alphabetically (in an ascending or descending order). The same task can be carried out by clicking *Data* → *Organize worksheet A-Z* or *Organize worksheet Z-A*.



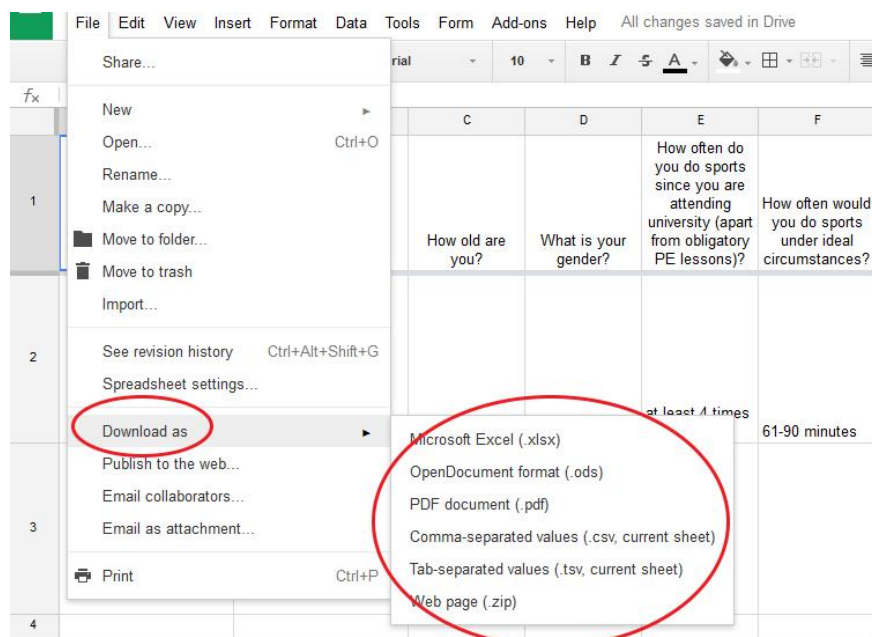
**Figure 4/24. Setting of the database**

*Save as* is an important item, by clicking on this the researcher can download the file in the chosen format and extension.

**The database can be downloaded as:**

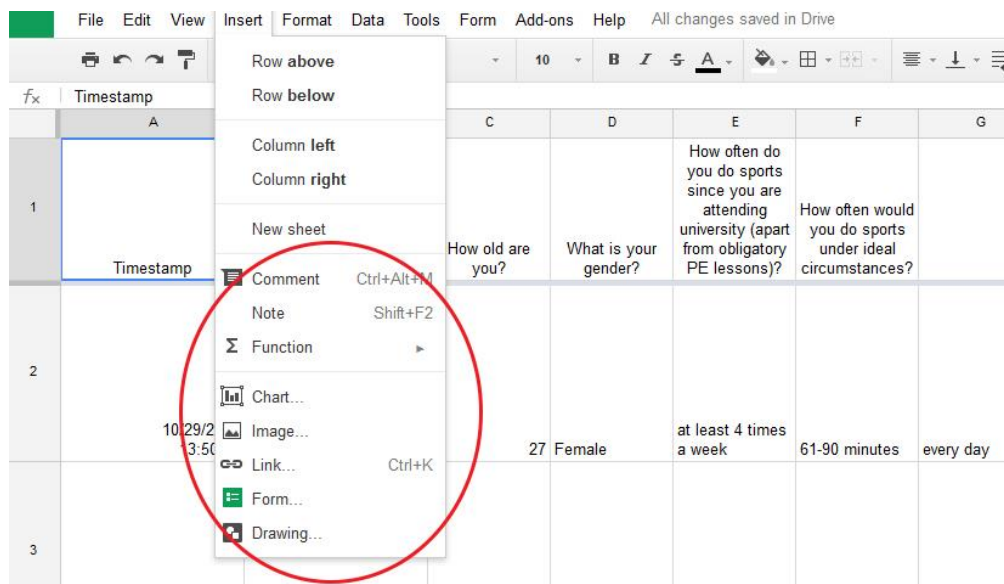
- Microsoft Excel format (.xlsx)
- OpenDocument format (.ods)
- PDF document (.pdf)
- Comma-separated values (.csv)
- Tab-separated values (.tsv)
- Web (.html)

For further editing and analysis we suggest choosing the Microsoft Excel (.xlsx) format.



**Figure 4/25. *Save as* menu**

It is also worth mentioning the *Insert* menu, which provides several options to insert extra items directly into the database, such as notes, separate notes, functions, etc. to separate cells.



**Figure 4/26. Insert menu**

The *Formatting* menu contains the same options for setting style and fonts as any other general text editing software.

Finally, another useful item can be accessed by clicking *Data* → *Filter*, as there is an opportunity also in the online database to apply various filters (e.g. to compare differences according to gender).

#### **4.5. Questionnaire editing – a practical example**

In the previous paragraphs we looked at the basic steps and options of questionnaire editing, and also discussed methodological principles required to assemble a professionally acceptable survey and extraction of a database that can form the basis of successful scientific research. After this we introduced Google Forms, its items and editing options. This chapter will describe the compilation of a short, real-life questionnaire, step by step.

The topic and field we would like to explore using a questionnaire is the changes (and reasons for these changes) in bodyweight during and after pregnancy.

Questions have been chosen to represent most items listed in the methodological description and also to fit principles described there, and thus our aim is to show a realistic situation. Our questionnaire contains 14 questions, in the following order:

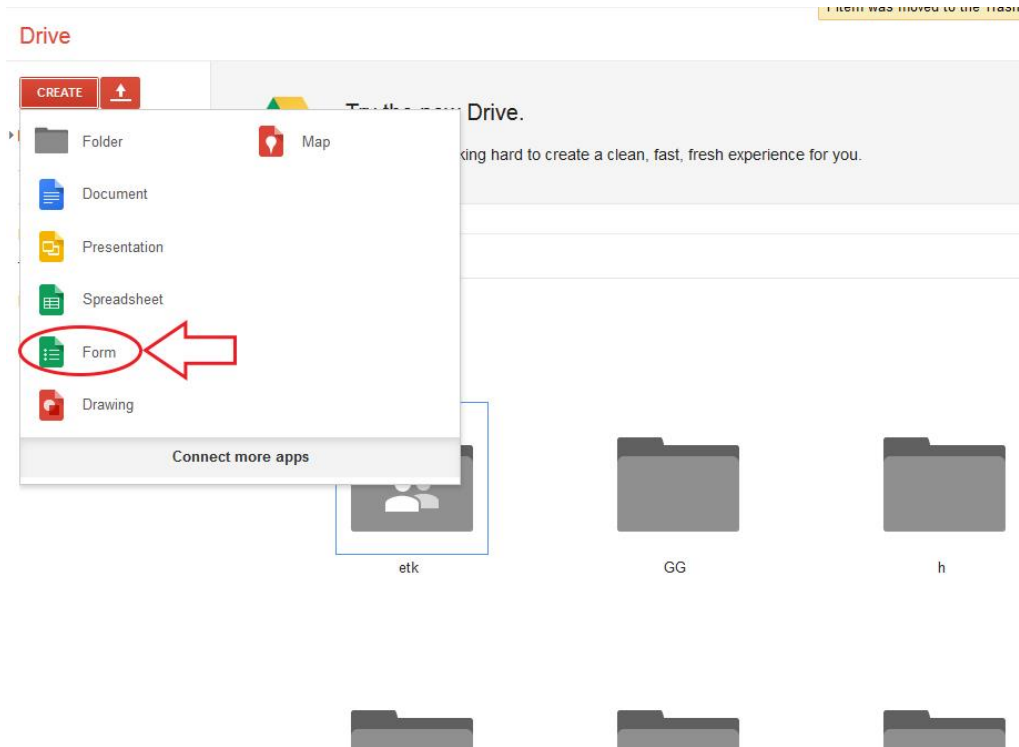


- Code of patient
- Age
- What is your highest level of qualification?
- Where are you from?
- Body height?
- Body weight before labour?
- Body weight on the day of labour?
- Body weight of the baby born?
- Body weight 6 months after labour?
- Do you consider your knowledge on healthy nutrition to be up-to-date?
- On which week of pregnancy was your child born?
- Did you suffer from a chronic illness?
- Type of labour?
- Pains during labour (on a scale of 1-5)?

We have listed the questions in the following order, and now let us edit the questionnaire using Google Forms.

The first step is to access our Google account and continue to the Drive application, as described in the previous chapter.

Reaching the application Drive, we should create a new form according to the previously described process, by clicking *Create* and choosing *Form* from the drop-down list.



**Figure 4/27. Creating a new form in Drive  
(as described previously)**

When we are ready with this step and the editing interface of the form can be viewed, we should first choose a general topic and insert the title of the questionnaire which, in this case, will be “Changes in weight during and after pregnancy”.

The first question is the code of the patient, which was sent already to the respondents involved. Our only task is to click on *Add item* and insert a short question where the respondents will be able to provide this information. Bars are filled according to the figure below, and the type of the question will be given under the heading of *Text* as already mentioned. If this information is a crucial one in the questionnaire, we may tick the box *Obligatory question* (this step should be taken in every necessary case; this will not be mentioned again).

Page 1 of 1

### Changes in body weight during and after pregnancy

Form Description

Question Title: Patient code

Help Text:

Question Type: Text

Their answer

Advanced settings

Done  Required question

Add item

**Figure 4/28. Questionnaire item to edit code of patient**

When we have set and edited everything on the screen depicted above, by clicking the blue “Done” icon we have completed our first question.

Other questions can always be added by clicking *Add Item*, (step not highlighted again in the following description).

Our second question is age. This is a statistically important continuous variable, but as this feature differs in case of all our respondents, listing all possible options for the number of years would not make any sense, and therefore we are going to use the same type of question as in the previous case. Bars are going to be filled according to the figure below and type of the question is again *Text*. Also, the question can be finalized again by clicking on *Done* (this is an obligatory step for each question).

Page 1 of 1

### Changes in body weight during and after pregnancy

Form Description

Question Title: Your age:

Help Text:

Question Type: Text

Their answer

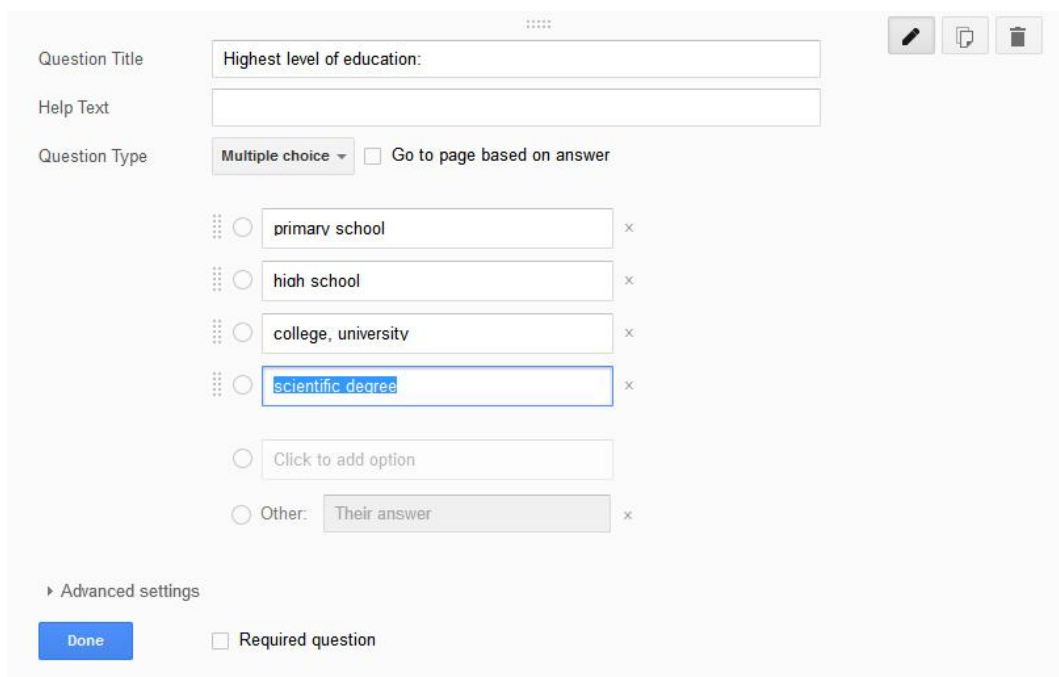
Advanced settings

Done  Required question

Add item

**Figure 4/29. Questionnaire item to edit age**

The next question refers to the highest level of qualification indicated by the respondent, and we would like to distinguish between five options: primary school, high school diploma, degree from a National Qualification Register programme, advanced vocational qualification, and finally college and university. In this case, we are applying a *Multiple Choice* question as we would like our respondents to provide only one answer. We may also apply the *Choose from a list* option by adding possible answers, as the methodology of answering is the same in the case of these two types of questions. We are going to describe the *Multiple Choice* option in our example, and the method for filling the bars is shown by the figure below.



The screenshot shows a question editor interface. At the top, there are three icons: a pencil, a document, and a trash can. Below these are three input fields: 'Question Title' with the text 'Highest level of education:', 'Help Text' (empty), and 'Question Type' with a dropdown menu set to 'Multiple choice' and a checkbox for 'Go to page based on answer' which is unchecked. Below the question type are five radio button options, each with a text input field and a small 'x' icon to its right. The options are: 'primary school', 'high school', 'college, university', 'scientific degree' (which is highlighted with a blue border), and 'Click to add option'. Below these is an 'Other:' option with a text input field containing 'Their answer'. At the bottom left, there is a blue 'Done' button and a checkbox for 'Required question' which is unchecked. A link for 'Advanced settings' is also visible.

**Figure 4/30. Item for the highest level of qualification**

The next question concerns the place of living. In this case we provide two options: city or village. As we framed the question as “Where are you from?” and do not intend to make any confusion, we add an extra help-box for the question which will help clarify what the questions actually focuses on. We would like the respondents to give only one answer so, similarly to the previous question, we can apply a *Multiple Choice*- or *Choose from a list* question. For better demonstration and considering the number of answers we are going to choose the latter option this time.

The screenshot shows a questionnaire editor interface. At the top right, there are three icons: a pencil (edit), a document (copy), and a trash can (delete). The main form has the following fields:

- Question Title:** "Where are you from?"
- Help Text:** "What type of settlement do you currently live in?"
- Question Type:** A dropdown menu set to "Choose from a list" and a checkbox for "Go to page based on answer" which is unchecked.
- Options:** A list of three options:
  1. "village" (with a delete 'x' icon)
  2. "city" (with a delete 'x' icon)
  3. "Click to add option" (with a plus icon)
- Advanced settings:** A section with a "Done" button and a checkbox for "Required question" which is unchecked.

**Figure 4/31. Questionnaire item to edit place of living**

As questions 5, 6, 7, 8, 9 all refer to continuous variables such as weight and height, we would like to demonstrate editing all these questions with a specific example, as in each case we need to apply *Text* questions due to answers being continuous (editing is demonstrated for questions 1 and 2). However, it is important to mention that for continuous variables the measurement should also be indicated in the help box to avoid the pitfall of diverse interpretations (for example cm instead of meter, etc.) of the question and, consequently, the resulting data.

The screenshot shows a questionnaire editor interface. At the top right, there are three icons: a pencil (edit), a document (copy), and a trash can (delete). The main form has the following fields:

- Question Title:** "Your height"
- Help Text:** "cm"
- Question Type:** A dropdown menu set to "Text".
- Placeholder:** A dashed box containing the text "Their answer".
- Advanced settings:** A section with a "Done" button and a checkbox for "Required question" which is unchecked.

**Figure 4/32. Questionnaire item to edit height**

Question 10, inquiring whether mothers consider their knowledge on healthy nutrition to be up-to-date, the options of “yes” and “no” are offered. In this case, we may apply either a *Multiple Choice* or a *Chose from a list* type of question, as similarly to questions 3 and 4 we would like the respondent to provide only one answer.

The screenshot shows a questionnaire editor interface. At the top right, there are three icons: a pencil (edit), a document (copy), and a trash can (delete). The main form has the following fields:

- Question Title:** "Do you consider your knowledge on healthy nutrition to be up-to-date?"
- Help Text:** An empty text box.
- Question Type:** A dropdown menu set to "Multiple choice". To its right is a checkbox labeled "Go to page based on answer".
- Options:** Two radio button options are visible. The first is "yes" and the second is "no". Each option has a small 'x' icon to its right. Below these is a third option: "Click to add option" with a radio button and the text "or Add 'Other'" to its right.
- Advanced settings:** A section with a right-pointing arrow and the text "Advanced settings".
- Buttons:** A blue "Done" button and a checkbox labeled "Required question".

**Figure 4/33. Questionnaire item to edit to edit knowledge on healthy eating**

Question 11, which says “On which week of pregnancy was your child born?” also refers to a continuous variable, and thus we need to apply the previously described (for questions 5, 6, 7, 8, 9) short, text-based question type the details of which can be checked above. The figure below demonstrates the information in connection with the method of filling the section in.

The screenshot shows a questionnaire editor interface. At the top right, there are three icons: a pencil (edit), a document (copy), and a trash can (delete). The main form has the following fields:

- Question Title:** "On which week of pregnancy was your child born?"
- Help Text:** An empty text box.
- Question Type:** A dropdown menu set to "Text".
- Preview:** A dashed box containing the text "Their answer".
- Advanced settings:** A section with a right-pointing arrow and the text "Advanced settings".
- Buttons:** A blue "Done" button and a checkbox labeled "Required question".

**Figure 4/34. Questionnaire item to edit week of labour**

Questions 12 and 13, which asks the following: “Did you suffer from a chronic illness (“yes” or “no”) and Type of labour (“vaginal” or “caesarean”)” – both provide two options for answers, so we proceed similarly to questions 3, 4 and 10, applying a *Multiple Choice*-type question. If requested, the *Choose from a list*-type question can also be applied, as the methodology of choice is the same in these two cases. To demonstrate both, we will show the previous question as a *Multiple choice*-type, and the next one as a *Choose from the list*-type.

The screenshot shows a questionnaire editor interface. At the top right, there are three icons: a pencil (edit), a document (copy), and a trash can (delete). The main form has the following fields:

- Question Title:** A text input field containing "Did you suffer from a chronic illness?".
- Help Text:** An empty text input field.
- Question Type:** A dropdown menu set to "Multiple choice" with a small square icon to its right. Below it is a checkbox labeled "Go to page based on answer" which is unchecked.
- Options:** Three radio button options are listed:
  - Option 1: "Yes" (radio button is unselected).
  - Option 2: "No" (radio button is unselected, the text "No" is highlighted in blue).
  - Option 3: "Click to add option" (radio button is unselected).
 To the right of the third option is the text "or Add 'Other'".
- Advanced settings:** A section with a right-pointing arrow and the text "Advanced settings".
- Buttons:** A blue "Done" button and a checkbox labeled "Required question" which is unchecked.

**Figure 4/35. Questionnaire item to edit previous chronic illnesses**

The screenshot shows a questionnaire editor interface. At the top right, there are three icons: a pencil (edit), a document (copy), and a trash can (delete). The main form has the following fields:

- Question Title:** A text input field containing "Type of labour".
- Help Text:** An empty text input field.
- Question Type:** A dropdown menu set to "Choose from a list" with a small square icon to its right. Below it is a checkbox labeled "Go to page based on answer" which is unchecked.
- Options:** Three numbered options are listed:
  - Option 1: "1. vaginal" (radio button is unselected).
  - Option 2: "2. cesarean" (radio button is unselected, the text "cesarean" is highlighted in blue).
  - Option 3: "3. Click to add option" (radio button is unselected).
- Advanced settings:** A section with a right-pointing arrow and the text "Advanced settings".
- Buttons:** A blue "Done" button and a checkbox labeled "Required question" which is unchecked.

**Figure 4/36. Questionnaire item to edit type of labour**

The 14<sup>th</sup> and last question refers to how much pain the mother experienced during labour, which will be demonstrated through the *Scale*-type question, used for the first time in this study.

This item provides a scale by giving a from-to measure to be evaluated from 1 to 5 depending on how strong the pain the mother felt during labour was. The two end points are “not at all” and “completely”. The required settings are demonstrated by the figure below.

The screenshot shows a questionnaire editor interface. At the top right, there are three icons: a pencil (edit), a document (copy), and a trash can (delete). Below these are four input fields: 'Question Title' with the text 'How painful was your labour?', 'Help Text' with 'Please define the amount of pain you felt during labour?', 'Question Type' with a dropdown menu set to 'Scale', and 'Scale' with a range of '1' to '5'. Below the scale range, there are two input fields for labels: '1: not at all' and '5: absolutely'. At the bottom left is a blue 'Done' button, and at the bottom right is a checkbox labeled 'Required question' which is currently unchecked.

**Figure 4/37. Questionnaire item to edit pains during labour**

This is the end of editing the sample questionnaire. The following task is to send the questionnaire to the respondents, embedding the link, and receiving and analysing answers.

We have not showed the completed questionnaire items for demonstration as the steps of editing had a priority, the following pages, however, will show the final and complete questionnaire.

Page 1/1.

**Title of the form**

**Changes in weight during and after pregnancy**

**Code of patient:**

**Your age:**

**What is your highest level of qualification?**

- primary school
- high school diploma
- degree from a National Qualification Register programme



- advanced vocational qualification
- college or university

**Where are you from?**

Where do you currently reside?

**Your height: cm**

**Your body weight before labour: kg**

**Your body weight on the day of labour: before the labour, kg**

**The body weight of your baby at birth: kg**

**Your body weight 6 months after giving birth: kg**

**Do you consider your knowledge on healthy nutrition to be up-to-date?**

- yes
- no

**On which week of pregnancy was your child born?**

**Did you suffer from a chronic illness? (**

- yes
- no

**Type of labour:**

**How much pain did you experience during labour:** Please define the strength of pain you felt during labour!

1 2 3 4 5

---

nothing at all      completely

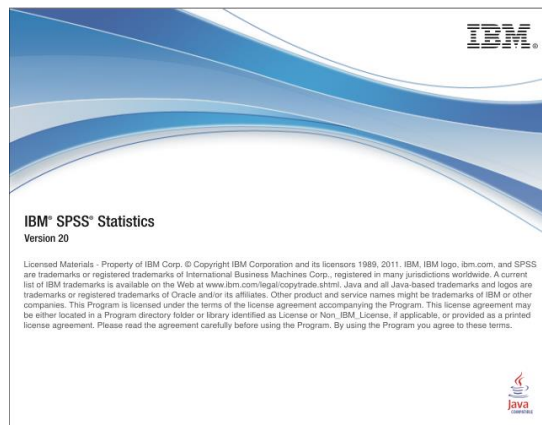
---

## **5. THE SPSS USER INTERFACE, IMPORTING AND EXPORTING DATA, BECOMING FAMILIAR WITH MENUS (Pongrác Ács)**

### **5.1. The SPSS user interface**

The software SPSS (Statistical Package for the Social Sciences) - with American and Canadian roots - was first introduced around 1968. At the moment, it is probably the most renowned market-leading statistical software in the world, partly owing to its early introduction. It was basically aimed to support research carried out in the field of social sciences but it has become a basic tool of scientific research and has also been integrated into higher education curricula. Its additional advantages include its user-friendly interface and the proper alignment and visualization of statistical analysis. As it is a programme created to carry out statistical analysis, it poses no difficulties to handle and work with mass data – not like in the case of using Excel. The programme is not available in Hungarian but as it is clearly arranged, the user does not need to have a very high command of English. There are numerous setting options but it is easy to get accustomed to its usage very quickly. The teaching and help modules are easy to use and are very detailed. Its graphics are not very spectacular but offer a lot of options. Until 2005, Hungarian higher education had had the right to use the software – with a data limit and only for education and research purposes – free of charge. In the autumn of 2005, however, this system was terminated and completely restructured but it is still available for universities at a reduced price. A new version of SPSS becomes available almost every year with the same basis and with some modifications in the design or with separate methodologies integrated. One of its most important advantages is that the databases created by other programmes and saved in different formats (Excel, dBase, Lotus, SAS, Stata, etc.) can be converted, opened and saved by SPSS. One can find more data on the software and the company itself on webpage [www.spss.com](http://www.spss.com) and [www.spss.hu](http://www.spss.hu).

From now on, we are going to use version SPSS 20, so statistical analyses will be introduced via its interface. Again, we would like to emphasize that these options are available in all the versions of the programme; one may find differences only in the menus.



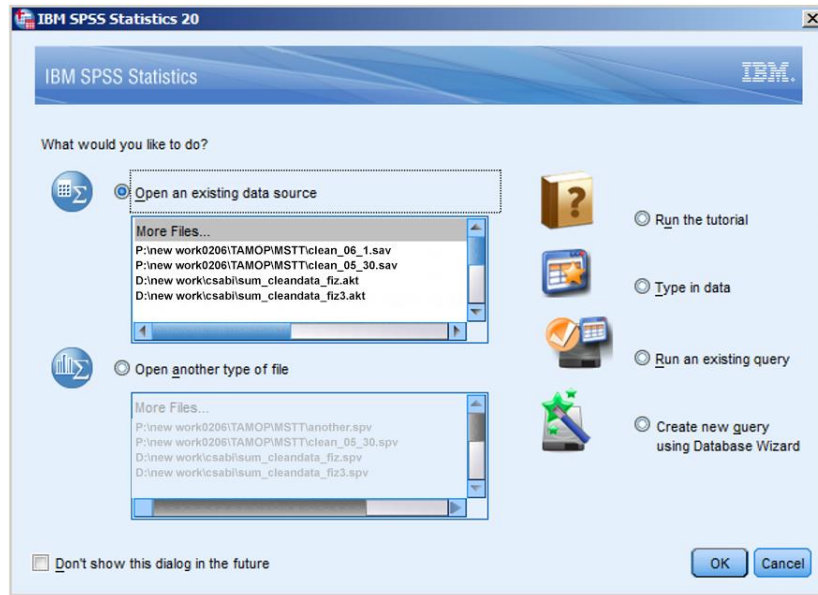
**Figure 5/1. Information received when logging in**

SPSS was created by a multi-windows technique, similarly to Windows-based programmes. This means that it handles the database and the output in a separate window, and there is a different window for editing syntaxes which is the internal language of SPSS, which help users store and run commands. This book does not concentrate on the language and the commands but one can read about these options in more detail in the book entitled “Túlélőkészlet az SPSS-hez” [A survival kit for SPSS] (Székelyi – Barna 2005). This book introduces the basic statistical methods via the description of menus (*File, Edit, Data, Analyse*), dialogue boxes and options. We will first make an attempt to introduce the most important options in the menu – without aiming at being exhaustive due to text length limits. The options not mentioned here can be checked in the *HELP* menu of SPSS.

At the start, SPSS offers four options to take the first step:

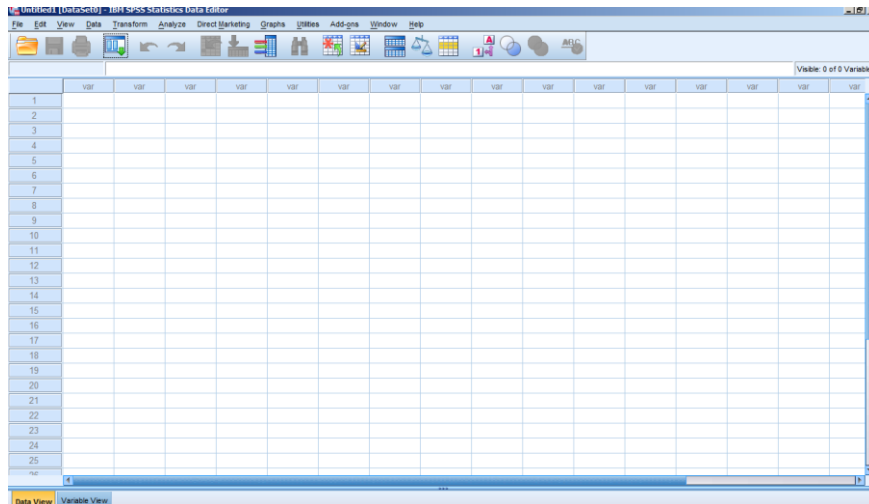
- Run the tutorial
- Type in data
- Run an existing query
- Open an existing data source.

By choosing the option *Open an existing data source* one can open a database used before. The file extension of our databases used in the current version of SPSS is \*.sav, while the *OUTPUT* created by SPSS has the file extension \*.spv.

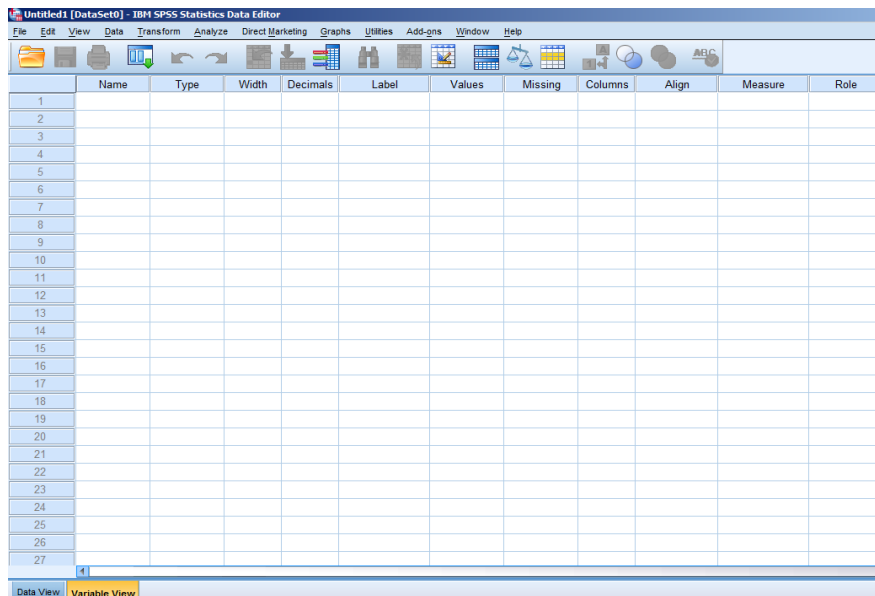


**Figure 5/2. Dialog box after logging in**

If one clicks on *Cancel*, a new blank database will appear, the menu structure of which is similar to the ones we know from Microsoft Office software. The most well-known processes (e.g. cut, insert, copy, etc.) are found and used here, too. One can find two types of view (*DATA VIEW* and *VARIABLE VIEW*) in the *DATA EDITOR* module. One can switch from one particular type to another by clicking on their name in the left bottom corner of the window, or by pressing CTRL+T. Uploading primary data into the database can be carried out in *DATA VIEW* (see Figure 5/3a.), similarly to Excel. The database contains records in rows (*CASES*) and columns (*VARIABLES*). It can mean that one item in the population is represented by the rows, and the variables are the columns. One can switch from one type of view to the other by clicking on their name in the left bottom corner of the window, or by pressing CTRL+T.



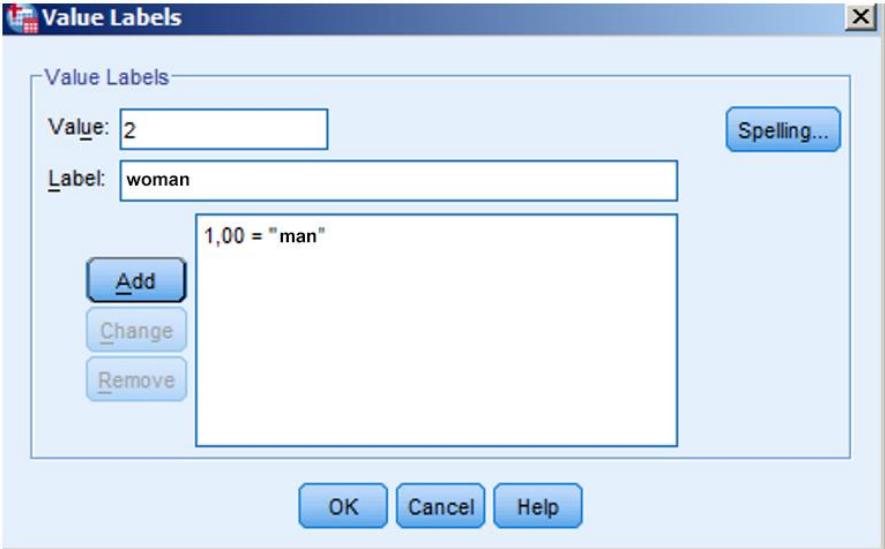
**Figure 5/3a. Data view**



**Figure 5/3b. Variable view**

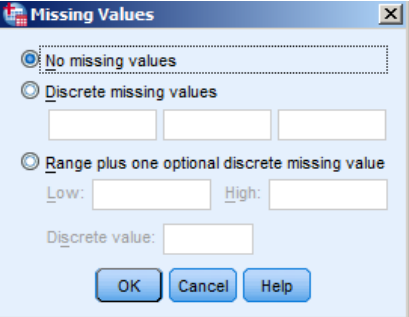
In variable view (See Figure 5/3b.), one can add, change, and edit parameters of the attributes. The rows denote variables, while the columns denote the parameters of the variables. *NAME*: the shortened version of the name of the variable. It starts with a letter and can be a maximum of 64 characters long. It is recommended to avoid accents and special characters in names. If we do not define the variable, the software will do it automatically, beginning with VAR0000, and increasing in number. For transparency reasons, it is useful to give short names that can be described in more detail in the *LABEL* option. In the column labelled *TYPE*, we can define the type of the variable. They need to be chosen one by one, by clicking on the right corners of the cells. The most common types are *NUMERIC* and *STRING* but one can also choose from date, financial and other options as well. In the column *WIDTH*, the number of characters are added, which can be a maximum of 255 in case of string, and max. 40 for

numbers. In the column *DECIMALS*, the number of decimals has to be indicated. In the column *LABEL*, we can define the meaning of the variable in more details. This information will also be displayed under the heading *OUTPUT*. Here, we can even type in the questions of the survey. In the column *VALUES*, we can give a title to nominal and ordinal data. E.g. naming the 1-5 values of the Likert scale. If one particular question was related to gender, value labels can be named here.



**Figure 5/4. Naming and labelling values**

In the cell labelled *VALUE*, a specific value can be added, while in the cell *VALUE LABEL*, we can name and label the value. After adding a name, press *ADD*. The remarks made can be changed by pressing *CHANGE*, and press *REMOVE* if you want them to disappear. In the column *MISSING*, missing values can be defined after adding the appropriate constraints. It is important to indicate missing or wrong data in order to get a valid statistical process and output. We have to define values that cannot be included.

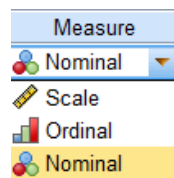


**Figure 5/5. Missing values**

The dialog box offers three options:

- *No missing values*: it means that no missing values can be added. In this case, a point appears at the place of the missing values. If we have no special reason to define it differently, it is recommended we use this default setting.
- *Discrete missing values*: here, we can add a maximum of three specific values, or missing data values (999 no answer).
- *Range plus one optional discrete missing value*: we can determine an interval or discrete data.

With *COLUMNS* we can set the displayed column width in *DATA VIEW*. The column *ALIGN* can be used to adjust the values in the cells to right, left or centre. In the column *MEASURE*, one has to choose the types of measures (scale, ordinal, nominal) already introduced above (See Figure 5/6.). The software does not differentiate between interval and ratio scales; it handles them equally as *SCALE*.



**Figure 5/6. Choosing the type of measure**

The *OUTPUT VIEW* in SPSS is applied to display, save and edit outputs. This view has its own menu and tools. There are two parts in display. The left-hand side contains a structure-tree with analysis, while the right-hand side contains tables and figures of output data. Clicking twice on the table or figure, we can format them individually. In order to copy the objects containing the outputs, press (CTR+C) to copy and (CTR+V) to paste.

## 5.2. Importing data

There are two ways of importing data in SPSS:

- *Primary data import*: write the data into the cells. First, define the type of variables in variable view, and then type data in the data view.
- *Secondary data import*: importing already existing data into SPSS.

We will now show the essence of *primary data import* through a simple example. Let us assume that one has to add data of new patients at a hospital department, based on some highlighted attributes (gender, age, residence, systolic blood pressure at the time of admission, diastolic blood pressure at the time of admission).



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	sex	Numeric	8	2	patient's name	{1,00,man}...	None	8	Right	Nominal	Input
2	age	Numeric	8	2	patient's age	None	None	8	Right	Scale	Input
3	residency	Numeric	8	2	patient's residency	{1,00, megy...	None	8	Right	Ordinal	Input
4	diast	Numeric	8	2	diastole blood pres	None	None	8	Right	Scale	Input
5	systol	Numeric	8	2	systole blood pres	None	None	8	Right	Scale	Input
6											
7											
8											
9											
10											
11											
12											
13											
14											

**Figure 5/7. Defining variables for primary data import**

When defining parameters (first step), we name the variables “nem” (i.e. gender) and “lakhely” (i.e. residence) (*VALUES*). In the case of gender, we coded male as 1 and female as 2. There are three categories of residence, based on the number of inhabitants (1: “megyeszékhely”, i.e. county seat, 2: “város”, i.e. town, 3: “falv”, i.e. village). In this case we did not define missing values. Gender and residence are discrete variables, the measures of which are nominal (gender) and ordinal. The other continuous measures can have infinite variable, so we choose ‘scale’ to be the type of measure.

Once this has been done, switch to *DATA VIEW* and fill in the records.

	sex	age	residency	diast	systol	var	var	var	var	var
1	1,00	47,00	1,00	144,00	89,00					
2	2,00	65,00	2,00	140,00	79,00					
3	1,00	72,00	2,00	162,00	88,00					
4	1,00	51,00	3,00	120,00	80,00					
5	2,00	60,00	2,00	160,00	90,00					
6	2,00	63,00	3,00	125,00	82,00					
7	1,00	66,00	3,00	135,00	85,00					
8	2,00	49,00	3,00	125,00	82,00					
9	2,00	55,00	1,00	138,00	88,00					
10	2,00	75,00	2,00	145,00	78,00					
11	1,00	80,00	3,00	170,00	98,00					

**Figure 5/8. Uploading data**

If one chooses *VALUE LABELS* in the menu called *VIEW*, then the labels we added before will replace the numbers.

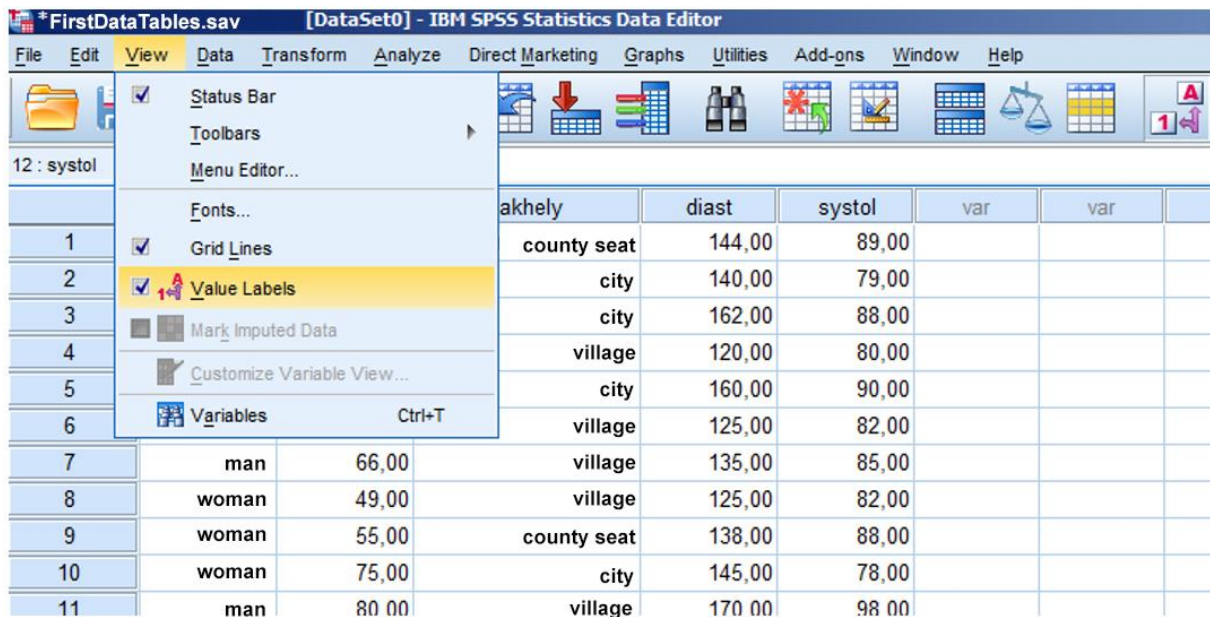


Figure 5/9. Choosing Value label option

If we want to import **secondary data**, an already existing database will be opened in SPSS. As mentioned before, not only databases saved by SPSS can be opened but also ones in other formats (typically Excel). If one would like to open an xls (Excel) file, the option *File*, then *Open Data*, then *File of type* has to be chosen where the proper format can be clicked on.

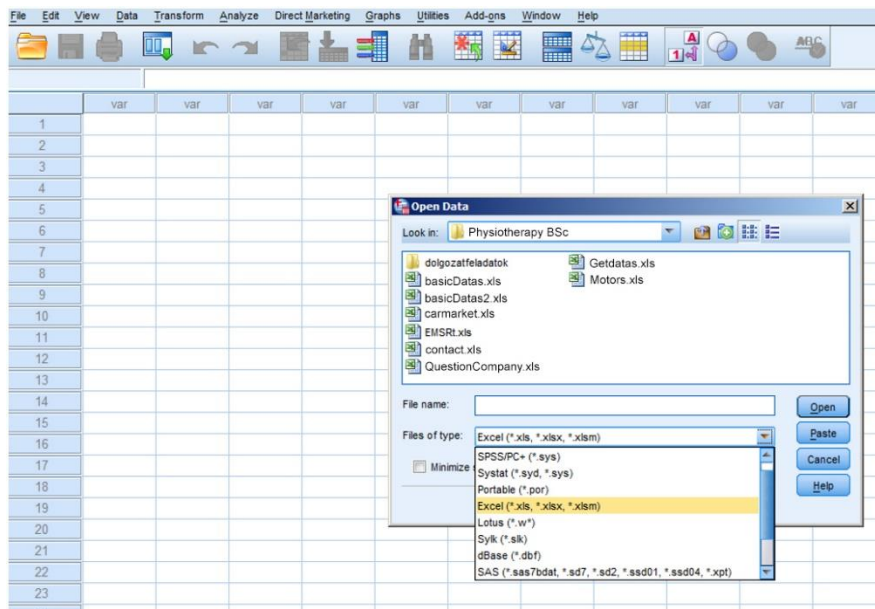
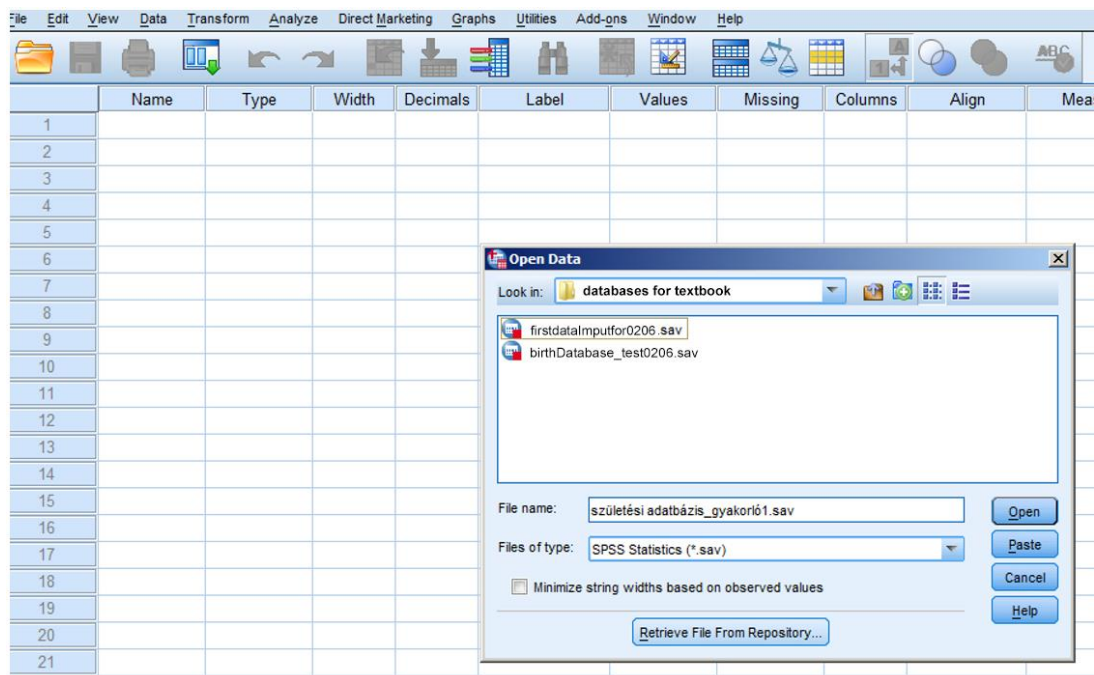


Figure 5/10. Settings of importing data

Before opening an Excel file, it is recommended to check whether the Excel programme is using the file or not. If it is, SPSS cannot open it. In order to use the database in both programmes at the same time, one has to open it first in SPSS and then in Excel. After choosing the file, it has to be decided the data of which sheet should be imported to SPSS. As default, the \*.sav extension of SPSS can be seen here.

Next, the database saved in \*.sav extension (születési adatbázis\_gyakorló1\_teljes\_67) has to be imported. We will use this database as an example from now on. The database is available on the webpage [www.etk.pte.hu](http://www.etk.pte.hu).



**Figure 5/11. Importing database születési adatbázis\_gyakorló1**

Pressing *OPEN* will import the database, and the properties of variables will be displayed in *VARIABLE VIEW*. The database contains 13 variables and the answers of 67 young mothers sixth month after giving birth. Data include the code of the patient (kód), age (év), education (isk), residence (lakhely), type of labour (szülés) and other related information.

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role	
1	K_code	Numeric	8	0	Patient code	None	None	8	Right	Nominal	Input
2	K_year	Numeric	8	0	How old are you?	None	None	8	Right	Scale	Input
3	K_school	Numeric	8	0	Highest level of education	{1, university...}	None	8	Right	Nominal	Input
4	K_residency	Numeric	8	0	Where are you from?	{1, city }...	None	8	Right	Nominal	Input
5	K_labour	Numeric	8	0	Type of labour	{1, vaginal}...	None	8	Right	Nominal	Input
6	K_lab_week	Numeric	8	2	On which week of pregnancy was your child born?	None	None	8	Right	Scale	Input
7	K_birth_weig	Numeric	8	2	Birth weight	None	None	8	Right	Scale	Input
8	K_wbef_preg	Numeric	8	2	Weight before pregnancy	None	None	8	Right	Scale	Input
9	K_hbef_preg	Numeric	8	2	Height before pregnancy	None	None	8	Right	Scale	Input
10	K_pday_weig	Numeric	8	2	Weight at labour's day	None	None	8	Right	Scale	Input
11	K_know	Numeric	8	0	Do you consider your knowledge on healthy nutritio	{1, Yes }...	None	8	Right	Nominal	Input
12	K_Chronical	Numeric	8	0	Did you suffer from a chronical illness?	{1, Yes }...	None	8	Right	Nominal	Input
13	K_w6months	Numeric	8	2	Weight six months after labour	None	None	8	Right	Scale	Input

Figure 5/12. Parameters of variables in database születési adatbázis\_gyakorló1\_teljes\_67

### 5.3. Becoming familiar with menus

This chapter provides a shortcut to the most important menus available.

**File:** the standard tasks such as opening, saving and renaming new and existing databases (data, output, text file) can be carried out here. The option *DISPLAY DATA FILE INFORMATION* provides a summarizing table on the *OUTPUT* about the variables of the file saved in \*.sav format. Figure 5/13 shows variable information on the currently used (születési adatbázis\_gyakorlól1\_teljes\_67.sav) database, owing to the option *WORKING FILE*.

The screenshot shows the SPSS Statistics Viewer interface. The main window displays a table titled "Variable Information" with the following data:

Variable	Position	Label	Measurement Level	Role	Column Width	Alignment	Print Format	Write Format
K_year	1	How old are you?	Scale	Input	8	Right	F8	F8
K_school	2	Highest level of education	Nominal	Input	8	Right	F8	F8
K_residecy	3	Where are you from?	Nominal	Input	8	Right	F8	F8
K_labour	4	Type of labour	Nominal	Input	8	Right	F8	F8
K_lab_wee	5	On wich week when the child was born?	Scale	Input	8	Right	F8.2	F8.2
K_birth_we	9	Birth weight	Scale	Input	8	Right	F8.2	F8.2
K_wbef_pr	7	Weight before pregnancy	Scale	Input	8	Right	F8.2	F8.2
K_hbef_pr	8	Height before pregnancy	Scale	Input	8	Right	F8.2	F8.2
K_know	10	Do you consider your knowledge on healthy nutriti	Nominal	Input	8	Right	F8	F8
K_chronical	11	Do you suffer from chronicl illness?	Nominal	Input	8	Right	F8	F8
K_w6monts	12	Weight six months after labour	Scale	Input	8	Right	F8.2	F8.2

Variables in the working file

Figure 5/13. Concluding meaning of variables of the database

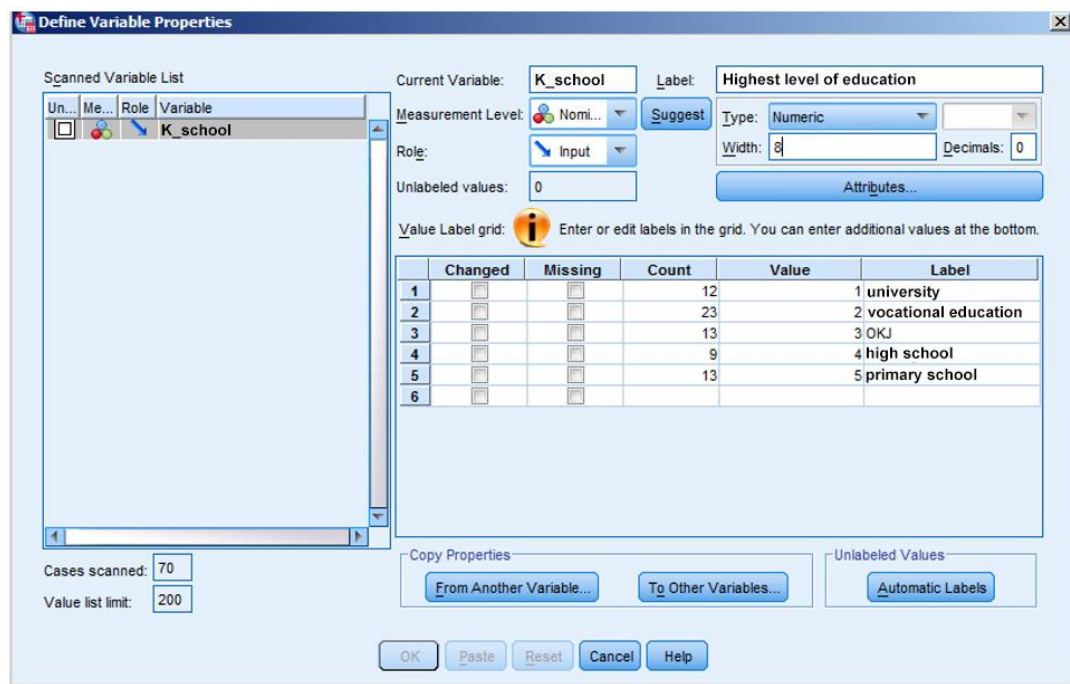
The summary provides information on the name of variables (*VARIABLE*), *POSITION*, *LABEL*, *MEASUREMENT LEVEL*, etc.

**Edit:** here you can find options for editing data. The options we know from Microsoft Office such as : copy, paste, cut, delete, search, etc. are also available here. The option *INSERT VARIABLE* makes it possible to insert a new variable (column), while the *INSERT CASE* means inserting a new case (row). We can set windows and variables under the menu item

labelled *OPTIONS*. The most important one is *GENERAL* settings where we can decide if the outputs should be displayed in the form of the name (*DISPLAY NAME*) or the complete meaning (*DISPLAY LABELS*) of the variable. It is recommended to apply the *DISPLAY LABELS* option if you do not know the correct meaning of the variable.

**View:** with the help of the settings we can calibrate the active window suiting our needs. The user can here display the icons and labels most often used.

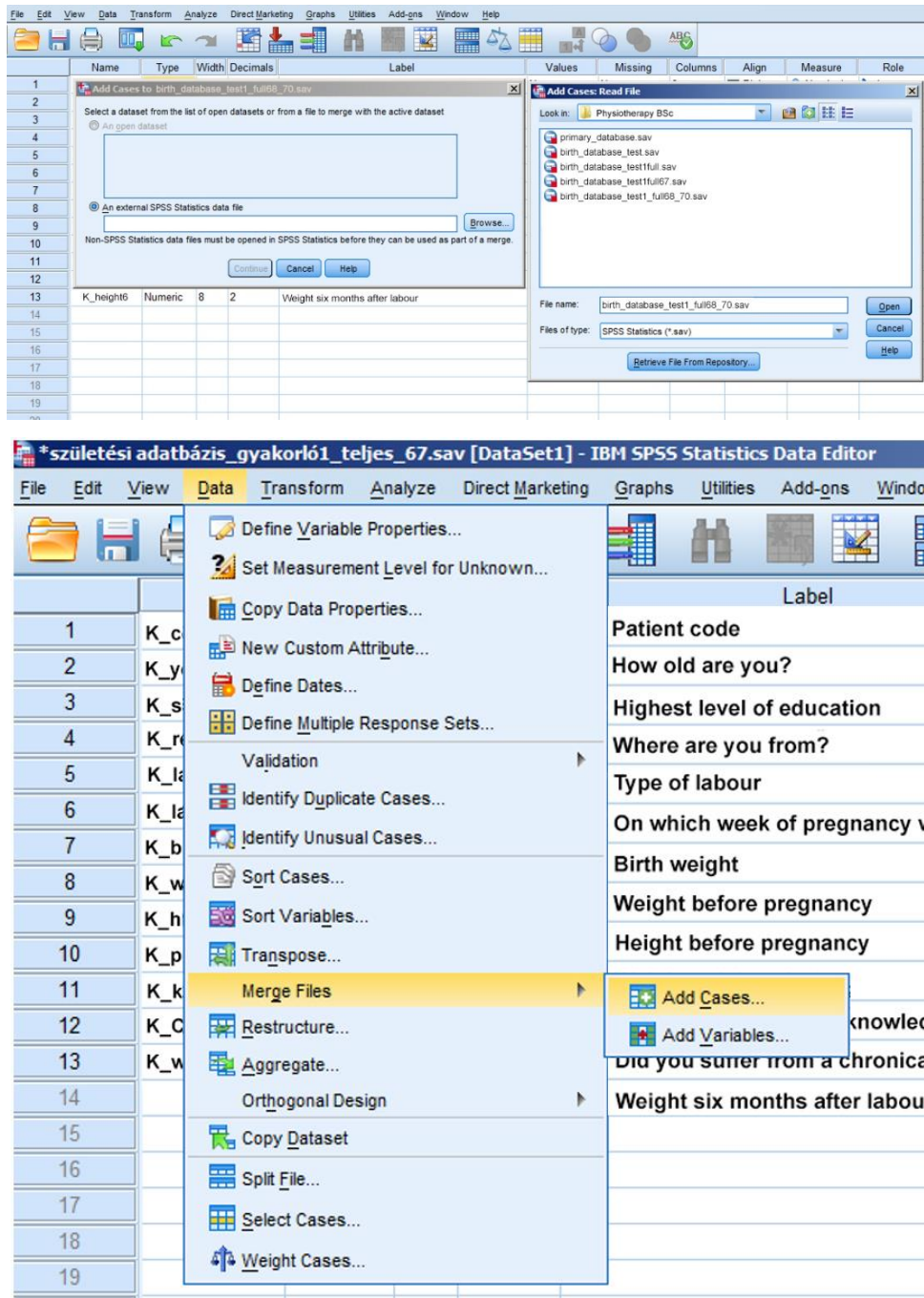
**Data:** Supports the user to handle and structure data. With the option *DEFINE VARIABLE PROPERTIES* we can determine the values and the number (*COUNT*) of the variables.



**Figure 5/14. Defining variable properties**

*SORT CASES* can sort cases one or more variables in an ascending or descending order. *TRANSPOSE* allows the switching of columns and rows in the database. *MERGE FILES* can merge and extend several databases or survey cases and variables. If the variables are the same –only cases have to be added to the database –choose the *ADD CASES* option. If the cases are the same but you want to add new variables, the new variables will need to be added, not the cases.

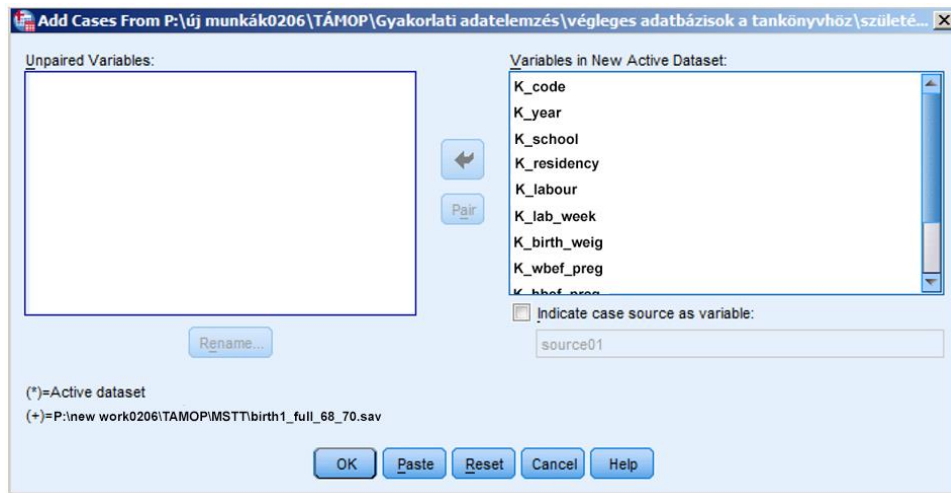
**For practice**, let us add data of three young mothers to the database (születési adatbázis\_gyakorló1\_teljes\_68\_70).



**Figure 5/15. Merging databases (cases)**

As the variables of the three young mothers are the same, we choose the *ADD CASES* option. The database we need to add the information to will appear in the cell labelled *FILE NAME*. Then, the common variables of the databases will appear in the window<sup>7</sup>.

<sup>7</sup> It is very important that the names of the variable have to be exactly the same. If there is a single character difference, the software cannot merge them.



**Figure 5/16. The common variables of the new database**

After pressing OK, the new common database will emerge, containing data of 70 young mothers.

	K_code	K_year	K_School	K_residency	K_labour	K_labweek	K_birth_w	K_wbef_p	K_hbef_p	K_pday_w	K_know	K_chronic	K_w6mont
46	46	37	OKJ	village	caesarean	39,00	3500,00	65,00	156,00	75,00	No	Yes	70,00
47	47	37	OKJ	village	caesarean	38,00	3300,00	64,00	155,00	76,00	No	Yes	74,00
48	48	38	vocational...	village	caesarean	38,00	3200,00	63,00	154,00	73,00	Yes	Yes	65,00
49	49	39	vocational...	village	caesarean	38,00	3400,00	62,00	154,00	75,00	Yes	Yes	70,00
50	50	33	vocational...	village	caesarean	38,00	3560,00	61,00	149,00	70,00	Yes	Yes	70,00
51	51	32	vocational...	city	caesarean	39,00	3460,00	60,00	166,00	75,00	Yes	Yes	73,00
52	52	22	vocational...	village	vaginal	40,00	3470,00	59,00	166,00	78,00	Yes	Yes	65,00
53	53	23	primary sc...	city	vaginal	37,00	3740,00	58,00	168,00	80,00	Yes	Yes	62,00
54	54	24	primary sc...	city	vaginal	37,00	3850,00	57,00	170,00	80,00	Yes	Yes	75,00
55	55	25	primary sc...	city	vaginal	38,00	3900,00	56,00	170,00	66,00	Yes	No	60,00
56	56	26	primary sc...	city	vaginal	39,00	4050,00	54,00	171,00	70,00	Yes	No	58,00
57	57	27	primary sc...	city	vaginal	39,00	2900,00	52,00	171,00	72,00	Yes	No	60,00
58	58	28	university	city	vaginal	36,00	2850,00	53,00	165,00	68,00	Yes	No	60,00
59	59	29	university	village	vaginal	37,00	2650,00	51,00	165,00	65,00	Yes	No	55,00
60	60	30	university	city	caesarean	37,00	2400,00	50,00	166,00	68,00	Yes	No	65,00
61	61	30	university	village	caesarean	38,00	2300,00	49,00	166,00	60,00	Yes	No	50,00
62	62	31	vocational...	village	caesarean	39,00	2550,00	48,00	165,00	60,00	Yes	No	105,00
63	63	32	vocational...	village	caesarean	40,00	2910,00	100,00	166,00	110,00	Yes	No	102,00
64	64	33	vocational...	village	caesarean	40,00	3000,00	110,00	168,00	112,00	Yes	No	111,00
65	65	29	high school	village	caesarean	40,00	3000,00	120,00	168,00	122,00	Yes	No	120,00
66	66	25	high school	village	caesarean	39,00	4000,00	119,00	169,00	123,00	Yes	No	121,00
67	67	26	high school	village	caesarean	39,00	4100,00	118,00	167,00	125,00	Yes	No	119,00
68	68	25	high school	village	caesarean	40,00	3900,00	99,00	168,00	102,00	Yes	No	100,00
69	69	27	OKJ	village	caesarean	39,00	3800,00	98,00	168,00	105,00	Yes	No	103,00
70	70	31	OKJ	village	caesarean	39,00	2770,00	65,00	165,00	78,00	Yes	No	70,00

**Figure 5/17. The new, “merged” database**

It is advised to save the new database with a different name (**születési adatbázis70.sav** in our example).

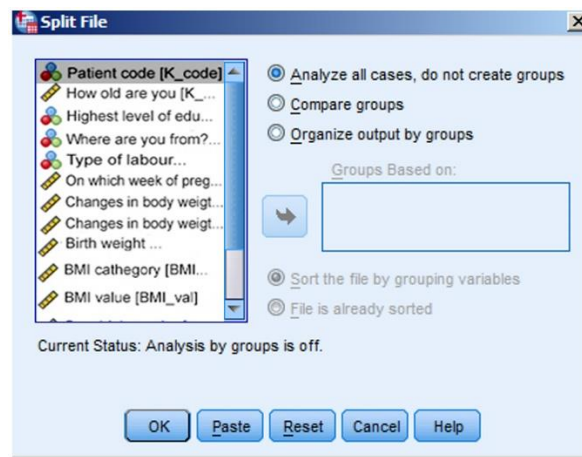
**AGGREGATE:** one can aggregate data based on a chosen function (sum, average, etc.).



**SPLIT FILE:** the database can be split into groups according to some important aspects. A statistical analysis can also be carried out for these groups. There are three options:

- Examining all cases, and not creating groups
- Comparing groups
- Displaying outputs by groups

Applying this option will display “*SPLIT FILE ON*” in the right bottom corner of the data view until we change the settings.

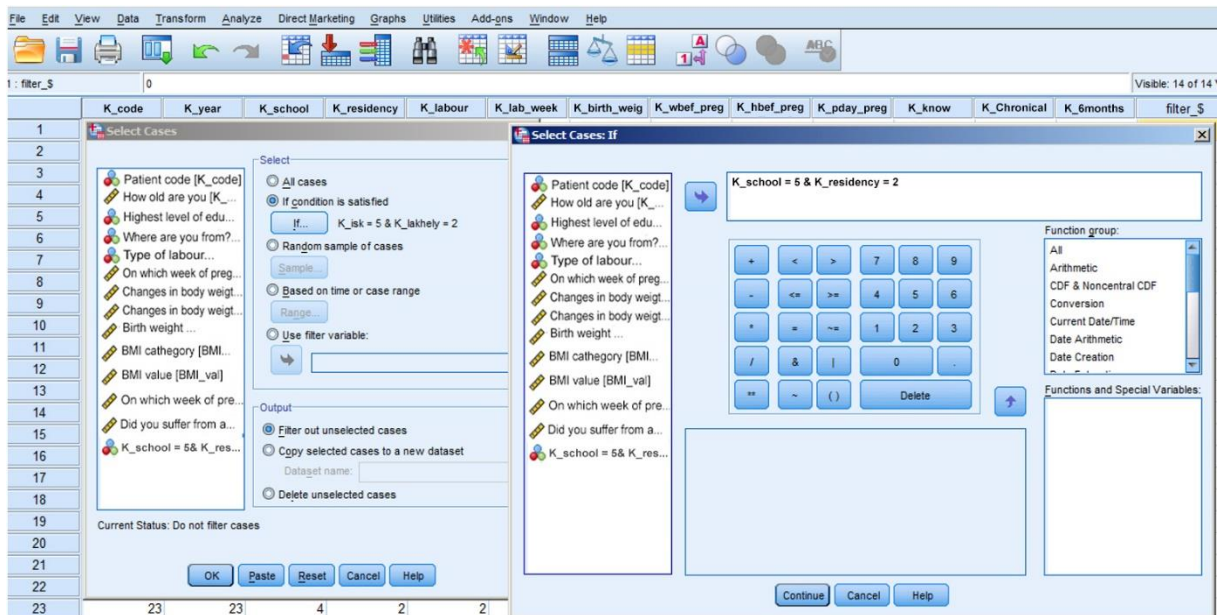


**Figure 5/18. Splitting data into groups**

**SELECT CASES FILE:** one can filter (select or exclude) cases, and analyse the remaining data. The sequence number of the improper cases or records will be assigned a black arrow, and the label “*Filter ON*” will appear. There are four setting options:

- Each case is involved in the analysis
- Only those that meet the requirement(s)
- Random sample of the cases
- Assigning particular requirements of the cases (labelling first and last cases)

**For practice,** let us filter those young mothers who live in a village and whose education level is higher than primary school. This means that we will only consider young mothers who meet these selection criteria.



**Figure 5/19. Filtering**

We selected cases in the dialog box according to education (K\_isk, selecting value 5) and residence (K\_lakhely, where village has the value 2).

Data view will display all data but crosses out the sequence number of improper cases with a diagonal black arrow.

	K_code	K_year	K_school	K_residency	K_labour	K_lab_week	K_birtheig	K_wbef_preg	K_hbef_preg	K_pday_wei	K_know	K_Chronical	K_w6months	filter_\$
1	1	40	1	1	2	38,00	2500,00	55,00	155,00	68,00	1	2	60,00	0
2	2	25	1	1	1	35,00	2600,00	58,00	160,00	62,00	1	2	60,00	0
3	3	23	2	1	1	36,00	3000,00	59,00	160,00	70,00	1	2	59,00	0
4	4	18	2	1	1	38,00	3200,00	56,00	164,00	70,00	1	2	56,00	0
5	5	22	2	1	1	37,00	3300,00	54,00	165,00	70,00	1	2	58,00	0
6	6	25	2	1	1	35,00	3400,00	57,00	165,00	71,00	1	2	59,00	0
7	7	26	2	1	1	34,00	3500,00	58,00	166,00	69,00	1	2	60,00	0
8	8	33	2	1	2	36,00	4000,00	59,00	166,00	70,00	1	2	59,00	0
9	9	35	2	1	2	37,00	3800,00	53,00	163,00	75,00	2	2	55,00	0
10	10	36	2	1	2	38,00	3500,00	60,00	163,00	78,00	2	2	62,00	0
11	11	24	3	1	2	39,00	3400,00	60,00	163,00	80,00	2	2	62,00	0
12	12	28	3	1	2	40,00	3300,00	60,00	162,00	85,00	2	2	80,00	0
13	13	29	4	1	2	41,00	2900,00	64,00	162,00	70,00	2	2	70,00	0
14	14	27	4	1	2	38,00	2200,00	65,00	162,00	75,00	1	2	75,00	0
15	15	19	2	2	2	35,00	2150,00	80,00	165,00	89,00	1	2	89,00	0
16	16	44	5	2	1	37,00	3000,00	82,00	164,00	95,00	1	1	95,00	1
17	17	42	5	2	1	36,00	3400,00	89,00	168,00	92,00	1	1	90,00	1
18	18	41	3	2	1	37,00	3700,00	97,00	169,00	100,00	1	1	97,00	0
19	19	39	3	2	1	38,00	3800,00	94,00	169,00	100,00	2	1	95,00	0
20	20	28	3	2	1	39,00	3900,00	95,00	169,00	110,00	2	2	100,00	0
21	21	25	3	2	12	38,00	4000,00	96,00	169,00	100,00	2	2	98,00	0
22	22	26	4	2	2	39,00	4100,00	93,00	170,00	98,00	2	2	95,00	0
23	23	23	4	2	2	36,00	4000,00	92,00	170,00	96,00	2	2	96,00	0
24	24	21	4	1	2	40,00	4200,00	91,00	171,00	95,00	2	1	92,00	0
25	25	20	5	2	2	41,00	3980,00	89,00	171,00	92,00	2	1	92,00	1

**Figure 5/20. Results of filtering**

Besides the sequential numbers, a new column appears where the cases filtered out are denoted by 0, while those cases meeting the filter criteria are denoted by 1. In our case, young

mothers with code 16, 27, and 25 meet the selection (filter) criteria. In order to go back to the complete database, the original settings should be set, i.e. analysing all data.

**WEIGHT CASES:** weighing cases is often used in order to make sample size more accurate. The underrepresented groups can get higher weight, while applying lower weights can mitigate overrepresentation.

**TRANSFORM:** is employed to transform or modify data. New variables can be created with or without the help of already existing variables. Variables may be recoded, and the ranking of the cases can be calculated as well. Most of the options found here are to clarify, or “manipulate” data. These will later be introduced in more detail.

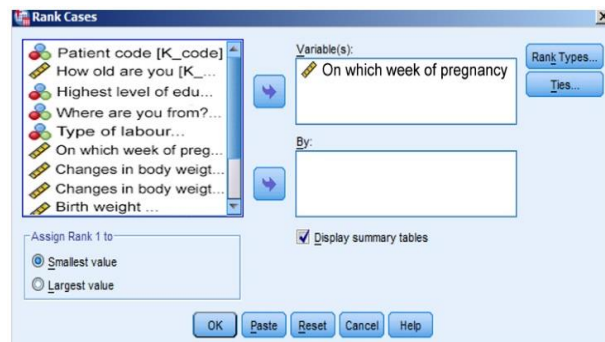
**COMPUTE VARIABLE:** computing new variable; typically used to create derivative data and measures.

**RECODE:** there are three recoding options in this menu, including:

- **INTO SAME VARIABLES:** no new variable appears but the existing one will be recoded. It is often used to change the orientation of scales; here the original variable can be discarded.
- **INTO DIFFERENT VARIABLES:** a new variable is created while keeping the former one. It is most frequently used when categorizing, or creating groups. We recommend using this option rather than the former one because the original variables can be deleted at any later time. The codes are introduced to group the quantity values of variables (e.g. in order of size) into groups defined verbally. The nominal variables can be named but it is also possible to assign nominal labels to numerical values (Jánosa 2008).
- **AUTOMATIC RECODE:** variables can be recoded automatically. We select the variable to be decoded, add the *NEW NAME*, and press OK in order to recode.

**RANK CASES:** cases can be ranked, and the ranking numbers can be represented in a new variable. This can be interpreted as a scale transformation since a variable measure scale becomes measure ordinal which may result in loss of information, and so particular types of statistical analyses cannot be carried out or have to be modified. The target variable is moved to the window *VARIABLES*. If a ranking in the specific groups is required to perform, the box displayed at the bottom will need to be used (*BY*). A sequence can be assigned to this window (the highest or the lowest value becomes rank one). Special ranking settings are also available (*RANK TYPES*), and the modification of equal ranking values is possible, too (*TIES*). The new variable will be retained with the original name with an “R” indicated at the beginning.

**For practice**, let us rank the dates of birth based on the weeks of pregnancy. The earliest birth shall be ranked first.



**Figure 5/21. Ranking options**

The options *DIVISUAL BINNING* and *OPTIMAL BINNING* organise the variables into categories. This means that the continuous quantitative variables are rated in discrete and unique categories. Both processes result in creating new variables.

*CREATE A TIME SERIES* creates times series variables from other types of time series variables. One can set the time series methodologies (seasonality, moving average, etc.) by *FUNCTION*. The analysis of time series falls out of the scope of this book but one can read more about it in Pintér – Rappai (2001).

*REPLACE MISSING VALUES* can be used when *MISSING VALUES* need to be replaced in order to carry out an analysis. There are different *METHODS* available to replace the average, the average or median of neighbouring values, linear trend values, etc.

The *RANDOM NUMBER GENERATOR* creates random numbers.

*ANALYSE* is one of the most important and most complex menu items of SPSS as it contains the statistical methodologies. Data analysis starts at this point. A detailed description of this menu will be provided in following chapters.

*DIRECT MARKETING* contains methods used in marketing research. This option was not available in former versions of SPSS. It includes techniques used in marketing campaigns and market segmentation. For more details read Jánosa (2011).

*GRAPHS* contains figures, graphs, diagrams applied in data presentation. The graphs aim to illustrate the research outputs convincingly. The programme provides a wide variety of the above options. Selecting the proper graph is influenced by the basic data and the purpose of the analysis. The basic figures originate from a specific geometric item (point, column, rectangular, circle, etc.) or their combination. There are two methods to create diagrams (*CHART BUILDER* and *LEGACY DIALOGS*). It is important to name the graphs very clearly,

displaying the measurement units and the period analysed. A detailed introduction of graphical illustration will be offered in the forthcoming chapters.

## 6. DATA CLEANING, SIMPLE ANALYSES WITH PRIMARY DATA, DATA MANIPULATION (Pongrác Ács)

Before starting to carry out a statistical analysis it is strictly important to get hold of a database that meets all requirements. Only research goals and hypotheses can determine and modify variables. This can mean that hypotheses require the aggregation of some variables (e.g. BMI index), or some sort of data transformation. A lot of methodologies expect normality as a precondition, and often the outlier values have to be examined in advance, too. This chapter will attempt to describe these processes in practice.

**For practice,** let us examine the weight loss (súlyvesztés) of young mothers in the last 6 months. We only consider mothers registered in the database who claimed they had no chronic illness. It may be interesting to know whether they can gain back their original weight in six months.

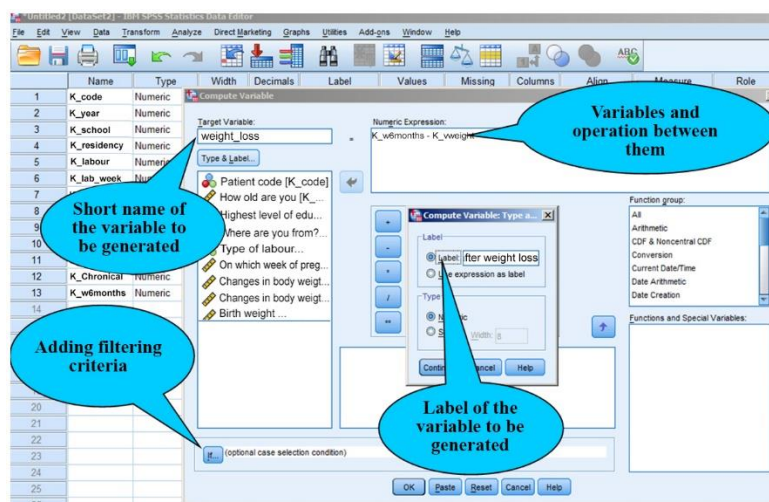


Figure 6/1. Generating the new variable

This data manipulation can be carried out using the option *COMPUTE VARIABLE* available in the *TRANSFORM* menu item. First, the new variable has to be named (sulyvesztes), then the necessary variables will need be moved to the window *NUMERIC EXPRESSION*. The mathematical and logical connections between them have to be given, of course. A number of built-in functions can be used (*FUNCTIONS AND SPECIAL VARIABLES*) but they are not required in the current example. Clicking on the bow *TYPE&LABEL*, the type and label of the new variables will need to be added.

The next item we need is the qualifying criterion, i.e. only young mothers without a chronic illness can be considered.



**Figure 6/2. Adding qualifying criteria**

The database does now only contain the data of young mothers who answered no (code: 2) to the question if they are aware of having a chronic illness („Volt-e krónikus betegsége”). Clicking on *CONTINUE* and *OK*, the new variable is created.

1:	weightloss	5,00								
	it	K_birth_we	K_weight	K_tmag	K_wbef_preg	K_know	K_chronical	K_w6months	K_weightloss	
1	00	2500,00	55,00	155,00	68,00		1	2	60,00	5,00
2	00	2600,00	58,00	160,00	62,00		1	2	60,00	2,00
3	00	3000,00	59,00	160,00	70,00		1	2	59,00	,00
4	00	3200,00	56,00	164,00	70,00		1	2	56,00	,00
5	00	3300,00	54,00	165,00	70,00		1	2	58,00	4,00
6	00	3400,00	57,00	165,00	71,00		1	2	59,00	2,00
7	00	3500,00	58,00	166,00	69,00		1	2	60,00	2,00
8	00	4000,00	59,00	166,00	70,00		1	2	59,00	,00
9	00	3800,00	53,00	163,00	75,00		2	2	55,00	2,00
10	00	3500,00	60,00	163,00	78,00		2	2	62,00	2,00
11	00	3400,00	60,00	163,00	80,00		2	2	62,00	2,00
12	00	3300,00	60,00	162,00	85,00		2	2	80,00	20,00
13	00	2900,00	64,00	162,00	70,00		2	2	70,00	6,00
14	00	2200,00	65,00	162,00	75,00		1	2	75,00	10,00
15	00	2150,00	80,00	165,00	89,00		1	2	89,00	9,00
16	00	3000,00	82,00	164,00	95,00		1	1	95,00	
17	00	3400,00	89,00	168,00	92,00		1	1	90,00	
18	00	3700,00	97,00	169,00	100,00		1	1	97,00	
19	00	3800,00	94,00	169,00	100,00		2	1	95,00	
20	00	3900,00	95,00	169,00	110,00		2	2	100,00	5,00
21	00	4000,00	96,00	169,00	100,00		2	2	98,00	2,00
22	00	4100,00	93,00	170,00	98,00		2	2	95,00	2,00
23	00	4000,00	92,00	170,00	96,00		2	2	96,00	4,00
24	00	4200,00	91,00	171,00	95,00		2	1	92,00	
25	00	3980,00	89,00	171,00	92,00		2	1	92,00	

**Figure 6/3. Data of the new computed variable**

The data view shows that the first young mother is still five kilograms of discrepancy in comparison with her weight before giving birth.

Besides mathematical operations, logical criteria can also be added in this option. It most typically comes up when not all types of questions are necessary from a scale of several items.

**Table 6/1. The most commonly used logical operations**

<i>Notation</i>	<i>Meaning</i>
&	„And”
	“Or”
~	“Not”
<	“Less than...”
>	“Greater than...”
<=	“Less than or equal to...”
>=	“Greater than or equal to ...”
=	“Equal to”
~=	“Inequality”

Source: author

The next example has two objectives. First, a new variable (BMI INDEX) will be generated using the method referred to above. Second, the new continuous variable will be converted into a discrete one. The calculation of the BMI INDEX is simple; the weight has to be divided by the square of the height (kg/m<sup>2</sup>).

BMI categories are defined as follows:

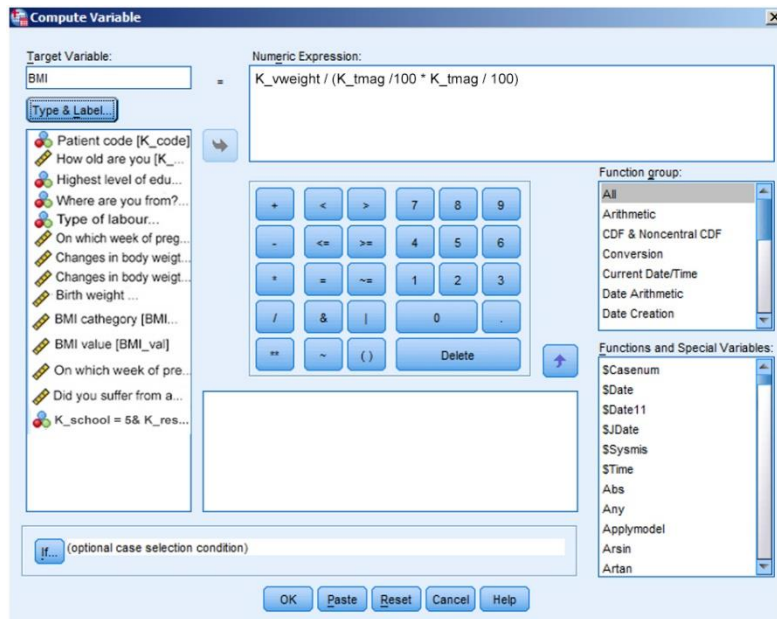
**Table 6/2. BMI categories**

Value	<i>Meaning</i>
<18.49	“underweight”
18.5- 24.9	“normal weight”
25-29.9	“overweight”
>30	“obesity”

Source: author

First, click on the first option (*COMPUTE VARIABLE*) in the main menu called *TRANSFORM*.





**Figure 6/4. Settings of the BMI index calculation**

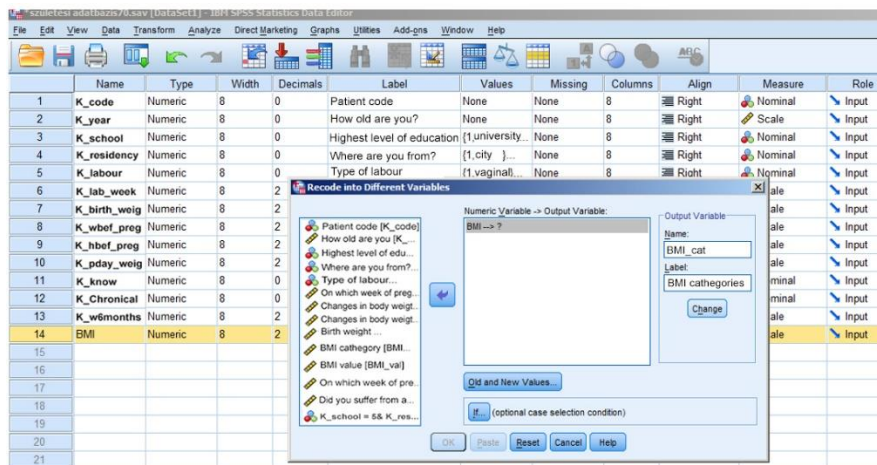
First, name the new variable (BMI) in the window *TARGET VARIABLE*, then define the function at *NUMERIC EXPRESSION* ( $K\_vs\acute{u}ly / (K\_tmag / 100 * K\_tmag / 100)$ ). This can be carried out directly, or in the calculator panel. The variables to be used can be moved with the help of the arrow in the middle. Next to the window with the numerical buttons, there is the window labelled *FUNCTION*. It contains built-in functions, formulas, and commands. Our calculation does not require other filters (IF). After changing these settings, press *OK*.

	K_birth	K_preg_week	K_birth_weight	K_vweight	K_tmag	K_wbef_preg	K_knowledge	K_chronical	K_w6months	BMI
1	2	38,00	2500,00	55,00	155,00	68,00	1	2	60,00	22,89
2	1	35,00	2600,00	58,00	160,00	62,00	1	2	60,00	22,66
3	1	36,00	3000,00	59,00	160,00	70,00	1	2	59,00	23,05
4	1	38,00	3200,00	56,00	164,00	70,00	1	2	56,00	20,82
5	1	37,00	3300,00	54,00	165,00	70,00	1	2	58,00	19,83
6	1	35,00	3400,00	57,00	165,00	71,00	1	2	59,00	20,94
7	1	34,00	3500,00	58,00	166,00	69,00	1	2	60,00	21,05
8	2	36,00	4000,00	59,00	166,00	70,00	1	2	59,00	21,41
9	2	37,00	3800,00	53,00	163,00	75,00	2	2	55,00	19,95
10	2	38,00	3500,00	60,00	163,00	78,00	2	2	62,00	22,58
11	2	39,00	3400,00	60,00	163,00	80,00	2	2	62,00	22,58
12	2	40,00	3300,00	60,00	162,00	85,00	2	2	80,00	22,86
13	2	41,00	2900,00	64,00	162,00	70,00	2	2	70,00	24,39
14	2	38,00	2900,00	66,00	162,00	75,00	1	2	75,00	24,77

**Figure 6/5. The generated BMI variable in data view**

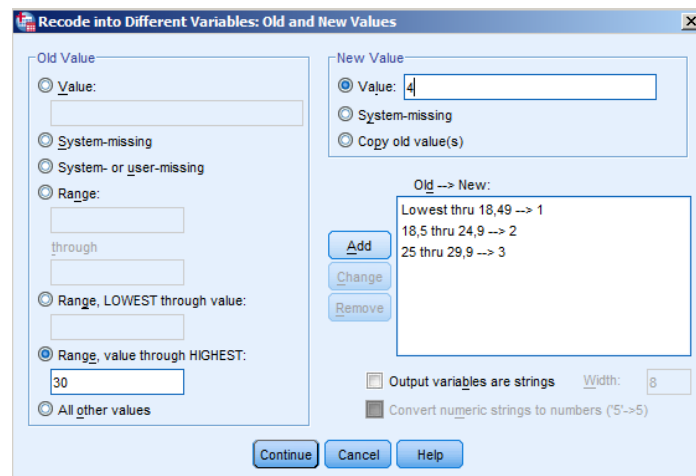
A new continuous variable called BMI has been generated. Afterwards, the process of recoding will take place whereby a new variable will be generated from an already existing one. This is practical because the original variable stays in the database besides the variable containing discrete categories.

When clicking on *TRANSFORM / RECODE INTO DIFFERENT VARIABLES*, the following window appears:



**Figure 6/6. Settings of recoding**

All the variables appear on the left-hand-side of the window,. BMI has been moved to the window *OUTPUT VARIABLE* with the help of the arrow in the middle. Then, a short variable name (BMI\_kat) and a label (BMI categories) are added. By clicking on *CHANGE*, the new name of coding becomes valid. By way of confirmation, the names of both the old and the new variables will appear in the window labelled *INPUT VARIABLE→ OUTPUT VARIABLE*. Afterwards, the values of the old variables will need to be added besides the calibration of the new item. The new window will have information on both the old and the new values.



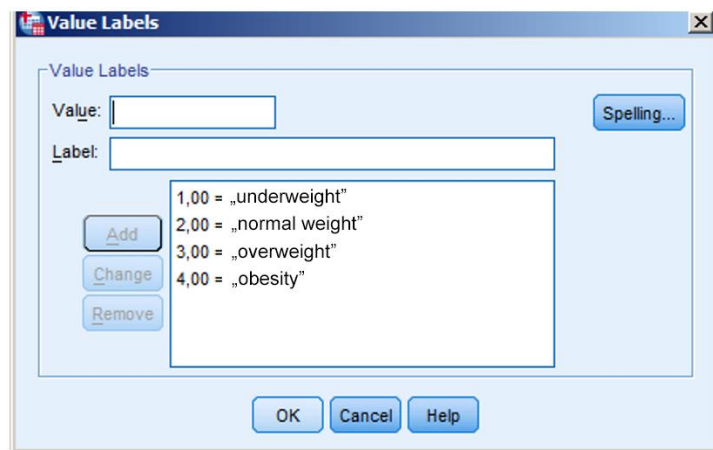
**Figure 6/7. Determining the categories of the new discrete variable**

Selecting the option *VALUE*, one can add the old values one by one. The options *SYSTEM-MISSING* or *SYSTEM -OR USER- MISSING* can exclude items that do not meet the relevant

requirements. The *RANGE* option sets up strings for different group intervals. The first alternative is to give the lower and the higher limit of group intervals by providing value sets using *RANGE... THROUGH...* (e.g. group 2 from the items between 18.5 and 24.9). The *RANGE LOWEST THROUGH* is for intervals without a lower limit, while *RANGE THROUGH HIGHEST* is for intervals without an upper limit (e.g. above 30, items get code 4). One can add new values on the right-hand-side if the old values have already been given, as explained before. New values will have to be added in the text box labelled *VALUE*, and then need to be finalized by clicking on *ADD*. The calibration of the variable appears then on the window as *OLD*→*NEW*. In order to modify categories, click on *CHANGE*, to delete, and click on *REMOVE*. If the categories are in a text format, click on *OUTPUT VARIABLES ARE STRINGS* in the checkbox.

The four categories will become valid after clicking on *CONTINUE* and *OK*.

As the four categories only contain one numerical value, their names will be the following:

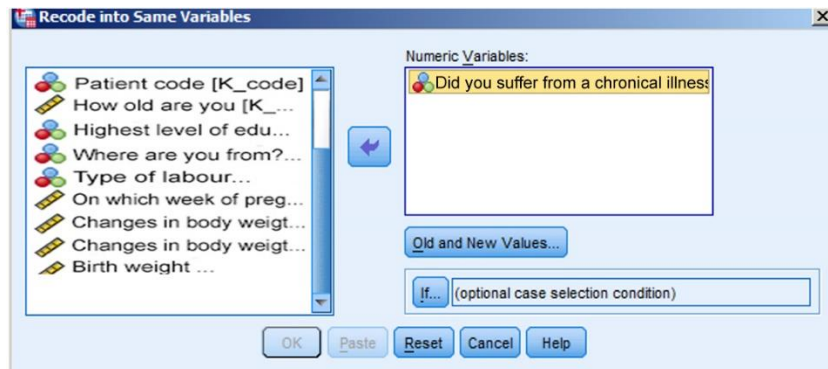


**Figure 6/8. Naming and labelling the new groups**

When modifying data, it is often not necessary to generate a new variable but change the existing one (e.g. if one would like to modify the orientation of the scale). Settings can be set under *TRANSFORM / RECODE INTO SAME VARIABLES*.

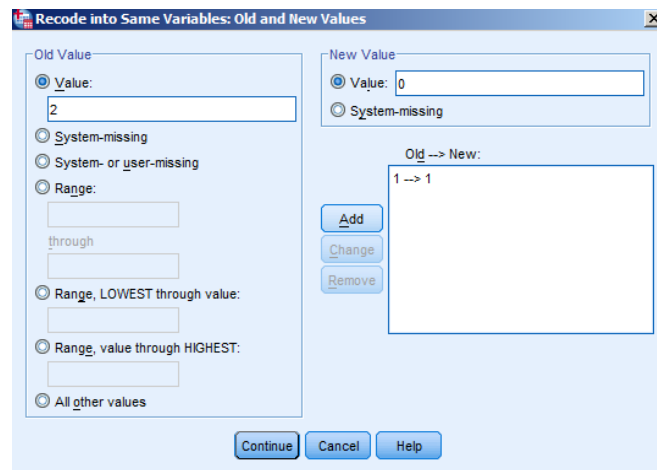
**For practice**, let us use recode the groups based on chronic illness. Originally, young mothers with a chronic illness belonged to group value 1, while those without one were assigned value 2. Let us modify the value of those cases to 0 where the answer was no, and retain all the remaining values (i.e. 1).

By clicking on *TRANSFORM / RECODE INTO SAME VARIABLES* you will see the following window:



**Figure 6/9. Settings of recoding**

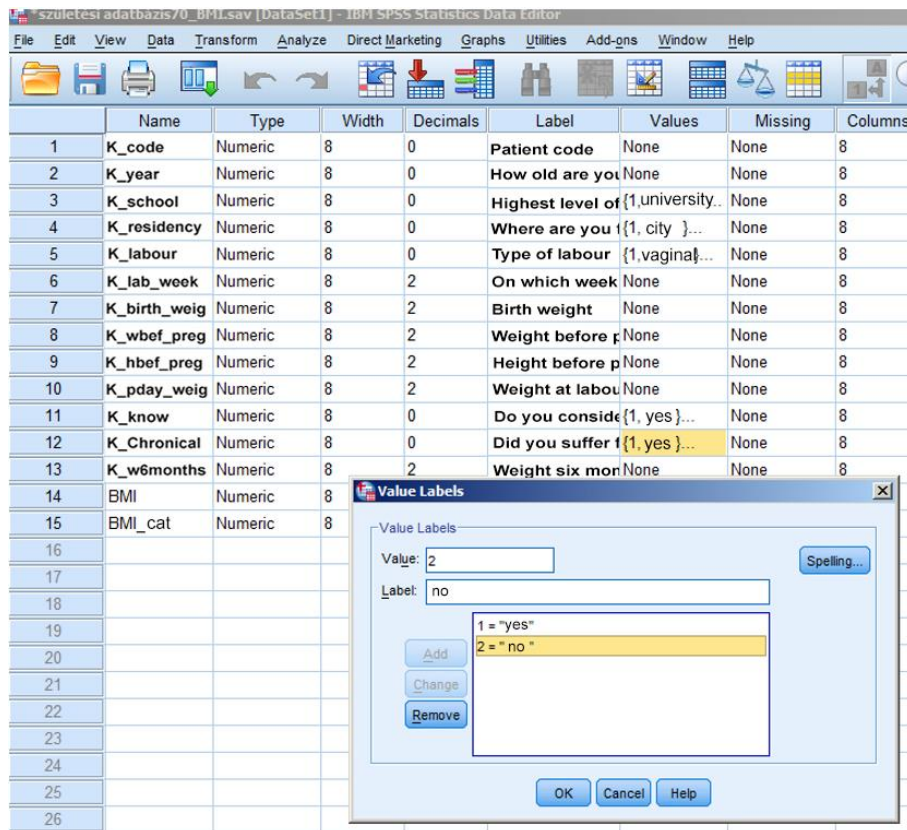
On the left-hand-side, the list of variables is seen, from which we chose the question related to the chronic illness (i.e. „Volt-e krónikus betegsége?”), and moved it with the arrow in the middle to the window *NUMERIC VARIABLES*. After that, the calibration was processed.



**Figure 6/10. Changing group codes**

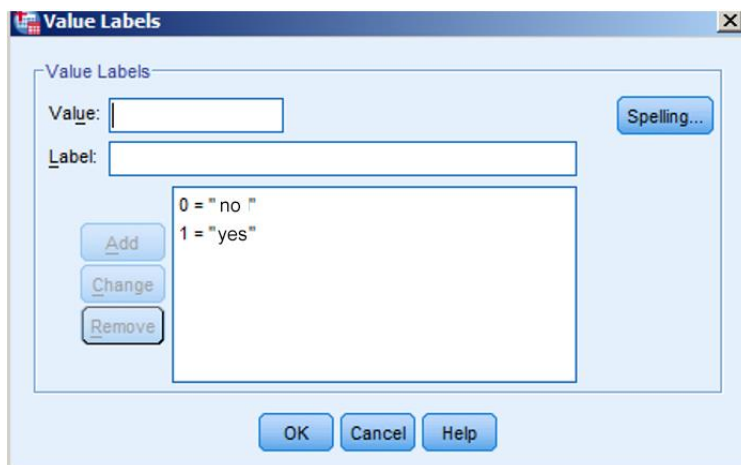
In this case, we select *VALUE*, and add the old, then the new values one by one. First, we add value 1 both in the box *OLD VALUE* and *NEW VALUE*, then click *ADD* and it will appear in the box in the middle. Then we modify the *OLD VALUE* (2), and leave the new one at 0 (*NEW VALUE*). After pressing *ADD*, *CONTINUE* and *OK*, recoding is finished, which can be checked in the *OUTPUT* table as well.

The last step is to rename labels in the variable view, since this process is not automatic.



**Figure 6/11. Deleting old variable codes**

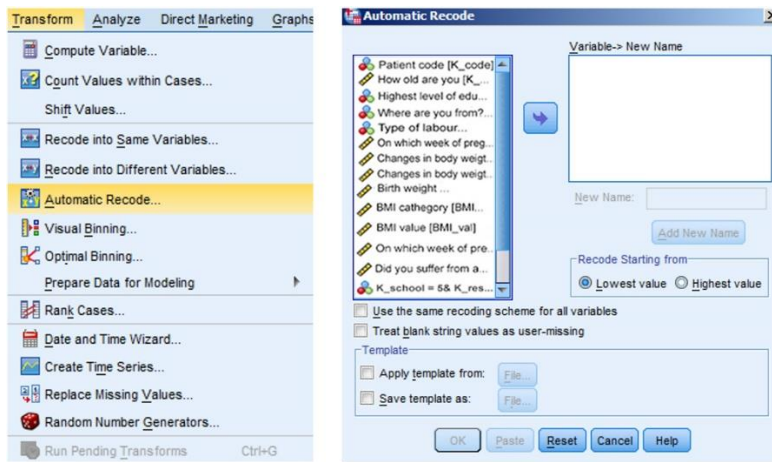
In the *VARIABLE VIEW* to remove codes 2 select *REMOVED* and click on *VALUES* belonging to the question. After pressing *OK*, the old button and its label will disappear. The new code and category name can be now added.



**Figure 6/12. Adding the new variable code**

In the *VALUE* box, we add value 0, and name it “nem” (i.e. no) in the *LABEL* window. After pressing *OK*, the new category code becomes valid.

A much simpler but not always applicable method of recoding is automatic recoding which is available at *TRANSFORM / AUTOMATIC RECODE*.



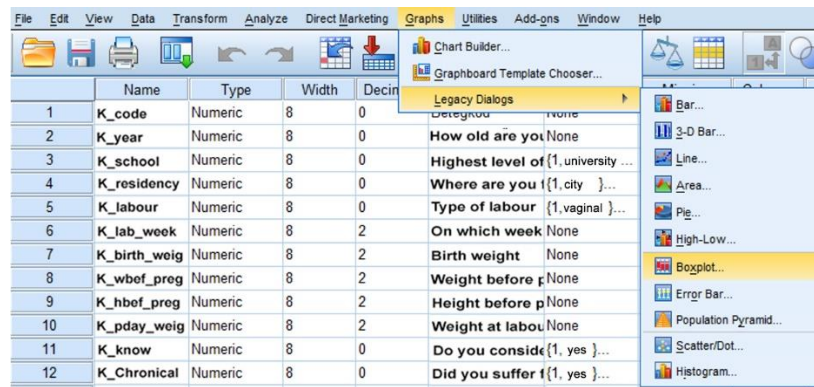
**Figure 6/13. Modules for automatic recoding**

On the right-hand side, select the variable to be recoded, move it to the empty window with the arrow, and name it in the box *NEW NAME*. Decide if the recoding should start from the *LOWEST VALUE* or the *HIGHEST VALUE* and set accordingly.

It often happens that one has to examine data and decide if the outlier values are valid or are just consequences of typing mistakes. This data-cleaning method is called *outlier test*. The researcher has to decide if he/she will exclude or retain the outliers as part of the analysis. It is vital to examine extreme values for continuous variables. The decision is based on the knowledge and experience of the researcher. In practice, researchers usually exclude the values that are outliers because of typing and coding mistakes or they are consequences of incidences that one has no objective explanations for. It is important to determine the exact value above which the data is considered to be extreme. It is considered to be an error if one retains an outlier that has a distorting effect (e.g. in case of a normality test), and also if one excludes the value, although it represents a real data that would support generality. One of the most commonly used methods to detect outliers is the so-called *Boxplot* diagram. The diagram makes it clear if there is an extreme value and identifies which case it belongs to.

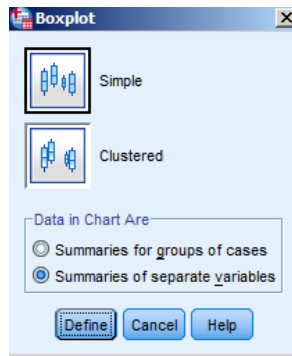
**For practice**, let us examine the age and BMI of young mothers in the database (születési adatbázis70\_BMI.sav) with the help of a Boxplot.

The option is available under *GRAPHS / LEGACY DIALOGS / BOXPLOT*.



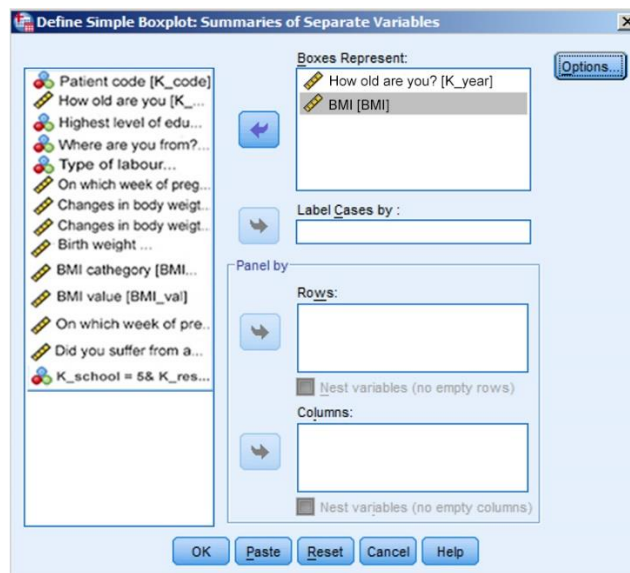
**Figure 6/14. The option Boxplot**

There are two options under the menu *BOXPLOT*: *SIMPLE* or *CLUSTERED*. Let us select the simple form. Now, under *DATA IN CHART ARE*, the type *SUMMARIES FOR GROUPS OF CASES* or *SUMMARIES OF SEPARATE VARIABLES* can be selected. Here we will choose the second option. During the calibration of this example, we advise the display of variables in the simple form since this makes it possible to introduce the distribution of one or more variables. If one decides to plot according to *CASES*, then the given variable (e.g. age) can be plotted depending on categories of another variable (e.g. residence; city, village). If we choose summaries of separate variables, the option *CLUSTERED* displays at least two continuous variables (e.g. age, BMI) according to the categories of another variable (e.g. residence).



**Figure 6/15. Calibrating the type of Boxplot**

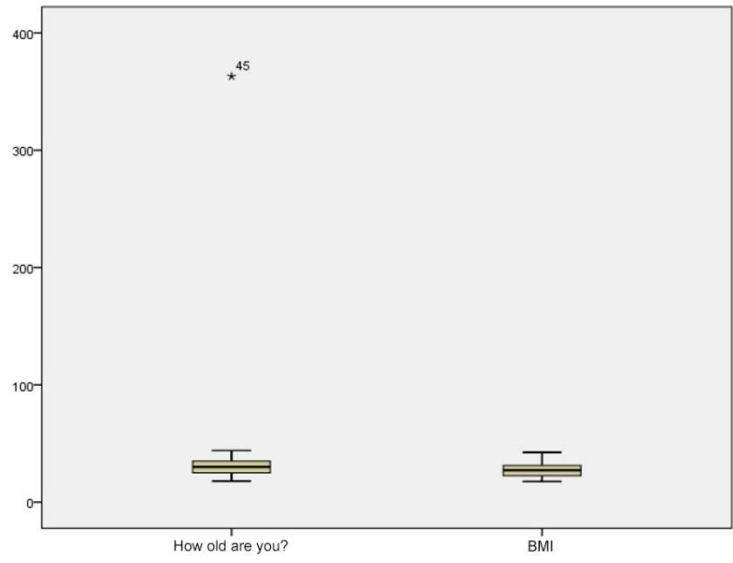
Now, we have to select the variables to include in the analysis. In the window *BOXES REPRESENT*, the age and BMI variables will need to be moved by the arrow in the middle. No other settings will need to be set, only press *OK*.



**Figure 6/16. Selecting the variables**

Edges of the boxes on the output figure will show the difference between the lower (25) and the upper (75) quartile, while the line in the middle is the median (50). The length of the lines extending the box upwards and downwards is one and a half times the length of the interquartile (the difference between the lower and the upper quartile). Ideally, values are at this interval (normal distribution), which is stressed by the horizontal sign at the ends of the lines. If the value is 1.5-3 interquartile away from the edge of the box, the programme denotes it as an outlier (notation: O). Values with even more differences are considered to be extreme and denoted by \*.





**Figure 6/17. The Boxplot of variables age and BMI value**

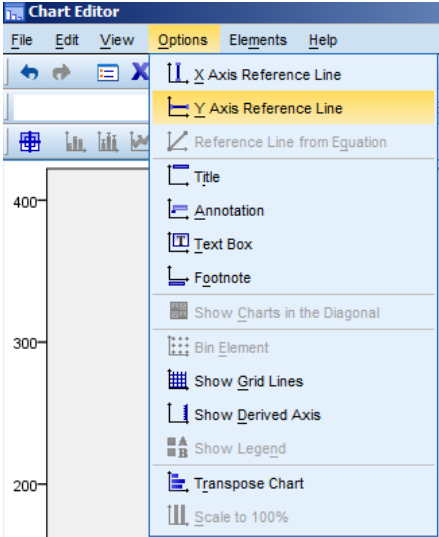
As shown by the figure, no outlier or extreme value appears for BMI, so its distribution can be considered to be normal. In case of age (“Hány éves Ön?”), the value of #45 is extreme. This value can be checked in data view.

	K_year	K_school	K_residency
25	20	primary scho	village c
26	22	primary scho	city c
27	24	primary scho	city c
28	25	primary scho	city c
29	26	primary scho	city c
30	31	primary scho	city c
31	32	vocational...	village c
32	31	vocational...	village c
33	33	vocational...	village
34	34	vocational...	village
35	35	vocational...	village
36	36	vocational...	village
37	37	university	city
38	38	university	city
39	39	university	city
40	40	university	city
41	31	university	city
42	32	university	city
43	34	OKJ	city c
44	35	OKJ	village c
45	363	OKJ	village c
46	37	OKJ	village c

**Figure 6/18. The age of a young mother #45 in data view**

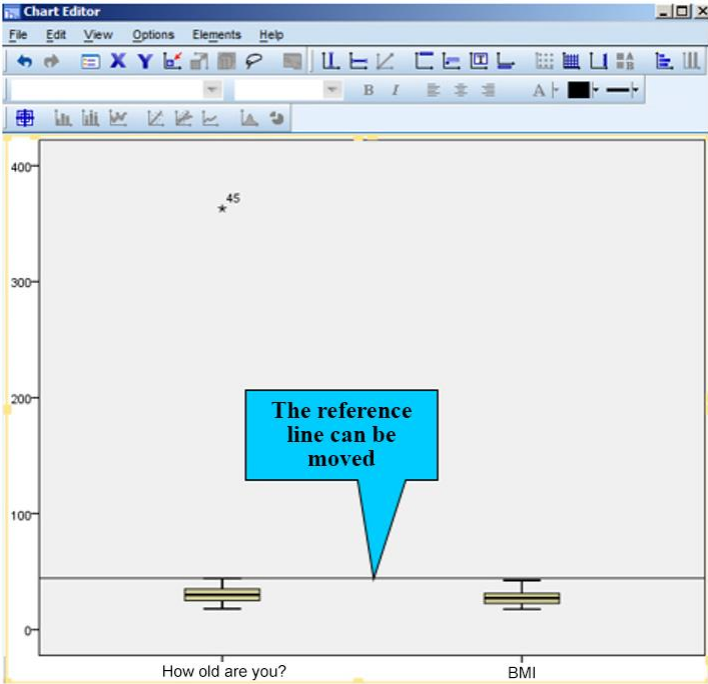
The age of young mother #45 is 363 which is extreme and also not possible, so probably a typing mistake occurred here. In these cases a *missing* value interval may be defined, or a data filter may be applied (*DATA / SELECT CASES*). To illustrate our point, we will do this latter one.

Clicking twice on the output figure, the *CHART EDITOR* appears, and clicking on the option *ELEMENTS* makes it possible to *ADD REFERENCE LINE TO THE Y AXIS*.



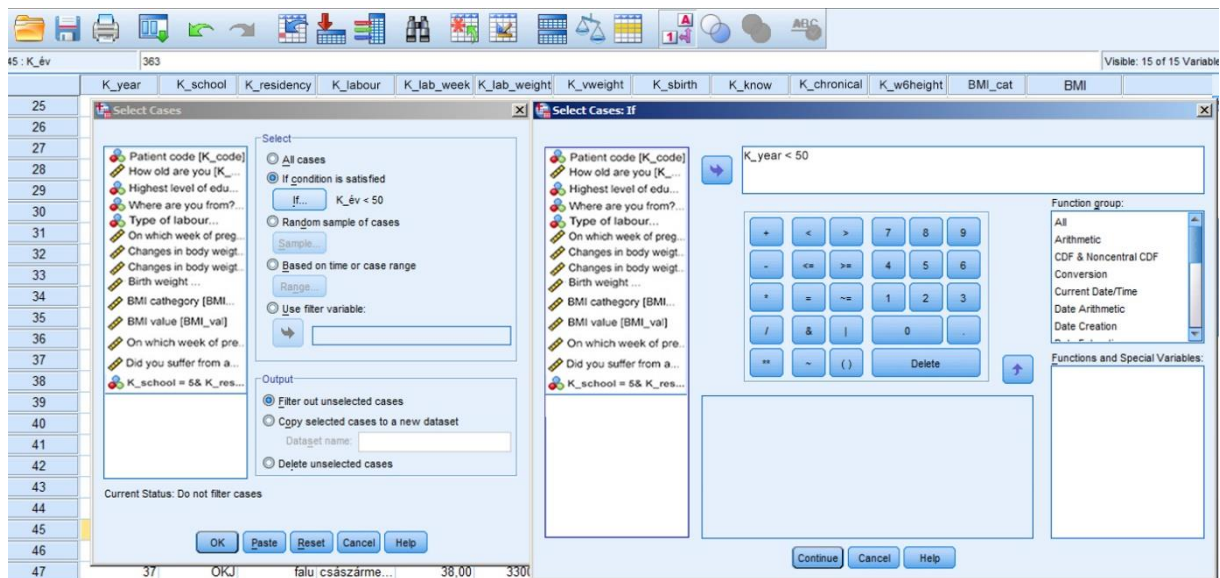
**Figure 6/19. Drawing a reference line**

A reference line at 200 appears which can be moved vertically by clicking on the line. We need to remove the reference line up to the horizontal line in the top, (50) which is the limit.



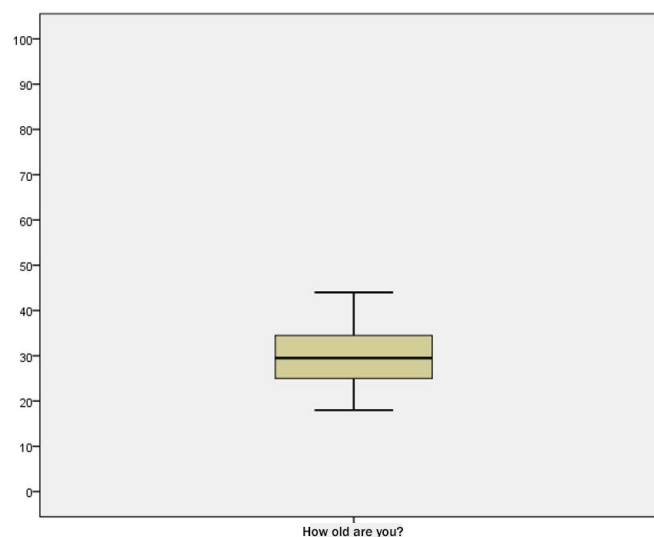
**Figure 6/20. Adding a reference line to the figure**

With option *DATA / SELECT CASES* data can be filtered, i.e. the variable age has to be calibrated in the IF window.



**Figure 6/21. Settings of filtering options**

Let us choose settings like on the figure and press *CONTINUE*, then *OK*. Now, the software filters and crosses out the values of the variable age that are greater than 50 years. This can be seen in data view. These data will be excluded from the analysis. The filter will stay active until one resets it. Plotting a new Boxplot will contain no outlier and extreme values of the variable age.



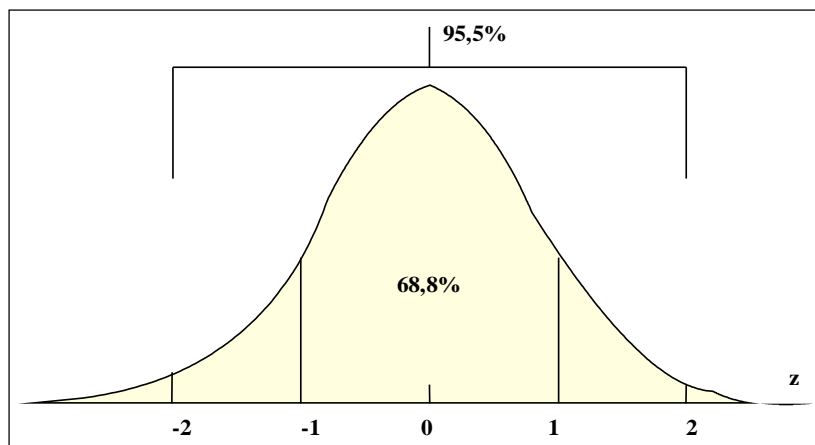
**Figure 6/22. The Boxplot free from outliers**

When examining outlying values, it is common to apply the methodology of **standardization**, **which** allows us to generate a standardized variable from the examined one. The method is also used if one would like to homogenise or standardise variables of different scales (and measurement units). There are statistical methods that require the standardization of variables with different measurement units.

**Standardization** means that the expected value is subtracted from the value of the random variable and the difference is divided by the standard deviation. The result is a **random variable with normal distribution** (sign: z). In formula:

$$z = \frac{x - \mu}{\sigma}$$

The expected value of this new variable is zero, and its standard deviation is one i.e. N (0,1). Both variables – the one of normal and the one of standard normal distribution - have a density function called the **Gauss curve (Figure 6/23)**. To show a standard normal distribution, both the random variables and their probabilities can be ordered in a table, with the help of which the resulting values can be used to solve problems quickly and easily.



**Figure 6/23. Most important probability values depending on z**

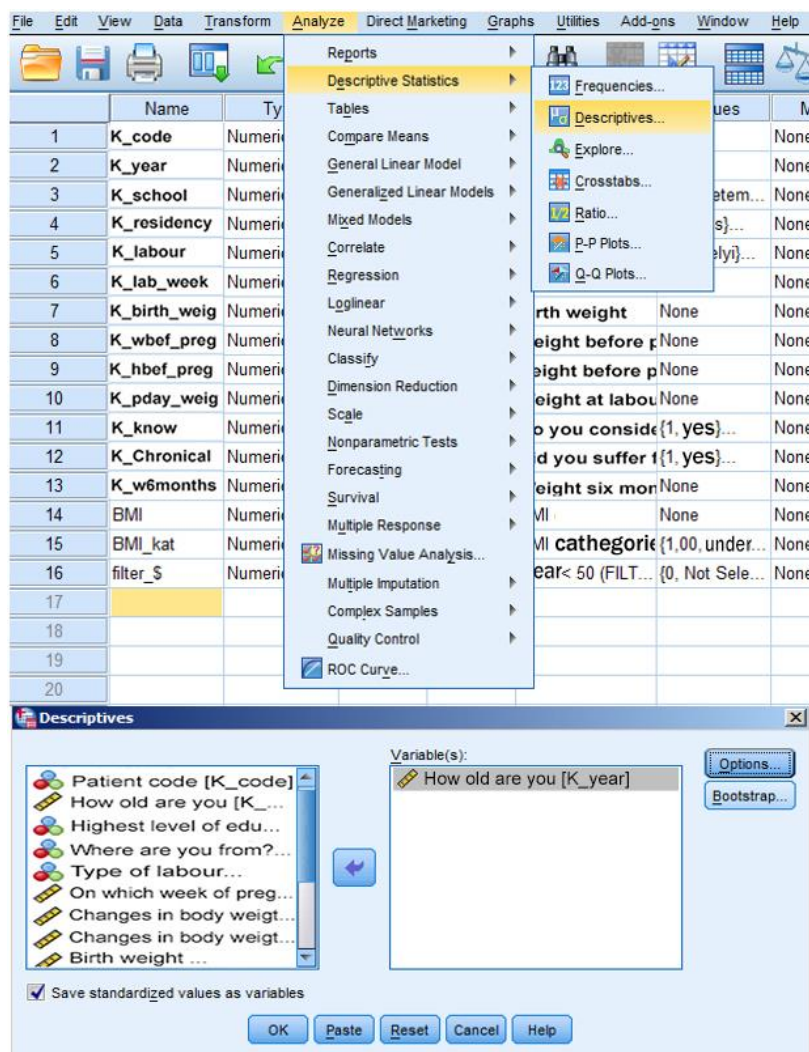
The area between the interval plus and minus one standard deviation from the expected value and the probability curve represent 68.8% probability – this is true for both the normal and the standard normal distribution. The same value for the interval of plus and minus two standard deviations represents 95.5%, while three standard deviations stand for 99.9%. As the density function is symmetric, it is sufficient to determine the probability value between zero and positive infinity.

In practice, it can mean that the standardized value has to be between -3 and +3 (in order to achieve normal distribution). Based on the theory of standard normal distribution, if the sample size is 80 or less, then cases with greater than 2.5 standardized values are considered to be outliers. For greater sample size this limit is 3 (Sajtos – Mitev 2007).

**For practice,** let us examine the outliers of the variable age with the method of standardization.

First, the filters will need to be reset to the default settings (*DATA/SELECT CASES/ALL CASES*).

Then, the standardization of the variable can be carried out in *ANALYSE / DESCRIPTIVES*.



**Figure 6/24. Settings of standardization**

On the left-hand side of the window, all variables are listed. Move the variable age to the right-hand side with the arrow, then click *SAVE STANDARDIZED VALUES AS VARIABLES*. Now, the software generates a new variable, called ZK\_év, which contains the standardized values.

In order to examine the standardized values, a table of frequencies will be required. This can be generated in menu *ANALYSE / FREQUENCIES*.

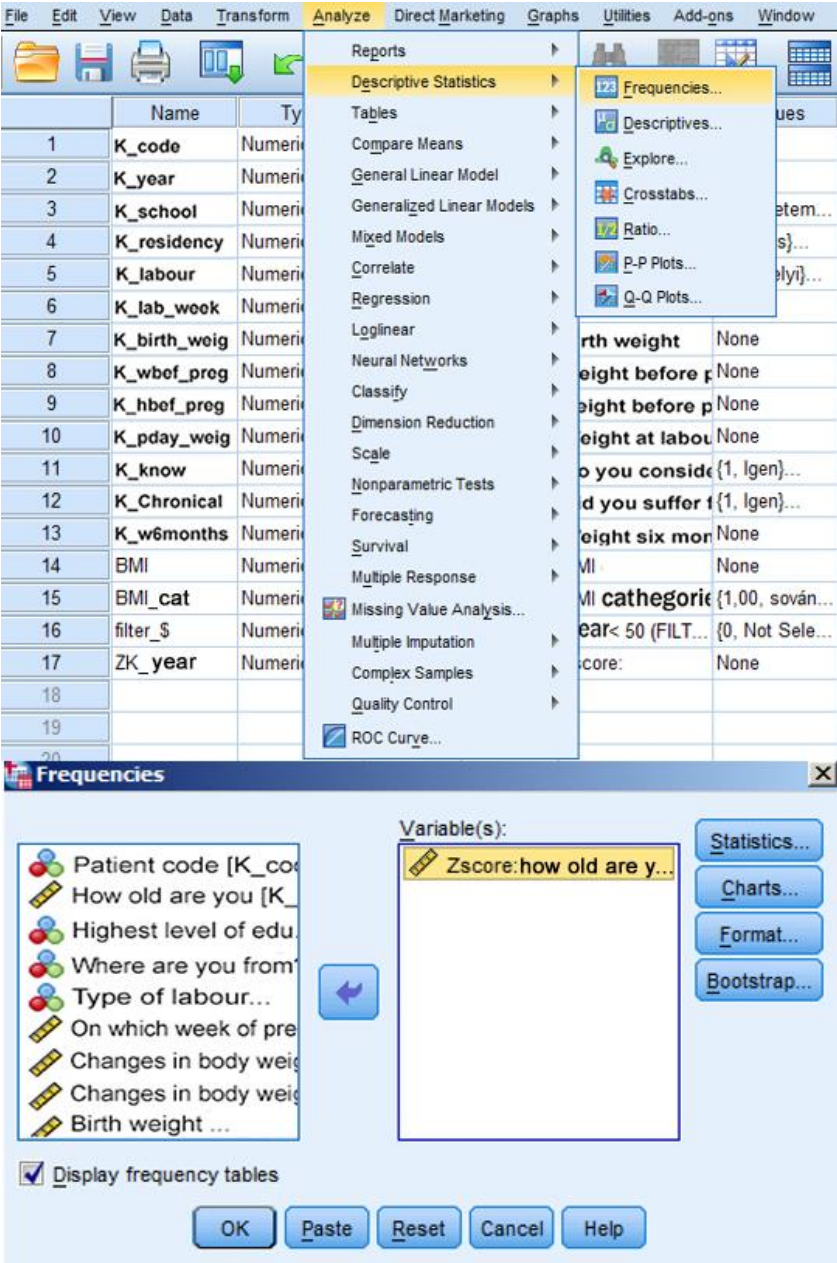
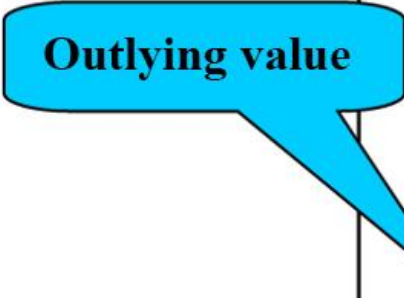


Figure 6/25. Settings of table of frequencies

In the settings, the standardized variables will have to be moved now from the right to the left-hand side window. *DISPLAY FREQUENCY TABLES* is a default setting, so after clicking on OK, the table will become available.

**Zscore: How old are you?**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-.41721	1	1,4	1,4	1,4
	-.39238	1	1,4	1,4	2,9
	-.36754	1	1,4	1,4	4,3
	-.34271	1	1,4	1,4	5,7
	-.31787	3	4,3	4,3	10,0
	-.29304	3	4,3	4,3	14,3
	-.26821	3	4,3	4,3	18,6
	-.24337	7	10,0	10,0	28,6
	-.21854	5	7,1	7,1	35,7
	-.19370	3	4,3	4,3	40,0
	-.16887	3	4,3	4,3	44,3
	-.14404	3	4,3	4,3	48,6
	-.11920	2	2,9	2,9	51,4
	-.09437	5	7,1	7,1	58,6
	-.06954	4	5,7	5,7	64,3
	-.04470	4	5,7	5,7	70,0
	-.01987	2	2,9	2,9	72,9
	,00497	3	4,3	4,3	77,1
	,02980	2	2,9	2,9	80,0
	,05463	3	4,3	4,3	84,3
	,07947	2	2,9	2,9	87,1
	,10430	3	4,3	4,3	91,4
	,12914	2	2,9	2,9	94,3
	,15397	1	1,4	1,4	95,7
	,17880	1	1,4	1,4	97,1
	,22847	1	1,4	1,4	98,6
	8,15051	1	1,4	1,4	100,0
	Total	70	100,0	100,0	



**Table 6/26. Frequencies of standardized variable of age**

As there are 70 items in the database (n=70), standardized values greater than 2.5 are outliers. There is only one outlier in the table (8.15), which will be excluded. In practice, it is very common that the distribution of a variable is not normal due to outliers, but filtering these values will make switch the distribution back to normal.

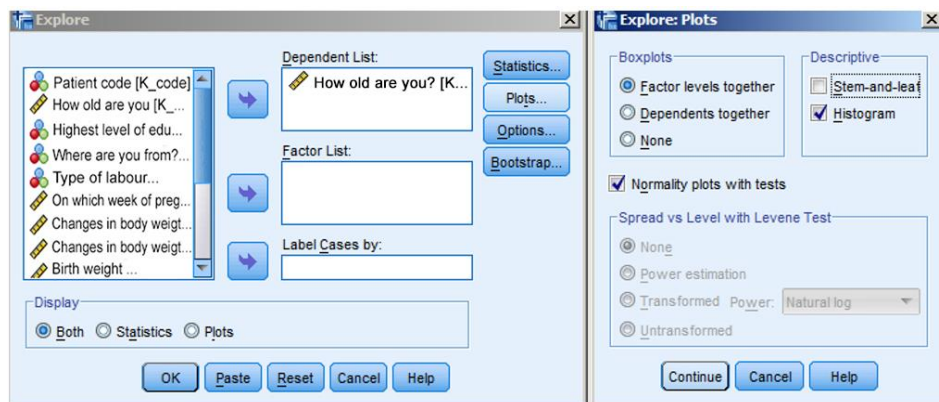
There is a method mentioned previously that can be used when examining outliers and transforming data. It is called **normality test**. This method is applied when comparing the

distributions of two random variables, or to check if the distribution of a random variable originates from an assumed (normal) distribution.

There are quantitative (numeric) and graphical methods to test normality. The most commonly used graphical methods are the histogram with the normal distribution curve, and the Q-Q (quantile-quantile) plot.

Based on graphs, a variable can be considered to have normal distribution if the shape of distribution is very similar to the histogram of the normal distribution, and it also fits the line of the hypothetic normal distribution line in the Q-Q plot. We introduce two numeric methods, including the Kolmogorov-Smirnov and the Shapiro-Wilk tests. The latter is sensible to apply if the sample is relatively small, i.e. less than 50 items. If the significance value is greater than 5%, the variable follows a normal distribution. Through data transformation, variables with different distributions can be transferred to follow normal distribution.

**For practice**, let us examine whether the original variable age has a normal distribution or not. The graphical and numerical tests are available at *ANALYSE / DESCRIPTIVE STATISTICS / EXPOLRE*.



**Figure 6/27. Numerical and graphical settings of normality test**

The variable age (K\_év) has to be moved to the dependent list with the top arrow, and after clicking on Plots, we need to select *HISTOGRAM* in the box labelled *DESCRIPTIVE*, and also select *NORMALITY PLOTS WITH TESTS*. Press *CONTINUE* and *OK*.

Now, the *OUTPUT* contains more tables and figure of results. We introduce only the ones that are relevant from our point of view.



**Table 6/2. Results of normality test**

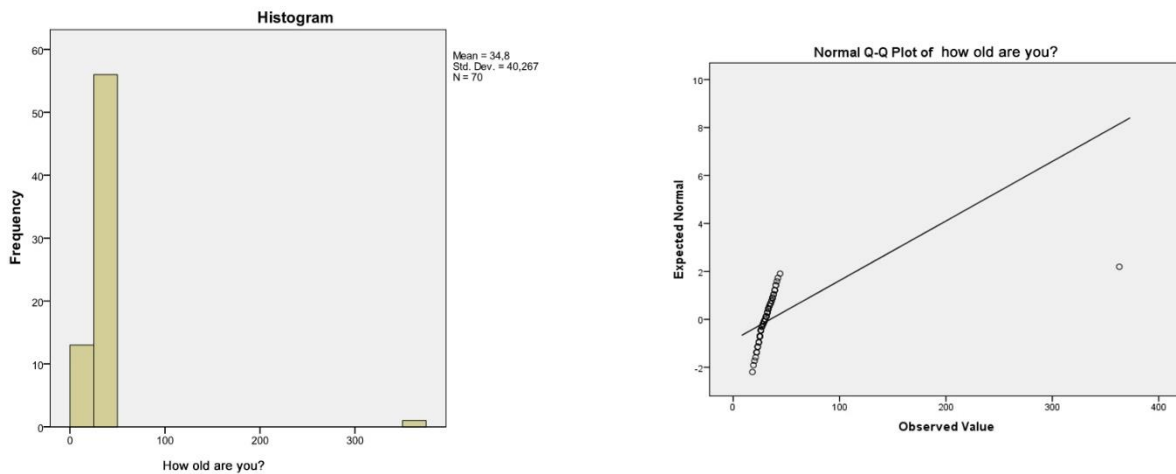
**Tests of Normality**

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
How old are you?	,400	70	,000	,205	70	,000

a. Lilliefors Significance Correction

Source: author

Significance for both tests (Kolmogorov- Smirov and Shapiro- Wilk) are lower than 0.05, thus the normality requirement is not complied with.



**Figure 6/28. Figures of normality test of age (histogram, Q-Q plot)**

The figures also state that the distribution is not normal, since the histogram does not fit a Gauss curve, and the Q-Q plot does not fit the line, either. In order to find the reason, the Boxplot may give adequate information since it is the one that shows an outlier value in the databox. Running the normality test after excluding the outlier (*DATA/SELECT CASES/ IF/ K\_év<50*) will give us new results.

**Table 6/3. Results of normality test after filtering outliers**

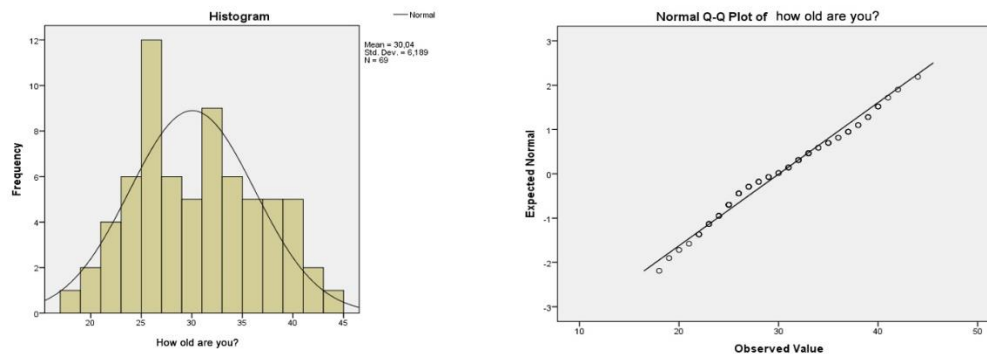
**Tests of Normality**

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
How old are you?	,106	69	,055	,977	69	,220

a. Lilliefors Significance Correction

Source: author

The P-value (sign.) for both tests is greater than 0.05, so the distribution is normal as shown in the figures as well.



**Figure 6/29. Figure of normality test after filtering outliers**

Based on the graphical methods, all variables have a normal distribution, their shape fits the Gauss-curve and the hypothetical line of the Q-Q plot (Figure 6/29)<sup>8</sup>.

We hope that this illustration has showed that during data analysis, most researchers examine normality after filtering the *outliers*.

---

<sup>8</sup> As in the case of the variable, values of skewness and kurtosis are almost zero, these indices prove the presence of normal distribution.

## 7. DESCRIPTIVE STATISTICS, TABLES AND GRAPHS (Pongrác Ács, László Bence Raposa)

### 7.1. Theoretical background, descriptive statistical indicators

Today, descriptive statistics means a form of evaluation when variables and survey questions are analysed. International literature of statistics differentiates between *descriptive statistics*, *inferential statistics* and *statistical decision theory*.

Descriptive statistics include the methods of gathering, summarizing and offering a compact description of numerical information. The most important fields of descriptive statistics include:

- collecting data
- plotting data
- grouping and clustering data
- doing simple arithmetic operations with data
- displaying outputs.

One of the prior aims of *descriptive statistics* is to provide a situation report or a summary on the properties of the sample. It is often referred to as *basic statistics* or univariate analysis. The first step of almost every database analysis is to examine the variables independently from one other. The terms of examination are the number of items, the mean value, and the variability of data.

**Table 7/1. The most commonly used descriptive measures**

Measures of central tendency	Measures of variability	Measures of shape	Other measures
Mean	Range	Kurtosis	Sum
Median	Standard deviation	Skewness	Number of cases
Mode	Variance		Minimum
			Maximum

Source: author, based on Sajtos – Mitev (2007)

### MEASURES OF CENTRAL TENDENCY

The quantitative measures and numerical data of the variables are very important properties of the statistical population. They provide useful information on the phenomenon or process to be examined. Ordering and clustering data will enhance understanding. Numerical data are

also a good basis of providing a general and compact description, to summarize their important property in one numerical data. This item of numerical data is called **mean value**. Mean value is **the numerical characteristic of data of the same type**. Two groups of mean values may be differentiated based on calculation (average, mean) and their place in the line of data (median, mode).

## MEAN

The **arithmetic mean** is the number that makes the sum if all values are replaced by this number.

From the definition it directly follows that the arithmetic mean ( $\bar{x}$ ) is the proportion of the sum of the observed values ( $x_1, x_2, x_3, \dots, x_n$ ) and the number of the items (n):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Here, frequency needs to be calculated with, in the formula of the **weighted arithmetic mean**.

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i} = \frac{\sum_{i=1}^k f_i x_i}{n}$$

The arithmetic average is the most commonly used mean value but it is not always applicable, usually because of its sensitivity to outliers. Its application is most relevant in the case of ratio and interval scales.

## MEDIAN

1. The median denotes or relates to a value or quantity laying at the midpoint of a frequency distribution of observed values or quantities, such that here is an equal probability of falling above or below it. The first step in determining the median is ranking our numerical data and
  - if n is odd, then the value of item number  $(n+1)/2$  is the median

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

- if n is even, then the arithmetic mean of values of items number  $n/2$  and number  $(n/2)+1$  is the median:

$$Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

The median is most often applied to for ordinal scales.

## MODE

Mode is the most typical value of data values. The mode of a quantitative variable with discrete values is the value most commonly found in the population. The mode of a continuous numerical variable is found at the place where the values become more dense, i.e. at the maximum of the frequency curve. It is most properly used in case of discrete, categorized variables but it also applicable for scale type variables.

The different scale types have their own most appropriate mean value, for nominal scales it is the mode, for ordinal scales it is the median. All mean values are applicable for ratios scales.

## MEASURES OF VARIABILITY

As it can be seen, none of the mean values meet all the requirements and there for a more sophisticated analysis is required. Dispersion in statistics means the deviation of (mostly numerical) data from one another, or from a particular value that is characteristic for the population. Examining dispersion is very important in statistical methodology, almost every method is linked to it.

Measures that use significant numerical data by definition are applicable to measure dispersion rapidly. Data that is different from one another (i.e. dispersing data) deviate from mean values, so they differ from the arithmetic average. This also means that most measures edited to measure dispersion make use of this property.

All measures of dispersion have to be zero if there is no deviation but their values need to be positive if there is some sort of deviation.

The most widely used measures of dispersion include:

- range (R),
- interquartile range (TQ),
- standard deviation ( $\sigma$ ), variance ( $\sigma^2$ )
- relative standard deviation (V)

**RANGE:** the difference between the greatest and the lowest value:

$$R = x_{\max} - x_{\min}$$

It is only applicable for ratio and interval scales. There is another thing we need to highlight here: the *interquartile range* is the interval, where the medial 50% of all the values is found. It can be calculated as:

$$TQ = Q_3 - Q_1$$

The most commonly applied measure of dispersion is *STANDARD DEVIATION* which is the square root of the average square deviation from the mean.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

When calculating standard deviation from frequency, the weighted form of the index has to be applied:

$$\sigma = \sqrt{\frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{\sum_{i=1}^k f}} = \sqrt{\frac{1}{n} \sum_{i=1}^k f_i (x_i - \bar{x})^2}$$

The square of standard deviation is called *variance* ( $\sigma^2$ ). It has no special meaning but is a key index in several statistical methods and we can apply them for ratio and interval scale.

The measures of dispersion are expressed with the measurement unit of the quantitative variable. It is often suggested that we disregard measurement units, and so make dispersion of different phenomena with different measurement units comparable. *Relative standard deviation* is eligible for this purpose, and it measures the percentage of average deviation from the mean.

## MEASURES OF SHAPE

**SKEWNESS:** an index to describe the horizontal shape of a distribution. Its value is positive in case of right-skewed distribution, and is negative for left-skewed distribution. In practice, right-skew is more common.

**KURTOSIS:** an index to describe the shape of distribution numerically. Ideally, its value is zero for normal distribution, positive for peaked, and negative for flat distribution.

## OTHER MEASURES

We only mention additional measures featuring in the descriptive module of SPSS.

**SUM:** the sum of items, i.e. their accumulated value.

**NUMBER OF CASES:** the number of cases in the analysis.

**MINIMUM:** the lowest value from values of all the cases.

**MAXIMUM:** the highest value from values of all the cases.

Typing mistakes may be detected by checking the minimum and maximum values, since values outside of the acceptable interval will be recognized.

There are three ways in SPSS to access descriptive statistics: *ANALYSE/DESCRIPTIVE STATISTICS/DESCRIPTIVE* or *ANALYSE/DESCRIPTIVE STATISTICS/FREQUENCIES* or *ANALYSE/DESCRIPTIVE STATISTICS/EXPLORE*.

**For practice,** let us analyse the variable weight of new-born children with the help of descriptives.

There is no binding rule to decide which method should be used from descriptive statistics. The module labelled *DESCRIPTIVE* is mostly used for interval or ratio scale (in SPSS. scale), provided there is no need for a frequency table. The module *FREQUENCIES* is rather used for variables of nominal and ordinal scales where both a frequency table and a graphical display are required. Of, course ratio and interval scale variables can also be analysed in this option but the output may not be as spectacular as in the *DESCRIPTIVE* module. In the *EXPLORE* module, also descriptive indices can be calculated, where the sample can be separated into groups.

The next table was published in Kovács (2004), based on Sajtos and Mitev (2007), where the tools of descriptive statistics in SPSS were summarized in terms of proper scales.

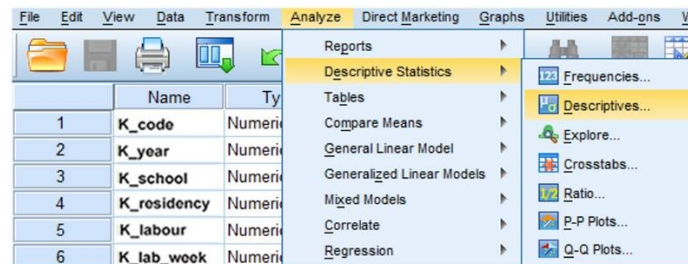
**Table 7/2. Connection between indices and scale types in descriptive statistics, modules applied in SPSS, based on Sajtos- Mitev (2007)**

Measures	Nominal scale	Ordinal scale	Interval or ratio scale
<b>Mean values</b>	Mode FREQUENCIES	Median, (Mode) FREQUENCIES	Averages (MEDIAN, MODE)
<b>Measures of dispersion</b>		Minimum, Maximum DESCRIPTIVES, FREQUENCIES	Minimum, Maximum DESCRIPTIVES, FREQUENCIES
<b>Indices of distribution</b>	Frequency, relative frequency FREQUENCIES	Range	Standard deviation, variance DESCRIPTIVES, FREQUENCIES

<b>Other measures</b>			Skewness, kurtosis
<b>Graphical display</b>	Bar chart and pie chart (for frequency) FREQUENCIES		Histogram FREQUENCIES

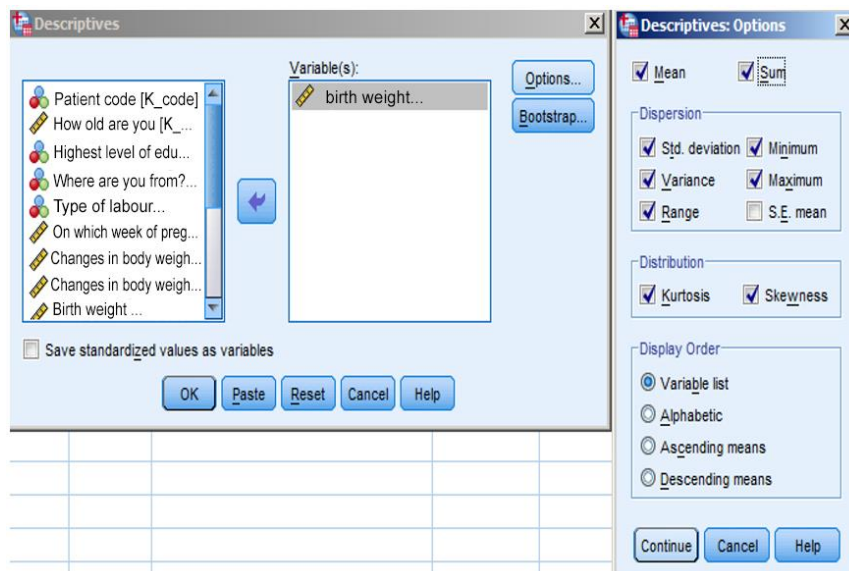
Source: author, based on Sajtos- Mitev (2007), Kovács (2004)

First, let us select the module *ANALYSE/DESCRIPTIVE STATISTICS/DESCRIPTIVE*.



**Figure 7/1. Access to DESCRIPTIVE**

As the next step, actual settings can be set. Variables will need to be chosen from the left-hand side; in this case it is the weight of new-born babies (“Gyermekének születési súlya”). It is also possible to examine more variables at the same time but in this case it is recommended to select variables with the same type of scale.



**Figure 7/2. Settings of DESCRIPTIVES**



After selecting the variable, press *OPTIONS* in order to be able to highlight the variables you need. All dispersion options except *STANDARD ERROR MEAN* should be selected here. Then press *CONTINUE* and *OK* for output.

**Table 7/3. Descriptive statistics of the weight of new-born babies**

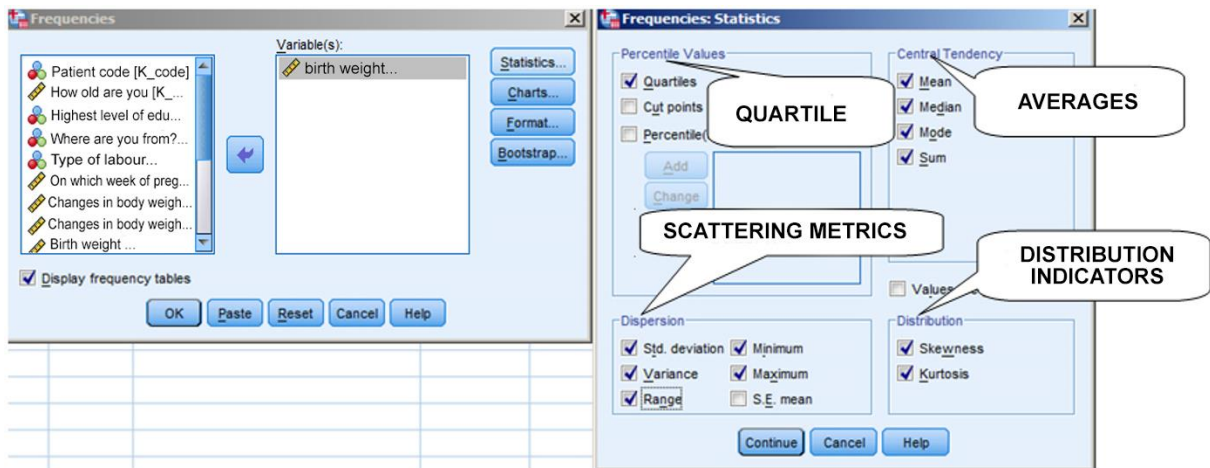
Descriptive Statistics												
	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Birth weight	70	2050,00	2150,00	4200,00	233080,00	3329,7143	547,43929	299689,772	-,294	,287	-,856	,566
Valid N (listwise)	70											

Source: author

The table contains the following outputs:

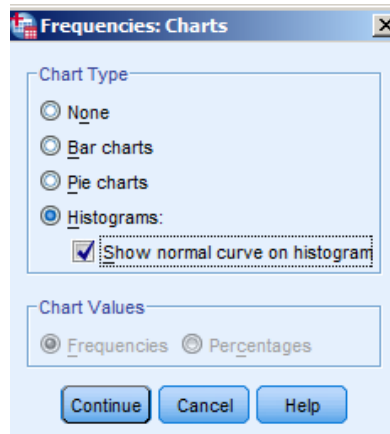
- data of 70 young mothers are gathered in the database (n=70).
- The *RANGE* of the variable is 2,050 grams, which is the difference between the maximum and the minimum values
- *MINIMUM* value belongs to the baby born with the lowest weight, i.e. 2,150 grams, while the *MAXIMUM* was 4,200 grams.
- The sum is 233,080 grams which is the total weight of new-born babies registered in the database.
- The average weight of the babies (*MEAN*) was 3,329.71 grams.
- The standard deviation of babies' weight at birth was 547,44 grams.
- Body weight variance was 299.689.77 grams.
- Value of *SKEWNESS*: -0.29 which means gentle left-skewed distribution (asymmetry).
- *KURTOSIS*: -0.86 i.e. flat data distribution.

Now, let us analyse the same variable with the help of frequencies (*ANALYSE/DESCRIPTIVE STATISTICS/FREQUENCIES*).



**Figure 7/3. Settings of FREQUENCIES**

Select the formerly introduced measures and quartiles. Then press *CONTINUE* and select the *CHARTS* module.



**Figure 7/4. Graphic chart module of the setting FREQUENCIES**

From the basic figure types (bar chart, pie chart, histogram), choose the histogram with the Gauss curve as the variable is continuous. Press *CONTINUE* and *OK* to get the results.

**Table. 7/4. Results of FREQUENCIES**

**Statistics**

Birth weight

N	Valid	70
	Missing	0
Mean		3329,7143
Median		3400,0000
Mode		3000,00 <sup>a</sup>
Std. Deviation		547,43929
Variance		299689,772
Skewness		-,294
Std. Error of Skewness		,287
Kurtosis		-,856
Std. Error of Kurtosis		,566
Range		2050,00
Minimum		2150,00
Maximum		4200,00
Sum		233080,00
Percentiles	25	2907,5000
	50	3400,0000
	75	3812,5000

a. Multiple modes exist. The smallest value is shown

Source: author

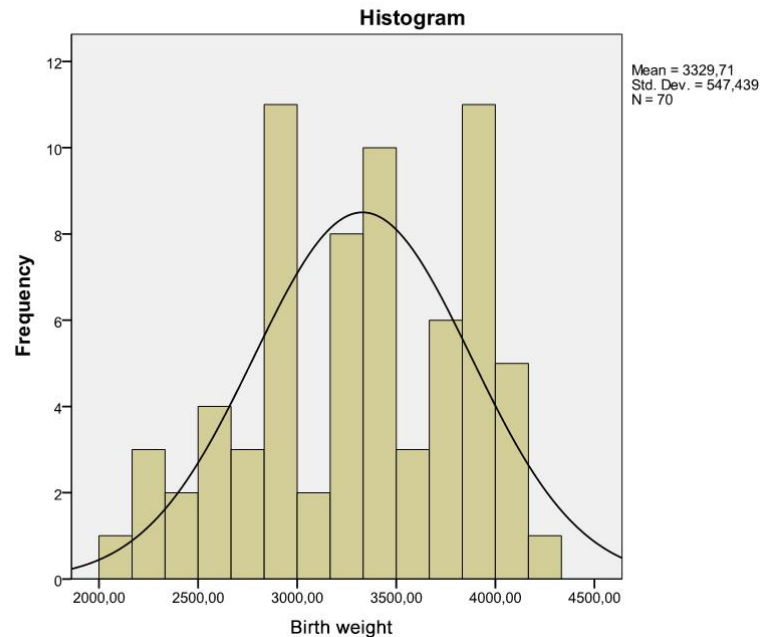
Data in the table contain data introduced above; here the quartiles will be the new data. The first quartile (25 percentile) states that 25 percent of young mothers gave birth to babies weighing less than 2.907.5 grams. The second quartile is the median, so in 50% of the cases the weight of the babies was below 3,400 grams, while in 50% of cases, the weight was higher than this. The third quartile, i.e. 75% of the babies weighed less than 3.812.5 grams. The new data also includes mode, which means that the most common birth weight was 3,000 grams but the sign at the value indicated that the variable has several modes. Then, the table of frequencies also appears.

**Table 7/5. Table of frequencies of birth weight**

Birth weight					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2150,00	1	1,4	1,4	1,4
	2200,00	1	1,4	1,4	2,9
	2300,00	2	2,9	2,9	5,7
	2400,00	1	1,4	1,4	7,1
	2500,00	1	1,4	1,4	8,6
	2550,00	1	1,4	1,4	10,0
	2600,00	1	1,4	1,4	11,4
	2640,00	1	1,4	1,4	12,9
	2650,00	1	1,4	1,4	14,3
	2700,00	1	1,4	1,4	15,7
	2770,00	1	1,4	1,4	17,1
	2800,00	1	1,4	1,4	18,6
	2850,00	1	1,4	1,4	20,0
	2900,00	3	4,3	4,3	24,3
	2910,00	1	1,4	1,4	25,7
	2980,00	1	1,4	1,4	27,1
	3000,00	5	7,1	7,1	34,3
	3100,00	1	1,4	1,4	35,7
	3120,00	1	1,4	1,4	37,1
	3200,00	4	5,7	5,7	42,9
	3210,00	1	1,4	1,4	44,3
	3300,00	3	4,3	4,3	48,6
	3400,00	4	5,7	5,7	54,3
	3450,00	1	1,4	1,4	55,7

Source: author

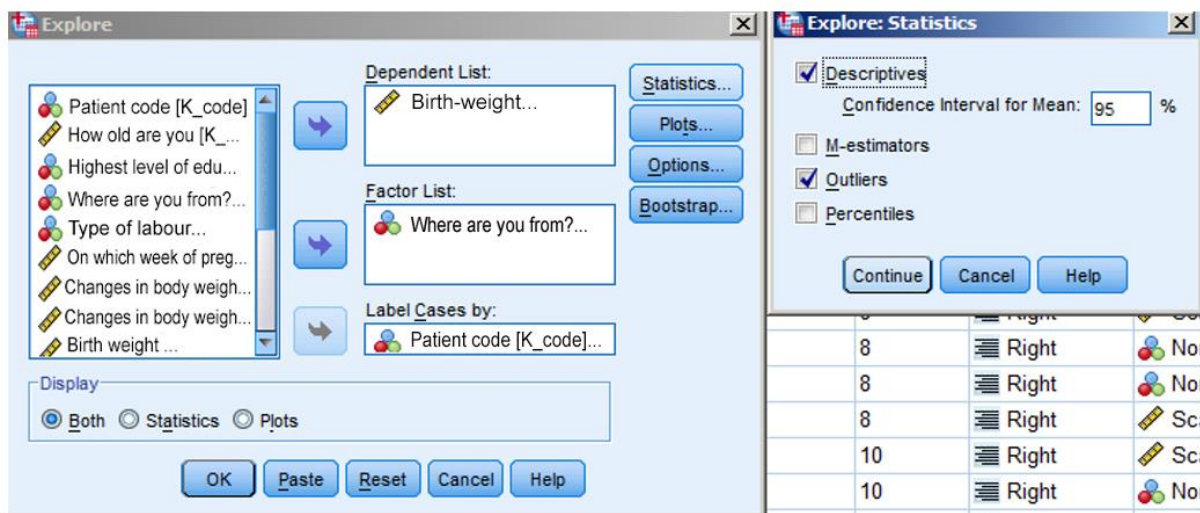
The first column (*VALID*) includes actual values of the continuous variable. The second column contains related *FREQUENCY*. In the third column (*PERCENT*), one can find the percentage of variables. The fourth one (*VALID PERCENT*) contains relative frequencies. *CUMULATIVE PERCENT* refers to accumulated frequency. Based on the table, there were five babies born with the weight of 3,000 grams which represents 7.1% of all the babies in the database. The 34.3% cumulative frequency means that 34.3% of babies were born with 3,000 grams or with less weight.



**Figure 7/5. Histogram of weights of new-born babies**

The histogram shows the normal distribution of the variable. It is a bar chart without gaps, on which continuous and discrete variables are indicated.

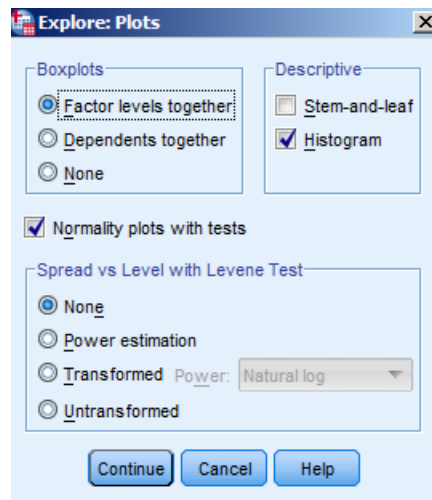
Now, we will examine the same variable with the help of frequencies in the module labelled *ANALYSE/DESCRIPTIVE STATISTICS/EXPLORE*.



**Figure 7/6. Settings of the module EXPLORE**

We will have to select the birth weight to be the dependent variable and the place the person comes from (“Honnan származik Ön”) to be the grouping or factor variable. These cases will be examined on the basis of their code. These settings also generate additional information such as descriptive statistics for different categories, containing data in terms of the category

chosen before. Of course, it is possible to carry out the normality test both in a graphic and numeric version.



**Figure 7/7. Graphical settings of the menu EXPLORE**

**Table 7/6. Results of the module FREQUENCIES**

Descriptives				Statistic	Std. Error
Birth weight	City	Where are you from?			
		Mean		3260,8571	83,32822
		95% Confidence Interval for Mean	Lower Bound	3091,5138	
			Upper Bound	3430,2005	
		5% Trimmed Mean		3266,4286	
		Median		3300,0000	
		Variance		243025,714	
		Std. Deviation		492,97638	
		Minimum		2200,00	
		Maximum		4200,00	
		Range		2000,00	
		Interquartile Range		750,00	
		Skewness		-,106	,398
		Kurtosis		-,569	,778
	Village	Mean		3398,5714	100,76943
		95% Confidence Interval for Mean	Lower Bound	3193,7833	
			Upper Bound	3603,3595	
		5% Trimmed Mean		3425,3968	
		Median		3470,0000	
		Variance		355406,723	
		Std. Deviation		596,15998	
		Minimum		2150,00	
Maximum		4100,00			
Range		1950,00			
Interquartile Range		980,00			
Skewness		-,532	,398		
Kurtosis		-,872	,778		

Source: author

As shown by the table, the population has been divided into two groups, based on the place one comes from, i.e. birth weights are displayed for young mothers coming from a town (város) or village (falu). There are new descriptives we did not define before but they are still

vital for us. The 95% *CONFIDENCE INTERVAL FOR MEAN LOWER BOUND/UPPER BOUND* contains the results of *statistical estimation*.

*Statistical estimation* is the approximate determination of a constant parameter of an unknown population. These parameters can be the expected value (for a finite population, average), the standard deviation and the ratio. A theoretical and practical introduction into the topic of statistical estimation will be offered in Chapter 9.

In our example, 95% of *CONFIDENCE INTERVAL FOR MEAN LOWER BOUND/UPPER BOUND* means that there is 95% probability that the birth weight of babies with the mother coming from a town will be between 3,091.51 and 3,430.20 grams.

It is often useful to display these results in graphics, the options of which we have shown above.

The *EXPLORE* module – as introduced before – can be used for graphical analysis of outliers and normality. To examine outliers, one may also use the table in which the five highest and lowest values of the categories are listed, naming specific cases (*LABELS*, code (“betegkód”)).

**Table 7/7. Table of the five highest and lowest values**

Extreme Values						
		Where are you from?		Case Number	Patient code	Value
Birth weight	City	Highest	1	24	24	4200,00
			2	56	56	4050,00
			3	8	8	4000,00
			4	55	55	3900,00
			5	54	54	3850,00
		Lowest	1	14	14	2200,00
			2	60	60	2400,00
			3	1	1	2500,00
			4	2	2	2600,00
			5	37	37	2700,00
	Village	Highest	1	22	22	4100,00
			2	33	33	4100,00
			3	67	67	4100,00
			4	34	34	4080,00
			5	21	21	<sup>a</sup>
		Lowest	1	15	15	2150,00
			2	61	61	2300,00
			3	31	31	2300,00
			4	62	62	2550,00
			5	36	36	2640,00

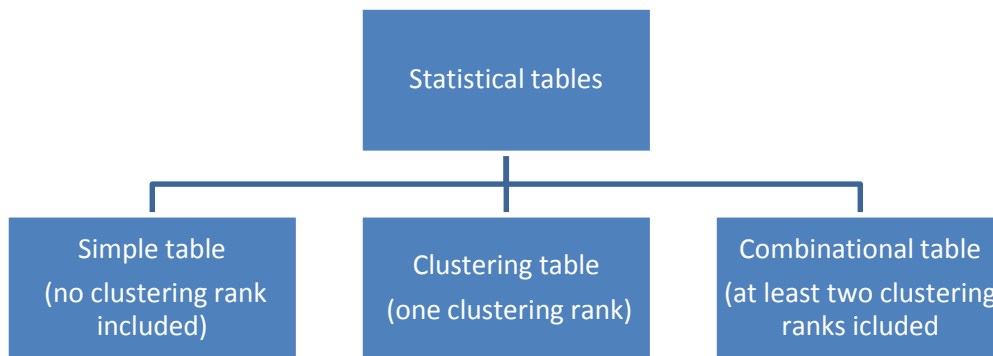
Only a partial list of cases with the value 4000,00 are shown in the table of upper extremes.

Source: author

Of course, this table helps us detect *outlier* values. The interpretation of these results was discussed in the chapter entitled data transformation, so now we shall not discuss the issue of normality test and outliers once again.

## 7.2. Statistical tables

Statistical tables are very commonly used data-presenting tools as their objective is to organise and compress data. They are supposed to show a system connection of the data they include and give a complex picture on the phenomenon examined. A statistical table is an order of statistical ranks where data is listed according to one or more aspects (variable). Statistical tables contain statistical ranks (time, area, qualitative, quantitative). Tables are usually classified according to two aspects. Based on the number of dimension, there are two or three dimension tables. Deciding which one we are facing depends on the number of variables indicated in the table. The other classification is based on the aim of listing variables, since it can be listed either for comparison or for clustering. Consequently, we have the following types:



**Figure 7/8. Clustering statistical tables**

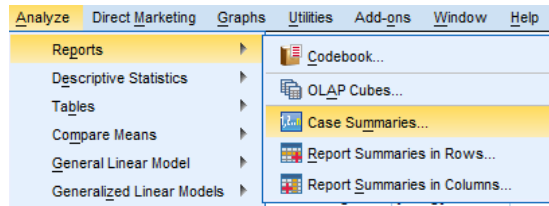
Source: Ács (2009)

Most of the statistical tables belong to the category of combinational tables. Tables containing the frequency distribution of variables are called contingency tables. There are formal requirements for statistical tables and figures to be met, the lack of which can lower the professional level of one's research (e.g. thesis). These requirements for format include the title, the source and an explanation. Requirements of content (i.e. entirety, classifiability) include that each item must have a place where it can be placed according to their corresponding data.

There are several options in SPSS to generate statistical tables.

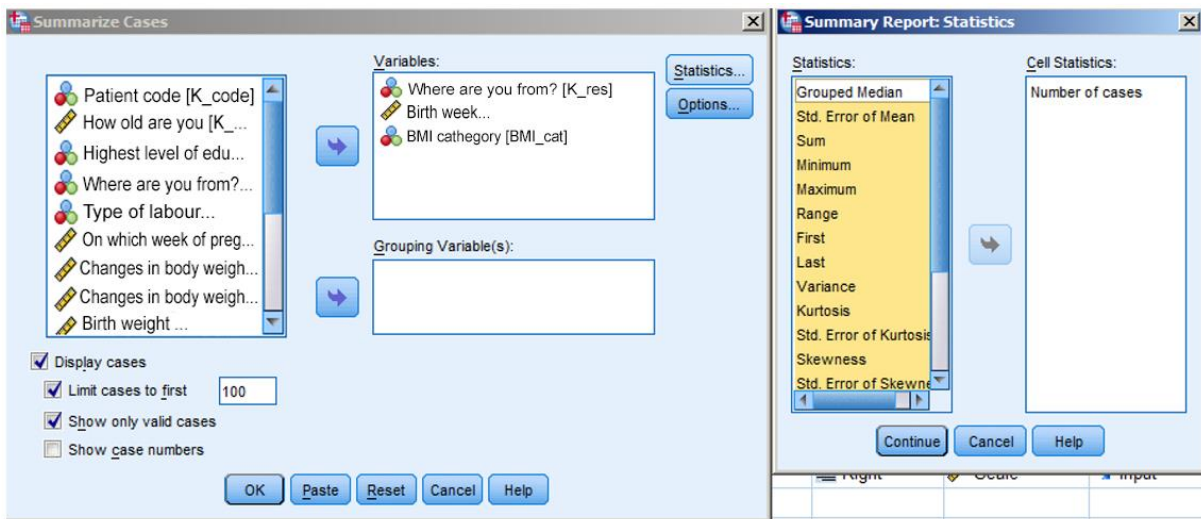
One of the simplest methods is when we would like to display data in the correct order. This function is available under *ANALYSE / REPORTS / CASE SUMMARIES*. Then, the user has to set the specific variables.





**Figure 7/9. Introducing cases according to variables**

Let us introduce all cases according to the place of origin (K\_lakhely), the week of birth (“Hányadik terhességi héten szült?”), and BMI categories (BMI\_kat).



**Figure 7/10. Setting of summary report**

Variables can be selected on the left-hand side, then one can also specify the *GROUPING VARIABLES*. Pressing *STATISTICS* will make it possible to list statistical indices. The table can be given a title (*TITLE*) and remarks can also be added (*CAPTION*) in *OPTIONS*. Pressing OK will generate the table.

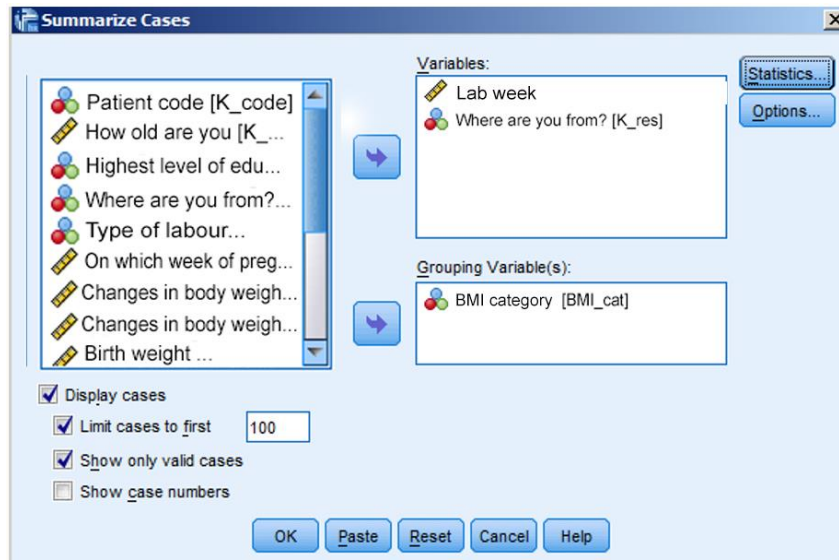
**Figure 7/8. Table of case summaries according to variables)**

**Case Summaries<sup>a</sup>**

	Where are you from?	Birth week?	BMI categories
1	City	38,00	normal weight
2	City	35,00	normal weight
3	City	36,00	normal weight
4	City	38,00	normal weight
5	City	37,00	normal weight
6	City	35,00	normal weight
7	City	34,00	normal weight
8	City	36,00	normal weight
9	City	37,00	normal weight
10	City	38,00	normal weight
11	City	39,00	normal weight
12	City	40,00	normal weight
13	City	41,00	normal weight
14	City	38,00	normal weight

Source: author

For practice, let us display data assigned to specific BMI categories.



**Figure 7/11. Selecting grouping variables**

In the settings window, OK has to be clicked on after selecting the BMI category as the grouping variable.

**Table 7/9. Grouping according to BMI categories**

			On which week of pregnancy was your child born?	Where are you from?	
BMI categories	underweight	1	39,00	City	
		2	39,00	City	
		3	37,00	City	
		4	38,00	Village	
		5	39,00	Village	
		Total	N	5	5
	normal weight	1	38,00	City	
		2	35,00	City	
		3	36,00	City	
		4	38,00	City	
		5	37,00	City	

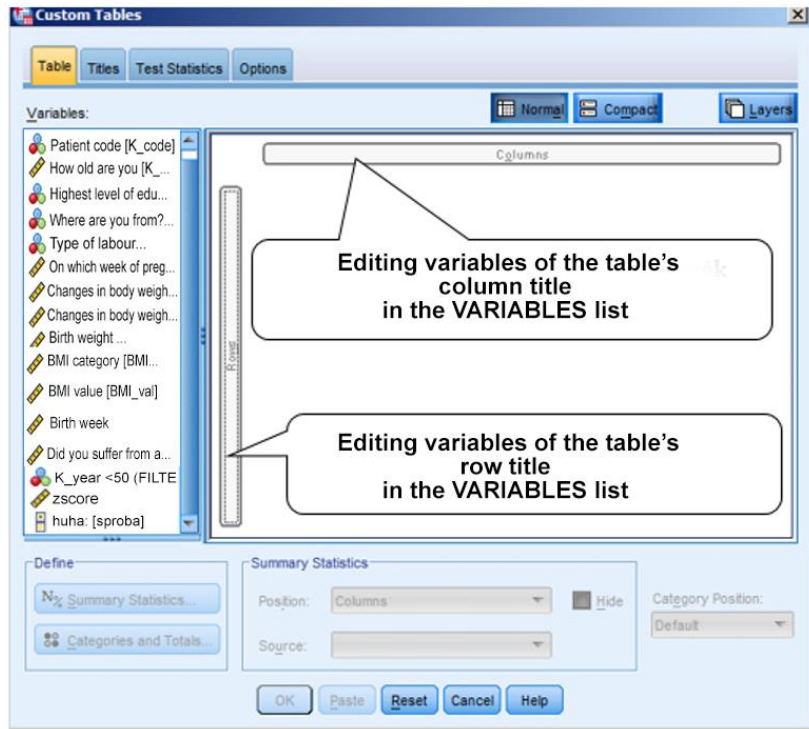
Source: author

Chosen variables are displayed now in the order of BMI categories.

A typical table generating module is available at *ANALYSE /TABLES, CUSTOM TABLES*.

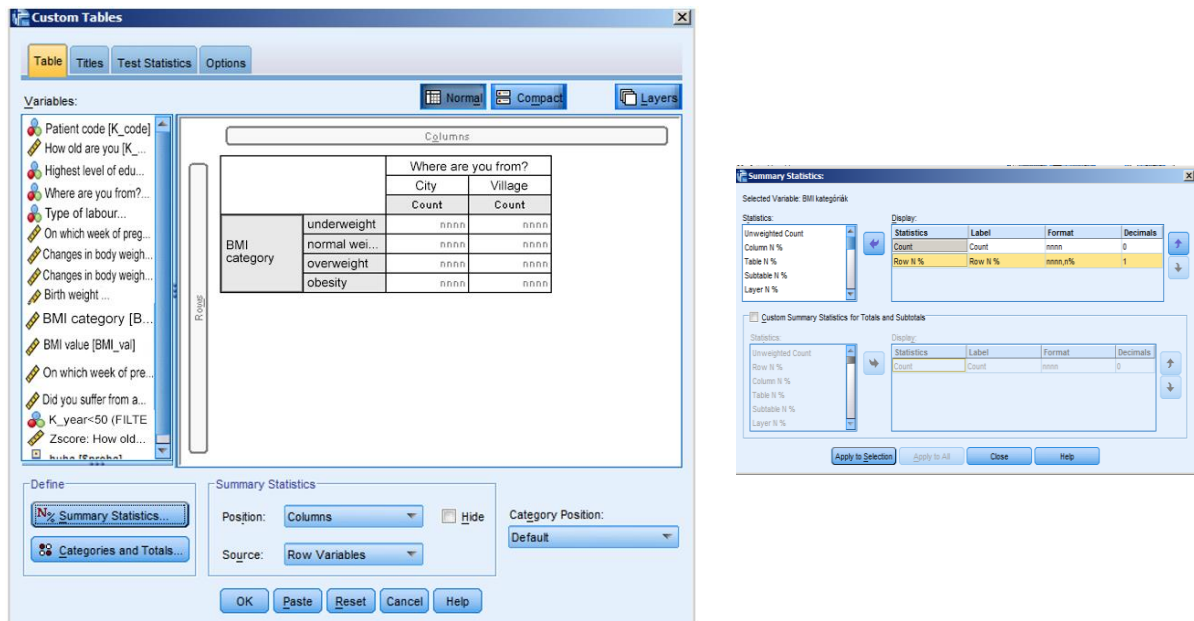
Press OK to go on the new panel to change settings.

The basic structure in the new window is a two-dimensional cross tab, in which the frequency of cells (*COUNT*) is the default setting.



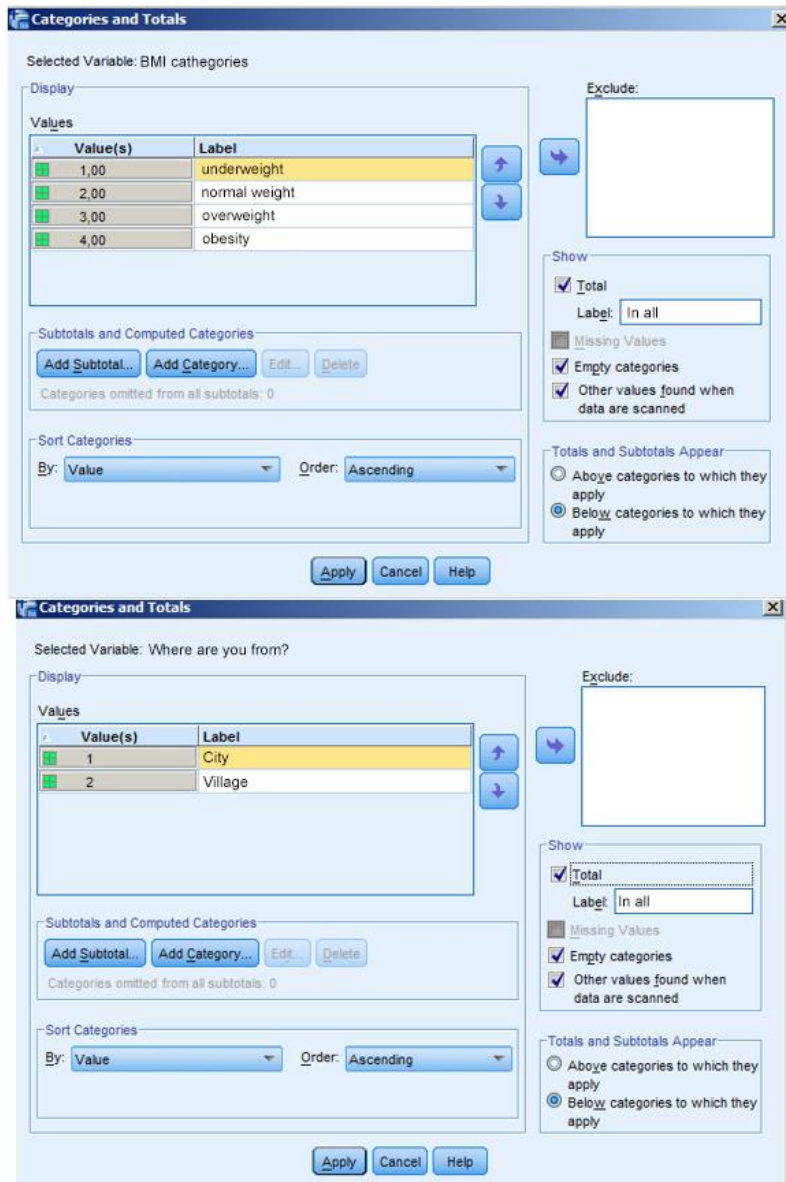
**Figure 7/12. Settings module of statistical table**

**For practice,** let us plot units in the BMI category according to their place of origin, in a contingency table. Let us also display their percentage besides their number of units. First, the variable BMI category will need to be moved to the heading of the row, and then the variable place of origin to the heading of the column.



**Figure 7/13. Calibrating variables and indices**

The display of percentages can be set in *DEFINE / SUMMARY STATISTICS*. Here, *ROW N%* option will need to be chosen with the arrow in the middle of the panel *STATISTICS*. Now, percentages will appear row by row. New statistics can be selected in *APPLY* to *SELECTION*. Then, it can be transferred to a combinational table in *CATEGORIES* and *TOTALS*.



**Figure 7/14. Editing categories and totals**

Since the combinational table consists of two sums, rename *LABEL* in the box *TOTAL* to sum (“Összesen”). Press *APPLY* then *OK* to get new table in *OUTPUT* view.

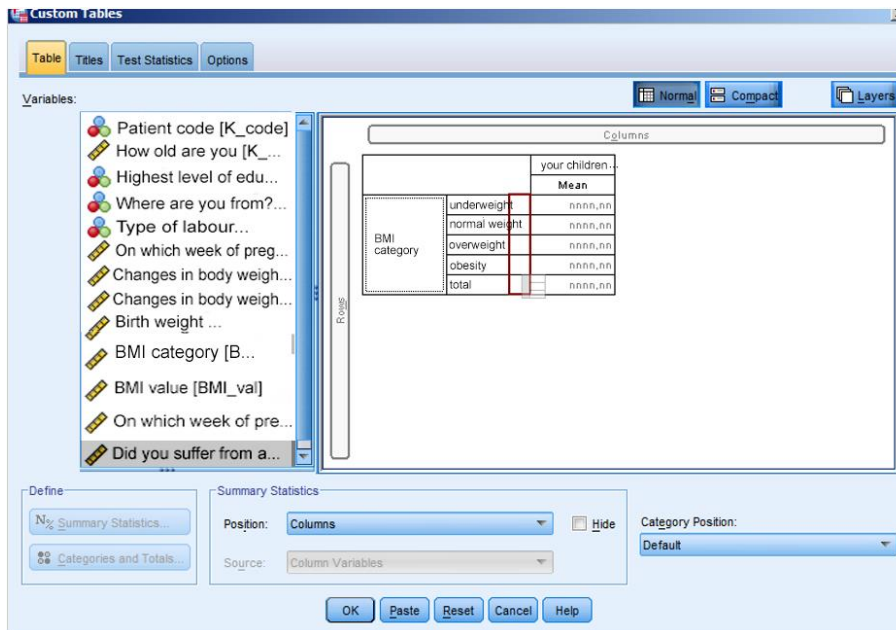
**Table 7/10. Making the combinational table**

		Where are you from?					
		City		Village		In all	
		Count	Row N %	Count	Row N %	Count	Row N %
BMI categories	underweight	3	60,0%	2	40,0%	5	100,0%
	normal weight	19	82,6%	4	17,4%	23	100,0%
	overweight	8	38,1%	13	61,9%	21	100,0%
	obesity	5	23,8%	16	76,2%	21	100,0%
	total	35	50,0%	35	50,0%	70	100,0%

Source: author

As indicated in the table, there were only five young mothers with an obesity issue, coming from a town (21), which represents 23.8% of those living in a town. Of course, the table can be further moderated by clicking twice on the left-hand side or twice on the right-hand side of the mouse. The number of dimensions and categories can be extended, too.

In the next table, we will show the average of birth weight in terms of BMI categories and chronic illnesses as grouping variables. To do so, the user will have to move the birth weight into the variable column, and then move the variable of chronic illness to the right edge of the variable BMI category.



				birth weight ...
				Mean
BMI categories	underweight	Did you suffer from a chronic illness?	yes	nnnn,nn
			no	nnnn,nn
	normal weight	Did you suffer from a chronic illness?	yes	nnnn,nn
			no	nnnn,nn
	overweight	Did you suffer from a chronic illness?	yes	nnnn,nn
			no	nnnn,nn
	obesity	Did you suffer from a chronic illness?	yes	nnnn,nn
			no	nnnn,nn
	total	Did you suffer from a chronic illness?	yes	nnnn,nn
			no	nnnn,nn

**Figure 7/15. BMI and chronic illnesses**

Continuous variables display averages as default, which can be modified by integrating statistics in option *SUMMARY STATISTICS*.

Pressing *OK* will generate the next statistical table.

**Table 7/11. The extended statistical table**

				Birth weight
				Mean
BMI categories	underweight	Did you suffer from a chronic illness?	yes	.
			no	2840,00
	normal weight	Did you suffer from a chronic illness?	yes	3630,00
			no	3202,63
	overweight	Did you suffer from a chronic illness?	yes	3453,33
			no	3128,67
	obesity	Did you suffer from a chronic illness?	yes	3680,00
			no	3585,33
total		Did you suffer from a chronic illness?	yes	3582,50
			no	3254,81

Source: author

### 7.3. Statistical graphs, diagrams

The essence of statistics and theoretical works is the message, and the illustration. For someone lacking expertise in the field we must stress that the “infinite” number of data is not to be interpreted. That is why it is important to present results in a convincing and interpretable way. Graphical display is probably the most illustrative tool for this purpose. It is not only illustrative but also helps recognize the relevance, since it does not only give information about the absolute measure of the data but also supports comparison by showing ratios. Graphs may be displayed with or without a coordinate system. A coordinate system may provide additional information but there are cases when it is not necessary to apply this method.

It is important to choose the proper type and tool of graphs so that it serves our purpose. The type of graphs needs to fit the specific phenomenon or characteristic one would like to illustrate. The tools of graphs include basic geometric items like points, lines, squares, rectangulars, circles and their variations (Székelyi – Barna 2005).

Graphs can illustrate changes, differences in distribution or time, comparison of measures, etc. (Jánosa 2011; Sajtos – Mitev 2007).

When editing graphs, one should consider the following factors:

- the titles and names should be meaningful and they should correspond to what they are representing
- the time interval of the chart needs to be mentioned



- the measurement unit of the data has to be mentioned since the figure cannot be interpreted without this piece of information
- to ensure understanding, a list of notations will need to be added

Before introducing the access to and the examples of graphical illustrations in SPSS, it is important to review the list of tools available, just as the fact what these can be used for (emphasis will be placed on simple graphical tools, so the various types of analyses will be mentioned, but not explained in detail) (Sajtos – Mitev 2007).

**”Bar Chart”**: this type of diagrams plots the structure of the population in terms of a categorical variable;

**”3D Bar Chart”**: the same option for application as that of the bar chart, the only difference is in their shape;

**“Line Chart”**: primarily applicable for time series where data points are bound by a continuous line;

**“Area Chart”**: illustration of area covered by data curves;

**“Pie Chart”**: illustrates composition and structure of data population graphically;

**“High Low Chart”**: applied to compare data pairs that belonging together;

**“Pareto Chart”**: a combined tool since and a combination of the bar chart and the line chart containing the accumulated sum;

**“Boxplot Chart”**: most descriptive characteristics are displayed on this chart such as the median, the minimum, the maximum or the quartiles

**“Control Chart”**: one can check if the variable has only values between the given limits or not;

**”Scatter Chart”**: a tool for examining the connection between quantitative criteria; it is similar to the tool, ”High Low Chart” in the sense that this tool also displays value pairs (here point cloud imaging is used for the display)

**“Error Bar Chart”**: as implied in the name itself, this tool shows standard errors of a variable, its standard deviation and its difference from the confidence interval;

**“Histogram”**: displays the distribution of a variable but is only applicable in case of metrical scales;

**“P-P Plot Chart”**: a tool for comparing two distribution functions where the function of a given variable will be compared to the normal distribution;

**“Q-Q Plot Chart”**: in case of a variable, quartiles can be examined between the theoretical quartiles of the normal distribution;

“*Sequence Chart*”: plotting different cases in chronological order; most appropriate for time series;

”*Autocorrelations Chart*”: a tool for examining autocorrelation between different time series;

“*Cross-correlation Chart*”: a correlation of two time series can be examined with this tool;

“*Spectral Chart*”: a complex type of line charts, applied for spectral estimation;

“*Population pyramid Chart*”: in epidemiology, its output is often referred to as age pyramid as it plots two opposite bar charts or histograms;

“*Mixed Chart*”: a mixed type of diagram in which two different types of charts can be combined in one chart (Jánosa 2011, Sajtos – Mitev 2007).

### Graphical illustration in practice

The majority of graphic tools are found under the menu **GRAPHS** in SPSS. After clicking on it, select the tool that fits the research goals the most. There are two options to plot a diagram in this menu; these include the *CHART BUILDER* and the *LEGACY DIALOGS* options.

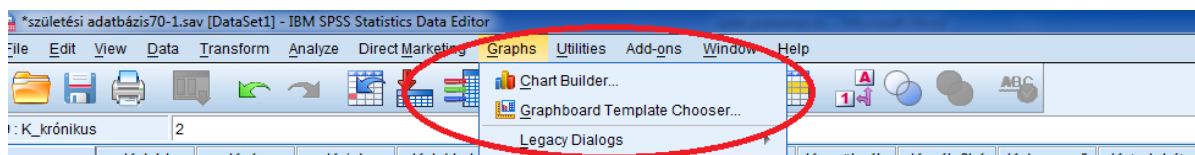


Figure 7/15. Options in menu **GRAPHS**

The *CHART BUILDER* provides a flexible and quite modern mode of editing. Running it for the first time, the programme will display a dialog box in which settings of metrical scales are listed for the variables. Pressing *OK* will make this box disappear but the option “*Don’t show me again*” will hide it.



**Figure 7/16. Editing panel of CHART BUILDER**

Variables in the database are listed in the box labelled *VARIABLES*. Under the tab *GALLERY*, the list of basic charts types are offered.. There are two options to choose from this list. Move the variable to the cell ,”*Chart preview uses example data*”, then select the proper type by clicking on it twice, and add the dependent and independent variables in relation to which the analysis will need to be made. The variable can be placed on the X (independent) and Y (dependent) axis of the coordinate system, or you can build the chart from the basic element with the help of the tab “*Basic elements*”.

Tab *TITLESS/FOOTNOTES* is used to add title or footnote to the graph.

The other option to insert a chart is labelled *LEGACY DIALOGS*, clicking on which the proper type can be selected. Then the chart is generated through determining the variable and the axes (Jánosa 2011).

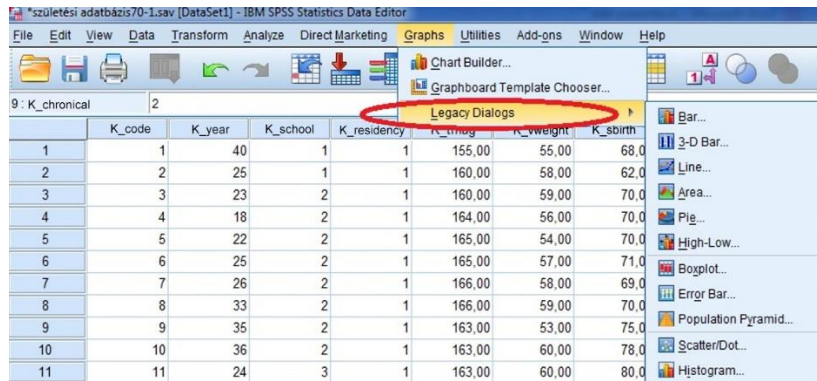


Figure 7/17. The editing panel of *LEGACY DIALOGS*

In the followings, an explanation will be provided for the more frequently used types such as line charts, bar charts, pie charts, and histogram).

### “Line Chart”

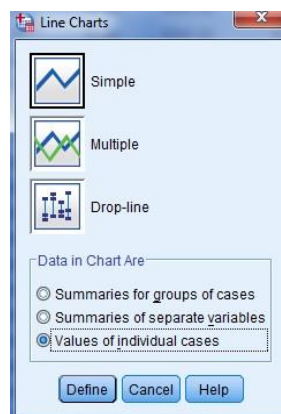
If the transition between our various data has a meaning, the value points can be connected, resulting in a line chart. In these cases, ratios and measures of changes during time intervals can be seen.

The database contains data on nutrition habits during pregnancy and its consequences in terms of change in body weight. The database is based on the survey of the following questions:

- Code of patient? “Betegkód”
- Age? “Hány éves Ön?”
- What is your highest level of qualification? “Mi az Ön legmagasabb iskolai végzettsége?”
- Where are you from? “Honnan származik Ön?”
- Body height? “Testmagassága?”
- Body weight before labour? “Szülés előtti testsúlya?”
- Body weight on the day of labour? “Szülés napján mért testsúlya?”
- Body weight of the baby born? “Gyermekeének születési súlya?”
- Body weight 6 months after labour? “Szülés után 6 hónappal mért testsúlya?”
- Do you consider your knowledge on healthy nutrition to be up-to-date? “Korszerűnek ítéli meg az egészséges táplálkozással kapcsolatos ismereteit?”
- On which week of pregnancy was your child born? “Hányadik terhességi héten szült?”
- Did you suffer from a chronic illness? “Volt e krónikus betegsége?”
- Type of labour? “Szülésének kimenetele?”
- Pains during labour (on a scale of 1-5)? “Szülés közben érzett fájdalom (1-5 skálán)?”

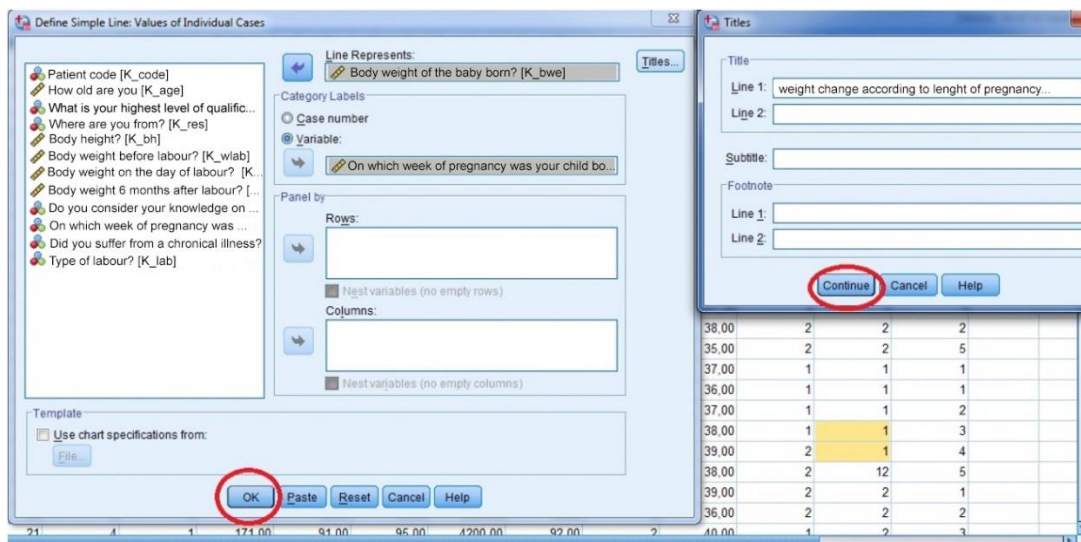
As we mentioned, a line chart can be applied to illustrate changes in time, the first example includes variables of birth weight and the week of pregnancy when the child was born.

There are several options to generate a line chart in SPSS. One of them is *GRAPH* → *LEGACY DIALOGS* → *LINE CHART* where one has to choose the type of line chart. Here we select *SIMPLE* and *VALUES OF INDIVIDUAL CASES*, then press *DEFINE*. The option *SIMPLE* is for plotting one data series, while the option *MULTIPLE* is for two or more. In order to examine the differences between connecting elements of multiple data series, the option *DROP-LINE* will need to be selected (Jánosa 2011, Sajtos – Mitev 2007).



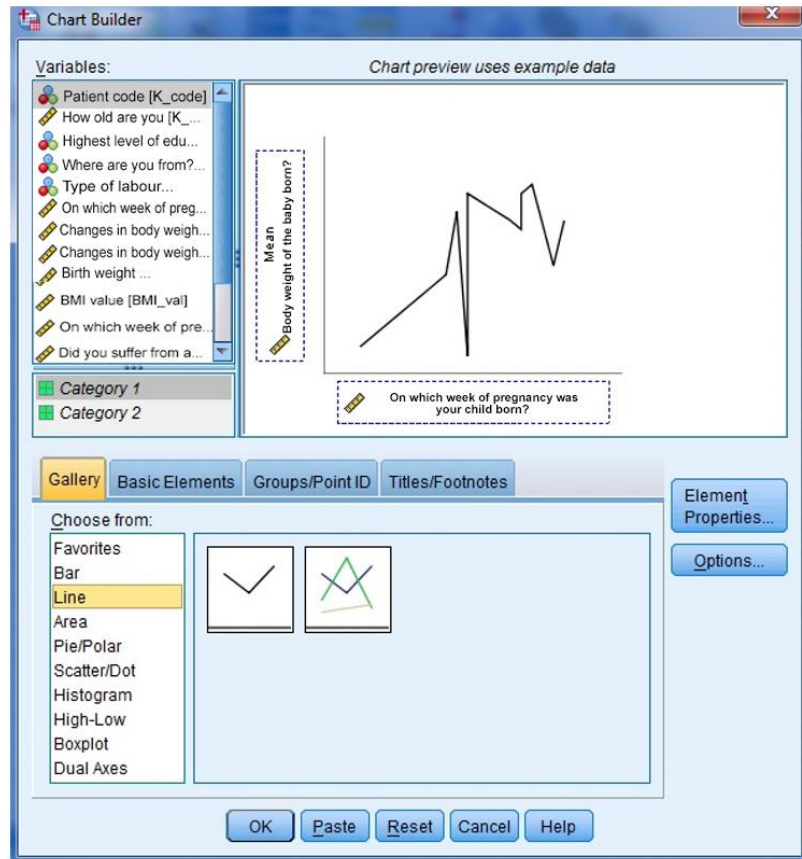
**Figure 7/18. Box for selecting the type of line chart**

After selecting the proper type, we get the following edit panel where we can add variables and give a title to the figure at *TITLES*. The process is finalized by pressing *CONTINUE* and *OK*.



**Figure 7/19. Edit panel for line charts (menu option *LEGACY DIALOGS*)**

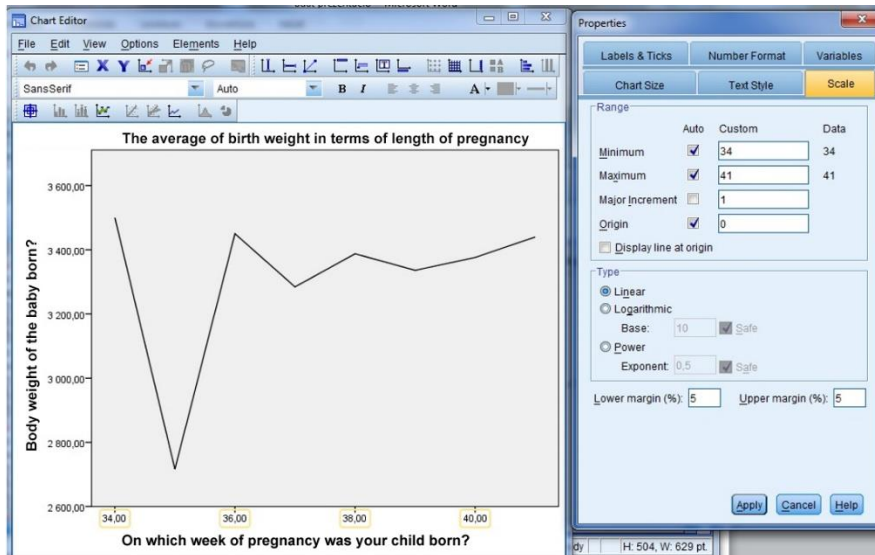
The second option is using the tool *CHART BUILDER*. It is available at *GRAPH* → *CHART BUILDER* where we choose *GALLERY* and then select the type of *LINE*, then the line chart will be generated by settings shown in the figure.



**Figure 7/20. Edit panel for line chart with *CHART BUILDER***

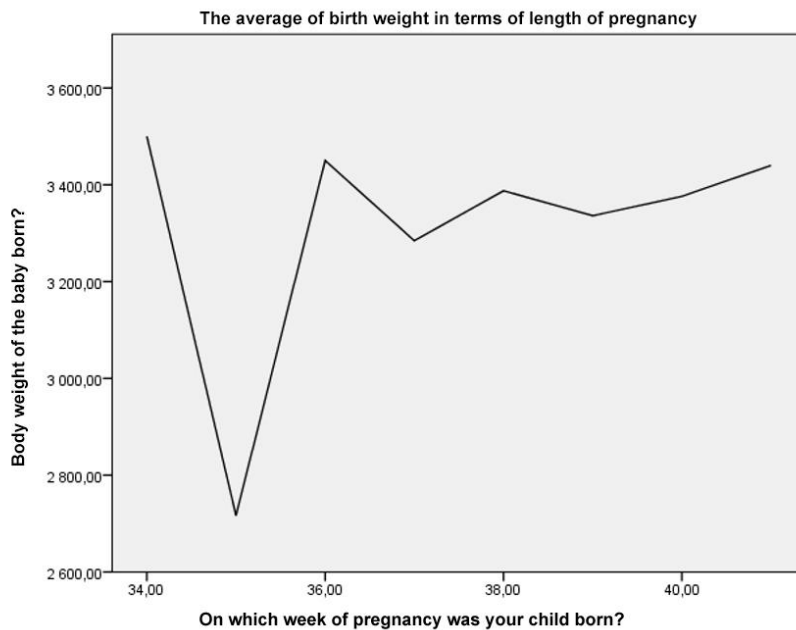
The title of the graph and footnotes can be added under the tab *TITLES/FOOTNOTES*. The information to be published can be typed in at *CONTENT*, then click *APPLY* to accept it. Finalize the process by pressing *OK*.

If there is something in the graph does not meet with our needs on the graph (e.g. the classification of scale of pregnancy length), then we can modify it by clicking twice on the diagram in the *OUTPUT VIEW*, and then the edit panel appears. Clicking twice on the X axis (length of pregnancy), the panel *PROPERTIES* will open up. Choosing here the tab *SCALE* makes it possible to plot the minimum and maximum, but in this example the classification can be set in *MAJOR INCREMENT* (Jánosa 2011).



**Figure 7/21. Modification of line chart settings**

It does not matter which version we chose, the output will be the following line chart (without modification of scale classification).



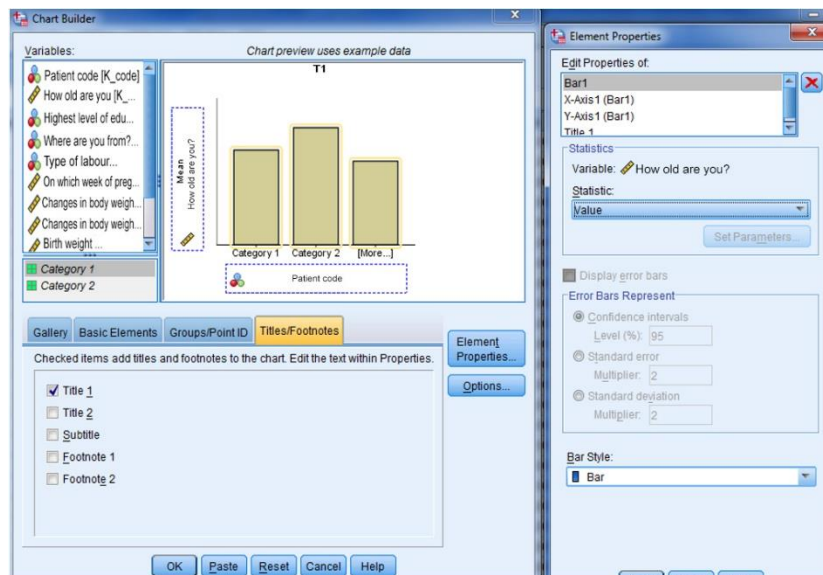
**Figure 7/22. The average birth weight in terms of length of pregnancy**

## “Bar Chart”

In case of frequencies belonging to categories, the bar chart is the most frequently applied tool to compare data and identify their ratio. Inter alia, it makes it possible to compare measures. It is a frequently applied and popular graphical tool.

In our example, we aim to illustrate and compare the age of young mothers (“Édesanyák életkora”). The X axis represents the code of patient (“Betegkód”), while Y stands for age (“Hány éves Ön?”).

SPSS offers two options to plot this type of diagram. One can use the menu *CHART BUILDER*, and then choose the most proper type of bar chart under option *GALLERY / BAR*. Variables have to be moved to the axes. After giving a title, press *OK*, and the graph will be completed.



**Figure 7/23. Edit panel for bar chart (*CHART BUILDER*)**

Of course, it is also possible to edit the figure in *LEGACY DIALOGS*. Following the access path *GRAPHS → LEGACY DIALOGS → BAR*, one can select the type. Choose the type *SIMPLE* and *VALUES OF INDIVIDUAL CASES*, then press *DEFINE* to make the edit panel appear. Note that you can apply the type of *SIMPLE* can be applied grouped according to one specific aspect or by applying several data series so that no grouping variable can be added. Type *CLUSTERED* is for plotting a given data series categorized by a primary and a secondary aspect. In case of type *STACKED*, several data series can be plotted, which will be stacked in one column. Options at the bottom refer to data groups to be displayed, meaning that the options are included in their titles which are labelled as *Summaries for group of cases*,



Summaries of separate variables, Values of individual cases (Jánosa 2011, Sajtos – Mitev 2007).

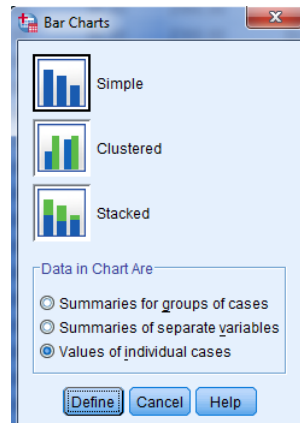


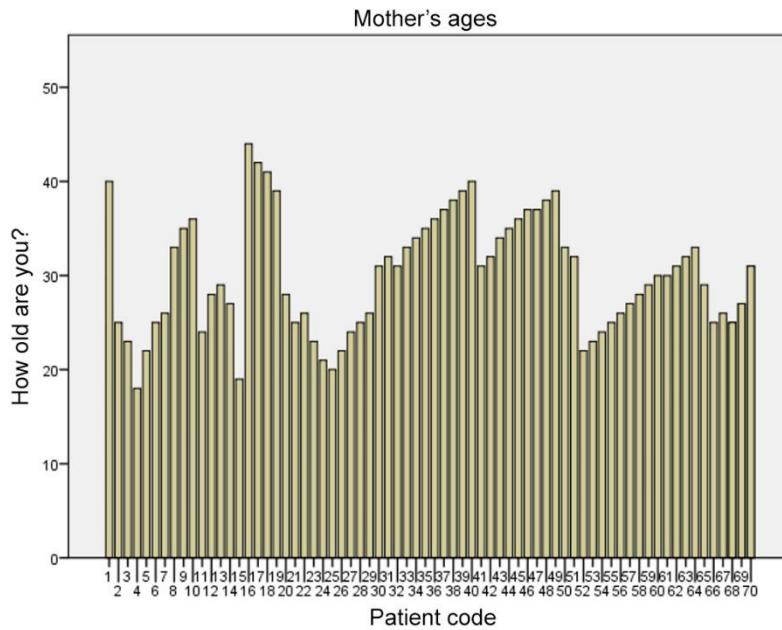
Figure 7/24. Selecting the type of bar chart

In the edit panel, the cell *BARS REPRESENT* should contain age (“Hány éves Ön?”), while in the *CATEGORY LABELS / VARIABLE* we add the code of the patient (“Betegkód”). In *TITLES*, one can add the title of the bar chart and also footnotes, if necessary. Press *OK* to generate the chart.



Figure 7/25. Edit panel for bar charts (in *LEGACY DIALOGS*)

Both options will result in the same bar chart, comparing the age of young mothers, based on their code.

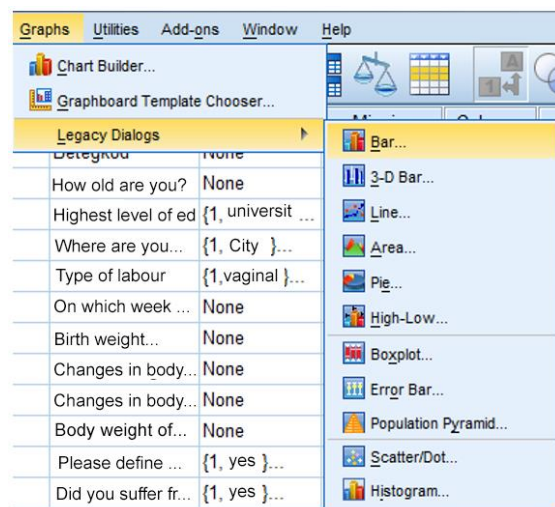


**Figure 7/26. The age of young mothers based on the code of patients**

Confidence interval (CI) has an important role when dealing with bar charts. Its calculation is carried out while making statistical estimates.

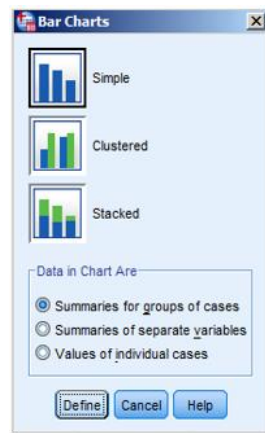
**As a practice exercise**, let us display the expected birth weight with 95% confidence interval, grouped according to residence.

Related settings are available at *GRAPHS / BAR*.



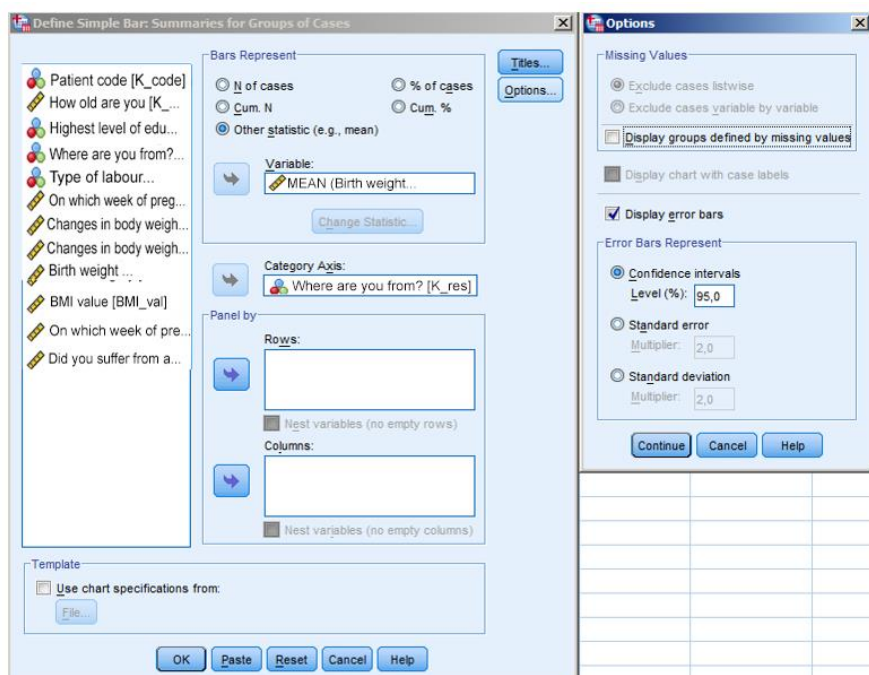
**Figure 7/27. Access path to the bar chart**

Afterwards, select the type of bar chart; here we will choose in this instance *SIMPLE* (one we may add a grouping variable) and *SUMMARIES FOR GROUP OF CASES*, then click on *DEFINE*.



**Figure 7/28. Defining the basic type of bar chart**

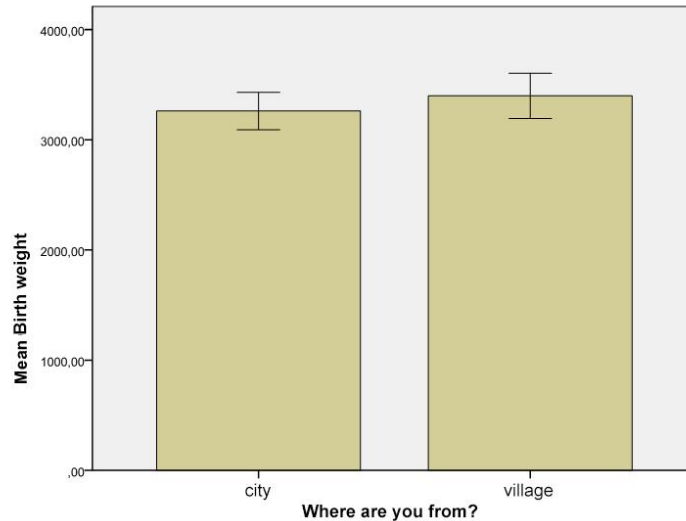
Now, the variable to be examined and its index will have to be selected. We would like to analyse the average of birth rate, and use the place of origin as grouping variable.



**Figure 7/29. Bar chart and confidence interval calculation**

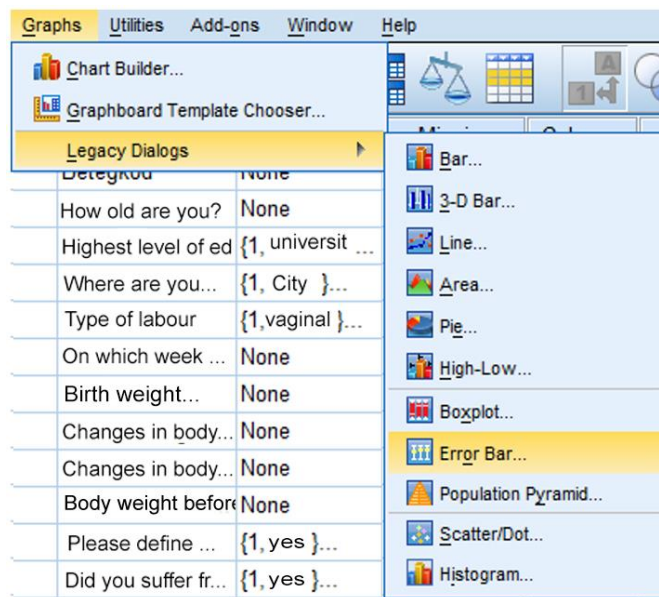
In the module *OPTIONS*, a confidence interval can be added to the bar chart by clicking on *DISPLAY ERROR BARS*, then *CONFIDENCE INTERVALS LEVEL(%)*: 95.

The diagram is calibrated.

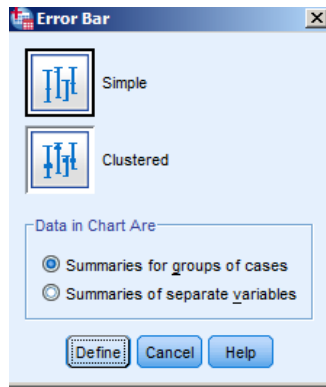


**Figure 7/30. Bar chart and confidence intervals (statistical estimation)**

The other display option for the confidence interval is the *ERROR BAR* which may show the confidence interval, the standard error or the standard deviation of a point (average) estimation.



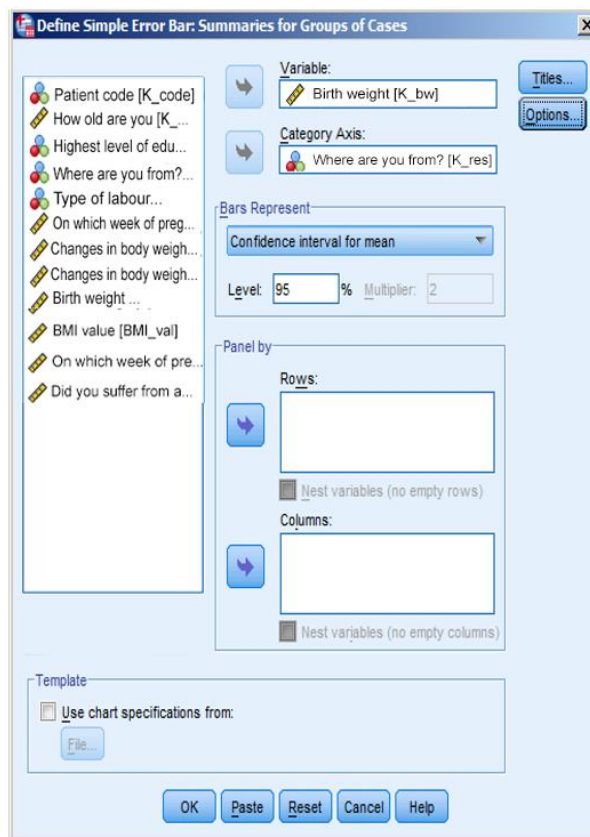
**Figure 7/31. Access path to confidence interval and margin of error**



**Figure 7/32. Defining the basic type of the diagram**

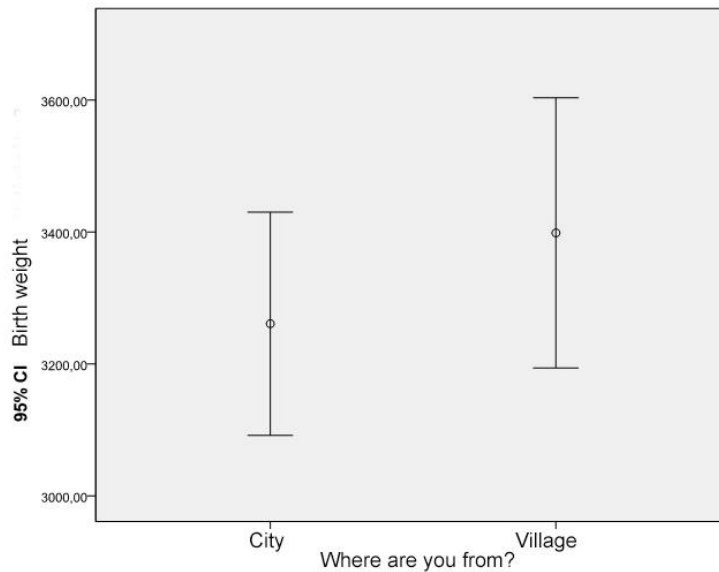
Afterwards, hawse will need to select the type of bar chart, here we will choose *SIMPLE* (we may add a grouping variable) and *SUMMARIES FOR GROUP OF CASES*, and click on *DEFINE*.

As the next step, the variable will need to be examined and the grouping variable (place of origin) set. In *BARS REPRESENTS*, the form of display can be selected where we choose the confidence interval that belongs to the average.



**Figure 7/33. Diagram settings**

When pressing *OK*, the chart will appear. It includes confidence intervals grouped according to values of the grouping variable.



**Figure 7/34. Confidence intervals (*ERROR BAR*)**

### **“Pie Chart”**

Pie charts display the composition of the population. They plot the parts of a round, and show their ratio. There are two types of chart to be distinguished: charts with one or multiple data series.

#### **Charts with one data series**

These charts consider a time unit, an area or an organizational unit to be a round, and display their parts (slices), their ratios, and their share.

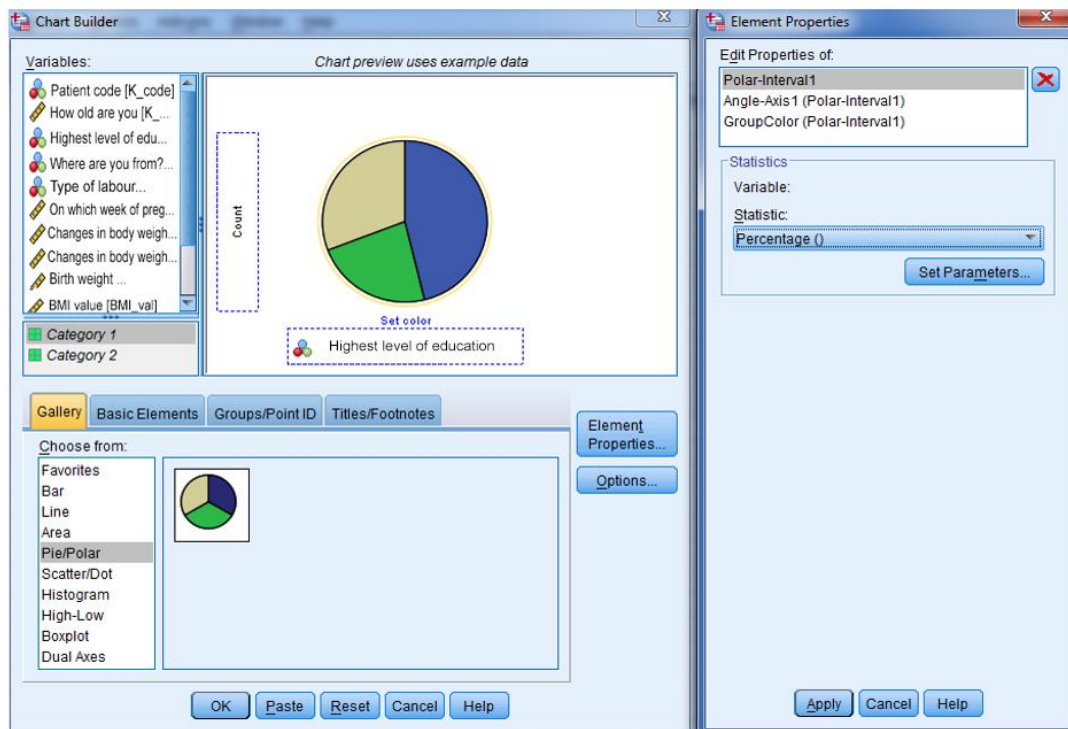
#### **Charts with multiple data series**

More changes have to be displayed at a time. The measure of the round and the measure and ratios of the parts (slices) change, too.

One of the most frequent charts displaying shares is the chart with one data series. The reason for this is that its aesthetics, and the fact that it is easy to analyse and understand.

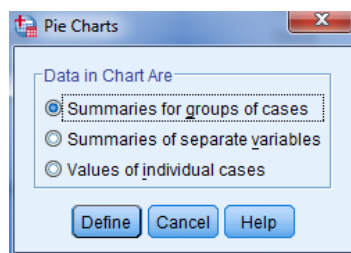
Our example includes a pie chart with one data series. We consider the highest level of qualification of young mothers compared to all others as representing the round. No other variables will be considered here.

In *CHART BUILDER*, go to *GALLERY*, then choose *PIE/POLAR*. Select and move the variable of qualification (“Az Ön legmagasabb iskolai végzettsége”) to the proper axis. Here, we only consider one variable since the point of reference is the whole population. Add the title of the chart in *TITLES / FOOTNOTES*, and then choose *PERCENTAGE (%)* under *ELEMENT PROPERTIES / STATISTIC*. Finalize settings by clicking on *APPLY*, and generate chart by pressing *OK*. (Jánosa 2011, Sajtos – Mitev 2007).



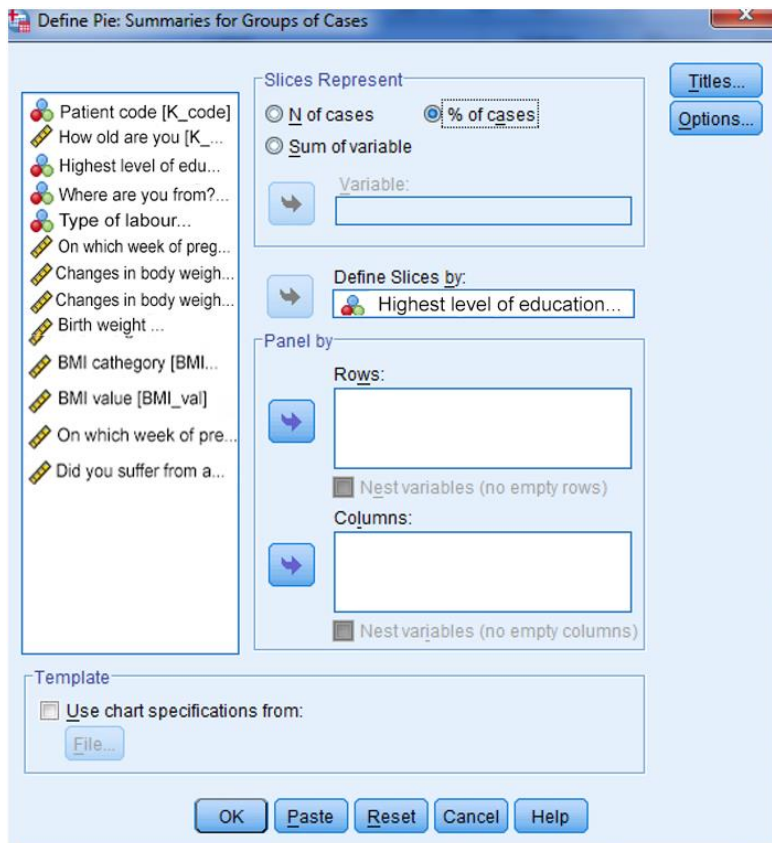
**Figure 7/35. Edit panel for pie charts (*CHART BUILDER*)**

The other option for editing is to follow the access path *GRAPHS* → *LEGACY DIALOGS* → *PIE* and because there is only one variable, choose *SUMMARIES FOR GROUPS OF CASES*, then press *DEFINE* to get the edit panel. Now, *SUMMARIES FOR GROUPS OF CASES* means categorization between values of one variable where the software puts data together and displays them. If we choose *SUMMARISE OF SEPERATE VARIABLES*, the summary and the shares will be displayed. The *VALUES OF INDIVIDUAL CASES* displays all values separately.



**Figure 7/36. Selecting the type of pie charts**

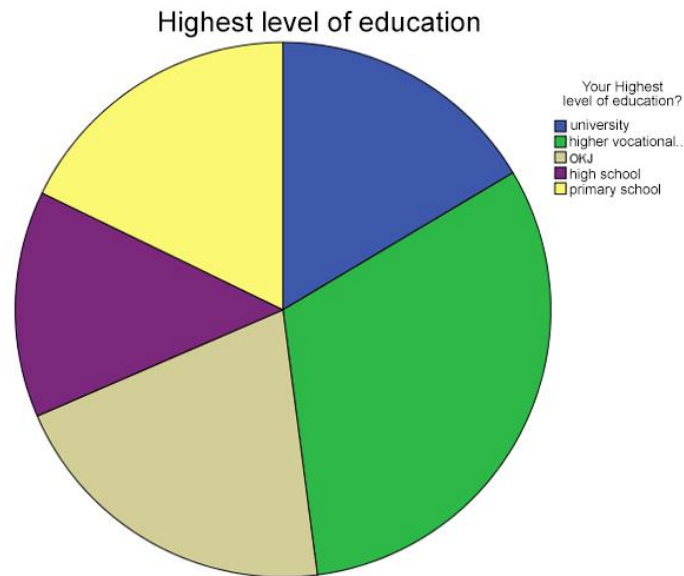
As we would like to see shares displayed in percentages, we have to select the option *% OF CASES*, and then put the variable to be examined into the cell *DEFINE SLICES BY* and press *OK*.



**Figure 7/38. Edit panel for pie chart (with menu option *LEGACY DIALOGS*)**

Both options result in the following pie chart.





**Figure 7/39. Breakdown by qualification**

### “Histogram”

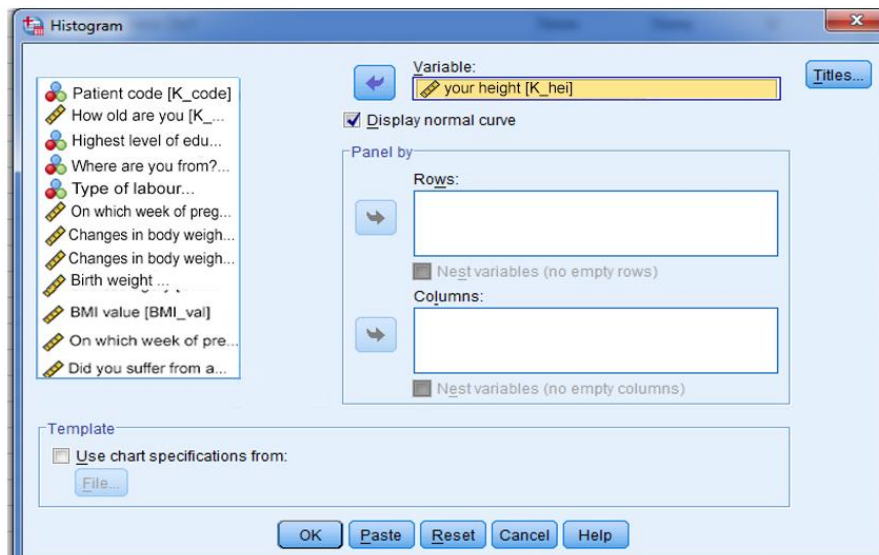
Histogram is a useful analytical tool, providing information on the distribution of the population, and supporting graphical illustration of distribution of variables. The bins of the histogram represent an interval of values in the same way that heights of columns represent the frequency of bins.

The first thing to be checked when looking at the histogram is whether the distribution is symmetric or not. These consequences can be drawn by simply looking at the figure besides the most common and least common measures.

When displaying the histogram in a coordinate system, the X axis represents value classes and the Y axis stands for their frequencies. Binding the columns will give information on the distribution of the population.

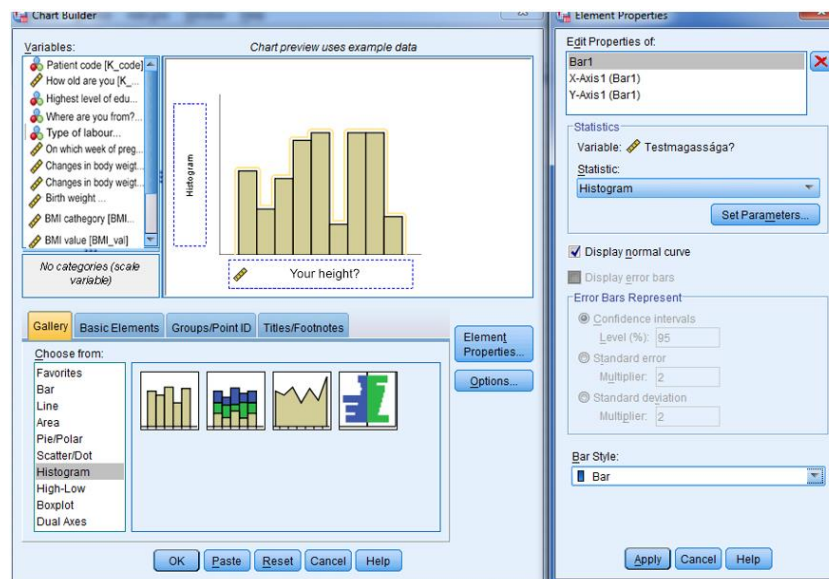
Our example is about the distribution of pregnant women’s body height.

Again, there are again two options to display this type of diagram. One of them has the access path *GRAPH / LEGACY DIALOGS / HISTOGRAM*. First, the variable to be displayed (measured on a ratio scale) has to be selected, and put into the cell *VARIABLE*. By ticking *DISPLAY NORMAL CURVE* the distribution curve will appear. In *TITLES*, the title and footnotes can be added if necessary. Finally, pressing *OK* will bring us to the end of the process (Jánosa 2011, Sajtos – Mitev 2007).



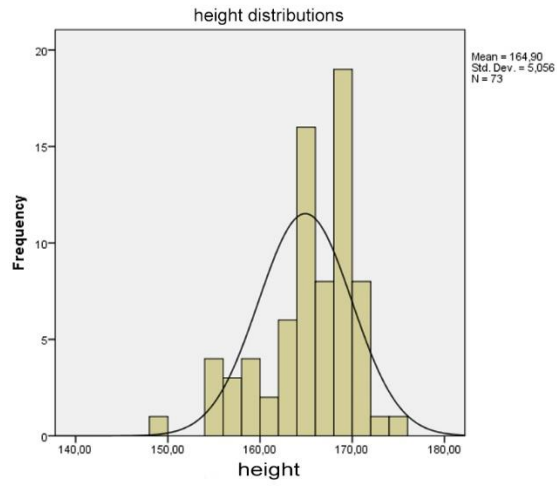
**Figure 7/40. Edit panel for histogram (with menu option *LEGACY DIALOGS*)**

The other option is *CHART BUILDING*, available under *CHART BUILDER / GALLERY / HISTOGRAM* where one can select the most appropriate type of histograms. We move the variable to be examined (body height), and tick *DISPLAY NORMAL CURVE* to display the distribution curve. Editing can be finalized by clicking on *APPLY*, then *OK*.



**Figure 7/41. Edit panel for histogram (*CHART BUILDER*)**

We will get the following figure, independent of the method chosen.



**Figure 7/42. Distribution of pregnant women's body height**

## 8. ASSOCIATION AND CORRELATION ANALYSIS (Pongrác Ács)

### 8.1. A theoretical background for statistical relationships

In order to gain more and more information about the world, it is important to become familiar with connections and relationships. Information gathered this way will play a significant role in making decisions and measuring the effects of decisions. Phenomena can be recognized in more detail if one does not only look at them as separate identities but examines their relationships with other phenomena as well.

The major characteristics of correlation analysis are as follows:

- When we conduct primary research with multiple variables, the research objective is to discover relationships and connections. These analyses are supported by hypotheses.
- Relationships can be calculated for at least two variables, belonging to the same type of data (scale, ordinal, nominal).

Phenomena and processes can be classified as follows:

- variables are *independent* from one another if there can be no consequences made from one to the other,
- for *stochastic relationships*, there is a relationship that is probabilistic like a trend. This case is the most interesting one since it means that there is a high probability for some sort of relationship, a “common organization”,
- for *deterministic* relationships, one variable determines the value of the other one.

Relationships can be classified according to the types of their variables. The three types include:

- All (both) variables are qualitative (categorical variables, since the values belong to categories), i.e. nominal ones, the relationship is called *association*. Data analysis in practice is called *crosstab* analysis and also *chi-squared test* (named after the most commonly used statistical measure).
- In case of *mixed association*<sup>9</sup>, the cause is a qualitative, while the effect is a quantitative variable.
- If all (both) variables are quantitative then it is called *correlation*.

---

<sup>9</sup> This expression does not exist in the English literature; it was translated from the Hungarian term “vegyes kapcsolat”. – translator’s note

All three relationships have to be expressed in numbers with the help of a measure. There are measures expressing the intensity of the relationship, the general scheme of which can be written as (measure of intensity in general is denoted by T):

$$0 \leq T \leq 1$$

In general, the above interval has to be set for the absolute value of T but in particular cases – especially in case of correlation – the sign represents relevant information as well, since it shows the positive or negative direction of the relationship. Of course, in these cases the interval is:  $[-1;1]$ .

Interpretation of measures always depends on the problem; one has to be aware of the nature of the relationship. The following scheme provides assistance for general interpretation:

**Table 8/1. General interpretation of association/correlation coefficient**

T = 0	no association/correlation
$0 < T < 0.3$	weak association/correlation
$0.3 \leq T \leq 0.7$	moderate association/correlation
$0.7 < T < 1$	strong association/correlation
T = 1	deterministic (perfect association/correlation)

Source: author

## 8.2. Association and crosstab analysis

For association, all the variables have to be quantitative. Data will have to be ordered in a combinational table – if it contains frequencies, it is called contingency table.

The most commonly applied measures can be classified on the basis of their symmetry or asymmetry. Symmetry means that we do not know which variable is the cause and the effect, i.e. the relationship has no direction.

**Table 8/2. Measures of nominal scales**

	Symmetric	Asymmetric
Only in case of 2*2 table	$\Phi$ (phi) coefficient	
In case of every type of tables (2*2 included)	Contingency coefficient, Cramer coefficient	Lambda, Goodman and Kruskal tau (uncertainty coefficient)

Source: author

If the variables are dichotomous or there are only two possible answers (alternative) that exclude each other e.g. male-female, yes-no, etc., then the general form of the two-dimensional contingency table is the following:

**Table 8/3. General form of a two-dimensional contingency table**

Versions of variable A	Versions of variable B		<b>Sum:</b>
	B <sub>1</sub>	B <sub>2</sub>	
A <sub>1</sub>	f <sub>11</sub>	f <sub>12</sub>	S <sub>1</sub>
A <sub>2</sub>	f <sub>21</sub>	f <sub>22</sub>	S <sub>2</sub>
<b>Sum:</b>	O <sub>1</sub>	O <sub>2</sub>	n

Source: author

n – the number of elements,

f<sub>11</sub> – frequency of first group of variables A and B (the frequencies of the other cells can be interpreted similarly!),

S<sub>1</sub> – the first row is the sum of frequencies (belonging to the first group of variable A),

O<sub>1</sub> – the first column is the sum of frequencies (belonging to the first group of variable B).

The following equation can be written:

$$S_1 + S_2 = O_1 + O_2 = n$$

We call the sum of rows and columns **marginal totals (frequencies)**.

In case of alternatives, one can apply **Yule's coefficient**, which can be calculated as the “cross-multiplication” of frequencies to be found in the table:

$$Y = \frac{f_{11} \times f_{22} - f_{12} \times f_{21}}{f_{11} \times f_{22} + f_{12} \times f_{21}}$$

The measure is always between -1 and +1 since it is the quotient of the sum and the difference of the same data.

The next example is based on the database, again.

Our data has been obtained from the following two questions:

**DO YOU CONSIDER YOUR KNOWLEDGE ON HEALTHY NUTRITION TO BE UP-TO-DATE?**

Yes

No

**DID YOU SUFFER FROM A CHRONICAL ILLNESS**

Yes

No

The results of the answers are summarized in the contingency table:

**Table 8/4.**

		Did you suffer from a chronic illness?		<b>Total</b>
		Yes	No	
Do you consider your knowledge on healthy nutrition to be up-to-date?	Yes	11	33	<b>44</b>
	No	5	21	<b>26</b>
	<b>Total</b>	<b>16</b>	<b>54</b>	<b>70</b>

Source: author

Calculating the coefficient, we get the following result:

$$a = \frac{11 \times 21 - 33 \times 5}{11 \times 21 + 33 \times 5} = 0,16$$

The absolute value is between 0 and 0.3, so we found a weak association between the two variables. When applying this coefficient, one has to be cautious that all elements in the diagonal need to be different from zero. If frequency is zero in a cell, then the coefficient indicates a deterministic relationship even if it this relationship has not been established.

When we have two or more variables, then another measure has to be applied. The **Cramer coefficient** resolves the dilemma of alternatives but is insensitive to extreme cases (zero in a cell). It is one of the most popular chi-square-based measures among researchers as it is applicable for almost all crosstabs. The basic idea is to examine how frequencies would change if there were no connection between the variables, i.e. they would be independent so a value of a variable would not attract the value of another variable. We start again from the contingency table.

The basic idea behind the calculation: if we detect differences between frequencies assuming independency and actual frequencies, then we may assume that there is a stochastic relationship, so the calculation of the expected frequencies means that we separate the element of the population according to the marginal totals. When filling in the contingency table with the expected frequencies, the distribution of all rows will be equal which means the same as the independence of two variables. Using a 2x2 contingency table, the frequencies

under the assumption of independence can be calculated with marginal totals, and marked with a sign \* :

$$\frac{S_1 \times O_1}{n} = f_{11}^* \quad \frac{S_1 \times O_2}{n} = f_{12}^*$$

$$\frac{S_2 \times O_1}{n} = f_{21}^* \quad \frac{S_2 \times O_2}{n} = f_{22}^*$$

First, the following relative differences will need to be calculated in all the cells:

$$\frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}$$

where  $f_{ij}$  means the frequency of row  $i$  and column  $j$ .

It suggests stochastic relationship if the actual and the frequencies under the assumption of independence are not equal. The differences between the two types of frequencies have to be expressed in a coefficient, which is the squared contingency measure, the so-called  $\chi^2$  (chi square) value.

$$\chi^2 = \sum \sum \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}$$

$\chi^2$  itself does not meet the requirements of measures of stochastic relationships. Its lower limit is zero but its upper limit may exceed 1 with a very high extent. This dilemma is solved by the Cramer coefficient (Cramer's V), which can be calculated as:

$$V = \sqrt{\frac{\chi^2}{n \times (s-1)}}$$

where  $s$  means the minimum of the versions of variables (the number of less versions).

The next example seeks to find out if there is an association between the type of labour and BMI categories.

**Table 8/5. Contingency table of type of labour and BMI categories**

		BMI categories				Total
		<i>underweight</i>	<i>normal weight</i>	<i>overweight</i>	<i>obesity</i>	
Type of labour	vaginal	2	13	7	8	30
	caesarean	3	10	14	13	40
	<b>Total</b>	<b>5</b>	<b>23</b>	<b>21</b>	<b>21</b>	<b>70</b>

Source: author



The frequencies under the assumption of independence with marginal totals:

$$\frac{30 \times 5}{70} = 2,14 \quad \frac{30 \times 23}{70} = 9,86 \text{ stb.}$$

Frequencies under the assumption of independence are listed in the following table:

**Table 8/6. Frequencies under the assumption of independence**

		BMI categories				Total
		<i>underweight</i>	<i>normal weight</i>	<i>overweight</i>	<i>obesity</i>	
Type of labour	vaginal	2,143	9,857	9,000	9,000	30
	caesarean	2,857	13,143	12,000	12,000	40
Total		5	23	21	21	70

Source: author

First, the relative frequencies need to be calculated in every cell:

$$\frac{(2 - 2.143)^2}{2.143} = 0.010 \text{ etc.}$$

Here is the newly generated contingency table:

**Table 8/7. Calculation of  $\chi^2$  values**

		BMI categories				Total
		<i>underweight</i>	<i>normal weight</i>	<i>overweight</i>	<i>obesity</i>	
Type of labour	vaginal	0,010	1,002	0,444	0,111	30
	caesarean	0,007	0,752	0,333	0,083	40
Total		5	23	21	21	70

Source: author

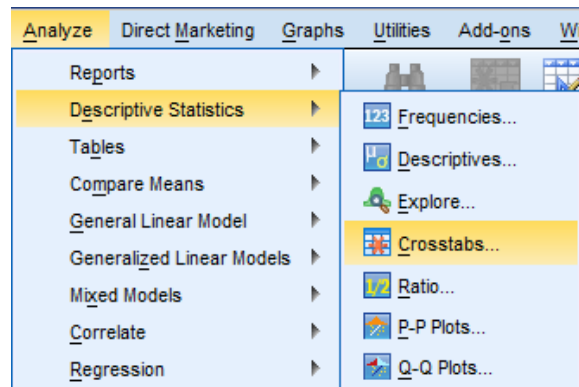
Cramer's V in our example can be calculated as follows:

$$V = \sqrt{\frac{2.743}{70 \times (2-1)}} = 0.188$$

for type of labour and BMI categories.

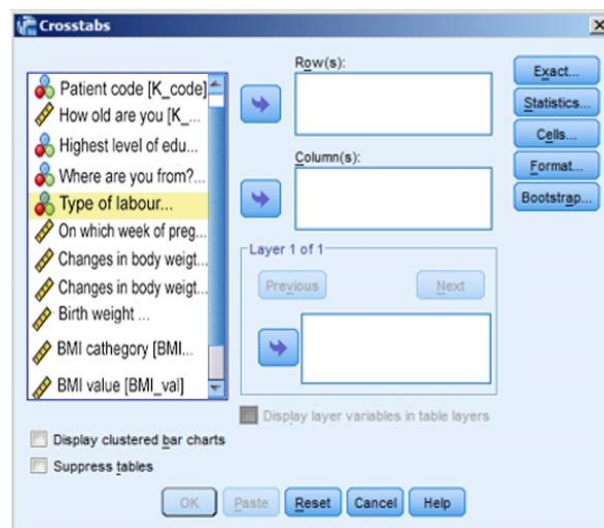
Based on the coefficient, the stochastic relationship between the BMI categories and the type of labour is weak. The measure  $V^2$  can also be interpreted. It shows that BMI categories determine the type of labour to an extent of 3.53%.

Association can be analysed in module *CROSSTABS* of SPSS, after choosing *ANALYSE / DESCRIPTIVE STATISTICS*.



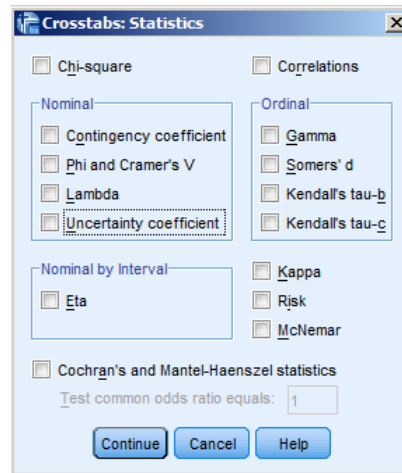
**Figure 8/1. Access path to CROSSTABS**

Now, the variables need to be selected by moving them to the windows labelled *ROW(S)* and *COLUMN(S)*. There are no obligatory rules to decide which variable should be the row of the column, so it is up to the researcher. To add a suggestion, in social sciences, *COLUMN(S)* are used to represent the dependent variable (whose distribution we need to find out), while *ROW(S)* will be used as independent variables (that we consider to have a significant effect on the dependent variable).



**Figure 8/2. Setting of variables of association (crosstab)**

Now, we present the module *STATISTICS* since it contains important settings; the coefficients of crosstab analyses can be accessed here.



**Figure 8/3. Measures of association**

As shown in the figure, statistical measures are listed in groups based on the scale types of variables. Now, we turn to measures to be applied when we deal with *nominal data*. Contingency (*CONTINGENCY COEFFICIENT*), phi (*PHI*) and Cramer (*CRAMER's V*) coefficients can be applied for symmetric relationships, while lambda (*LAMBDA*), Goodman and Kruskal tau, and the uncertainty coefficient (*UNCERTAINTY COEFFICIENT*) will need to be chosen in case of asymmetry. When working with association, especially with symmetry, measures belong to the chi-square value or one of its mutations. One can calculate the chi-square value by choosing *CHI-SQUARE* in the left-hand corner on the top.

- **Contingency coefficient (*CONTINGENCY COEFFICIENT*):** can be applied for crosstabs of any numbers of variables, even in the special case of 2x2. However, it is not commonly used due to difficulties in interpretation. Instead Cramer's coefficient

(*CRAMER V*) is suggested. The calculation contains the sample size. 
$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

- **Phi (*PHI*) coefficient ( $\Phi$ ):** used in the special case of crosstabs, in case of 2x2 tables since it is easy to be interpreted because the upper limit is 1. In the calculation, the chi-square value has to be modified by the sample size. It is not recommended to apply it in case of several variables since it has no upper limit in this case, so it is not easy to

interpret. 
$$\phi = \sqrt{\frac{\chi^2}{N}}$$
. If this particular measure is used, then the Yates continuity

correction (*CONTINUITY CORRECTION*) cannot be applied, which is the modification of the chi-square coefficient with 2x2 tables.

- **Cramer coefficient** (*CRAMER's V*): the most commonly used measure; it is easy interpreted. It has to be used for variables with two or more versions.  $V = \sqrt{\frac{\chi^2}{n \times (s-1)}}$ .

There are statisticians with the opinion that measures based on chi-square – and therefore the Cramer coefficient as well – are not applicable if more than 20% of the values in cells are under 5.

- **Lambda** (*LAMBDA*): asymmetric measure. When interpreting the coefficient, we get the percentage value that shows the extent to which the independent variable can forecast the dependent variable. The calculated value shows the percentage decrease of the forecast error if the expected cause is added as independent variable.

$$\lambda = \frac{SUM(f_i - f_d)}{N - f_d}$$

- The Goodman and Kruskal tau, and the uncertainty coefficients can be interpreted similarly to lambda. The maximal value is 1, which means that if values of the independent variables are known, then the value of the dependent variable can be estimated without error (100% certainty).

On the right-hand side of the module, the measures used for **ordinal scale** variables are listed.

**Table 8/8. Measures of association in case of ordinal scales**

Measure (for ordinal scales)	Types of tables
<i>Gamma</i>	For every type of tables and size (easy to interpret)
<i>Sommers' d</i>	For every type of tables and size (not easy to interpret)
<i>Kendall tau-b</i>	In case of symmetric tables
<i>Kendall tau-c</i>	In case of asymmetric tables

Source: author

Connections in variable orders are searched since here the order of categories is relevant, so the direction is also important besides the strength of association. The sign is positive if the increasing value of a variable causes an increase in the other variable. If it causes decrease, then the association is negative. In general, for ordinal scale variables, the goal

is to compare pairs. If all the variables of a pair member are higher than its pair's, then the pair is concordant (concordant). If the values are the same of both, then it is a tied pair. If one value is higher and the other is lower in the comparison, then it is called a discordant pair. The calculation is based on differences between the concordant and discordant pairs. Positive association means that most pairs are concordant, while in case of negative association, pairs are rather discordant.

- **Gamma (GAMMA) coefficient<sup>10</sup>**: to be applied in case of any kind of ordinal data and tables. Its values range between -1 and +1 where 0 means the independence of variables. It refers to the extent how possible it is to find concordance or discordance dominant in the phenomena on which the research is being carried out.  $\gamma = \frac{S - D}{S + D}$ . If concordance is dominant ( $\gamma > 0$ ), then the higher category of a variable causes higher category of the other variable (e.g. if examining the level of parents' education, the positive value can refer to the fact that the father's higher qualification level will result in the mother's higher level of qualification, i.e. he chooses someone with a higher level of education)
- **Somers's d coefficient**: it measures the association of ordinal variables between -1 and +1. To be applied for any kinds of tables, just like in the case of the coefficient gamma. The absolute value close to 1 means a strong relationship but its interpretation is more complex than gammas.
- **Kendall tau-b (KENDALL'S TAU-B)**: To be applied in case of symmetric tables, and variables. Its value can range between -1 and +1 where +1 means that the order of pairs is similar, equal (concordant), while -1 means that order of pairs are of the opposite directions (discordant).
- **Kendall's tau-c (KENDALL'S TAU-C)**: To be applied in case of asymmetric tables; its interpretation is the same as Kendall's tau-b.

The *STATISTICS / CORRELATION* option will be introduced later, when discussing the connections between numerical variables and also when calculating ETA.

- **Kappa (KAPPA)<sup>11</sup>** is a measure of agreement that measures the agreement of values (raters). Base on Landis-Koch (1977), it can be interpreted as:

---

<sup>10</sup> Is is often referred to as Goodman and Kruskal's Gamma.

<sup>11</sup> It is often referred to as Cohen's Kappa.

**Table 8/9. Interpretation of Kappa**

$\kappa$	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

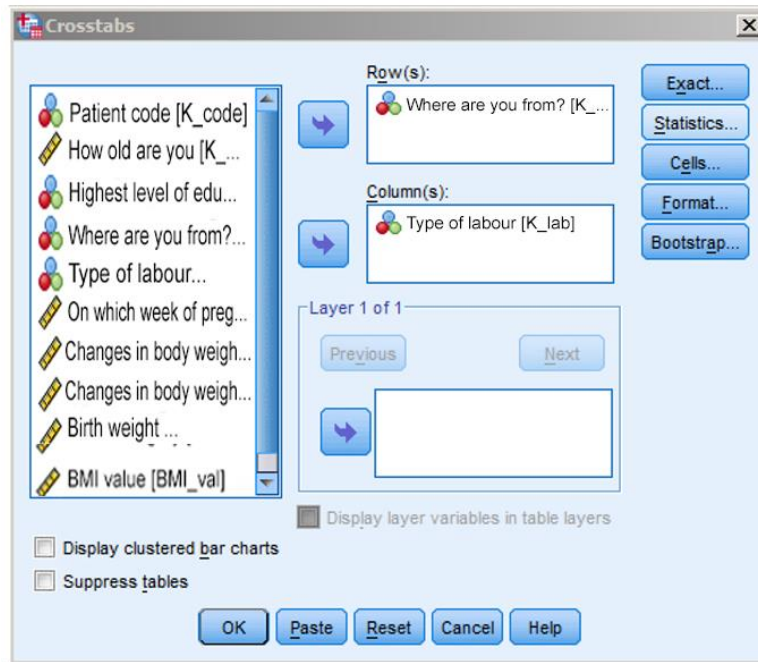
Source: author, based on Landis and Koch

If one would like to interpret the excessive number of categories together, then values under 0.4 can be considered as weak agreement; a 0.4-0.8 interval can stand for acceptable agreement, and above 0.8 it is excellent agreement. This can be applied for symmetric tables and if the opinions of raters are measured by the same scale.

- **Risk quotient (*RISK*)** to be applied in case of 2x2 tables. Besides the measure (with 0 as the lower limit and with no upper limit), a confidence interval is also part of the result. If the result is 1, and this value is included in the confidence interval, then there is no connection. If it is higher than 0 or 1, then we assume association. The method calculates relative risk and chance ratio for 2x2 tables with dichotomous variables. One of the variables can be interpreted as cause, while the other one as event.

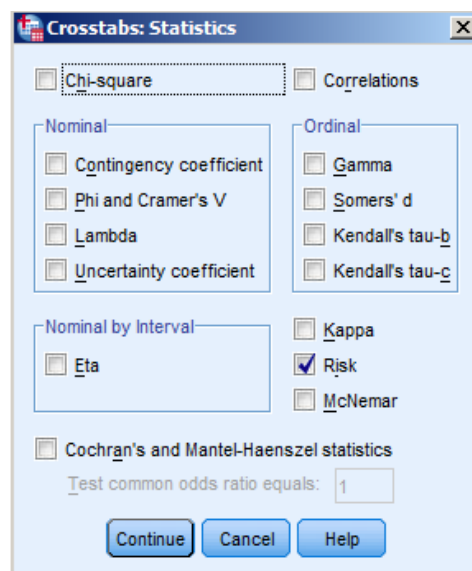
**As a practice exercise** let us find out if there is a causal connection between residence and type of labour. With this method, we can compare the probability of one case in the two groups.

First, the variables will need to be selected for analysis. Their type has to be dichotomous, i.e. variables with two outcomes.



**Figure 8/4. Selecting variables**

Add place of origin (“Honnan származik ÖN?”, town/village) as a row, which will be interpreted as cause, and type of labour (vaginal/caesarean) as column, interpreted as event. In module *STATISTICS*, it is only the option *RISK* to be selected.



**Figure 8/5. Selecting the risk quotient to be calculated**

Press *CONTINUE* and *OK* to get your results. In this particular case, we only present interpretation of the third table.

**Table 8/10. Results of relative risk quotients and chance ratios**

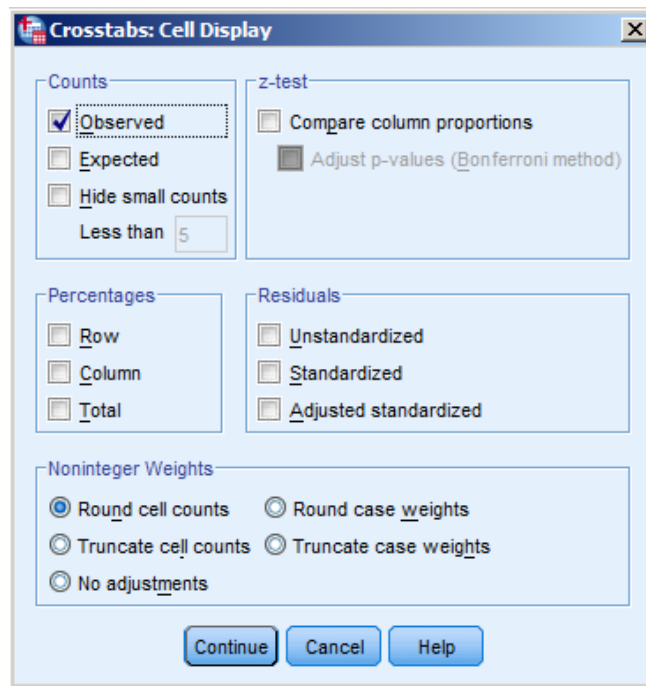
	Risk Estimate		
	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Where are you from? (City / Village)	2,029	,775	5,314
For cohort Type of labour = vaginal	1,500	,856	2,627
For cohort Type of labour = caesarean	,739	,487	1,121
N of Valid Cases	70		

Source: author

1. Odds Ratio:  $2.029 = \frac{1.500}{0.739}$  means that there is more than twice as many chances (2.029) that the type of labour of a young mother living in a town will be vaginal.
  2. Relative Risk  $_{\text{vaginal}} = 1.500$  means that risk of vaginal labour is 1.5 times higher in a town than in a village.
  3. Relative Risk  $_{\text{caesarean}} = 0.739$  compares the risk of caesarean labour in town vs. village.
  4. The *OUTPUT* contains *Odds Ratio* and *Relative Risk* and a confidence interval from which the existence of the relationship can be assumed. If value 1 belongs to the interval or the measure itself equals to 1, then no relationship can be detected. This is the situation in our example as well.
- **McNemar test (McNEMAR)** is a measure analysing the connection between dichotomous variables, measuring the change when the same measurement has been carried out. It represents the percentage of respondents who chose the same option in both surveys. In practice, it is used for comparing opinions on two different occasions (e.g. customers' opinion, elections, etc.).
  - **Cochran and Mantel-Haenszel statistics (COCHRAN AND MANTEL-HAENSZEL STATISTICS)** examines the connection of two dichotomous variables, assuming the joint effect of control variables. Its advantages include that it takes the effects of all control variables into consideration simultaneously.

In the followings, the interpretation of options in module *CELL* will be discussed.





**Table 8/6. Settings available in module *CELL***

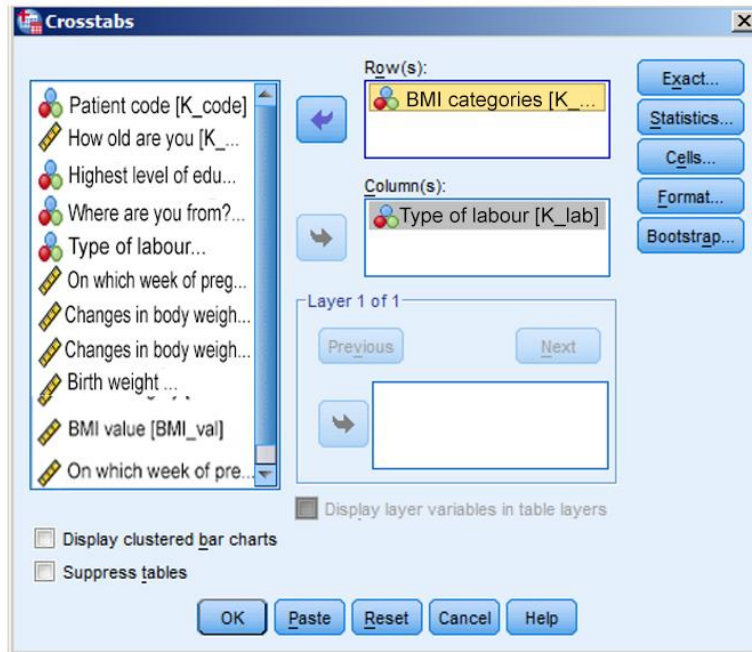
The left top corner contains the settings of data, where *OBSERVED* means unique (actual) data observed, while *EXPECTED* stands for frequencies expected to occur in the case of independence.

The box under includes ratios in percentage (row percentage=*ROW*; column percentage=*COLUMN*; total percentage=*TOTAL*).

Row stands for the percentage the frequency in the cell represents from the row. Column stands for the percentage the frequency in the cell represents from the column. Total frequency is the proportion of cell frequency total row, total column and the sample size.

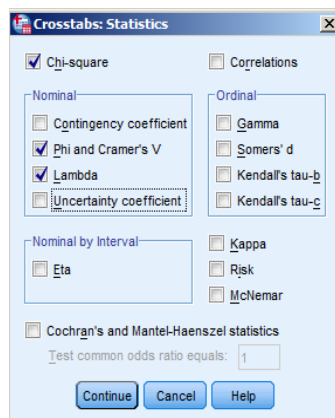
Measures found in the box of residual values (*RESIDUALS*) will be calculated from the differences of observed and expected frequencies. If the value is negative, then the observed frequency is smaller than it would be reasonable in the case of independence. From the three measures, probably *ADJUSTED RESIDUAL* is the most useful. It displays the categories that cause relationships. If the absolute value is greater than 2, then there is a significant connection between the two categories. If it is less than 2, then there is no significant relationship between them.

**As a practice exercise** let us figure out – on the basis of the example given above – if there is a relationship between the type of labour and BMI categories.



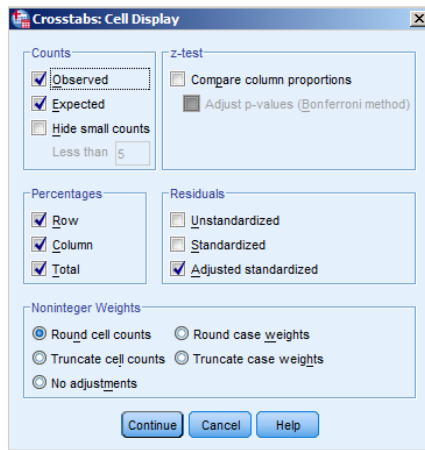
**Table 8/7. Selecting variables of the crosstab**

“Column” includes the type of labour (dependent variable, the distribution of which we are interested in), while “row” includes BMI categories (independent variable, it is the cause that we expect to affect the dependent variable).



**Figure 8/8. Calibrating measures**

As both variables are nominal, we can choose from the measures in the left top corner. Select Chi-square, Cramer’s V and Lambda and press *CONTINUE* for further settings.



**Figure 8/9. Settings of module *CELLS***

If we select the options *OBSERVED* and *EXPECTED* in the box labelled *COUNTS*, it will display us a table of observed and expected frequencies. Percentages will also be displayed if one ticks them under *PERCENTAGES*. To get residuals we should only chose *ADJUSTED STANDARDIZED*. Outputs will be displayed after pressing *CONTINUE* and *OK*.

The output view provides a *case processing summary*, containing information on the sample size (N), the number of *valid* and *missing* cases, and their *percent*.

**Table 8/11. Outputs of relative risk coefficients and chance ratio**  
**Case Processing Summary**

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
BMI categories *	70	100,0%	0	0,0%	70	100,0%
Type of labour						

Source: author

The sample size is 70, and there are no missing values, so 100% can be evaluated.

The next table contains the observed crosstab, which makes it easy to compare observed and expected (fictitious) data. The values are not equal but no big differences were detected which assumes the lack of a stochastic relationship. Calculated values in the table are the same as the ones above.

The table contains the following data:

- Observed frequencies (*COUNT*)
- Frequencies expected in the case of independence (*EXPECTED COUNT*)

- Row percentage (*% WITHIN BMI KATEGÓRIÁK*)
- Columns percentage (*% WITHIN SZÜLÉSÉNEK KIMENETELE*)
- Percentage of total sample (*% OF TOTAL*)
- Standardized adjusted residual (*ADJUSTED RESIDUAL*)

**Table 8/12. Summary of values in the crosstab**

BMI categories \* Type of labour      Crosstabulation

BMI categories		Type of labour		Total
		vaginal	caesarean	
	Count	2	3	5
	Expected Count	2,1	2,9	5,0
	% within BMI categories	40,0%	60,0%	100,0%
	% within Type of labour	6,7%	7,5%	7,1%
	% of Total	2,9%	4,3%	7,1%
	Adjusted Residual	-,1	,1	
	Count	23	10	23
	Expected Count	9,9	13,1	23,0
	% within BMI categories	43,3%	43,5%	100,0%
	% within Type of labour	43,3%	25,0%	32,9%
	% of Total	18,6%	14,3%	32,9%
	Adjusted Residual	1,6	-1,6	
	Count	7	14	21
	Expected Count	9,0	12,0	21,0
	% within BMI categories	33,3%	66,7%	100,0%
	% within Type of labour	23,3%	35,0%	30,0%
	% of Total	10,0%	20,0%	30,0%
	Adjusted Residual	-1,1	1,1	
	Count	8	13	21
	Expected Count	9,0	12,0	21,0
	% within BMI categories	38,1%	61,9%	100,0%
	% within Type of labour	26,7%	32,5%	30,0%
	% of Total	11,4%	18,6%	30,0%
	Adjusted Residual	-,5	,5	
Total	Count	30	40	70
	Expected Count	30,0	40,0	70,0
	% within BMI categories	42,9%	57,1%	100,0%
	% within Type of labour	100,0%	100,0%	100,0%
	% of Total	42,9%	57,1%	100,0%

Source: author

The next table contains the Pearson Chi-Square and a measure equal to the square contingency we calculated before.

**Table 8/13. Calculated chi-square values**

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,743 <sup>a</sup>	3	,433
Likelihood Ratio	2,741	3	,433
Linear-by-Linear Association	,967	1	,325
N of Valid Cases	70		

a. 2 cells (25,0%) have expected count less than 5. The minimum expected count is 2,14.

The degree of freedom is denoted by df, and it can be calculated as  $df=(row-1)*(column-1)$ . This value plays an important role in determining the theoretical value. The observed value is the squared contingency measure ( $\chi^2$ ), and it has to be compared to the theoretical one in order to decide if the null hypothesis should be accepted or rejected.<sup>12</sup> The table of  $\chi^2$  distribution provides a basis for comparison (7.815; in Excel: =inverz.khi (0.05;3)). As the observed value is smaller than the theoretical one, the null hypothesis is accepted so there is no relationship between the two variables. This means that the type of labour is not determined by the BMI categories. The same consequences can be made based on the tables about significance (*DIRECTIONAL MEASURES, SYMMETRIC MEASURES*) since the value is greater than the 5% we picked. The very last table lists the symmetric measures (*SYMMETRIC MEASURES*).

**Table 8/14. Calculated measures of correlation**

**Directional Measures**

			Value	Asymp. Std. Error <sup>a</sup>	b	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,091	,103	,860	,390
		BMI categories	,085	,100	,820	,412
		Dependent				
	Goodman and Kruskal tau	Type of labour	,100	,152	,627	,530
		Dependent				
		BMI categories	,018	,021		c
	Type of labour	,039	,047		c	
	Dependent					

Not assuming the null hypothesis.  
Using the asymptotic standard error assuming the null hypothesis.  
Based on chi-square approximation

**Symmetric Measures**

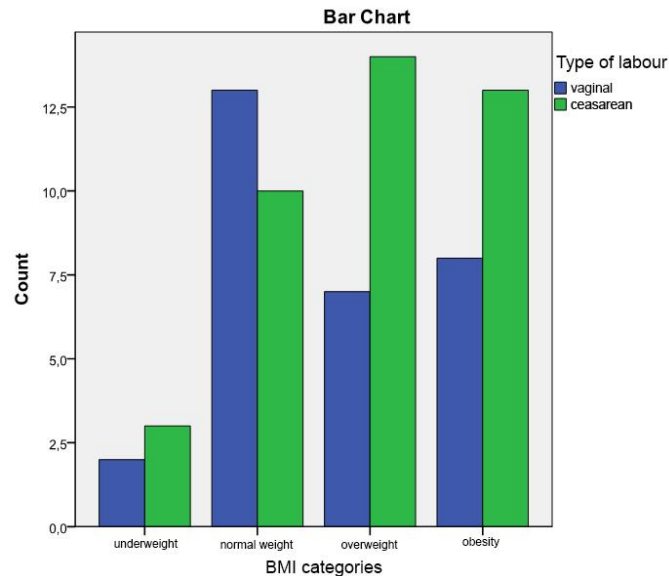
		Value	Approx. Sig.
Nominal by Nominal	Phi	,198	,433
	Cramer's V	,198	,433
N of Valid Cases		70	

Not assuming the null hypothesis.  
Using the asymptotic standard error assuming the null hypothesis.

Source: author

<sup>12</sup> For more details see following chapters.

The Cramer coefficient suggests that there is a weak relationship between the two variables. When generating crosstabs, the programme offers the opportunity to display charts (*DISPLAY CLUSTERED BAR CHARTS*) as well. As default, it illustrated the association on a bar chart.



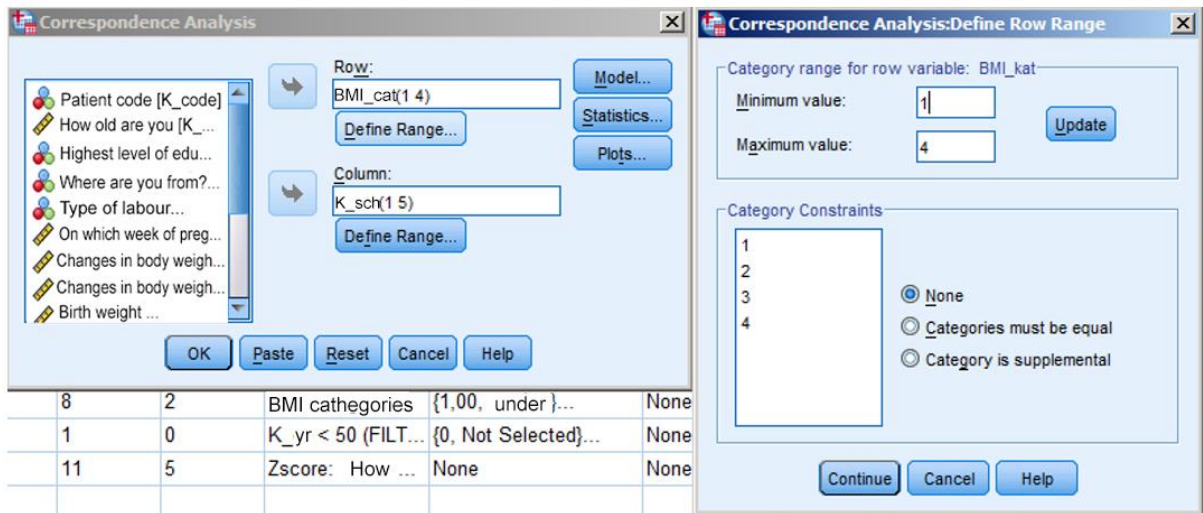
**Table 8/10. Graphic illustration of the crosstab**

Note that the example above was presented only as an illustration only, since obviously it is not reasonable to apply chi-squared measures if more than 20% of the values are under 5.

### 8.3. Correspondence analysis

In order to display association, a *correspondence analysis* can be applied, a method becoming increasingly popular. „Correspondence analysis makes it possible to display the relationship between two nominal variables in a multidimensional space consisting of a small number of dimensions (mostly two) in order to make it easy to interpret. Categories similar to one other will be located close to one another also in the graphical illustration. Interpretation of the results depends on the method of normalization. Default type of normalization in SPSS analyses the relationship between the row and column variables.” (Ketskeméty – Izsó 2005, p. 417). For illustration, let us decide if there is a relationship between the BMI categories and the level of qualification.

The access path to this method is the following: *DATA REDUCTION / CORRESPONDENCE ANALYSIS*.



**Figure 8/11. Settings of a correspondence analysis in SPSS**

First, the *ROW* and *COLUMN* variables will need to be selected. Then, all the variables will have to be defined according to the number of versions available.

In this case, we defined the type of BMI categories from 1 to 4, depending on the number of versions. After determining the two variables, no other settings should be modified. Results will be available after pressing *OK*.

The first table (*CORRESPONDENCE TABLE*) contains observed frequencies just like a crosstab, and the second one summarizes the results. The relationship is significant ( $p=0.03$ ), the chi-square value is high ( $\chi^2=22.78$ ) and the two dimensions are possible to display since the values account for 94.2% of dispersion. The next two tables contain coordinates of the variables in the two default dimensions. Probably *BIPLOT* fits our needs the best which display values that belong together. It actually displays values of *ADJUSTED RESIDUALS*.

**Table 8/15. Results of the correspondence analysis**

BMI categories	Highest level of education					Active Margin
	university	vocational education	OKJ	high school	primary school	
underweight	2	1	0	0	2	5
normal weight	4	11	3	2	3	23
overweight	4	9	5	0	3	21
obesity	2	2	5	7	5	21
Active Margin	12	23	13	9	13	70

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
1	,490	,240			,737	,737	,099	-,009
2	,258	,067			,205	,942	,111	
3	,137	,019			,058	1,000		
Total		,325	22,780	,030 <sup>a</sup>	1,000	1,000		

a. 12 degrees of freedom

**Overview Row Points<sup>a</sup>**

BMI categories	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		Total
					1	2	1	2	
underweight	,071	-,375	1,811	,065	,020	,907	,075	,925	1,000
normal weight	,329	-,389	-,206	,038	,102	,054	,648	,096	,743
overweight	,300	-,549	-,182	,056	,184	,038	,794	,046	,839
obesity	,300	1,064	-,024	,167	,694	,001	,999	,000	,999
Active Total	1,000			,325	1,000	1,000			

a. Symmetrical normalization

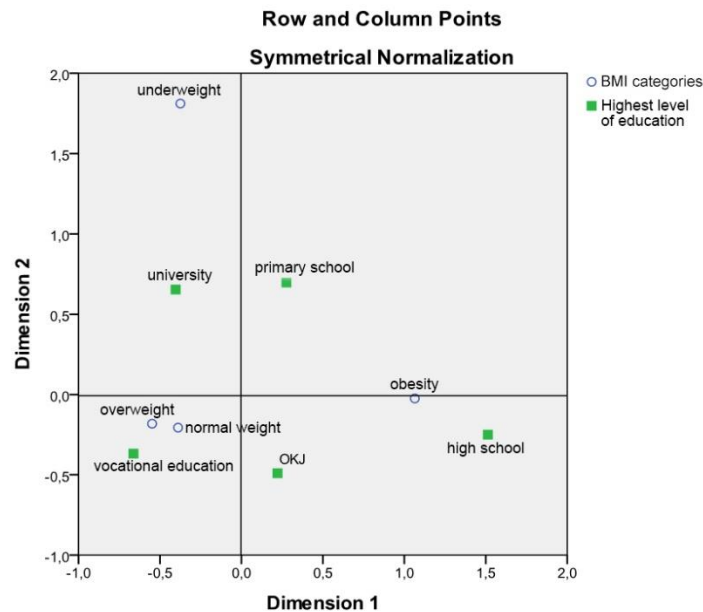
**Overview Column Points<sup>a</sup>**

Highest level of education	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		Total
					1	2	1	2	
university	,171	-,404	,653	,033	,057	,283	,420	,580	1,000
vocational education	,329	-,663	-,359	,085	,295	,164	,833	,129	,962
OKJ	,186	,221	-,490	,028	,019	,173	,162	,419	,581
high school	,129	1,513	-,250	,150	,601	,031	,961	,014	,975
primary school	,186	,276	,697	,030	,029	,349	,227	,764	,991
Active Total	1,000			,325	1,000	1,000			

a. Symmetrical normalization

Source: author

In graphical illustration, it is a key factor to name the dimensions in a way that supports understanding. This is the researcher's responsibility.



**Figure 8/12. Graphic illustration of the correspondence analysis**

More details on correspondence analysis are available in Jánosa (2011).



#### 8.4. Mixed association

In mixed association, the cause is always the qualitative variable, while effects are represented by the quantitative variable(s). The focus of the analysis of a mixed association is to assess the extent to which the information included in the quantitative variable(s) can be determined by grouping according to the qualitative variable. The quantitative variable makes it possible to expand methods of calculations. The measurement of the strength of mixed association is based on the partition of the standard deviation. The square total standard deviation is the sum of the square of internal and external standard deviation.

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

Internal and external squares of standard deviation can be calculated as follows:

$$\sigma_B^2 = \frac{\sum_{j=1}^m n_j \sigma_j^2}{n}$$

$$\sigma_K^2 = \frac{\sum_{j=1}^m n_j (\bar{x}_j - \bar{x})^2}{n}$$

Let us divide both sides of the equation by the square of the total standard deviation:

$$1 = \frac{\sigma_B^2}{\sigma^2} + \frac{\sigma_K^2}{\sigma^2}$$

The effect of the grouping (qualitative) variable – which is the cause in the stochastic relationship at the same time – is represented by the external standard deviation. If it is zero, then the qualitative variable has no measurable effect; the two variables (the qualitative and the quantitative) are independent. In the extreme case of the opposite – when the inner standard deviation is zero -, the external standard deviation equals to the total standard deviation, so the association is deterministic. Based on this, the following **standard deviation ratio** can be calculated from the external and total standard deviations:

$$H = \frac{\sigma_K}{\sigma} = \sqrt{\frac{\sigma_K^2}{\sigma^2}} = \sqrt{1 - \frac{\sigma_B^2}{\sigma^2}}$$

It is also true for the sum of square that the total is the sum of the internal and the external sums:

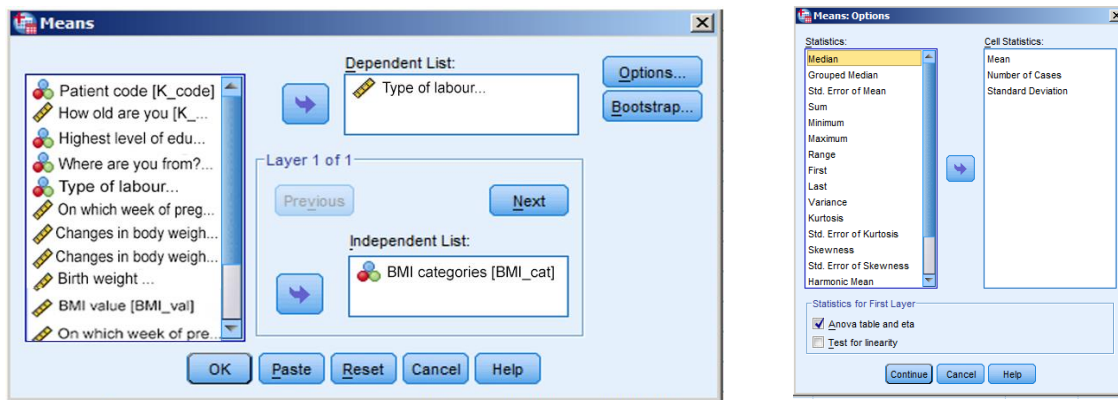
$$SS = SS_B + SS_K = \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 = \sum_{j=1}^m f_j (\bar{x}_j - \bar{x})^2 + \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

Therefore the following is also true:

$$H^2 = \frac{SS_K}{SS}$$

Here is a concrete example: we would like to find out if the BMI category before giving birth influences birth weight, i.e. if there is a relationship between the mother's birth weight and the BMI category.

Using SPSS makes our task easy since we only have to calibrate variables and interpret results. Go to *ANALYSE / COMPARE MEANS / MEANS*, add dependent and independent variables, then click options to select statistics to be calculated.



**Figure 8/13. Settings of mixed association in SPSS**

The most important thing is to select *ANOVA TABLE AND ETA*. Eta ( $\eta$ ) can be calculated the following way:

$$\eta = \sqrt{\frac{\sigma^2_K}{\sigma^2}}$$

Press *CONTINUE* and *OK* to get the results which will first list the summarizing tables, as usual. The next table (*REPORT*) contains means, numbers of elements and standard deviations.

**Table 8/16. Basic data of the categories**

birth weight

BMI categories	Mean	N	Std. Deviation
underweight	2840,0000	5	713,61754
normal weight	3276,9565	23	515,78215
overweight	3221,4286	21	529,69128
obesity	3612,3810	21	446,66436
Total	3329,7143	70	547,43929

Source: author

It is followed by the ANOVA table, which lists the internal (within groups) and external (between groups) sums of squares and information on significance.

**Table 8/17. Results of mixed association**

**ANOVA Table**

		Sum of Squares	df	Mean Square	F	Sig.
birth weight * BMI categories	Between Groups (Combined)	3187269,234	3	1062423,078	4,009	,011
	Within Groups	17491325,05	66	265020,077		
	Total	20678594,29	69			

Source: author

On the basis of the significance value ( $p < 0.05$ ), the variables show a significant relationship, so the results can be generalized (the phenomenon is not random). The measures of associations are calculated, too.

**Table 8/18. Table of the measures of association**

**Measures of Association**

	Eta	Eta Squared
birth weight * BMI categories	,393	,154

Source: author

The standard deviation ratio (H) shows that there is a medium strength association between the mother's BMI category and their birth weight. The type of BMI category determines the birth weight by 15.4 %. Probably, the other 85.6 % can be explained by other factors (e.g. quantity and quality of food consumed during pregnancy, etc.)

## 8.5. Correlation analysis

If both the causes and the effects are quantitative variables, then the relationship is called **correlation**. This book primarily focuses on measuring the strength of correlation between a *factor* or an *independent variable* (X) and a *dependent variable* (Y). It must be highlighted though that in reality most phenomena and processes are the results of complex effects of multiple factors. During the measurement of correlation, a simultaneous analysis of multiple causes can be carried out relatively easily.

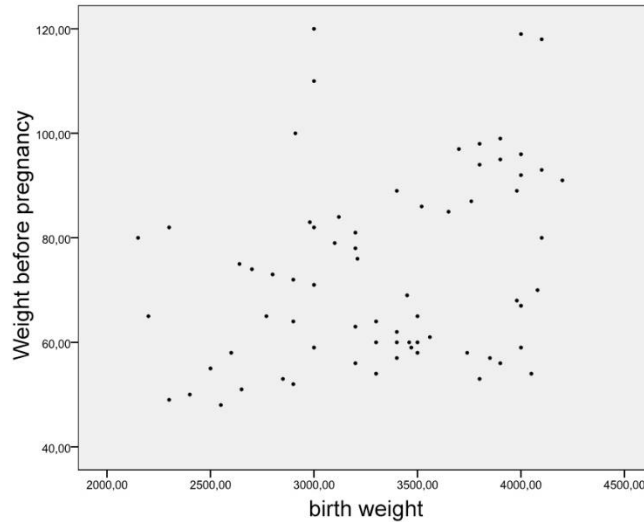
According to the nature of correlation, the following relationships between variables can be interpreted: monotonic correlation, and as part of this, linear relationship. Both correlations can be positive or negative, and the graphic display helps decide which of the two we have. The correlation between two quantitative variables can be plotted in a coordinate system in the form of a point chart. For further details on the topic consult Pintér – Rappai (2007): *Statisztika*.

The most popular measure in this field is the **linear correlation coefficient** (noted as: **r**). It can be applied under the assumption that there is a linear relationship between the variables, and linearity is conceivable problem examined. The correlation coefficient can be calculated with the measure of covariance (expressing how much the variables change together) and the standard deviation of variables in the following algorithm:

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}, \text{ where covariance is } C_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum d_x d_y}{n} = \frac{\sum xy}{n} - \bar{x}\bar{y}$$

where  $\sigma_x$  and  $\sigma_y$  are the standard deviations of the variables.

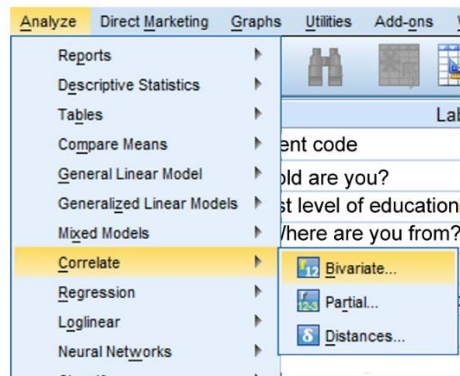
Now, we examine the strength of correlation in an example on the body weight of the mother before giving birth and the birth weight of the baby. First, we plot the result in a scatter chart.



**Figure 8/14. Scatter chart**

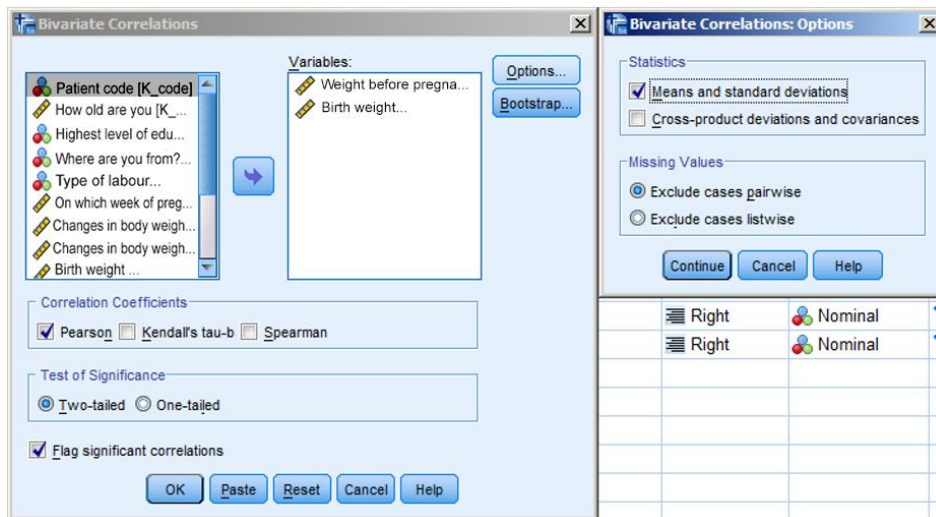
Based on the chart, a positive linear correlation can be expected. Both the weight of the mother before giving birth and the birth weight of the baby are increasing.

The successive settings in SPSS can be accessed as *ANALYSE / CORRELATE / BIVARIATE*.



**Figure 8/15. Access path to correlation**

Here, the variables will need to be selected by clicking on the arrow in the middle. The body weight before giving birth (“Várandósság előtti testsúly”) and the birth weight of the baby (“Gyermekének születési súlya”) will have to be moved into the box of variables.



**Figure 8/16. Adding variables to the correlation analysis**

The Pearson product-moment correlation coefficient is offered as default setting. We can request further statistics under *OPTIONS*. Here select the module labelled *MEANS AND STANDARD DEVIATIONS*. Press *CONTINUE* and *OK* to receive results.

**Table 8/19. Table of descriptive statistics**  
Descriptive Statistics

	Mean	Std. Deviation	N
Weight before pregnancy	73,5286	18,18800	70
Birth weight	3329,7143	547,43929	70

Source: author

The first table lists the mean, standard deviation and sample size. The following table contains the correlation coefficient.

**Table 8/20. Table of correlations**  
Correlations

		Weight before pregnancy	Birth weight
Weight before pregnancy	Pearson Correlation	1	,309**
	Sig. (2-tailed)		,009
	N	70	70
Birth weight	Pearson Correlation	,309**	1
	Sig. (2-tailed)	,009	
	N	70	70

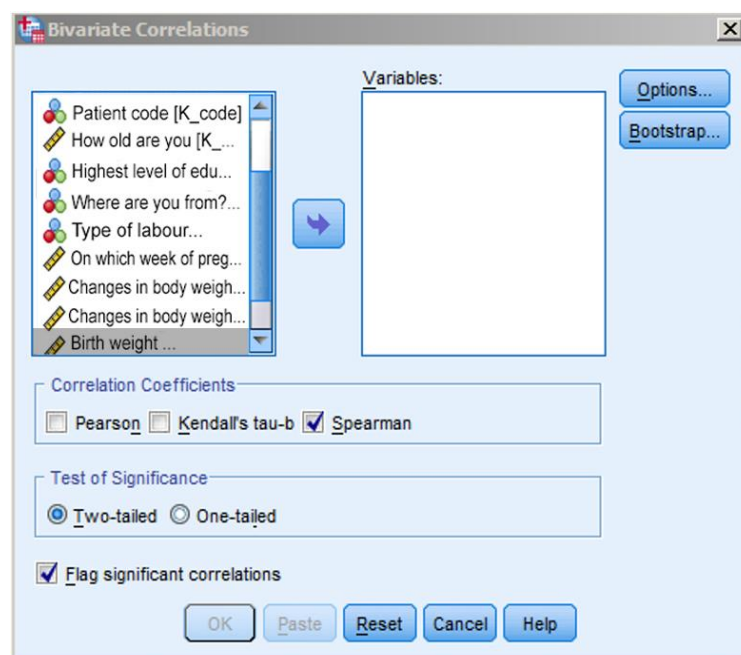
\*\* . Correlation is significant at the 0.01 level (2-tailed).

Source: author

The results show a moderate correlation ( $0.3 < R < 0.7$ ). Based on significance ( $p < 0.05$ ), the relationship is not random, so it can be generalized. The square of the correlation coefficient is the coefficient of determination, based on which it can be stated that the weight before giving birth determines birth weight by 9.55%. Of course, there are also possibilities for calculation if data are ordered (ranked), i.e. listed on an ordinal scale. In the case of monotonic relationship<sup>13</sup> the strength will be measured by Spearman's rank correlation coefficient, which is a more robust measure, i.e. not very reactive to outliers since it uses the ordinal scale instead of the interval or the ratio scale. This means that data can be transformed from a higher order scale to a lower one. The formula of the rank correlation coefficient is:

$$\rho = 1 - \frac{6 \sum_{i=1}^n [R(y_i) - R(x_i)]^2}{n(n^2 - 1)} = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)}.$$

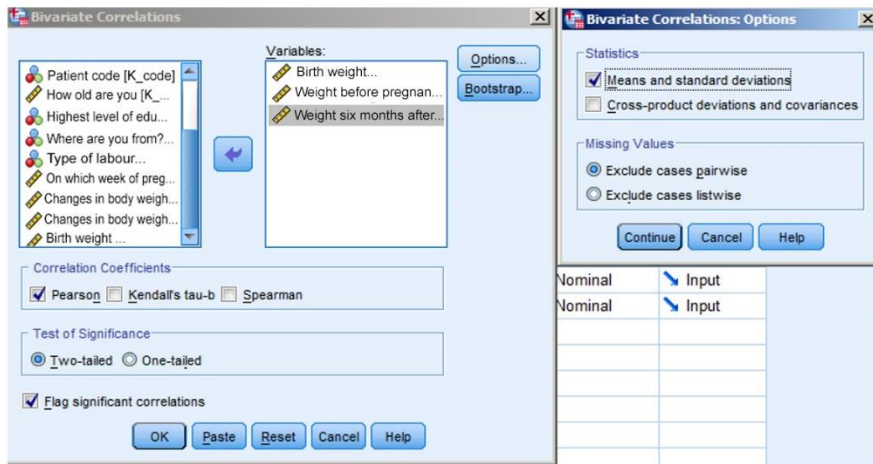
The access path to its calculation in SPSS is *ANALYSE / CORRELATE / BIVARIATE*. The coefficient *SPEARMAN* needs to be selected instead of the default, labelled *PEARSON*. Interpretation is the same as it was in the case of the linear Pearson coefficient.



**Figure 8/17. Settings of rank correlation**

In the next example, a correlation matrix will be generated, and the system of two or more variables (quantitative, continuous, scale) will be analysed. The analysis will be extended by adding the weight of the mother 6 months after giving birth.

<sup>13</sup> In case of monotonic relationship, the measure of the change in Y is most constant for a unit change in X.



**Figure 8/18. Expanding the number of variables**

After running the analysis, the results will be listed in the correlations matrix.

**Table 8/21. The correlations matrix**

		Birth weight	Weight before pregnancy	Weight six months after labour
Birth weight	Pearson Correlation	1	,309**	,237*
	Sig. (2-tailed)		,009	,049
	N	70	70	70
Weight before pregnancy	Pearson Correlation	,309**	1	,904**
	Sig. (2-tailed)	,009		,000
	N	70	70	70
Weight six months after labour	Pearson Correlation	,237*	,904**	1
	Sig. (2-tailed)	,049	,000	
	N	70	70	70

\*\* . Correlation is significant at the 0.01 level (2-tailed).  
 \* . Correlation is significant at the 0.05 level (2-tailed).

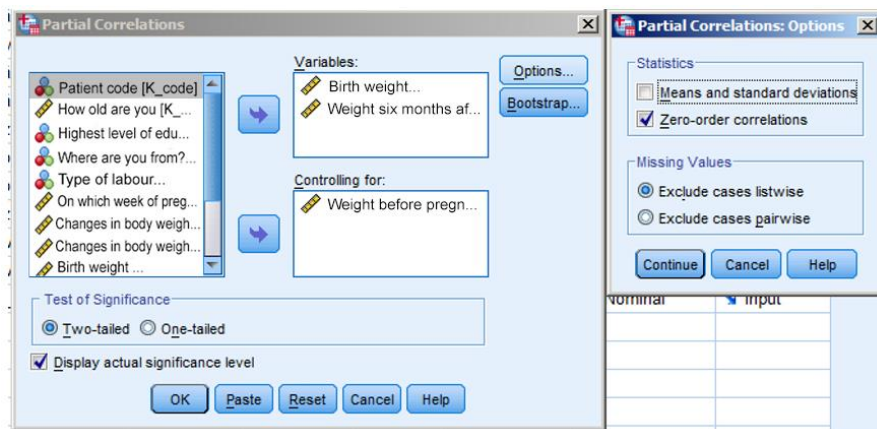
Source: author

Values in the diagonal are 1 by definition, since the variables determine themselves. The new result shows that there is a positive correlation between the birth weight and the mother's weight 6 months after giving birth. This correlation is significant ( $p=0.049$ ) but weak (0.24). The correlation between the mother's weight before and after giving birth is significant ( $p=0.00$ ) and strong (0.9). Stars after values show significance, their interpretation will be discussed later.

Interesting results may be obtained when calculating partial correlations. In SPSS, this calculation can be done quickly under *ANALYSE / CORRELATE / PARTIAL*. The *partial*



*correlation coefficient* measures the correlation between two variables, filtering the effect of one or more variables. Relationships that are not latent can also be detected by this method.



**Figure 8/19. Calculating partial correlation in SPSS**

The box of *VARIABLES* will need to contain the variables the correlation of which we would like to examine. The box *CONTROLLING FOR* includes variables the effect of which one would like to filter out. Under *OPTIONS*, select *ZERO-ORDER CORRELATIONS*. One of the three settings will give interesting results:

**Table 8/22. Table of partial correlation coefficient**

			Correlations		
Control Variables			Birth weight	Weight six months after labour	Weight before pregnancy
-none. <sup>a</sup>	Birth weight...	Correlation	1,000	,237	,309
		Significance (2-tailed)	.	,049	,009
		df	0	68	68
	Weight six months after labour	Correlation	,237	1,000	,904
		Significance (2-tailed)	,049	.	,000
		df	68	0	68
Weight before pregnancy	Correlation	,309	,904	1,000	
	Significance (2-tailed)	,009	,000	.	
	df	68	68	0	
Weight before pregnancy	Your children's birth weight	Correlation	1,000	-,104	
		Significance (2-tailed)	.	,394	
	Weight six months after labour	Correlation	-,104	1,000	
		Significance (2-tailed)	,394	.	
		df	67	0	

a. Cells contain zero-order (Pearson) correlations.

Source: author

The lower part of the table will be new information to us since here we analyse the correlation between the birth weight and the mother's body weight after 6 months, where weight before giving birth has been filtered out. The new results are not significant (sig=0.394) and the weak positive correlation has disappeared, and the direction has become negative. This means that we also need the body weight before giving birth. We did not find any latent correlations with the other two options.

## 9. INFERENCE STATISTICS (Pongrác Ács)

### 9.1. Introduction, theoretical background

In everyday life, it is very common that we do not have all the relevant information about an event or phenomenon.

*Inferential statistical methods* provide results to make inferences on all elements of the population, after observing a part of it. The database created this way is not comprehensive and will only concern a specific sub-population selected by a particular sampling technique. That is why “uncertainty” is always present when applying inferential statistical methods, i.e. there is a risk of error. Errors, however, can be kept to a minimum by choosing the proper statistical methodology and providing a correct interpretation.

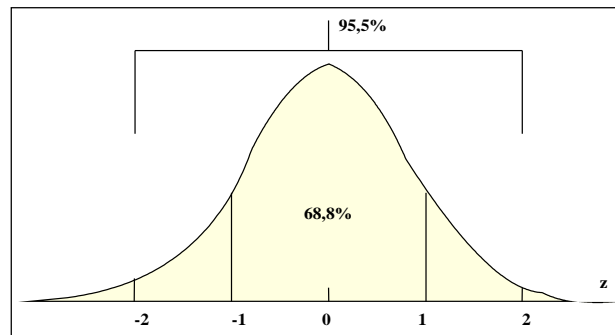
Two main groups of inferential statistical methods, *estimation and hypothesis testing* will be presented, mostly from the aspect of their practical application.

A significant proportion of social and economic phenomena, or even sport performances and results are assumed to be continuous variables with a normal distribution. Continuous variables can have an infinite number of values in a given interval; and the probability that variable X equals the value x is zero. Important measures defining probability distributions include the expected value ( $\mu$ ) and the variance, i.e. the square of standard deviation ( $\sigma^2$ ). Normal distribution is easy to be identified on the basis of its expected value and standard deviation, denoted as:  $N(\mu, \sigma)$ . We assume normality for e.g. weight, volume, height, length, BMI and performance.

Depending on the subject of analysis, expected values and standard deviations can have a lot of different values, which can make difficulties since their extent depends on the dimension of the variables. This problem can be solved by *standardization*, which means that we subtract the expected value from the value of the random variable, and divide this difference by the standard deviation. This way we get a *standard normal deviate* (noted by z). In formula:

$$z = \frac{x - \mu}{\sigma}$$

The result of standardization is a random variable of standard normal distribution with zero as the expected value and one unit as standard deviation:  $N(0,1)$ . The density functions of both random variables (with normal and standard normal distribution) have the shape of a so-called bell curve; the *Gauss curve* (Figure 9/1). In the case of standard normal distribution, both the random variables and their probabilities can be ordered in a table, with the help of which the resulting values can be used to solve the problems quickly and easily.



**Figure 9/1. Most important probability values depending on z**

The area between the interval plus and minus one standard deviation from the expected value and the probability curve represent 68.8% probability – this is true for both the normal and the standard normal distribution. The same value for the interval of plus and minus two standard deviations represents 95.5%, while three standard deviations stand for 99.9%. As the density function is symmetric, it is sufficient to determine the probability value between zero and positive infinity.

Mean values – especially arithmetic average – play a special role in inferential statistics. The question presents itself what sort of relationship we have between the means and the standard deviations of the samples, and the mean and standard deviation of the population. It is important to note that according to the *central limit theorem*, the average of a sample (simple random sampling) from a population of any kind of distribution is a random variable since its values differ from sample to sample but the *means are random variables with a normal distribution*. Of course, this considerably improves the practicability and popularity of normal distribution.

Inferential statistics require the detection of relationships between sample means, their standard deviations and the means and standard deviation of the population. It can be easily proven that if the means of all samples are known, then their mean will equal to the mean of the population. The standard deviation of the sample means, however, will differ from that of the population.

There is a formula on the relationship between the standard deviation of the population (variance) and the standard deviation of the sample means (variance):

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \left[ \frac{N-n}{N-1} \right]$$

where  $n$  is the sample size and  $N$  is the size of population.

Let us note that the second factor  $\left[ \frac{N-n}{N-1} \right]$  is called correction factor or finite multiplier. It plays an important role for samplings without replacement<sup>14</sup> but does not appear in the formula of sampling with replacement. The correction factor can be ignored in samples without replacement, i.e. simple random sampling if the measure of population ( $N$ ) differs very much from the sample size ( $n$ ) since its value is around 1 in this case.

The square of the sample mean's standard deviation  $\sigma_{\bar{x}}^2$  is a mean squared error, which is a consequence of replacing the expected value by the sample mean. The standard deviation of the sample mean ( $\sigma_{\bar{x}}$ ) is of great importance, and is called the sample mean's **standard error**. If the standard deviation of the population is known, the standard deviation of sample means is easy to calculate. Elements of the random sample are random variables, that is why any of their transformations – similarly to their arithmetic average – will be random variables, too.

## 9.2. Statistical estimation

Statistical estimation is the approximate determination of a constant parameter of an unknown population. These parameters can be the mean value (for finite population, the average), the standard deviation, and the ratio.

As seen before, there is a relationship between the mean of the population and the sample means, and the standard deviation of them. The standard error, i.e. the standard deviation of the sample means plays a particularly important role. It gives the opportunity to add an interval to the estimation where the occurrence of an event can be guaranteed with a probability.

In the formula  $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \left[ \frac{N-n}{N-1} \right]$ , the standard deviation of the population has to be known. If there is only the sample available, then the corrected sample standard deviation will be used which can be calculated as follows:

---

<sup>14</sup> Sampling without replacement (e.g. simple random sampling) is quite popular in practice since its application causes no waste of information.

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

With the corrected sample standard deviation, the formula of standard error – which can be applied in practice – where the finite multiplier

$$\left(1 - \frac{n}{N}\right)$$

is used if the sample size exceeds 5% of the population size:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Let us note that the standard error formula above only characterises the dispersion of means. Standard errors can also be defined for other parameters, e.g. total value, ratio.

Sample statistics that are applied for the approximate determination of population parameters are called estimators. The concrete value of the estimator for a given sample is called *point estimate*. The average measure of a random error that can occur in the estimation is represented by the standard error (the standard deviation of the estimator). The next table contains properties of the most commonly used estimators of population parameters.

**Table 9/1. Estimators of the most relevant population parameters**

Population parameter	Unbiased estimator	Standard error	Estimator's distribution
expected value	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$	$S_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n(n-1)}}$	small sample (n<50) t-distribution large sample (n≥50) normal distribution
ratio	$p = \frac{k}{n}$	$S_p = \sqrt{\frac{p(1-p)}{n}}$	small sample (n<50) binomial large sample (n≥50) normal distribution

Source: Pintér – Rappai (2001)

One can gain practically relevant information by interval estimation. In case of *interval estimation* one can rely on the fact that the sample parameters are random variables of a known distribution so an interval can be determined *with a given level of confidence* based on the value of the given distribution. This interval is referred to as *confidence interval*. The critical value to determine the intervals is located symmetric to zero because of the symmetry of normal distribution. The confidence interval can be determined based on the point estimation, the standard error and the type of distribution (because it is a point estimation to which the error limit is added in both directions). The error limit contains the tolerated

“imprecision” both in negative and positive directions. Confidence interval for mean estimation is calculated as:

$$\bar{x} \pm z \times \sigma_{\bar{x}}$$

where  $z$  is the given value of standard normal distribution from which the following ones are applied the most often:

**Table 9/2. Often applied critical values**

$\alpha$	$1-\alpha$	$Z_{(\alpha/2)}$	$Z_{(1-\alpha/2)}$
0.01	0.99	-2.576	2.576
0.05	0.95	-1.96	1.96
0.1	0.9	-1.645	1.645

Source: author

To sum up, estimation consists of six steps. The first three are theoretical tasks (sampling, establishing an estimator, assessing the estimator), while the other three are practical (point estimation, calculating the standard error, interval estimation)<sup>15</sup>.

Let us have a look at an example based on the above:

**As a practice exercise** let us consider the birth weight (gram) of the 70 babies involved in the database and estimate their average value with 95% confidence interval.

Calculating the (sample) mean:

$$\mu = \frac{\sum x}{n} = 3329.71$$

and the standard deviation (corrected standard deviation because we have data from a sample):

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = 547.44$$

based on the result, the standard error is (The finite multiplier cannot be applied here since the sample size does not exceed 5% of the population size):

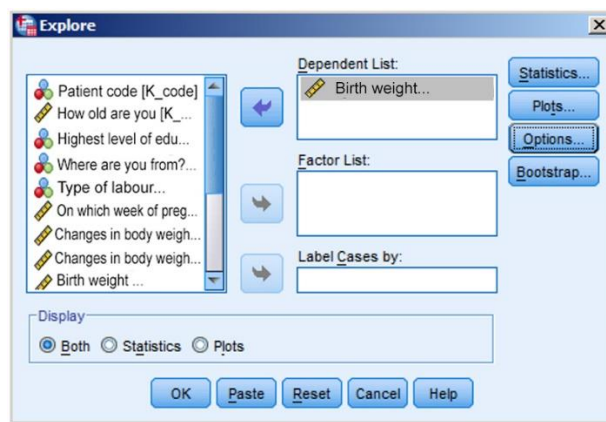
$$s_x = \frac{s}{\sqrt{n}} = 65.43$$

As the sample is large ( $n > 50$ ), the critical value from the table of normal distribution (see Table 9/2) will need to be identified which is:  $z=1.96$ . ( $Z(1-\alpha/2)$ ). The value of the error limit is ( $z \times \sigma_{\bar{x}}$ ): 128.25, from which the following result can be obtained:  $3329.71 \pm 128.25$ .

<sup>15</sup> Pintér – Rappai (2001)

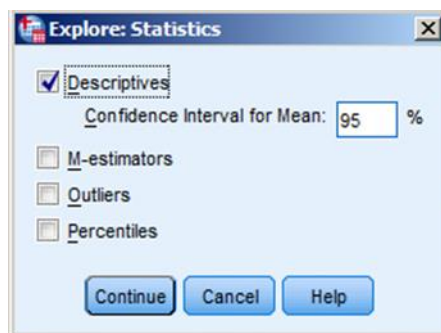
Note that SPSS does not consider sample size, and calculates results from the t-distribution. This leads to a small difference in the final results:  $z=1.9949$ , i.e. the error limit is 130.53 so the final result is  $3329.71 \pm 130.53$ . It means that we can state with 95% confidence that the birth weight of babies will be between 3199.18 and 3460.25 grams (interval estimation). Even additional continuous variables (expected week of birth, weight after 6 months, etc.) may be estimated using a similar method. Of course, this calculation and its graphical illustration can also be displayed in Excel as in Oláh 2008.

SPSS can also carry out estimations relatively quickly, see *ANALYSE / DESCRIPTIVE STATISTICS / EXPLORE*.



**Figure 9/2. Module for selecting variables**

Let us calculate the example above in SPSS. First, move the dependent variable to the box *DEPENDENT LIST*, then choose the module *STATISTICS* where we can select the estimation to be done.



**Figure 9/3. Settings of statistical estimations**

The confidence interval can be added in *DESCRIPTIVES*. Here we chose 95% based on the example above. Press *CONTINUE* and *OK* to get the results we are already familiar with.

**Table 9/3. Results of the estimation**

Descriptives

			Statistic	Std. Error
Birth weight	Mean		3329,7143	65,43151
	95% Confidence Interval for Mean	Lower Bound	3199,1820	
		Upper Bound	3460,2466	
	5% Trimmed Mean		3346,5079	
	Median		3400,0000	
	Variance		299689,772	
	Std. Deviation		547,43929	
	Minimum		2150,00	
	Maximum		4200,00	
	Range		2050,00	
	Interquartile Range		905,00	
	Skewness		-,294	,287
	Kurtosis		-,856	,566

Source: author

The result is the same as above but the programme calculated with t-distribution.

The ratio of the population in terms of a variable (partition coefficient) can be estimated similarly. Let us consider an element with a particular property, and denote its ratio in the population by  $P$ . **Point estimation of  $P$  ratio:**

$$p = \frac{k}{n}$$

where  $k$  is the number of elements having the given property and  $n$  is the sample size.

The standard error of the ratio in the sample is:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

$$\sigma_p = \sqrt{\frac{p(1-p)}{n} \left(1 - \frac{n}{N}\right)}$$

When working with a large sample, we can expect that the distribution of  $p$  is (approximately) normal, so we will use the values of standard normal distribution to calculate the confidence interval.

The confidence interval is:

$$p \pm z \times \sigma_p$$

**As a practice exercise** let us calculate the ratio of babies born with a caesarean section, with a confidence level of 95.5%.

$$p = \frac{k}{n} = \frac{40}{70} = 0,5714 = 57,14\%$$

$$0.5714 \pm 2 \times \sqrt{\frac{0.5714 \times 0.4286}{70}}$$



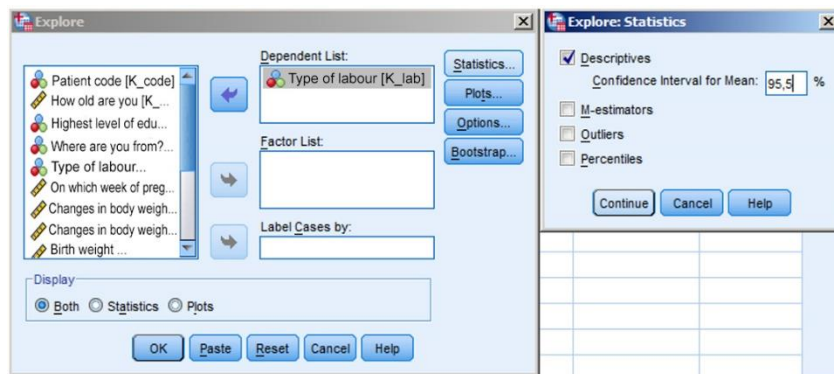
0.5714±0.1182

45.31%

68.97%

We can be 95.5% certain that the ratio of caesarean section will be between 45.31% and 68.97%.

In ratio estimation, we have to select the type of labour as a dependent variable and require 95.5% confidence level in module *STATISTICS*.



**Figure 9/4. Settings of ratio estimation**

Press *CONTINUE* and *OK* to get the same results in SPSS.

**Table 9/4. Results of ratio estimation**

			Descriptives	
Type of labour	Statistic		Statistic	Std. Error
Mean			1,57	,060
95.5% Confidence Interval for Mean	Lower Bound		1,45	
	Upper Bound		1,69	
5% Trimmed Mean			1,58	
Median			2,00	
Variance			,248	
Std. Deviation			,498	
Minimum			1	
Maximum			2	
Range			1	
Interquartile Range			1	
Skewness			-,295	,287
Kurtosis			-1,970	,566

Source: author

There is 95.5% certainty that the ratio of caesarean birth is minimum 45% and maximum 69%.

### 9.3. Difference tests

Difference tests are mathematical-statistical methods that help decide whether there is a significant difference between one or more sub-samples in terms of a given variable. Difference tests analyse different types of variables. These tests belong to the group of hypothesis testing but are one of the most popular methodologies so we would like to give a short review here. Of course, detailed practical examples will be presented in hypothesis testing.

If one also performs a significance analysis when doing a difference test, then useful results will be obtained not only for samples but also for the differences of the populations they represent.

It can also be figured out if the difference detected at the given confidence level can also be detected in other cases as well. If it can be proved that it does not occur randomly, then the difference is significant.

Usually, two samples are considered to be significantly different above 95% level of probability, i.e. when the chance of error does not exceed 5%. This significance level is denoted as  $p < 0.05$ . (From 100 cases, there are fewer than 5 errors.)

The most popular methods of difference tests are the following:

1. **Self-control test:** at least one variable of a sample or subsample is recorded in two different times. It is often referred to as paired sample test.
2. **Control group test:** comparing two independent samples with the same tool (e.g. comparing the smoking habits of boys and girls).
3. **Complex control group test:** testing if there is a difference in case of two or more subsamples and the same variable.

**Table 9/5. A possible grouping of tests**

		types of variables		
		data measured on interval scale (parametric tests)	ranked, ordinal data (non-parametric tests)	measurable, nominal data
number of samples	one sample (two dates)	one-sample t-test, paired sample t-test (self-control)	Wilcoxon test (self-control)	Chi squared test
	two samples	two-sample t-test with F test (control group)	Mann-Whitney test (control group)	Chi squared test
	more than two	Variance analysis (ANOVA, complex control group)	Kriska-Wallis test (complex control group)	Chi squared test

Source: author

#### 9.4. Hypothesis testing (parametric and non-parametric test in practice)

Hypothesis testing is the common term referring to the most often applied statistical methods. Hypothesis testing is a statistical method to decide if an assumption should be accepted or rejected, based on a chosen statistical test. These assumptions (hypotheses) contain a measure (e.g. mean, ratio) or parameter (e.g. expected value) of the population, or the distribution of the population (e.g. normal distribution) in a rather exact mathematical-statistical form. That is why it is possible to test the hypothesis and accept or reject it based on the results.

Hypothesis testing always refers to a hypothesis system which always includes a null hypothesis ( $H_0$ ) – basic assumption – and an opposing alternative hypothesis ( $H_1$ ). The result of hypothesis testing is always a yes-no (accept-reject) decision which is valid with a probability of error. Note that the decision always refers to the null hypothesis. Let us denote the population's unknown value by  $\Theta$  (theta) and the expected value by  $\Theta_0$ .

In most fields of science it supports an experiment, which in practice often means that it has to be decided if a new or modified method has an effect on entities taking part in the experiment. Two or more groups are separated and these groups receive different methods to follow, so e.g. it can be examined what changes have been caused by different methods of training. Two groups of patients are generated and the groups receive different methods of physiotherapy. The null hypothesis states that there is no essential difference between the effects of the two methods. The basic *null hypothesis* can be written as:

$$H_0: \Theta = \Theta_0$$

Of course, this expression itself is not enough to be interpreted; its opposition also needs to be drawn up. The *alternative hypothesis* makes the hypothesis system complete since it covers the entire “space of possible events”. About the example above: if there is a difference between results of the two groups (the effect is not random) but it is not clear which one is better, then the alternative hypothesis is *two-tailed*:

$$H_1: \Theta \neq \Theta_0$$

and it is *one-tailed* if it is clear which result are more favourable:

$$H_1: \Theta < \Theta_0$$

vagy

$$H_1: \Theta > \Theta_0$$

Hypothesis testing is carried out with the help of mathematical functions, a so-called *test statistic*. This makes comparison with theoretical values of types of distribution possible. Comparing the theoretical and the calculated value, the hypothesis will be accepted or

rejected, considering a given *significance level*. This is how the statement on the population is tested.

Hypothesis testing consists of four steps:

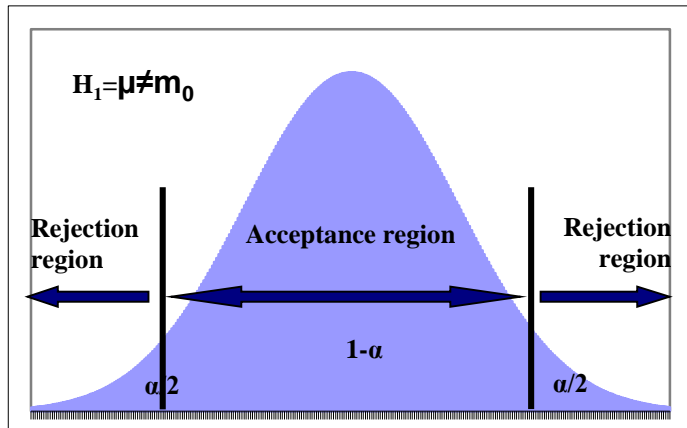
1. Set up the hypothesis system (Define  $H_0$  and  $H_1$ ).
2. Select the proper test statistic.
3. Calculate the value of test statistic from the sample (empirical data).
4. Make a decision.

When selecting the test statistic, the distribution of the populations, the type of sampling and the sample size will have to be considered. In most cases, independent, identically distributed sample (IID) is expected but there are only assumptions about the distribution of the population (this can be tested by fit tests).

The value region of the test statistic is separated into two regions, excluding each other: the acceptance region and the rejection (critical) region. One has to calculate the probability of the test statistics value to be between the limits of the acceptance region. If the test statistic exists in the rejection region, then the null hypothesis has to be rejected. Otherwise, it needs to be accepted. The probability of the test statistic belonging to the rejection region is called significance level.

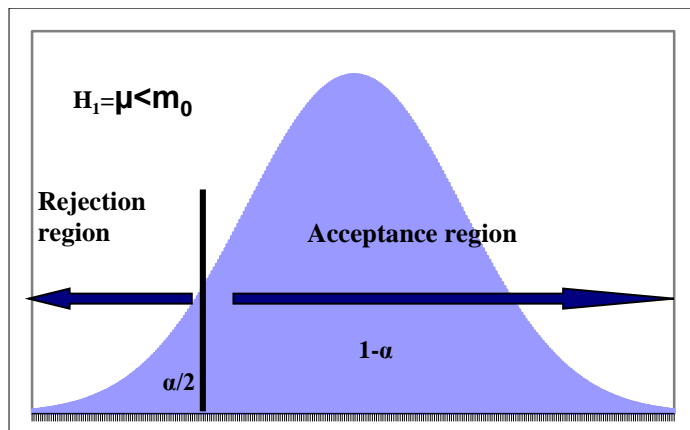
The other method to make a decision is based on the *significance value (p value)*, which represents the probability of error by rejecting the null hypothesis. If the p value is small, then there is a small probability of error Type I, so it is reasonable to reject the null hypothesis. On the other hand, if the p value is large, then the null hypothesis has to be accepted. Regions can be located in several different ways, depending on the alternative hypothesis. If the probability for the test statistic to be within the rejection region is  $\alpha$ , then the probability to be in the acceptance region is:  $1-\alpha$ .

Let us assume that the hypothesis is about the equality of the expected value ( $\mu$ ) and a value assumed ( $m_0$ ). One of the most common types of hypothesis tests is to decide if the expected value is equal to a constant given in advance. This type is the *one-sample expected value* test. In this case, it is tested if the expected value of the population is equal to a given number, and there can be different alternative hypotheses.

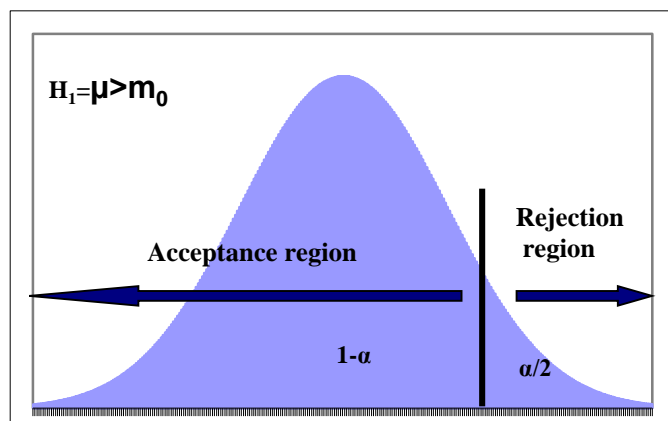


**Figure 9/5. Acceptance and rejection region in case of a two-tailed hypothesis**

The probability that the test statistic is within the rejection region is  $\alpha$ . As the rejection region consists of two parts of the same size, both represent the probability of  $\alpha/2$ . If the alternative hypothesis does not only state that the expected value is not equal to the test statistic but is larger or smaller, then it is referred to as the right-tailed or the left-tailed hypothesis.



**Figure 9/6. Acceptance and rejection region in case of a left-tailed alternative hypothesis**



**Figure 9/7. Acceptance and rejection region in case of a right-tailed alternative hypothesis**

Tests are called one- or two-tailed, and they can refer to expected values of the population, standard deviation or a ratio. Test statistics of the most common one-tailed tests:

Null hypothesis	Large sample (100≤n)	Small sample (n<100)
$H_0 : \mu = \mu_0$	$\tilde{z} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}   H_0 \sim N(0;1)$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}   H_0 \sim_{n-1} t$
$H_0 : P = P_0$	$\tilde{z} = \frac{P - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}   H_0 \sim N(0;1)$	
$H_0 : \sigma^2 = \sigma_0^2$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}   H_0 \sim_{n-1} \chi^2$	

Let us now consider a test in practice.

When consulting literature for his/her thesis, a student reads in an article that the average of BMI before giving birth was 25.6 with young Swedish mothers. The average BMI of young mothers in our database of 70 elements is 26.99 and the corrected standard deviation is 6.08. Let us test if there is a difference in this case between Hungarian and Swedish mothers. Can we accept the statement that the BMI value of Hungarian mothers cannot exceed that of Swedish mothers?

$$H_0: \mu=25.6$$

$$H_1: \mu>25.6$$

The null hypothesis states that the expected mean value of young Hungarian mothers' BMI is equal to the same of the Swedish mothers (difference from the Swedish mean is random). The alternative hypothesis says that this value is greater than 25.6, which means that there is a systematic reason for the difference<sup>16</sup> (e.g. unhealthy nutrition, sedentary lifestyle).

In practice, there is usually a lack of large samples, so the researcher has to rely on a small sample. In case of small samples, a standard normal distribution cannot be applied, so ***Student's t-distribution*** and its table will need to be applied. The so-called ***degrees of freedom*** will have to be taken into account here, which is sample size minus 1. A system's degree of freedom is the number of values to be chosen freely in the system (for t- and  $\chi^2$  distribution it is one, in case of F distribution d.f. is two).

$$t = \frac{26.99 - 25.6}{\frac{6.08}{\sqrt{70}}} = 1.91$$

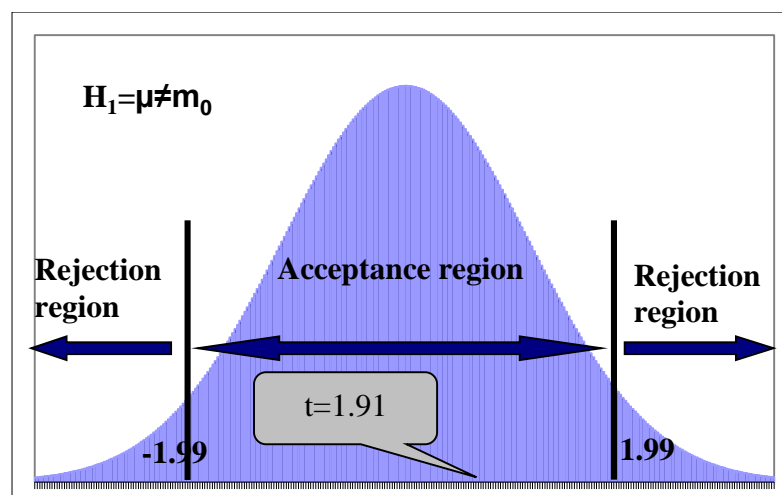
<sup>16</sup> This assumption is based on our professional opinion, the literature review of which is not included in this book.

The critical value of the t-distribution in case of 5% significance level and 69 (n-1) degrees of freedom is 1.99 as written in the table of t-distribution.

As the calculated value is smaller than the one in the table, the null hypothesis has to be accepted (there is no reason for rejection), i.e. the alternative hypothesis is rejected (the acceptance region of the chart is valid). This means that the difference between the sample value and the value expected is random (in case of 5% significance level).

If we are only interested to learn if the value in the sample is equal to the one for the Swedish sample, then a two-tailed alternative hypothesis has to be drawn up:

$$H_1: \mu \neq 25,6$$



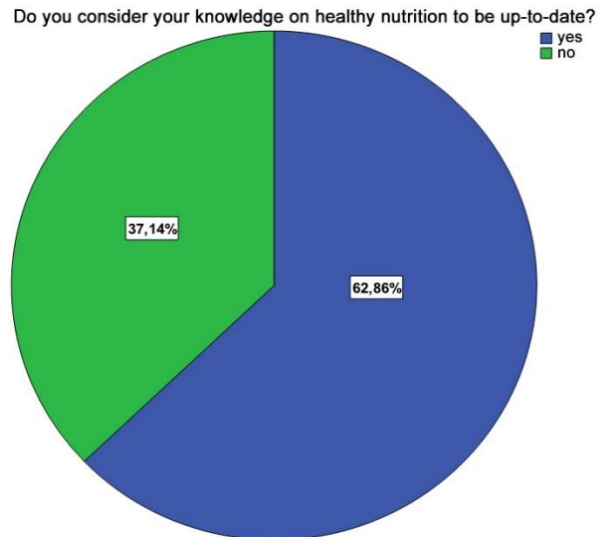
**Figure 9/8. Illustration supporting the decision in case of two-tailed alternative hypothesis**

Considering the alternative hypothesis above, both tails of the density function will have to be taken into account so the critical value (with 5% significance level) is:  $\pm 1.99$ . Comparing the t-value to these, the hypothesis on the BMI value of 25.6 has to be accepted, once again.

The method is similar if the assumption does not refer to the population mean but the ration. The standard normal distribution is again only applicable in case of a large sample.

$$z = \frac{p - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}$$

As an example, let us test if it can be expected that 60% of young mothers have up-to-date knowledge on healthy nutrition. First, we have to examine the ratio of the answers. (The question was: Do you consider your knowledge on healthy nutrition to be up-to-date?) This is displayed by the pie chart below:



**Figure 9/9. Pie chart on the ratio of answers**

44 young mothers said yes, and 26 said no to the question. Let us test if it can be expected that 60% of young mothers have the proper knowledge.

$$H_0: P=0.6$$

$$H_1: P<0.6$$

The alternative hypothesis states that the ratio of mothers with the proper knowledge is less than 60%.

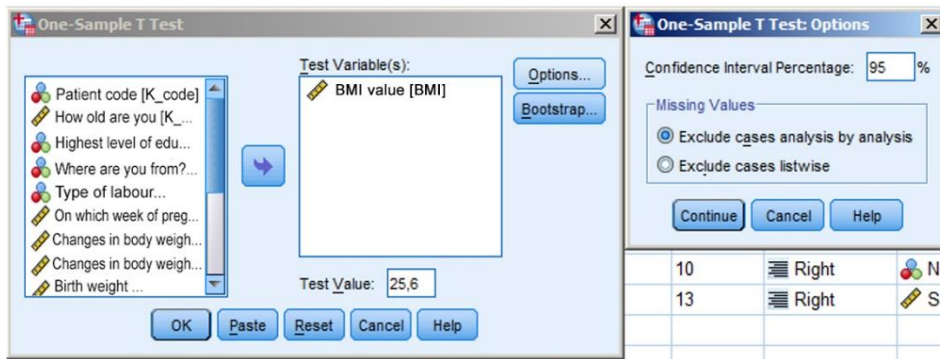
$$p = \frac{44}{70} = 0,628$$

$$u = \frac{0,628 - 0,6}{\sqrt{\frac{0,6 \times (1 - 0,6)}{70}}} = 0,93$$

The value in the table (considering the negative tail) is -1.645. The empirical value is located in the acceptance region so the null hypothesis has to be accepted. This means that the ratio of young mothers with up-to-date nutrition knowledge is lower than 60%.

It is quicker to test the hypothesis with the help of SPSS. To do so, one has to go to *ANALYSE / COMPARE MEANS / ONE- SAMPLE T TEST*.





**Figure 9/10. One-sample t-test in SPSS**

First, the user has to select the variable BMI as test variable, and then write the value in the null hypothesis (25.6) into the *TEST VALUE* box. This constant will be the basis of comparison. Add significance level (confidence interval percentage, 95%) under *OPTIONS*. Press OK and go to *OUTPUT VIEW* to get the results.

The results consist of two tables. The first contains descriptive statistical data.

**Table 9/6. Descriptive statistics**

**One-Sample Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
BMI value	70	26,9891	6,08032	,72674

Source: author

The second one helps to decide about the acceptance or rejection of the hypothesis.

**Table 9/7. Table containing t-value and significance**

**One-Sample Test**

	Test Value = 25.6					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
BMI value	1,911	69	,060	1,38907	-,0607	2,8389

Source: author

The t-value and the degrees of freedom are displayed but not the critical value. That is why the decision has to be made based on the significance of the calculated value of t. In general, null hypothesis is usually rejected under Sig=5% (0.05). As its value is larger in our example (more than 6 from 100 cases), we accept the null hypothesis. The confidence interval shows limits between which 95% of the difference is situated. Note that the significance value is close to the limit so one has to be careful with drawing general conclusions.

It often happens in practice that random and independent samples are collected from two different populations. In these cases, the same parameters of the two populations will have to be compared, their differences and common properties tested. In practical applications, the identity of the expected values of the two populations is often tested. As in the examples above, the general statement is drawn up by the null hypothesis, while the concrete form is stated in the alternative hypothesis.

$$H_0: \mu_1 - \mu_2 = \delta$$

Drawing up the alternative hypothesis in different ways will make it possible to make a decision on the measures and relations of the expected values:

$$H_1: \mu_1 - \mu_2 < \delta \quad ; \quad H_1: \mu_1 < \mu_2 \text{ (left-tailed)}$$

$$H_1: \mu_1 - \mu_2 > \delta \quad ; \quad H_1: \mu_1 > \mu_2 \text{ (right-tailed)}$$

$$H_1: \mu_1 - \mu_2 \neq \delta \quad ; \quad H_1: \mu_1 \neq \mu_2 \text{ (two-tailed)}$$

Now, we consider the most popular *two-sample t-test*, the application of which has two requirements to be met: distributions of both populations will need to be normal (external, additional information needed), and the squares of standard deviations of the populations are expected to be equal.

If the researcher has a small sample from a population with a normal distribution, and the standard deviation of the population is not known but their identity is expected, then a t-test can be applied (it is also applicable for bigger samples as well<sup>17</sup>):

The test statistic to be applied:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Degrees of freedom:  $n_1 + n_2 - 2$

Where the squared formula of the common standard deviation ( $s_p$ ):

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 1}$$

If the squares of standard deviations of the population are not known, the researcher should tests them. The quotient of the corrected variances in the sample (random variable) follows an F-distribution with  $n_1-1; n_2-1$  degrees of freedom.

---

<sup>17</sup>If one compares the values of t-distribution with higher degrees of freedom to the similar data of standard normal distribution, the similarity is quite obvious.

$$F = \frac{s_1^2}{s_2^2} \Big|_{H_0} \approx F_{n_1-1; n_2-1}$$

Differences of the two expected values are interpreted in the following example. From the database we know the descriptive statistics of BMI according to residence. The BMI average of young mothers from a town (35 people) is 24.14 with a standard deviation of 4.38, while mothers from a village (35 people) have an average of 29.84 kg, and a standard deviation of 6.25.

**Practice exercise:** Is there a difference between the BMI of young mothers in the two samples at 5% significance level?

Since it is a two-sample t-test, the researcher first has to examine if standard deviations can be considered identical, and test normality (see later).

Then the two-sample t-test follows:

$$H_0: \mu^1 = \mu^2$$

$$H_1: \mu^1 \neq \mu^2$$

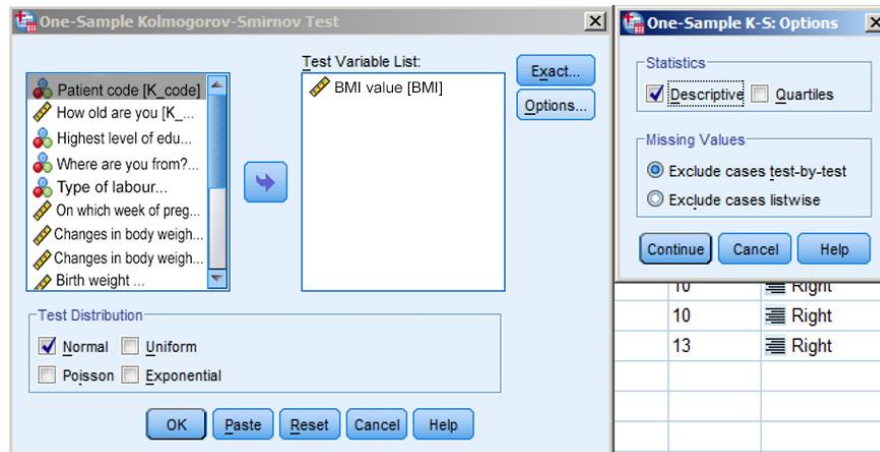
$$s_p^2 = \frac{(35-1) \times 4,38^2 + (35-1) \times 6,25^2}{35 + 35 - 2} = 29,12$$

$$s_p = \sqrt{29,12} = 5,396$$

$$t = \frac{24,14 - 29,84}{5,396 \times \sqrt{\frac{1}{35} + \frac{1}{35}}} = -4,41$$

The table value of the t-distribution with 68 degrees of freedom is 1.99 since the hypothesis is two-tailed. The calculated value is in the rejection region so at 5% significance level there is a difference in the BMI of young mothers living in towns and villages ( $p < 0.05$ ).

The same test can be carried out in SPSS. First, the user has to check if the BMI follows a normal distribution. The Kolmogorov-Smirnov test (presented before) is offered to support this decision. The access path is *ANALYSE / LEGACY DIALOGS / 1-SAMPLE K-S* where BMI has to be selected as *TEST VARIABLE*. In *OPTIONS*, descriptive statistics can be requested (*DESCRIPTIVE*). Finally, *CONTINUE* and *OK* will have to be pressed.



**Figure 9/11. Settings of normality test of the variable BMI**

The programme displays two tables from which the second contains the most relevant information.

**Table 9/8. Normality test of BMI**  
One-Sample Kolmogorov-Smirnov Test

			BMI value
N			70
	a,b	Mean	26,9891
		Std. Deviation	6,08032
Most Extreme Differences		Absolute	,084
		Positive	,084
		Negative	-,062
Kolmogorov-Smirnov Z			,707
Asymp. Sig. (2-tailed)			,700

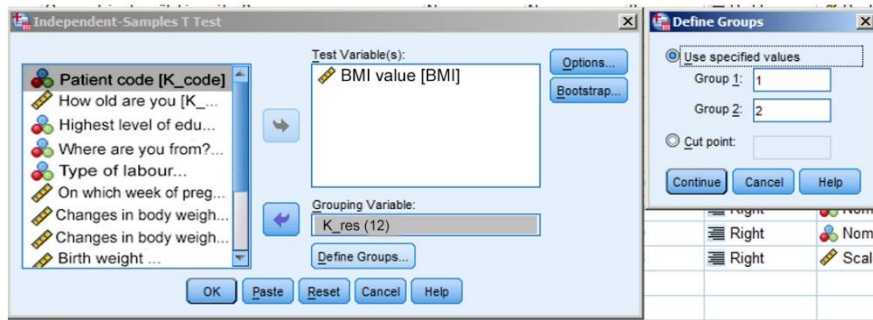
Test distribution is Normal.

Calculated from data.

Source: author

From the significance level at the bottom of the table ( $p=0.7$ ) we presume that the BMI follows a normal distribution ( $p<0.05$ ) so parametric methods can be applied.

The proper method can be found in *ANALYSE / COMPARE MEANS / INDEPENDENT-SAMPLES T TEST*. BMI has to be moved to the box of test variable(s), and where residence should be indicated as the grouping variable. Here, the codes of the two groups will need to be specified, too.



**Figure 9/12. Settings of the two samples t test**

The first table of the results contains descriptive statistics:

**Table 9/9. Descriptive statistics**

Group Statistics					
Where are you from?		N	Mean	Std. Deviation	Std. Error Mean
BMI value	City	35	24,1393	4,38158	,74062
	Village	35	29,8389	6,25372	1,05707

Source: author

The prerequisite for the application of the two-sample t-test is the equality of the standard deviations, which can be tested by the Levene (Levin) test in SPSS. The test can be taken as a special kind of F-test and its interpretation method is the same as the examples above because the null hypothesis assumes that standard deviations are the same. As the significance level observed is greater than 0.05, the null hypothesis has to be accepted and the upper row of the table has to be consulted. If the null hypotheses were to be rejected (standard deviations are not equal), then the second row would be valid. The software carries out a two-tailed test, testing the alternative hypothesis.

**Table 9/10.**

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
BMI value	Equal variances assumed	1,804	,184	-4,416	68	,000	-5,69963	1,29071	-8,27519	-3,12407
	Equal variances not assumed			-4,416	60,899	,000	-5,69963	1,29071	-8,28064	-3,11862

Source: author

Further results are the same as the ones above, and thus the null hypothesis will also have to be rejected. The negative sign before the t-value suggests that the average of the first variance category is the smaller one. This means that there is a significant difference in the BMI in terms of residence.

With a lack of normality, the two-sample t-test cannot be applied and instead the Mann-Whitney test is appropriate from the non-parametric tests.

**As a practice exercise** let us figure out if there is a difference in the week of birth in terms of residence, i.e. if the place of home determines the week of pregnancy when the child is born. According to the normality test, the week of birth does not follow a normal distribution ( $p < 0.05$ ).

**Table 9/11. Normality test on the week of birth**  
One-Sample Kolmogorov-Smirnov Test

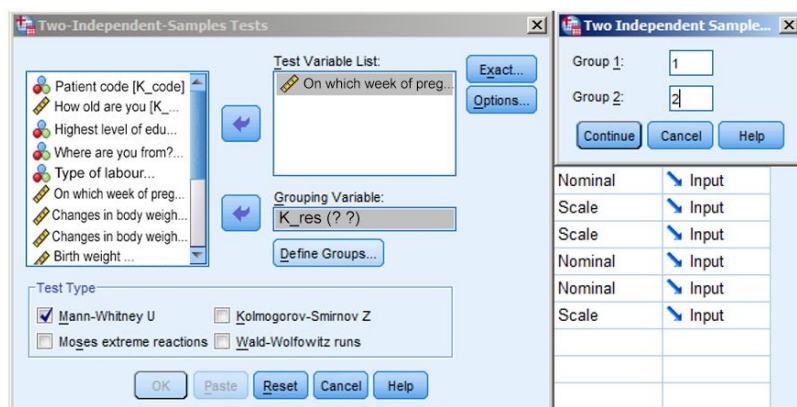
			On which week of pregnancy was your child born?
N			70
	a,b	Mean	38,2714
		Std. Deviation	1,49319
Most Extreme Differences		Absolute	,187
		Positive	,113
		Negative	-,187
Kolmogorov-Smirnov Z			1,566
Asymp. Sig. (2-tailed)			,015

Test distribution is Normal.

Calculated from data.

Source: author

If the precondition is not met, a nonparametric test will have to be carried out. The non-parametric pair of the two-sample t-test is the Mann-Whitney test, available at *ANALYSE / LEGACY DIALOGS / 2 INDEPENDENT SAMPLES*.



**Figure 9/13. Settings of the Mann-Whitney test**

The variable week of birth (“On which week of pregnancy was your child born?”) has to be put in the box *TEST VARIABLES* and the *GROUPING VARIABLE* has to be the residence. The codes of the variable values of the grouping variable need to be calibrated. In *OPTIONS*, descriptive statistics can be requested (*DESCRIPTIVE*). Finally, press *CONTINUE* and *OK*.

**Table 9/11. Results of the Mann-Whitney test**

	On which week of pregnancy was your child born?
Mann-Whitney U	529,500
Wilcoxon W	1159,500
Z	-1,000
Asymp. Sig. (2-tailed)	,318

Grouping Variable:  
Where are you from?

Source: author

Based on the significance in the table ( $p=0.318$ ), there was no significant difference between young mothers from towns and villages in terms of birth week. This means that the difference was random.

There are methods applicable in case of more than one population. First, we present a variance analysis. This method tries to measure subsample differences based on quantitative variables where the subsamples were generated based on one or more qualitative variables. The internationally applied abbreviation of the method is ANOVA. It is also referred to as “one-way ANOVA” or *one-way variance analysis*. The aim of the variance analysis (**Analysis Of Variance=ANOVA**) is to compare means but it is also a tool for examining variances. A variance analysis requires a normal distribution of the quantitative variable in the population and in all groups (sub-populations). The other precondition is the homogeneity of variance, i.e. that the standard deviations of the groups need to be equal (homoscedastic).

In statistical comparisons, dependent and independent variables can be defined. The independent variable provides the aspect of grouping – whether it is a grouping variable in the database or not. Values of independent variables are the samples themselves. The variable dependent from the samples is the continuous variable the means of which will be compared. This information can be important because statistical software may require the selection of a variable such as this. In cases when the database does not contain a grouping variable, then it will need to be generated.

There are versions of “more-way” variance analysis. A detailed description is included in Pintér – Rappai, 2007.

The three most typical cases of the one-way variance analysis are:

- testing a hypothesis if the expected values of more than two (sub)populations are equal;
- a homogeneity test;
- a significance test of mixed association (a relationship between a quantitative and a qualitative variable).

The model of variance analysis:

$$x_{ji} = \mu + \tau_j + \varepsilon_{ji}$$

where the  $i$ -th element of group  $j$  is  $(x_{ji})$  the sum of the expected value of the whole population  $(\mu)$ , the group effect of class  $j$   $(\tau_j)$  and the random effect  $\varepsilon_{ji}$ . The following hypothesis system is tested:

$$H_0 : \mu_1 = \mu_2 = \dots \mu_m = \mu$$

$$H_1 : \mu_j \neq \mu$$

The acceptance of the null hypothesis means that the expected values are the same, the population separated into parts is homogeneous and there is a lack of mixed association (i.e. independence). In terms of the grouped population, three sums of squares can be calculated from a sample, and the following connection between them is valid:

$$\sum \sum (x_{ij} - \mu)^2 = \sum n_j (\mu_j - \mu)^2 + \sum \sum (x_{ij} - \mu_j)^2$$

where the formula divides the total sum of square into external (between groups) and internal (within groups) sum of squares.

**Table 9/12.**

Sum of squares	Type of sum of squares
$SS = \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ji} - \bar{x})^2$	Total
$SS_K = \sum_{j=1}^m (\bar{x}_j - \bar{x})^2$	External (between groups)
$SS_B = \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)^2$	Internal (within groups)

Source: author



The test statistic made of sum of squares follows an F distribution where the degrees of freedom are m-1 in the numerator (m is the number of groups) and n-m in the denominator (n is the number of elements in the population). In the case of a larger one-tailed hypothesis, the test statistics is valid for variance analysis, i.e. if the calculated value of F is greater than the critical value, then the null hypothesis has to be rejected.

$$F = \frac{\frac{SS_K}{m-1}}{\frac{SS_B}{n-m}}$$

The following table summarizes the formulas and factors to support the decision:

**Table 9/13.**

Factors	Sum of squares (SS)	Degrees of freedom (df)	Mean of sum of squares (MS)	F
Between groups	$SS_K = \sum_{j=1}^m n_j (\bar{x}_j - \bar{x})^2$	m-1	$\frac{SS_K}{m-1}$	$\frac{MS_K}{MS_B}$
Within groups	$SS_B = \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)^2$	n-m	$\frac{SS_B}{n-m}$	
Total	$SS = \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ji} - \bar{x})^2$	n-1	$\frac{SS}{n-1}$	

Source: author

Let us consider the following example:

**As a practice exercise** let us figure out if the birth weight differs in the different BMI categories (underweight, normal weight, overweight, obesity), i.e. if the birth weights (in grams) of babies with the same BMI category mother are homogeneous or not.

First, the distribution of the continuous variable has to be tested since normality is a requirement for a variance analysis. Th normality test can compare the distribution of two random variables and can find out if a random variable follows the expected distribution. In this case, the researcher has to establish if the distribution of birth rate is normal. Normality is tested after checking the outlier values. If the calculated significance value is higher than 5%, then the variable follows a normal distribution. The size of the database is 70 so from the list of nonparametric tests, the Kolmogorov- Smirnov test can be applied as normality test.

The results in SPSS are summarized in the following tables:

**Table 9/14.**

	N	Mean	Std. Deviation	Minimum	Maximum
Birth weight	70	3329,7143	547,43929	2150,00	4200,00

		Birth weight
N		70
Normal Parameters <sup>a,b</sup>	Mean	3329,7143
	Std. Deviation	547,43929
Most Extreme Differences	Absolute	,091
	Positive	,069
	Negative	-,091
Kolmogorov-Smirnov Z		,758
Asymp. Sig. (2-tailed)		,614

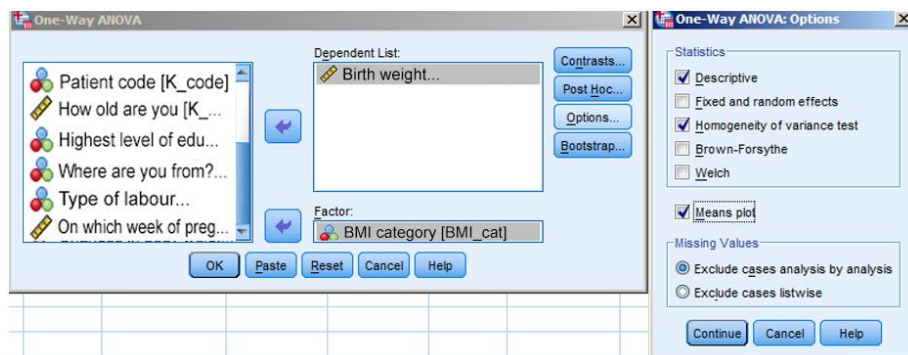
a. Test distribution is Normal.  
b. Calculated from data.

Source: author

Based on the z value of the Kolmogorov-Smirnov test and the significance level ( $p=0.614$ ), the variable follows a normal distribution so the variance analysis as a parametric test can be carried out. If the significance of the z value is less than 5%, then the nonparametric Kruskal-Wallis test will have to be applied (see later). Should the normality test be problematic, it should be presumed that the sample did not follow a normal distribution. Carrying out a nonparametric test on normal distribution data will give us the same result as the parametric test but this is not true vice versa.

An analysis can be started in *ANALYSE / COMPARE MEANS / ONE-WAY ANOVA*.

First, the variables will have to be selected. Select birth weight as the dependent variable (*DEPENDENT LIST*) and the BMI category as the independent (grouping) variable (*FACTOR*).



**Figure 9/14. Settings of variance analysis in SPSS**

Choose the following *OPTIONS: DESCRIPTIVE, HOMOGENEITY OF VARIANCE, MEANS PLOT*. The *HOMOGENEITY OF VARIANCE* should always be selected since it tests the precondition, i.e. if the variances are equal. Press *CONTINUE* and *OK* to see the results.

The first table contains descriptives (category size, mean, standard deviation, standard error, the lower and upper bound of confidence interval, and the minimum and maximum values):

**Table 9/15.**

**Descriptives**

Birth weight

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
underweight	5	2840,0000	713,61754	319,13947	1953,9268	3726,0732	2300,00	4050,00
normal weight	23	3276,9565	515,78215	107,54801	3053,9156	3499,9974	2200,00	4080,00
overweight	21	3221,4286	529,69128	115,58811	2980,3160	3462,5412	2150,00	4100,00
obesity	21	3612,3810	446,66436	97,47015	3409,0618	3815,7001	2910,00	4200,00
Total	70	3329,7143	547,43929	65,43151	3199,1820	3460,2466	2150,00	4200,00

Source: author

According to the table, there were 5 people in the category of underweight (“sovány”). Mean of the birth weight of babies with a mother in the category of obesity (“erősen túlsúlyos”) was 3612.38 grams with a standard deviation of 547.44. It is also clear that 95% confidence interval for birth weight was between 3409.06 and 3815.70 grams in the category.

The next table tests the homogeneity of variances by Levene’s test.

**Table 9/16.**

**Test of Homogeneity of Variances**

Birth weight?

Levene Statistic	df1	df2	Sig.
,210	3	66	,889

Source: author

If the significance level of the test is lower than 0.05, then the hypothesis has to be rejected, otherwise the variances are equal. If the variances are not equal, then the Brown-Forsythe and Welch’s test will have to be applied since the F-test does not provide relevant results in this case. In our example ( $p=0.889$ ), the variances can be considered to be normal. The next table is about variance analysis.

**Table 9/17.**

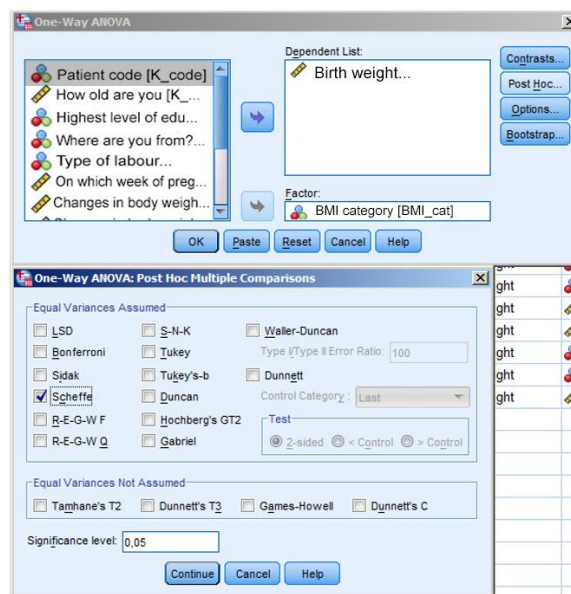
**ANOVA**

Birth weight?

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3187269,234	3	1062423,078	4,009	,011
Within Groups	17491325,05	66	265020,077		
Total	20678594,29	69			

Source: author

The first column contains sum of squares between and within groups and the total sum of squares. Degrees of freedom are listed in the second column. Dividing the sum of squares by them gives the mean squares between and within groups. F can be calculated ( $F=4.01$ ) by comparing the mean squares between and within groups ( $1062423.08/265020.08$ ). Significance is lower than 0.05 so the null hypothesis will be rejected which means that the birth weights of babies will differ according to their mother's BMI category. After making this statement, the difference of the means of different categories can be tested to find out which categories differ from one another. This can be done by the Post Hoc module, which requires at least three categories.



**Figure 9/15. The settings of Post Hoc tests**

The module is available as *ONE-WAY ANOVA / POST HOC*. There are several post hoc tests for testing differences between groups. There is no generally accepted method. Post hoc test are primarily clustered whether the requirements for equal variances are met or not. There are two important aspects to be considered when selecting the test: 1) how easy it is to

demonstrate a difference by the test (unresistingness), and 2) the degree of reliability. The first group of Post Hoc tests contains tests applicable in case of equal variances.

Some often applied **Post Hoc tests** will be introduced in this chapter. For equal variances, the Bonferroni and Scheffe test are often used. Bonferroni's test can be applied to test differences of mean pairs when the size of the two groups can also be different. It corrects the t-value belonging to the  $\alpha$ -error according to the number of independent comparisons. Test statistics of the Bonferroni test:

$$L = t(\text{Table}) \sqrt{S_p^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

**The Scheffe's test** belongs to the group of traditional ones, which tests null hypotheses. The F-test rejects  $H_0$  hypothesis if a vector  $a > 0$  exists where the confidence interval does not contain 0. If there are  $k$  number of groups to be compared, then  $k(k-1)/2$  number of comparisons have to be made. Statistics:

$$L = \sqrt{s_p^2 (k-1) F_{(\text{Table})} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

The **Dunnett's test** compares a given group (control) with all the others. Originally it was only valid for groups with equal numbers of elements but it was generalized later so that it can be applied for unequal sizes as well. Basically, it carries out pairwise comparison simultaneously, but an original control group has to be given and it compares means of other groups with the given one. The access path to the Dunnett's test is *ANALYSE / COMPARE MEANS / ONE-WAY ANOVA / POST HOC*. Before running the test, the control group has to be selected (*CONTROL CATEGORY*). Either the first or last group can be selected from the list. In addition, we will also need to specify if the comparison should be one or two-tailed. Default settings include a two-tailed symmetric comparison. In this case, the researcher has no preliminary information on the pairs to be compared; any group can be higher or lower than the control group. In case of a one-tailed test, the researcher has preliminary information whether the group to be compared can only be greater or lower than the control group. If there is no information available on the relation of the groups, then the two-tailed test has to be applied.

Statistics:

$$\bar{x}_i - \bar{x}_o \pm |d| s_p \sqrt{\frac{2}{n}}$$

$\bar{x}_o = \text{control group}$

If variances differ, then Tamhane-test and Dunnett’s T3 tests can be applied.

Let us present the interpretation of Scheffe’s post hoc test carried out on our database. In post hoc options, we chose Scheffe’s test to get the result in the next table.

**Table 9/18. Table of post hoc test of the variance analysis  
Post Hoc Tests**

**Multiple Comparisons**

Dependent Variable: Birth weight  
Scheffe

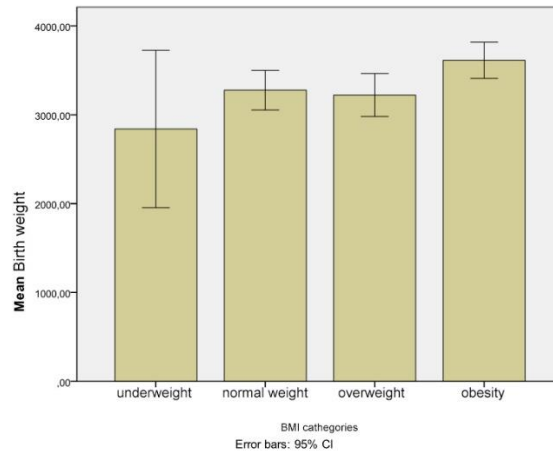
(I) BMI categories	(J) BMI categories	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
underweight	normal weight	-436,95652	254,02092	,405	-1165,7413	291,8282
	overweight	-381,42857	256,17185	,533	-1116,3843	353,5272
	obesity	-772,38095*	256,17185	,035	-1507,3367	-37,4252
normal weight	underweight	436,95652	254,02092	,405	-291,8282	1165,7413
	overweight	55,52795	155,37894	,988	-390,2535	501,3094
	obesity	-335,42443	155,37894	,209	-781,2058	110,3570
overweight	underweight	381,42857	256,17185	,533	-353,5272	1116,3843
	normal weight	-55,52795	155,37894	,988	-501,3094	390,2535
	obesity	-390,95238	158,87104	,120	-846,7526	64,8478
obesity	underweight	772,38095	256,17185	,035	37,4252	1507,3367
	normal weight	335,42443	155,37894	,209	-110,3570	781,2058
	overweight	390,95238	158,87104	,120	-64,8478	846,7526

\*. The mean difference is significant at the 0.05 level.

Source: author

The first column lists the basis of comparison (BMI categories, I), while the second one contains the subject of comparison (BMI categories, J). Scheffe’ post hoc analysis shows a difference between underweight (“sovány”) and obese (“erősen túlsúlyos”) (p<0.05). Mean Difference I-J is listed in the third column; significance is marked by a star.

In this case, it is recommended to present results in the form of bar chart with confidence intervals. The settings are presented above, and the figure is the following:



**Figure 9/15. Birth weight depending on the BMI categories before pregnancy**

If the normality requirement is not met, i.e. the continuous variable does not follow a normal distribution, the parametric test (variance analysis) above cannot be applied. If the distribution of the sample is not normal and it consists of more than three groups, then the nonparametric *Kruskal-Wallis* test will solve the problem. Nonparametric test do not require a specific distribution of the population – unlike parametric ones where the type of the distribution is an important precondition. That is why nonparametric methods are usually referred to as distribution free methods.

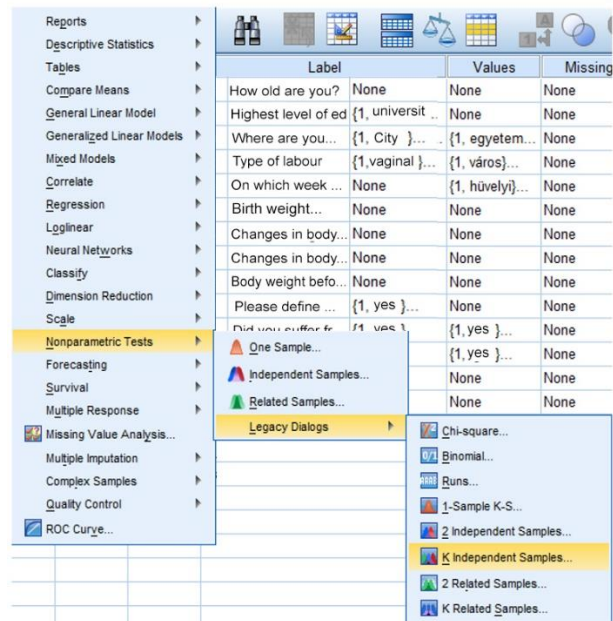
The *Kruskal-Wallis* test is often referred to as the nonparametric pair of the one-way variance analysis. Its methodology is similar to the Mann-Whitney U test, and when we have two independent groups, both tests can have the same results. Practically, the *Kruskal-Wallis* test is the general form of the Mann-Whitney test for three or more independent samples. The *Kruskal-Wallis* test unifies samples, calculates ranks, and then averages them by groups. If medians are equal, then rank averages do not differ significantly.

Let us consider the following example:

**As a practice exercise**, let us test if the birth weeks belonging to different BMI categories before pregnancy (underweight, normal weight, overweight, obese) differ. The question is if the birth weeks of different BMI categories are homogeneous or not.

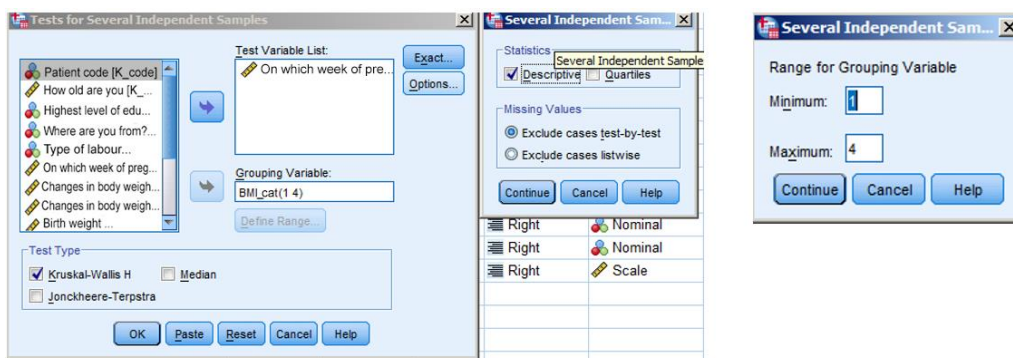
First, the normality of the variable birth week has to be tested. According to the Kolmogorov-Smirnov test's significance ( $p=0.015$ ), it is obvious that the distribution of the variable is not normal.

From the nonparametric methods, the *Kruskal-Wallis* test can be applied in this case. Access path is the following: *ANALYSE / LEGACY DIALOGS / K INDEPENDENT SAMPLES*.



**Figure 9/16. Settings of the Kruskal-Wallis test**

Select birth week (week of pregnancy when the child was born) as the target variable, and BMI category as the grouping variable. Select *OPTION / DESCRIPTIVE* to get descriptive statistics. The number of categories will need to be added as well, which is four in this case.



**Figure 9/17. Settings of the variables in the Kruskal-Wallis test**

The results show no significant difference in terms of birth week between the BMI categories. The first table lists the results of descriptive statistics. These include the sample size, the mean, the standard deviation, and the minimum and maximum values of the variables.



**Table 9/19. Descriptive statistics**

**Descriptive Statistics**

	N	Mean	Std. Deviation	Minimum	Maximum
On which week of pregnancy was your child born?	70	38,2714	1,49319	34,00	41,00
BMI cathegories	70	2,8286	,94748	1,00	4,00

Source: author

The second table contains the mean ranks and the number of elements in the different categories.

**Table 9/20. Number of elements and mean ranks of variable categories**

**Ranks**

BMI categories		N	Mean Rank
On which week of pregnancy was your child born?	underweight	5	35,80
	normal weight	23	26,59
	overweight	21	39,86
	obesity	21	40,83
Total		70	

Source: author

The third table contains the most important pieces of information from our point of view. It shows the significance value, on the basis of which the researcher can decide if there is a difference between the categories.

**Table 9/21.**

a,b

	On which week of pregnancy was your child born?
Chi-Square	7,167
df	3
Asymp. Sig.	,067

Kruskal Wallis Test

Grouping Variable:  
BMI categories

Source: author

## 10. AN INTRODUCTION TO REGRESSION ANALYSIS (Pongrác Ács )

### 10.1. Two-variable linear regression

Besides correlation analysis, regression analysis is the most commonly applied method to test the relationship between quantitative variables. Regression analysis examines tendencies of phenomena, and attempts to describe the nature of the correlation by a function. These functions are called regression functions. We will start this chapter by introducing a basic method called two-variable linear regression. In this case, changes (increase or decrease) in the target variable stochastically depend on the only independent variable. The correlation between variables is often not linear in practice. In cases like this, both the measurement of strength and the mathematical model requires relatively complex mathematical-statistical methods which are not presented in this book. If a linear stochastic correlation can be assumed, then a relatively simple mathematical model can be applied and a linear regression function  $\hat{c}Y = b_0 + b_1 X$  is:

An estimation of the constant parameters of the linear can be carried out by the method referred to as the ordinary least squares (OLS)<sup>18</sup>.

The two parameters can be calculated in the following way:

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$
$$b_0 = \bar{Y} - b_1\bar{X} = \frac{\sum Y}{n} - \frac{b_1 \sum X}{n}$$

In practice, the parameter of the independent variable ( $b_1$ ) has an especially important role, and is called regression coefficient, while the parameter  $b_0$  is called intersection or constant. The regression coefficient represents the expected change in the target variable, expressed in the original measurement unit due to one unit increase in the explanatory variable. A single change of unit in the value of the explanatory variable will change the target variable by  $b_1$  units. Having identified the measure of the regression coefficient makes it possible to quantify elasticity, which shows the relative measure of change (in percentages). The coefficient expresses the fact how many percentage change in Y dependent variable causes 1% change in X explanatory variable. The average elasticity can be determined by the means of variables in the following way:

---

<sup>18</sup>This book does not present a description of this estimation method only the formulas obtained via the application of the method.

$$E1 = b_1 \frac{\bar{x}}{\bar{y}}$$

There is a lot of literature available in the topic of regression analysis, see for example A Pintér – Rappai (2007), Mundruczó (1981) or Ramanathan (2003).

The two-variable linear regression will be presented first via a practical example. The method will be introduced with the help of a new database<sup>19</sup> which can be applied to practice other types of exercises.

The database was prepared by Dániel Kehl, associate professor, at the Faculty of Business and Economics (Pécsi Tudományegyetem Közgazdaságtudományi Kar), the University of Pécs, and has also been analysed in the book entitled “Sporttudományi kutatások módszertana” [Research Methodology in Sport Sciences]. Of course, the following illustrative and practice-oriented examples are new but the database offers numerous opportunities for students to practice.

The database contains the most significant motorcycle brands of the world and their most popular types. The final database contains fifty-three motorcycles. Data were collected from the publication “Motor katalógus” 2003 (a motorcycle catalogue).

The following data are available on different motorcycles:

1. manufacturer: most manufacturers have several products in the database,
2. type: type mark applied by the manufacturers,
3. origin: nationality of the manufacturer,
4. engine displacement (cm<sup>3</sup>)
5. performance (KW)
6. performance (horse-power, lóerő, LE)
7. weight (kg)
8. consumption (l/100km)
9. acceleration (0-100 km/h (s))
10. terminal velocity (km/h)
11. price (Ft).

**The method will be illustrated via a practical example.** First, we need to look at the correlation between performance (LE) and weight (kg). We have two continuous, metrical

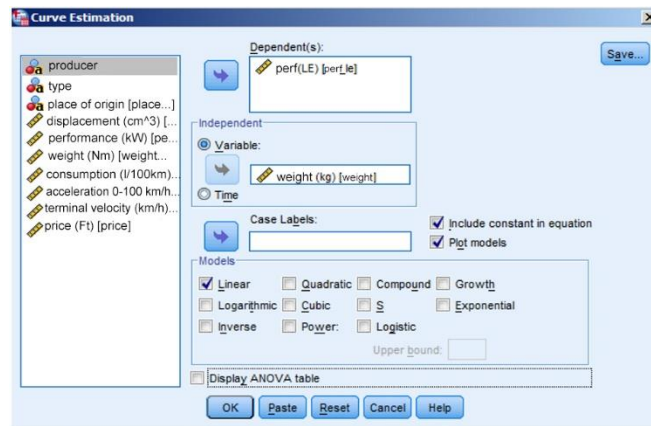
---

<sup>19</sup> The database (motor.sav) is available on the website of Pécsi Tudományegyetem Egészségtudományi Kar and the DVD supplement of the book entitled „Sporttudományi kutatások módszertana”. Multiple variable methods have been illustrated based on the database in the e-book of Ozsváth-Ács.

variables of which the performance (horsepower) is the *DEPENDENT* variable, and weight is the independent one. This means that we are going to test performance (LE) affected by weight.

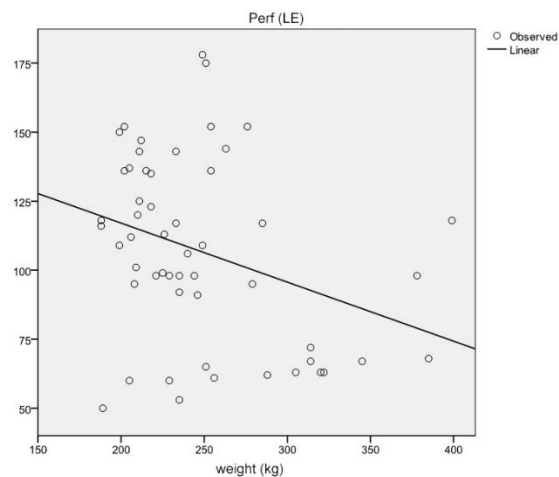
Researchers often display basic data in a scatter diagram to see the type and direction of correlation.

Access path in SPSS is *ANALYSE / REGRESSION / CURVE ESTIMATION* where graphical fit test can be displayed besides the scatter diagram.



**Figure 10/1. Graphical fit test of regression**

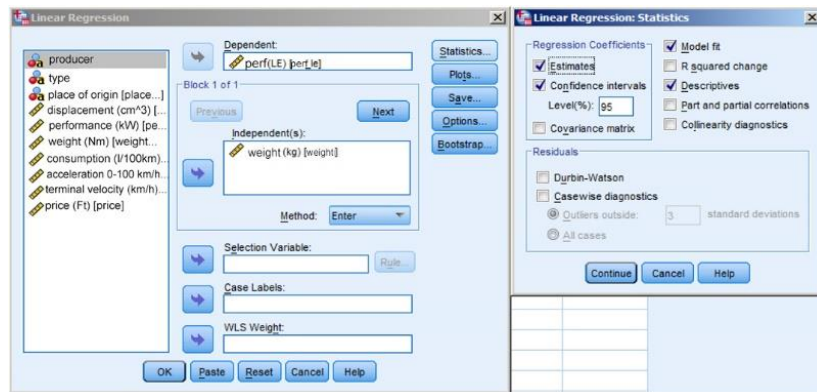
The variable performance (LE) has to be put into the box *DEPENDENT(s)*, while weight has to be defined as *INDEPENDENT* variable. Press *OK* to get the figure.



**Figure 10/2. Scatter diagram of regression**

Based on the diagram, one can state that the direction of correlation is negative but quantitative results are not displayed here.

The values of the regression line can be calculated at *ANALYSE / REGRESSION / LINEAR*. The settings are presented on the following screen view.



**Figure 10/3. Settings of linear regression**

Performance (LE) has to be defined as the dependent variable, and weight as the independent one, then by pressing *STATISTICS*, one has to choose the options *CONFIDENCE INTERVALS*, *MODEL FIT* and *DESCREPTIVES*. Press *CONTINUE* and *OK*, and see the results of calculation in the *OUTPUT VIEW*.

**Table 10/1. Summary data of the regression model**

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,327 <sup>a</sup>	,107	,089	31,932

a. Predictors: (Constant), weight (kg)

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6207,500	1	6207,500	6,088	,017 <sup>b</sup>
	Residual	52003,254	51	1019,672		
	Total	58210,755	52			

a. Dependent Variable: Perf (LE)

b. Predictors: (Constant), weight (kg)

Source: author

The first table contains the PEARSON correlation coefficient referring to moderate strength correlation ( $r=0.327$ ). The next result is the coefficient of determination (R Square) expressing the strength of the relationship ( $r^2 =0.107$ ). In this example, the independent variable explains almost 11% of the variation which means that the changes in performance are only 11% influenced by weight. The higher  $r^2$  is, the better the line fits the scatters. On the graphical figure, the coefficient of determination can be edited and required. It is followed by the standard error of the estimation which is the proxy of the precision of the forecast (the higher it is, the less capable the model). The next table is the ANOVA that we know from variance analysis, containing the value of the F-test and the significance value proving the existence of the correlation ( $p<0.05$ ). Based on the table, there is a correlation between the

two variables and it is not random. The table of coefficients also contains parameters of the regression line.

**Table 10/2. Parameters of the regression line**

Coefficients <sup>a</sup>								
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		
	B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	159,824	21,967		7,276	,000	115,725	203,924
	weight (kg)	-,214	,087	-,327	-2,467	,017	-,388	-,040

a. Dependent Variable:Perf(LE)

Source: author

Before interpreting the model, one has to see that the t values of both variables and their significance values ( $p < 0.05$ ) prove the existence of the model. At the end of the parameter estimation, there is also an opportunity to interpret the interval estimation of the parameters.

The regression equation contains the following non-standardized coefficients:

$$b_1 = -0.21$$

$$b_0 = 159.82$$

The regression function:

$$\hat{Y} = 159.82 - 0.21X$$

Based on the regression coefficient one can state that one unit increase in performance may be expected to decrease weight by 0.21 kg on average.

The performance of a 150 kb motorcycle can be estimated by the model as follows:

$$\hat{Y} = 159.82 - 0.21 \times 150 = 128.32$$

## 10.2. Multiple linear regression

Multiple linear regression is also a popular method, containing multiple independent variables. The two-variable linear regression presented above is not always appropriate since in practice there are not many occasions when a phenomenon can be described and explained by only two variables.

To summarize what linear regression can be applied for:

- To find out if independent variables influence the dependent variable: Does the relationship exist?

- To calculate to what extent the independent variables explain the variability of the dependent variable: the strength of the relationship.
- To define the form and the mathematical structure of the relationship.
- To forecast the values of the independent variable.

The method of linear regression can be extended to two or more independent variables. This is called multiple linear regression. The *multiple linear regression* is a statistical method to describe the relationship between a dependent variable ( $Y$ ) and two or more independent (explanatory) ones ( $X_1, X_2, \dots, X_i$ ). Similarly to linear regression, the procedure attempts to find out how one unit change in independent variables influences the dependent variable.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_i X_i + \varepsilon$$

- $Y$ : dependent variable
- $X_1, X_2, X_3, \dots, X_i$ : independent (or explanatory) variables
- $i$ : number of independent variables
- $\beta_0$ : (or  $\alpha$ ) constant, permanent value
  - Refers to the intersection of the regression line and the vertical axis of the coordinate system ( $y$ ).
- $\beta_1, \beta_2, \beta_3, \dots, \beta_i$ : constant coefficients of regression
  - Refers to the rise of the regression line.
  - Displays to what extent the dependent variable changes due to one unit change in the independent variables.
- $\varepsilon$ : error term

Preconditions:

- elements of the populations have to be independent from one another
- only populations with a normal distribution can be compared
- standard deviations are equal in the sample

Explanatory variables do not depend on one another (there is no multicollinearity between them).

Multicollinearity is the linear correlation between predictor variables. It can be tested several ways:

- Multiple coefficient of determination
- F-test ( $F > F_{crit}$ )
- Variance inflation factor (VIF)

Here we present the VIF method.

Variance inflation factor (VIF):

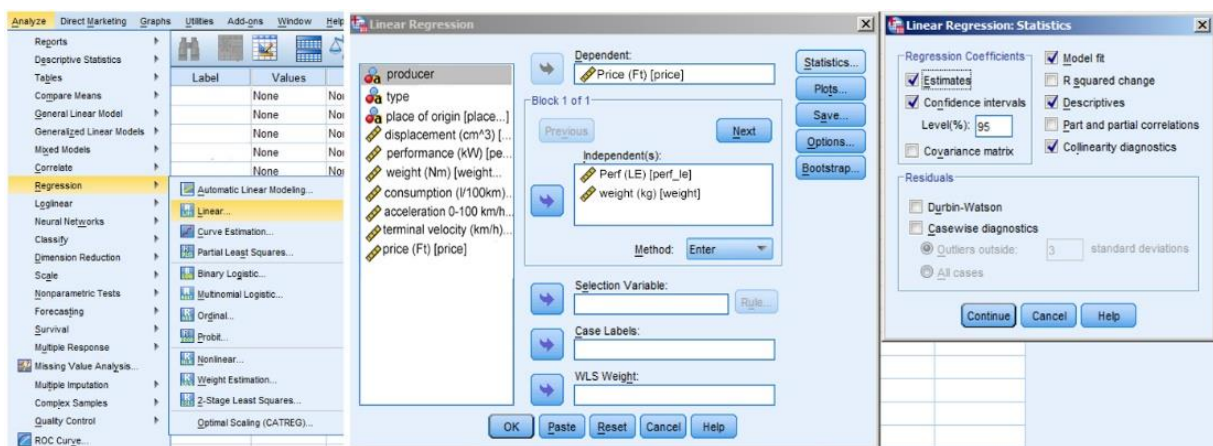
- $1 < VIF \leq \infty$
  - $VIF=1$  if  $R_j^2=0$  (jth independent variable does not correlate with the others)
  - $VIF \rightarrow \infty \Rightarrow R_j^2=1$  (jth independent variable is an exact linear combination of other independent variables)
- $1 < VIF \leq 2$  - weak multicollinearity
  - $2 < VIF \leq 5$  - strong, disturbing multicollinearity
  - $5 < VIF$  - very strong, harmful multicollinearity

Excluding multicollinearity: the correlation coefficient between two independent variables cannot exceed the value 0.7, the coefficient of determination cannot exceed 0.5.

**As a practice exercise**, let us find out if the performance and the weight of motorcycles correlates to the price (in HUF). The independent variables in our example are performance (LE) and weight (kg).

Settings:

Go to *ANALYSE / REGRESSION / LINEAR* (Figure 10/4). Select price (“ár”) as the dependent variable and select performance (“telj.”) and weight (“tömeg”) as the independent variables. In *STATISTICS*, select *ESTIMATES*, *CONFIDENCE INTERVALS*, *MODEL FIT*, *DESCRIPTIVES*, *COLLINEARITY DIAGNOSTICS*.



**Figure 10/4. Settings of multiple linear regression**



Press *CONTINUE* and *OK* to get the following results:

The first table consists of descriptive statistics of the three variables. The average price of motorcycles is 3 440 618 HUF, the average performance is 106 (LE), and the average weight is 250.5 kg. Besides the means, the standard deviations and the number of elements are also listed.

**Table 10/3.**

**Descriptive Statistics**

	Mean	Std. Deviation	N
Price (Ft)	3440618,00	1343598,867	50
Perf (LE)	106,16	34,164	50
weight (kg)	250,54	51,765	50

Source: author

The next table is a correlation matrix containing correlation measures and significance values. The correlation coefficient value in significant relationships between variables should not be higher than 0.7.

The next table lists information on the coefficient of determination (R, R square, adjusted R square, standard error of the estimate). Independent variables explain almost 50% (48) of variability in the dependent one, i.e. the price of motorcycles.

**Table 10/5.**

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,693 <sup>a</sup>	,480	,458	989521,817

a. Predictors: (Constant), weight (kg), Perf (LE)

Source: author

The next table contains values of the ANOVA test. The regression model explains variability of price ( $p < 0.05$ ) so it is applicable to estimate the dependent variable. The last two columns of the table contain the most relevant information. Here we need to reject the null hypothesis which means that the model can explain the dependent variable.

**Table 10/6.**

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4,244E+13	2	2,122E+13	21,670	,000 <sup>b</sup>
	Residual	4,602E+13	47	9,792E+11		
	Total	8,846E+13	49			

a. Dependent Variable: Price (Ft)

b. Predictors: (Constant), weight (kg), Perf (LE)

Source: author

Values belonging to the t-test in the last table show that the variables of the model are appropriate ( $p < 0.05$ ), and thus there is a linear relationship. VIF in the last column tests multicollinearity. Its value states that the extent of multicollinearity is not disturbing.

**Table 10/7.**

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-2267145,374	984129,116		-2,304	,026	-4246957,788	-287332,960		
	Perf (LE)	9004,631	4365,278	,229	2,063	,045	222,824	17786,438	,898	1,113
	weight (kg)	18966,360	2880,989	,731	6,583	,000	13170,557	24762,162	,898	1,113

a. Dependent Variable: Price (Ft)

Source: author

Based on B values in the last table, the following equation can be written:

$$\hat{ár} = -2267145.37 + 9004.63 * \text{teljesítmény (LE)} + 18966.36 * \text{tömeg (kg)}$$

The coefficients show that both performance (“teljesítmény”) and weight (“tömeg”) have an influence on price (“ár”); their direction is the same. Growth in performance and weight will lead to a price increase. The price of a motorcycle with 136 horsepower and 215 kg can be estimated as follows:  $(3035252 = -2267145.37 + 9004.63 * 136(\text{LE}) + 18966.36 * 215 (\text{kg}))$

## 11. LITERATURE

---

1. Ács P. (2009) Sporttudományi kutatások módszertana. Pécsi Tudományegyetem Természettudományi Kar Testnevelés- és Sporttudományi Intézet, Pécs
2. Alexin Z. – Lelovics Zs. (2009): Etikai kérdések a beavatkozással nem járó humán orvosi kutatásokban. Orvosi Hetilap 150. évfolyam 37. szám 1749-1752.o.
3. Babbie E. (2004): A társadalomtudományi kutatás gyakorlata. Balassi Kiadó, Budapest
4. Bajsz V. (2013) Egy multinacionális cég egészségfelmérése. Szakdolgozat. PTE ETK, Pécs
5. Bakonyi G. – Kokas K. (2006): Bevezetés a könyvtári informatika alapjaiba. Jatepress, Szeged 63-76. o.
6. Barton D. – Pál V. (2013): Open Access és osztályozás. Könyvtár figyelő. 59. évfolyam 4. szám 700–709. o.
7. Bauer András – Berács József – Kenesei Zsófia (2007): Marketing alapismeretek, Aula Kiadó, Budapest, o. 124
8. Berhidi A. (2008): Az orvosi szakirodalom adattárai 2. A tudomány hálójában: Web of Science. Nőgyógyászati Onkológia 13. évfolyam 3. szám 154-158.o.
9. Berhidi A. (2008): Az orvosi szakirodalom adattárai 3. Scopus: Elsevier „madárka” a tudomány „fészkeben”. Nőgyógyászati Onkológia 13 évfolyam 4. szám 186–190.o.
10. Bertáné Németh Á. – Farkas Fné. (2013): A PhD. disszertációval szemben támasztott formai követelmények. Online elérhető: <http://ktk.pte.hu/sites/default/files/mellekletek/2014/04/disszertacio.pdf> (2014-08-10.)
11. Betlehem J. – Boncz I. – Oláh A. (2010): Tudományos közlések az egészségtudományban. Nővér 23. évfolyam 6. szám 4-11.o.
12. Bíbor M. – Gulyás B. – Földes Zs. – Hegyi Á. – Kiss Farkas G. – Lacházi Gy. – Orosz A. – Parádi A. (2005): A magyar irodalom filológiája. Online elérhető: <http://www.tankonyvtar.hu/hu/tartalom/tkt/magyar-irodalom/ch07s11.html> (2014-08-01)
13. Bíróné Nagy E. – Bognár J. – Farkas J. – Gombocz J. – Hamar P. – Kovács A.T. – Mészáros J. – Ozsváth K. – Rétsági E. – Rigler E. – Salvára I. M. – Szabó B. – Tihanyiné Hős Á. – Vináné Kokovay Á. (2011): Sportpedagógia – Kézikönyv a testnevelés és sport pedagógiai kérdéseinek tanulmányozásához. Online elérhető:

[http://www.tankonyvtar.hu/en/tartalom/tamop425/0025\\_Birone\\_Nagy\\_Edit-Sportpedagogia/ch01s07.html](http://www.tankonyvtar.hu/en/tartalom/tamop425/0025_Birone_Nagy_Edit-Sportpedagogia/ch01s07.html) (2014-08-24)

14. Boncz I. – Döbrössy L.-Péntek Z. – Kovács A. – Budai A. – Imre L. – Vajda R. – Sebestyén A. (2013): A szervezett országos emlőszűrési program negyedik (2008-2009) szűrési körének részvételi arányai. Orvosi Hetilap 154. évfolyam 50. szám 1975-1983.o
15. Csajbók E. (2009): Az orvosi szakirodalom adattárai 4. Az orvostudomány határán. Nőgyógyászati Onkológia 14. évfolyam 3. szám 133-137. o.
16. Cserné Adermann G. (1999): A tanulás- és kutatómódszertan alapjai. JPTE-FEEFI, Pécs.
17. Demsey P.A. – Dempsey A.D. (1999): Kutatómunka az ápolásban. Medicina Könyvkiadó Rt, Budapest, 37-45.o.
18. Drótos L. – Martin Rebecca A. (2012): Értéknövelt szolgáltatás az olvasóknak: ingyenes és nyílt hozzáférésű források megtalálása. Tudományos és műszaki tájékoztatás 59. évfolyam 10. szám 437-439.o.
19. Martin Rebecca A. (2012): Értéknövelt szolgáltatás az olvasóknak: ingyenes és nyílt hozzáférésű források megtalálása. Tudományos és műszaki tájékoztatás 59. évfolyam 10. szám 437-439.o.
20. Falus I. (szerk) (2004): Bevezetés a pedagógiai kutatás módszereibe. Műszaki Könyvkiadó, Budapest
21. Falus I. (szerk.) (2000): Bevezetés a Pedagógiai kutatás módszereibe. Pedagógus Könyvek. Budapest. Műszaki Könyvkiadó. 540. o.
22. Farkas Livia (2012): A Google továbbra is a barátod: kérdőívek készítése
23. Fésüs L. (2014): Tudományetikai kihívások és válaszok hazánkban és Európában. Magyar Tudomány 175. évfolyam 6. szám 645-650. o.
24. Fidy J. – Makara G. (2005): Biostatisztika. Online elérhető: <http://www.tankonyvtar.hu/hu/tartalom/tkt/biostatisztika-1/ch11.html> (2014-06-30)
25. Gőcze I. (2011): A tudományos kutatás módszerei. Hadtudományi Szemle 4. évfolyam 3. szám 157-166. o.
26. Hajdu O. (1987): Sokváltozós statisztikai módszerek gyakorlati alkalmazása. Proinform Műszaki Tanácsadó Vállalat. Budapest
27. Héra G. – Ligeti Gy. (2006): Módszertan. Bevezetés a társadalmi jelenségek kutatásába. Osiris Kiadó, Budapest.

28. Hornyacsek J. (2013): A tudományos kutatás elméleti és gyakorlati kérdései 2. (A tudományos kutatás folyamata). Online elérhető: [http://hhk.uni-nke.hu/downloads/kiadvanyok/mkk.uni-nke.hu/kulonszam2013julius/eloadasokpdf/02hornyacsek\\_tudkut\\_final%20x.pdf](http://hhk.uni-nke.hu/downloads/kiadvanyok/mkk.uni-nke.hu/kulonszam2013julius/eloadasokpdf/02hornyacsek_tudkut_final%20x.pdf) (2014-08-24)
29. Hunyadi L. (2001): Statisztikai következtetéselmélet közgazdászoknak. KSH, Budapest
30. Hunyadi L. (2002): Grafikus ábrázolás a statisztikában, Statisztikai Szemle 2002/1. 22-53. o.
31. Istvánfi Cs. (2000): Gondolatok a sporttudományokról. Kalokagathia 1-2 szám 7-18. o.
32. Jackson P.A. – Reay J.L. – Scholey A.B. – Kennedy D.O. (2012): DHA-rich oil modulates the cerebral haemodynamic response to cognitive tasks in healthy young adults: a near IR spectroscopy pilot study. British Journal of Nutrition 107. évfolyam 8. szám
33. Jánosa A. (2005): Adatelemzés számítógéppel, Perfekt Kiadó. Budapest, 271. o.
34. Jobbágy Á. – Csordás P. – Mersich A. – Lupkovics G. – Sztaniszláv Á. (2008): Vérnyomás otthoni monitorozása. IME 7. évfolyam 10. szám 36-40.o.
35. Karamánné Pakai A. – Peterka G. – Dér A. – Bujtor A. – Dancsné Balogh K. – Czömpöl O. (2006): Egy kifejlesztés alatt álló otthon monitorozó készülék iránti igények felmérése Zala Megyében. Nővér, 19. évfolyam 5. szám 30-37. o.
36. Kecskeméty L. – Izsó L. (2005): Bevezetés az SPSS programrendszerbe, ELTE-Eötvös Kiadó, Budapest, 460. o.
37. Kehl D. – Rappai G. (2006): Mintaelem-szám tervezése Likert-skálát alkalmazó lekérdezésekben. Statisztikai Szemle 84. évfolyam 9. szám 848- 876. o.
38. Kerlinger F. (1980): Analysis of Covariance Structure Tests Of A Criterial Referents Theory Of Attitudes. Multivariate Behavioral Research 15 évfolyam 4. szám 403- 422. o.
39. Kerpel-Fronius S. (2008): A nürnbergi orvosper. A kényszereutanázia-program örök érvényű társadalmi tanulságai. LAM 18. évfolyam 2. szám 94-96. o.
40. Kopp M. – Kovács M.E. (2006): A magyar népesség életminősége az ezredfordulón. Semmelweis Kiadó, Budapest
41. Kovács J. (2007): Bioetikai kérdések a pszichiátriában és a pszichoterápiában. Medicina, Budapest
42. Lám J. – Balázsné Szelei E. – Rózsa E. – Bódi M. (2011): Gyógyszerosztással összefüggő gyógyszerelési hibák direkt megfigyeléses vizsgálatának szervezése, és a

- vizsgálat legfontosabb tanulságai. Semmelweis Egyetem, Egészségügyi Menedzserképző Központ. Online elérhető: [http://semmelweis.hu/emk/files/2012/02/7et\\_gyogyszereles.pdf](http://semmelweis.hu/emk/files/2012/02/7et_gyogyszereles.pdf) (2014-08-25.)
43. Lampek, K. – Kívés, Zs. (2012): Kutatásmódszertani és biostatistikai ismeretek (pp.177-208). In: Oláh A. (szerk.) Az ápolástudomány tankönyve. Medicina Könyvkiadó, Budapest, 177-210. o.
  44. Lógó Emma (2007): Kérdőíves technikák, módszerek (előadás)
  45. Majoros P. (2004): A kutatómódszertan alapjai: Tanácsok, tippek, trükkök - nem csak szakdolgozat-íróknak. Perfekt, Budapest
  46. Móczár Cs. – Borda F. – Faragó K. – Borgulya G. – Brauniczer F. – Vörös V. (2007): Egészséges életmód hatása túlsúlyos és elhízott betegeken. Orvosi Hetilap 148. évfolyam 2. szám 65–69. o.
  47. Mozaffarian D. – Hao T. – Rimm E.B. – Willett W.C. – Hu F.B. (2011): Changes in Diet and Lifestyle and Long-Term Weight Gain in Women and Men. The New England Journal of Medicine 364. évfolyam 25. szám 2392-2404. o.
  48. Mundruczkó Gy. (1981): Alkalmazott regressziószámítás. Akadémiai Kiadó, Budapest
  49. Oláh A. (szerk.) (2008): Az ápolástudomány tankönyve. Medicina Könyvkiadó, Budapest
  50. Online elérhető: <http://erg.bme.hu/oktatas/tleir/GT52AT02/kerdoiv.pdf> (2014-07-15)
  51. On-line elérhető: <http://www.informaworld.com/smpp/title~content=t775653673~db=all~tab=issueslist~branches=15 - v15> (2014-07-15)
  52. Online elérhető: <http://www.urban-eve.hu/2012/01/28/a-google-tovabbra-is-a-baratod-kerdoivek-keszítése/> (2014-07-12)
  53. Ozsváth Károly, Ács Pongrác (2011): Bevezetés a sporttudományos kutatásba [http://sek.nyime.hu/layouts/1038/Sport/ozsvath\\_acs\\_sportudkut121010.pdf](http://sek.nyime.hu/layouts/1038/Sport/ozsvath_acs_sportudkut121010.pdf) (Letöltés dátuma: 2014.09.01.)
  54. Pakai A. – Kívés Zs. (2013): Kutatásról ápolóknak 2. rész: Mintavétel és adatgyűjtési módszerek az egészségtudományi kutatásokban. Nővér 26. évfolyam 3. szám 20-43. o.
  55. Pakai A. – Tóth M. – Várady Horváth Á. – Oláh A. – Horváth Ö. P. (2013): Lehetséges protektív tényezők a sebgyógyulásban egy felmérés tükrében. Nővér 26. évfolyam 2. szám 8-15. o.
  56. Pálvölgyi M. (2011): Információkereső nyelvek III. Eger: Eszterházy Károly Főiskola. Online elérhető:

[http://www.tankonyvtar.hu/hu/tartalom/tamop425/0005\\_13\\_infkereso\\_nyelvek\\_iii\\_scor\\_m\\_05/531\\_online\\_katalogusok.html](http://www.tankonyvtar.hu/hu/tartalom/tamop425/0005_13_infkereso_nyelvek_iii_scor_m_05/531_online_katalogusok.html) (2014-08-12.)

57. Papp L. (2013): Kutatásról ápolóknak 1. rész: A kutatás tervezése. Nővér 26. évfolyam 2. szám 3-7.o.
58. Pintér J. – Rappai G. (2001): A mintavételi tervek készítésének néhány gyakorlati megfontolása. Marketing & Menedzsment 35. évfolyam 4. szám 4-11. o.
59. Ramanathan R. (2003): Bevezetés az Ökonometriába alkalmazásokkal. Panem Kft. Budapest
60. Rappai G. (2001): Üzleti statisztika Excellel. Központi Statisztikai Hivatal, Budapest
61. Reiczigel, J. (2005): Válogatott fejezetek a biostatistikából. SZIE ÁOTK, Online elérhető: <http://www.univet.hu/users/jreiczig/valfej/val-fej-jegyzet-2005-02-05.pdf> (2014-06-30)
62. Sajó A. (1999): Elektronikus folyóiratok az Interneten. Tudományos és Műszaki tájékoztatás 46. évfolyam 7. szám 275-280. o.
63. Sajtos L. – Mitev A. (2007): SPSS kutatási és adatelemzési kézikönyv. Alinea Kiadó, Budapest, 402. o.
64. Salavecz Gy. – Neculai K. – Rózsa S. – Kopp M. (2006): Az Erőfeszítés-Jutalom Egyensúlytalanság Kérdőív magyar változatának megbízhatósága és érvényessége. Mentálhigiéné és pszichoszomatika 7. évfolyam 3. szám. 231-246. o.
65. Siegrist J – Klein D – Voigt KH (1997): Linking sociological with physiological data: the model of effort-reward imbalance at work. Acta Physiol Scand 161. 112–116. o.
66. Szabó J. – Gerevich J. (2009): Kapcsolatok a felépülésben, felépülés a kapcsolatokban. A társas támogatottság mérése alkoholbetegek önéletrajzaiban. Lege Artis Medicinae 19. évfolyam 1. szám 67-72. o.
67. Szabó K. (2002): Kommunikáció felsőfokon. Kossuth Kiadó. Budapest. 2.Kiadás. 404 o.
68. Székelyi M. – Barna I. (2005): Túlélőkészlet az SPSS-hez. Typotex Kiadó, Budapest, 455. o.
69. Szepesi J.: HunKat, a HunTékát használó könyvtárak közös katalógusa. Online elérhető: <http://www.bdf.hu/konyvtar/tempus/HunKat.pdf> (2014-08-15)
70. Tomcsányi P. (2000): Általános kutatómódszertan. Szent István Egyetem, Gödöllő
71. Vajda R. – Karamánné Pakai A. – Éliás Zs. – Sélleyné Gyuró M. – Tamás P. – Várnagy Á. – Kívés Zs. (2014): A méhnyakrákkal kapcsolatos ismeretek és a

szűrővizsgálaton való részvételi mutatók vizsgálata. LAM 24. évfolyam 3. szám 118-125. o.

72. Vasas L. (2008): Az orvosi szakirodalom adattárai 1. MEDLINE: mindig és mindenhol
73. Vízvári D. (2010): Könyvek, folyóiratok, adatbázisok. Egészségügyi Stratégiai Kutatóintézet, Budapest



## 12. APPENDIX

### 1. Melléklet

**Forrás:** Egészségügyi Tudományos Tanács honlapján a [www.ett.hu](http://www.ett.hu)

**Hazai vonatkozásban a kutatás megkezdése előtt az alábbi törvényi szabályozás áttekintése ajánlott:**

- Egészségügyi törvény: **1997. évi CLIV törvény** VIII-IX. fejezet
- Genetikai törvény: a humán genetikai adatok védelméről, a humán genetikai vizsgálatok és kutatások, valamint a biobankok működésének szabályairól szóló **2008. évi XXI. törvény**
- Gyógyszertörvény: az emberi alkalmazásra kerülő gyógyszerekről és egyéb, a gyógyszerpiacot szabályozó törvények módosításáról szóló **2005. évi XCV. törvény**
- A humán reprodukciós eljárásokkal kapcsolatos, kötelezően nyilvánosságra hozandó érvényességi adatok, statisztikák köréről, a nyilvánosságra hozatal módjáról és helyéről, továbbá az ellenőrzés módjáról szóló **339/2008 (XII.30.) Korm. Rendelet**.
- Az emberen végzett orvostudományi kutatások, az emberi felhasználásra kerülő vizsgálati készítmények klinikai vizsgálata, valamint az emberen történő alkalmazásra szolgáló klinikai vizsgálatra szánt orvostechnikai eszközök klinikai vizsgálata engedélyezési eljárásának szabályairól szóló **235/2009. (X.20.) Korm. Rendelet**.
- Az emberi reprodukcióra irányuló különleges eljárások végzésére vonatkozó, valamint az ivarsejtekkel és embriókkal való rendelkezésre és azok fagyasztva tárolására vonatkozó részletes szabályokról szóló **30/1998. (VI.24.) NM rendelet**.
- **23/2002 (V.9.) EüM rendelet** az emberen végzett orvostudományi kutatásokról.  
/módosította: 31/2009. (X.20.) EüM rendelet/
- **28/2014. (IV. 10.) EMMI rendelet** Az Egészségügyi Tudományos Tanácsról.
- **35/2005. (VIII.26.) EüM rendelet** Az emberi felhasználásra kerülő vizsgálati készítmények klinikai vizsgálatáról és a helyes klinikai gyakorlat alkalmazásáról.  
/módosította: 32/2009 (X.20.) EüM rendelet/
- **33/2009. (X.20.) EüM rendelet** Az orvostechnikai eszközök klinikai vizsgálatáról.
- **34/2009. (X.20.) EüM rendelet** Az Állami Népegészségügyi és Tisztiorvosi Szolgálat egyes közigazgatási eljárásaiért és az igazgatási jellegű szolgáltatásaiért fizetendő díjakról szóló 1/2009 (I.30.) EüM rendelet a népjóléti ágazatba tartozó egyes

államigazgatási eljárásokért és igazgatási jellegű szolgáltatási díjakról szóló 50/1996 (XII.27.) NM rendelet, valamint az emberen végzett orvostudományi kutatásokról szóló 23/2002 (V.9.) EüM rendelet módosításáról.

- **ADATVÉDELEM 2011. évi CXII. tv.** Az információs önrendelkezési jogról és az információszabadságról. (Ezzel az 1992. évi LXIII. tv. "A személyes adatok védelméről és a közérdekű adatok nyilvánosságáról" hatályát veszítette.) **1997. évi XLVII. tv.** Az egészségügyi és a hozzájuk kapcsolódó személyes adatok kezeléséről és védelméről. /II. fejezet 21. §. tudományos kutatás céljából történő adatkezelés/
- **BÜNTETŐ TÖRVÉNYKÖNYV 2012. évi C. törvény XVI. fejezet** "az egészségügyi beavatkozás és kutatás rendje elleni bűncselekmények, valamint az egészségügyi önrendelkezési elleni bűncselekmények".
- **SZAKÉRTŐI TESTÜLET 33/2007. (VI.22.) IRM rendelet** Az Igazságügyi Szakértői Testületek szervezetéről és működéséről

### 13. SUPPLEMENT (TABLES)

#### 1. Standard normal distribution

##### Density function values

<b>z</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>0,0</b>	0,500	0,496	0,492	0,488	0,484	0,480	0,476	0,472	0,468	0,464
<b>0,1</b>	0,460	0,456	0,452	0,448	0,444	0,440	0,436	0,433	0,429	0,425
<b>0,2</b>	0,421	0,417	0,413	0,409	0,405	0,401	0,397	0,394	0,390	0,386
<b>0,3</b>	0,382	0,378	0,374	0,371	0,367	0,363	0,359	0,356	0,352	0,348
<b>0,4</b>	0,345	0,341	0,337	0,334	0,330	0,326	0,323	0,319	0,316	0,312
<b>0,5</b>	0,309	0,305	0,302	0,298	0,295	0,291	0,288	0,284	0,281	0,278
<b>0,6</b>	0,274	0,271	0,268	0,264	0,261	0,258	0,255	0,251	0,248	0,245
<b>0,7</b>	0,242	0,239	0,236	0,233	0,230	0,227	0,224	0,221	0,218	0,215
<b>0,8</b>	0,212	0,209	0,206	0,203	0,200	0,198	0,195	0,192	0,189	0,187
<b>0,9</b>	0,184	0,181	0,179	0,176	0,174	0,171	0,169	0,166	0,164	0,161
<b>1,0</b>	0,159	0,156	0,154	0,152	0,149	0,147	0,145	0,142	0,140	0,138
<b>1,1</b>	0,136	0,133	0,131	0,129	0,127	0,125	0,123	0,121	0,119	0,117
<b>1,2</b>	0,115	0,113	0,111	0,109	0,107	0,106	0,104	0,102	0,100	0,099
<b>1,3</b>	0,097	0,095	0,093	0,092	0,090	0,089	0,087	0,085	0,084	0,082
<b>1,4</b>	0,081	0,079	0,078	0,076	0,075	0,074	0,072	0,071	0,069	0,068
<b>1,5</b>	0,067	0,066	0,064	0,063	0,062	0,061	0,059	0,058	0,057	0,056
<b>1,6</b>	0,055	0,054	0,053	0,052	0,051	0,049	0,048	0,047	0,046	0,046
<b>1,7</b>	0,045	0,044	0,043	0,042	0,041	0,040	0,039	0,038	0,038	0,037
<b>1,8</b>	0,036	0,035	0,034	0,034	0,033	0,032	0,031	0,031	0,030	0,029
<b>1,9</b>	0,029	0,028	0,027	0,027	0,026	0,026	0,025	0,024	0,024	0,023
<b>2,0</b>	0,023	0,022	0,022	0,021	0,021	0,020	0,020	0,019	0,019	0,018
<b>2,1</b>	0,018	0,017	0,017	0,017	0,016	0,016	0,015	0,015	0,015	0,014
<b>2,2</b>	0,014	0,014	0,013	0,013	0,013	0,012	0,012	0,012	0,011	0,011
<b>2,3</b>	0,011	0,010	0,010	0,010	0,010	0,009	0,009	0,009	0,009	0,008
<b>2,4</b>	0,008	0,008	0,008	0,008	0,007	0,007	0,007	0,007	0,007	0,006
<b>2,5</b>	0,006	0,006	0,006	0,006	0,006	0,005	0,005	0,005	0,005	0,005
<b>2,6</b>	0,005	0,005	0,004	0,004	0,004	0,004	0,004	0,004	0,004	0,004
<b>2,7</b>	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003
<b>2,8</b>	0,003	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002
<b>2,9</b>	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,001	0,001	0,001
<b>3,0</b>	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001

## 2. Critical values for different significance levels

Significance level ( $\alpha$ )						
<b>One-tailed</b>	0.1000	0.0500	0.0250	0.0225	0.0100	0.0050
<b>Two-tailed</b>	0.2000	0.1000	0.0500	0.0450	0.0200	0.0100
<b>z</b>	1.280	1.645	1.960	2.000	2.330	2.587

### 3. Student's t-distribution

#### Critical values of Student's t-distribution for different significance levels

Degrees of freedom	Significance level				
	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
150	1.287	1.655	1.976	2.351	2.609
200	1.286	1.653	1.972	2.345	2.601

#### 4. $\chi^2$ -distribution

##### Critical values of $\chi^2$ -distribution for different significance levels

Degrass of freedom	Significance level					
	0.9900	0.9500	0.9000	0.1000	0.0500	0.0100
1	0.000	0.004	0.016	2.706	3.841	6.635
2	0.020	0.103	0.211	4.605	5.991	9.210
3	0.115	0.352	0.584	6.251	7.815	11.345
4	0.297	0.711	1.064	7.779	9.488	13.277
5	0.554	1.145	1.610	9.236	11.070	15.086
6	0.872	1.635	2.204	10.645	12.592	16.812
7	1.239	2.167	2.833	12.017	14.067	18.475
8	1.647	2.733	3.490	13.362	15.507	20.090
9	2.088	3.325	4.168	14.684	16.919	21.666
10	2.558	3.940	4.865	15.987	18.307	23.209
11	3.053	4.575	5.578	17.275	19.675	24.725
12	3.571	5.226	6.304	18.549	21.026	26.217
13	4.107	5.892	7.041	19.812	22.362	27.688
14	4.660	6.571	7.790	21.064	23.685	29.141
15	5.229	7.261	8.547	22.307	24.996	30.578
16	5.812	7.962	9.312	23.542	26.296	32.000
17	6.408	8.672	10.085	24.769	27.587	33.409
18	7.015	9.390	10.865	25.989	28.869	34.805
19	7.633	10.117	11.651	27.204	30.144	36.191
20	8.260	10.851	12.443	28.412	31.410	37.566
21	8.897	11.591	13.240	29.615	32.671	38.932
22	9.542	12.338	14.041	30.813	33.924	40.289
23	10.196	13.091	14.848	32.007	35.172	41.638
24	10.856	13.848	15.659	33.196	36.415	42.980
25	11.524	14.611	16.473	34.382	37.652	44.314
26	12.198	15.379	17.292	35.563	38.885	45.642
27	12.878	16.151	18.114	36.741	40.113	46.963
28	13.565	16.928	18.939	37.916	41.337	48.278
29	14.256	17.708	19.768	39.087	42.557	49.588
30	14.953	18.493	20.599	40.256	43.773	50.892
31	15.655	19.281	21.434	41.422	44.985	52.191
32	16.362	20.072	22.271	42.585	46.194	53.486
33	17.073	20.867	23.110	43.745	47.400	54.775
34	17.789	21.664	23.952	44.903	48.602	56.061
35	18.509	22.465	24.797	46.059	49.802	57.342
36	19.233	23.269	25.643	47.212	50.998	58.619
37	19.960	24.075	26.492	48.363	52.192	59.893
38	20.691	24.884	27.343	49.513	53.384	61.162
39	21.426	25.695	28.196	50.660	54.572	62.428
40	22.164	26.509	29.051	51.805	55.758	63.691
50	29.707	34.764	37.689	63.167	67.505	76.154
60	37.485	43.188	46.459	74.397	79.082	88.379
70	45.442	51.739	55.329	85.527	90.531	100.425
80	53.540	60.391	64.278	96.578	101.879	112.329
90	61.754	69.126	73.291	107.565	113.145	124.116
100	70.065	77.929	82.358	118.498	124.342	135.807
150	112.668	122.692	128.275	172.581	179.581	193.207
200	156.432	168.279	174.835	226.021	233.994	249.445
250	200.939	214.392	221.806	279.050	287.882	304.939

## 5. F-distribution

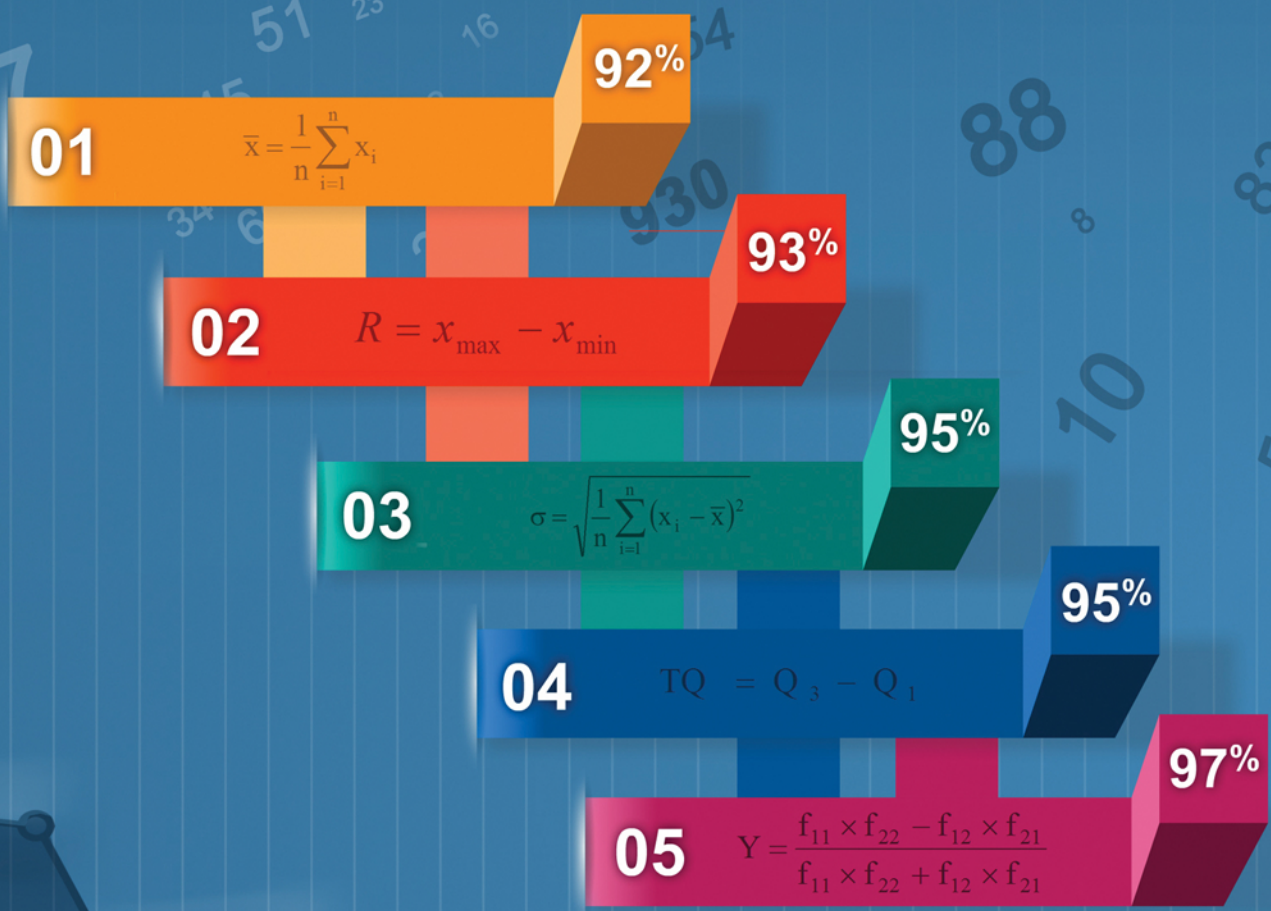
Critical values of F-distribution for one-tailed 5% significance level (two-tailed 10%)

Nevező szf.	Számológó szabadságfoka															
	1	2	3	4	5	6	7	8	9	10	15	20	25	30	50	100
2	18,5	19,0	19,1	19,2	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5
3	10,1	9,5	9,2	9,1	9,0	8,9	8,8	8,8	8,8	8,7	8,7	8,6	8,6	8,6	8,5	8,5
4	7,7	6,9	6,5	6,3	6,2	6,1	6,0	6,0	6,0	5,9	5,8	5,8	5,7	5,7	5,7	5,6
5	6,6	5,7	5,4	5,1	5,0	4,9	4,8	4,8	4,7	4,7	4,6	4,5	4,5	4,5	4,4	4,4
6	5,9	5,1	4,7	4,5	4,3	4,2	4,2	4,1	4,1	4,0	3,9	3,8	3,8	3,8	3,7	3,7
7	5,5	4,7	4,3	4,1	3,9	3,8	3,7	3,7	3,6	3,6	3,5	3,4	3,4	3,3	3,3	3,2
8	5,3	4,4	4,0	3,8	3,6	3,5	3,5	3,4	3,3	3,3	3,2	3,1	3,1	3,0	3,0	2,9
9	5,1	4,2	3,8	3,6	3,4	3,3	3,2	3,2	3,1	3,1	3,0	2,9	2,8	2,8	2,8	2,7
10	4,9	4,1	3,7	3,4	3,3	3,2	3,1	3,0	3,0	2,9	2,8	2,7	2,7	2,7	2,6	2,5
11	4,8	3,9	3,5	3,3	3,2	3,0	3,0	2,9	2,9	2,8	2,7	2,6	2,6	2,5	2,5	2,4
12	4,7	3,8	3,4	3,2	3,1	3,0	2,9	2,8	2,8	2,7	2,6	2,5	2,5	2,4	2,4	2,3
13	4,6	3,8	3,4	3,1	3,0	2,9	2,8	2,7	2,7	2,6	2,5	2,4	2,4	2,3	2,3	2,2
14	4,6	3,7	3,3	3,1	2,9	2,8	2,7	2,7	2,6	2,6	2,4	2,3	2,3	2,3	2,2	2,1
15	4,5	3,6	3,2	3,0	2,9	2,7	2,7	2,6	2,5	2,5	2,4	2,3	2,2	2,2	2,1	2,1
16	4,4	3,6	3,2	3,0	2,8	2,7	2,6	2,5	2,5	2,4	2,3	2,2	2,2	2,1	2,1	2,0

**6. Critical values of F-distribution for one-tailed 2.5% significance level (two-tailed 5%)**

Nevező szf.	Számológó szabadságfoka															
	1	2	3	4	5	6	7	8	9	10	15	20	25	30	50	100
2	38,5	39,0	39,2	39,3	39,3	39,3	39,3	39,3	39,3	39,3	39,3	39,3	39,3	39,3	39,3	39,3
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,4	14,4	14,3	14,2	14,1	14,1	14,0	14,0
4	12,2	10,7	10,0	9,6	9,3	9,2	9,0	8,9	8,9	8,8	8,6	8,5	8,5	8,4	8,3	8,3
5	10,0	8,4	7,7	7,3	7,1	6,9	6,8	6,7	6,6	6,6	6,4	6,3	6,2	6,2	6,1	6,0
6	8,8	7,2	6,6	6,2	5,9	5,8	5,7	5,6	5,5	5,4	5,2	5,1	5,1	5,0	4,9	4,9
7	8,0	6,5	5,8	5,5	5,2	5,1	4,9	4,9	4,8	4,7	4,5	4,4	4,4	4,3	4,2	4,2
8	7,5	6,0	5,4	5,0	4,8	4,6	4,5	4,4	4,3	4,3	4,1	4,0	3,9	3,8	3,8	3,7
9	7,2	5,7	5,0	4,7	4,4	4,3	4,2	4,1	4,0	3,9	3,7	3,6	3,6	3,5	3,4	3,4
10	6,9	5,4	4,8	4,4	4,2	4,0	3,9	3,8	3,7	3,7	3,5	3,4	3,3	3,3	3,2	3,1
11	6,7	5,2	4,6	4,2	4,0	3,8	3,7	3,6	3,5	3,5	3,3	3,2	3,1	3,1	3,0	2,9
12	6,5	5,1	4,4	4,1	3,8	3,7	3,6	3,5	3,4	3,3	3,1	3,0	3,0	2,9	2,8	2,8
13	6,4	4,9	4,3	4,0	3,7	3,6	3,4	3,3	3,3	3,2	3,0	2,9	2,8	2,8	2,7	2,6
14	6,3	4,8	4,2	3,8	3,6	3,5	3,3	3,2	3,2	3,1	2,9	2,8	2,7	2,7	2,6	2,5
15	6,2	4,7	4,1	3,8	3,5	3,4	3,2	3,2	3,1	3,0	2,8	2,7	2,6	2,6	2,5	2,4
16	6,1	4,6	4,0	3,7	3,5	3,3	3,2	3,1	3,0	2,9	2,7	2,6	2,6	2,5	2,4	2,4





**SZÉCHENYI 2020**



HUNGARIAN GOVERNMENT

European Union  
European Social Fund



INVESTING IN YOUR FUTURE

The book was prepared within the Social Renewal Operational Programme-4.1.2. E-13/1/KONV-2013-0012 tender.