UNIVERSITY OF PÉCS
FACULTY OF HEALTH SCIENCES

PONGRÁC ÁCS

RESEARCH METHODOLOGY IN SPORT SCIENCES

University of Pécs, Faculty of Health Sciences

Institute of Physiotherapy and Sport Science

RESEARCH METHODOLOGY IN SPORT SCIENCES

Pongrác ÁCS

Pécs, 2015

PÉCSI TUDOMÁNYEGYETEM
UNIVERSITY OF PÉCS

RESEARCH METHODOLOGY IN SPORT SCIENCES

Author & editor: Dr. Pongrác Ács

PUBLISHED BY

UNIVERSITY OF PÉCS

FACULTY OF HEALTH SCIENCES

The first edition was reviewed by

Dr. habil. Erzsébet Rétsági

Dr. Sándor Herman

Dr. habil. Gábor Rappai

The second, extended edition was reviewed by

Dr. habil Ferenc Ihász

Cover design and technical support by

Gábor Varga

Translated by

Kata Füge

Second, extended edition

ISBN 978-963-642-971-3

The manual has been produced in the framework of a project registered as TÁMOP-4.1.2. E-15/1/KONV-2015-0003

SZÉCHENYI 2020

MAGYARORSZÁG KORMÁNYA

Európai Unió
Európai Szociális Alap

BEFEKTETÉS A JÖVŐBE

TABLE OF CONTENTS

FOREWORD

All of the Hungarian higher education institutions at the field of sport have been considering the course 'Introduction to research in sport sciences' as highly important for several years now. The course was taught since the early 1940's at the predecessor of Semmelweis University Faculty of Physical Education and Sport Sciences as 'Basics of scientific research', and has the prospect of being continuously taught at the University of Physical Education. The course became even more important during recent years: as the *Bologna-process* changed the Hungarian highed education scene the course became an obligatory, basic subject in every Hungarian institution teaching sport sciences. It is included in the curriculum of the P.E. Teacher and Coach-, the Sport Management-, and ~~the~~ Recreation programs, both at the bachelor and master levels, and it will also be essential in case of the BSc- and MSc programs to be organised according to the new government decree on the vocational qualifications' register. We expect research- and leadership aspects are going to be gain special importance within the framework of MSc-training, where this type of knowledge is essential.

The importance of research in sport sciences is higlighted as sports are becoming more and more performance-focused, as no results can be achieved in international competitive sports without applying the scientific results.Besides, as people dispose of more free time, theoretical and practical applications of research on healthy living are becoming important as well.

It is evident that various sport results are an infinite mine for researchers, as well as for the sports' active participants, competitors, trainers and managers alike. As there is such abundant data source available, the information provided by all this data should be examined and analysed, with the results and conclusions published.

We may state that the Hungarian government has been treating sport as a strategic sector in the past four years, initiating major changes in the social, legal and economical environment of sport. Good examples for this are the introduction of daily P.E. lessons in schools, and the new system of corporate income tax, which changed the sport financing scene fundamentally. These measures have multiplicative factors in the sector.

It was in 2008, that inspired by all the abovementioned facts we considered for the first time the importance of writing a coursebook that – apart from the theoretical basics of research methodology – would provide practical aid for students to prepare their scientific works. In these past years we experienced the role and benefits of the book in practical education, and we also realised what the most challenging issues for students are. These

experiences and also the major changes that took place with softwares most often used in statistical data analysis (Excel, SPSS) provided a good reason to edit and extend the coursebook we compiled in 2008. Including requests from students, we wrote a completely new workbook, whichis useful for practice and checking the level of knowledge. We prepared the second, extended edition of the coursebook by using IBM SPSS version 22, and Microsoft Excel 2010.

Divided into two main parts, the book includes the complete list of topics covered by the BSc and MSc training in sport sciences research, and it may also be useful for students at PhD programs. Furthermore, it could also be useful for students at other scientific fields who must carry out their own research. The material of the book starts from simple methodological topics and ends with more complex ones, thus it is easy to understand and practice with. Our aim was to build a structure for the book where the particular methods are built upon each other, so that even those readers, who are only starting to learn about scientific research may learn new, hands-on knowledge. Our textbook, containingreal-life examples provide good source of knowledge also for experts of the field.

Backed up by student feedback, we believe that the structure of the book written in 2008 to be well-planned, consequently we did not undertake any further changes so we did not change that significantly. We found it important to reduce the number of databases included in the first book instead, the aim was to use real-life databases as much as possible, and to use a minimum number of real-life databases as examples. Thus in the first part of the book we used a database compiled of data gained from university student population collected by the Hungarian School Sport Federation, on the basis of fitness tests modelled for primary school student sample. We believe it is important to demonstrate statistical methods on a database which is easily accessible for everybody and is defined by standard measurements. Thus experts may practice these statistical methods based on their own calculations as well.

We provide a frame for the book by introducing the definitions and structure of science and sports sciences in the first chapter of the book. We describe how sport sciences became a science, provide details about its history and also about the basic models of sport sciences research.

In the first part of the second chapter we describe the preparation of the research plan, and illustrate topic choice, literature review, research hypotheses, designating research sample and other tools to complete the research itself. In this we strongly build on our experience in teaching, thus we emphasise the appropriate methods of preparing reference lists, one issue that seems to present particular challenge for current university and college students.

By describing research tools we provide detailed information about how to edit surveys, using up-to-date examples from sport.

In the second part of this chapter we discuss data processing and statistics in detail, trying to provide a peak into the complex world of descriptive-, inferential-, and multivariate statistical analysis.

Developments in informational technology and the extended use of PC-s in everyday life mean that besides the traditional statistical calculations with pen on paper we have to demonstrate measurement, description and modelling of mass phenomena by computer softwares (SPSS, Excel) that are popular in scientific research. At the examples we are expect basic Excel user's knowledge from readers – this should be no problem as the software is introduced to students in primary school, and also it is available for every university student.

At the end of the chapter we provide suggestions for the publication and presentation recent scientific results.

We aimed to find a good balance between theory and practice throughout the whole book. We tried to make theoretical sections more understandable by using real-life sport examples and databases that we published already. Unlike our first book, there is no DVD attached to this extended edition, as the book, the databases and the electronic workbook as well are free for download from the website of University of Pécs, Faculty of Health Sciences (www. etk.pte.hu).

I would like to thank László Harsányi†, Gyöngyvér Prisztóka, Zsuzsanna Pótó and Dániel Kehl for their advices, suggestions and useful tips during the writing of the first book. Many other colleagues helped me during the preparation of the second edition, for which I am incredibly grateful. I would like to thank the support and professional help in particular for József Betlehem, András Oláh, Bence Cselik and Gábor Varga.

I am grateful for my colleagues for their support and inspiration to edit and extend this book, and I am also particularly thankful for the students for their feedback on ways for improvement. I thank the precise reviewing work of *Sándor Herman, Gábor Rappai, Erzsébet Rétsági, and Ferenc Ihász*, who drew my attention to certain shortcomings of the book – by correcting these the material improved significantly.

I would like to dedicate this coursebook to Dr. Ferenc Farkas, a teacher whose support I enjoyed throughout my years in active sport and at the university alike.

I would like to invite my colleagues and students alike to help upgrading and actualising this work with their comments and suggestions, providing me with the opportunity to

correct any mistakes that might have been left in the material (for which only the author is responsible).

Finally I would like to express my hope that this book will support our nation to become a great sport nation again, an essential indicator of which are good quality researchin sport sciences.

Kozármisleny, 7th August 2015

Pongrác Ács
author

1. SCIENCE, THE PLACE OF SPORT SCIENCES WITHIN THE SYSTEM OF SCIENCES

> *„A word is only that of science if it travels the world. Thus if we wish to be true scientist and – a must – good Hungarians, we shall hold up the flag of science high enough to be seen and respected across many borders."* (Loránd Eötvös)

The term science comes up all the time in our everyday lives, and the scientific point of view and scientific basis is essential in different aspects of life. Science is as old as humanity itself, as even the first humans started to examine and store information on principles of nature (eg. cave paintings), and applied this information in favour of the community.

Afterwards it was *philosophy* that integrated new knowledge homogenously. In the beginning science (philosophy) included all activities of understanding: religious doctrines, mythology, arts, and ideological thinking as well as experiences, examinations and meditations. The first philosophers were therefore the first astronomers, mathematicians and physicists as well.

In the following we list a few definitions for science.

„The system of verifiable knowledge on the objective relationships within nature, society and thinking"[1]

„What we consider science today: activities to examine, define and verify disciplines and relationships; the system of verified knowledge; and the institutions and organs for storing, publishing, applying and organising science."[2]

„Science is the systematic scope of knowledge we have on a given subject; the respective pieces of facts must be organised systematically to provide the scope of science and help the understansing of related phenomena. Systematic nature is thus a significant requirement in all forms of sciences. By systematic nature we mean the objective order of

[1] Magyar Értelmező Kéziszótár (1973). 378.o.

[2] Hepp- Nádori (1971). 10. o.

principles on a matter, by which we may comprehend and understand principles according to their entitled co- and subordination and in their entirety fullness."[3]

„Science experiments and explores, measures and examines, and creates theories that explain the hows and whys. It thinks up technical methods and tools, comes up with and rejects suggestions. It creates and tests hypotheses, questions nature (and, should we add, social reality), forcing the answers, giving opinion, confuting, verifying and denying, distinguishing the true from the false and the the sensible from the senseless. It tells how to get where we want and how to do what we want. A scientist is a man like anybody, but also different form other for knowing how all this should be done. He received a strictly scientific education training, and became a stubborn, confident and sensible man... What is more: the scientist posesses the rare privilege of being able to think for himself – he may practice the great and solitary art of thinking individually. And still he belongs to that universal community that speaks a universial language."[4] (Wartofsky [1977] pg. 13)

There are several well-distinguished meanings for the term *science* in Hungarian language, many of which are depicted by the abovementioned definitions:

1. It may be one of the most important tools for getting to know the universe and our own selves, as an active process, a social activity, the *scientific research*, which is carried out by well-defined methodology.
2. We may call *science* the group of people who carry out the actual activity, the international scientific community. All committed people may be considered here, who carry out scientific research at their own fields and they publish their results at an official outlet. In Hungary those people belong to this category who completed their state-defined Ph.D education, whithin which they proved their knowledge in front of a specialised board. As a positive result the relevant accredited scientific board granted scientific title for them. The method for this in Hungary is the doctoral (Ph.D.) process legally defined by the higher education act, the awading the title by the university committees, the doctoral oath set by the state, and the official doctoral commencement ceremony.
3. However, what we mean by science most of the time is the end-product, created by the abovementioned community for the whole of humanity. [5]

---

[3] Pallas Nagy Lexikona. (http://www.mek.iif.hu/porta/szint/egyeb/lexikon/pallas)

[4] Szabó (2003)

[5] http://hu.wikipedia.org

Definition of scientific fields and disciplines according to the Hungarian Goverment's Decree Nr. 169/2000. [IX. 29.] és 154/2004. [X. 14.]:

Table 1/1.

Scientific field		Scientific discipline (piece)
1.	Natural sciences	7
2.	Engineering sciences	11
3.	Medical sciences	5
4.	Agricultural sciences	6
5.	Social sciences	10 (Sport sciences)
6.	Humanities	9
7.	Arts	7 (Various forms of arts)
8.	Theology	

We may see that the abovementioned govement decree places sport sciences within the category of social sciences with nine other disciplines. Table 1/2. lists these disciplines:

Table 1/2.

Social sciences	
1.	Management and business administration
2.	Economics
3.	Law and legal studies
4.	Sociology
5.	Psychology
6.	Pedagogy
7.	***Sport sciences***
8.	Political science
9.	Military science
10.	Multidisciplinary social sciences

When we want to define sport sciences, we must remember that the subject of sport sciences is an *examination and understanding a specific group of human activities*

"*The definition of sport science:* as a sub-field of humankind's universal culture, it is a theoretical sysrem representing the culture of the body by the evidence-based, systematical and generalised principles, themes, laws and rules, theories and methods. Its research aim

is to enrich values of the society's culture of the body (as a subculture of the universal culture), and thus support individual and eventually the totality of societal development. It is the examination of people as biological-psychological and social units, who consciously practice physical activity"' (Bíróné N. E. 2004. p. 18)[6].

Of course we may differentiate between the two types of reality when striving for for knowledge in sport sciences (Babbie 2000: 33). *Consensual reality* means we accept something becasue most people accept it, too. Some things are real because they are said to be real. In team sports we accept the coach's words (eg. the defenders are not so tall, we may score goals from headers), even if we never had a match with an opposing team. In *experiential reality* understanding comes from direxts experiences. During sports it is cruicial how and how fast we can react to the unexpected, "experienced" situations (e.g. defenders are not tall, but they jump twice as high as our forwards).

***The science of sport and physical education is interdisciplinaryinsterdisciplinary*:** it is a borderline field emerging from the similar methodological techniques of the two classic scientific fields (natural- and social sciences), and displays involvement with several scientific fields at the same time (Figure 1). A basic distinction within science refers to *living- and inert* **natural sciences**, and **social sciences.** Natural science is the broad group of studies examining the objects of living and inert nature. **Social sciences** examine the human being as a social creature, the society created by humans, and also the relationship between the two.

„*PhysicalPhisical education is one of the most complex sciences.* Within its horizon we find more or less all aspects of natural- and social sciences. In our point of view, *sport science is a social science considering its content and aims, while its methods are predominantly built on that of natural sciences.* Obviously, separating the two aspects is rather difficult, and such a distinction serves mainly the clarification of concepts only.." (Hepp F.- Nádori L. 1971: 20)

[6] Bíróné N. E. (2004). 18.o.

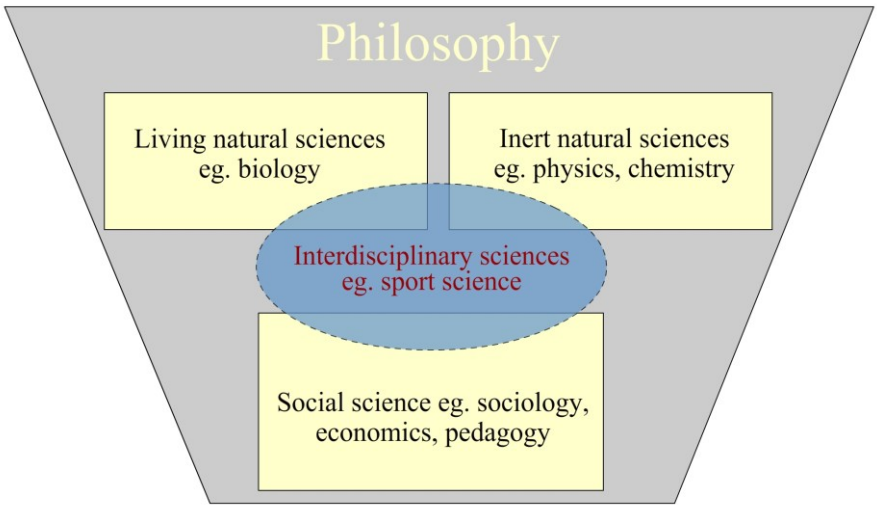


Figure 1/1. The place of sport sciences within the system of sciences

Source: edited by the authors

1.1 Development of sport science

The beginning of sport science's intensive development was in the 1950's, thanks to the professional sports rivalry of the world's two superpowers, the Soviet Union and the United States of America. Before this era there existedscientific health care examinations, focusing on physical education's and sport movements' influence on the human body. The difference between the two superpowers' research in sport existed from early on: while the Soviet Union almost solely focused on professional sports, the United States started to carry out researches in recreation, rehabilitation and specialised physical education as well, parallel to the professional sport researches. The field's international acceptance was enhanced by the conferences on sport sciences as accompanying programs, which have been organised side by side with the Games since the 1956 Olimpics. Regarding the field's institutions, the International Federation of Sports Medicine (FIMS) was founded in 1928, and the International Council of Sport Science and Physical Education (ICSSPE) was founded in 1960.

The first step of the development of sport science in Hungary was the foundation of the University of Physical EducationEducaiton (TF) in 1925, followed by the foundation of the National Institute of Physical Education and Sport Science (OTSI) in 1952. The Council for Physical Education was established at TF in 1954, and the independent Physical EducationEduacation Reserach Institute (TTKI) was organised in 1959, which continued its work from 1969 as a TF research institute. According to the decree of 1st September 1975 by the Presidential Council of the Hungarian People's Republic the TF operated first as a college and then as a college with university features, finally getting university status

in 1986. The decree by the Presidential Council of the Hungarian People's Republic defined the new name of TF as Hungarian Univeristy of Physical Education. The Hungarian Accreditation Committee accepted the Ph.D (Doctor of Phylosophy) program of the Hungarian University of Physical Education in 1997. Later on the Sport Sciences Subcommittee was organised within the Hungarian Academy of Sciences, V. Section of Medical Sciences, Committee of Preventive Medicine.[7] The institution, which celebrated its 75th anniversary in 2000 has served the cause of physical education and sport as the Semmelweis University's Faculty of Physical Education (TF).[8] Another milestone was 4th July 2014, when the National Assembly modified the national higher education act and founded a new institution: thus the Faculty separated from Semmelweis University and continued its work as an independentindependet institution, under the name of University of Physical Education.

There are several arguments proving the existence of Hungarian sport science:

- **Accredited higher educational institutional systems** (Budapest, Pécs, Szombathely, Eger, stb.)
- **Scientific association** (The Hungarian sport science is being organised by the Hungarian Society of Sport Science –MSTT–, in which specialized committees function.)
- **Scientific Journals in Hungarian:** Testnevelés (1928)[Gymnastics]; Testneveléstudomány (1950)[Sport science]; Sport és Tudomány (1956-1964)[Sport and Science]; TF Tudományos Közlemények, later Kalokagathia (1959)[Scientific Publications]; Testnevelés- és Sportegészségügyi Szemle, later Sportorvosi Szemle (1960)[Sport medicine review]; Sportélet (1965)[Sporting life]; A Sport és Testnevelés Időszerű Kérdései (1969-1982)[Current issues in sport and physical education]; Testnevelés Tanítása (1965)[Teaching physical education]; Mesteredző (1991)[Master coach]; Sporttudomány (1998)[Sport science]; Magyar Edző (1998)[Hungarian Coach etc.-

Sport science meets all classification criteria of the academic system of sciences as

- it has its own focus;
- it uses its own scientific research methods. Significant overlaps and adaptations can be found with the methods of other scientific fields;
- it has its own termonology and conceptual framework;

[7] Harsányi L. (1998)

[8] Istvánfi Cs. (2000)

- it developed and formulated its own theories;
- it has its own institutions.

We can state that sport science is a field in continuous development, the phases of which are described by Istvánfi (2000) in the followings:

- empirical phase
- disciplinary phase
- interdisciplinary phase

Empirical phase (1930-)

Recognition, interpretation and analysis of the cause and effect relationships within acitivities concerning the culture of the body and specific sports, as well as organising all this knowledge in specialised books and handbooks. Eg. Endre Kerezsi (1953): *Torna*[[Gymnastics]; Árpád Csanádi (1955): *Labdarúgás* [Football]

Understanding the decisive inner features of sport activities and endeavours to highlight specific characteristics. Development of subdisciplines also starts (seperating from the "mother-science"). E.g. Alfonz Kereszty (1954): *Élettan, sportélettan* [Physiology, Sport physiology]; Mihály Nemessuri (1960): *Sportanatómia* [Sport anatomy]

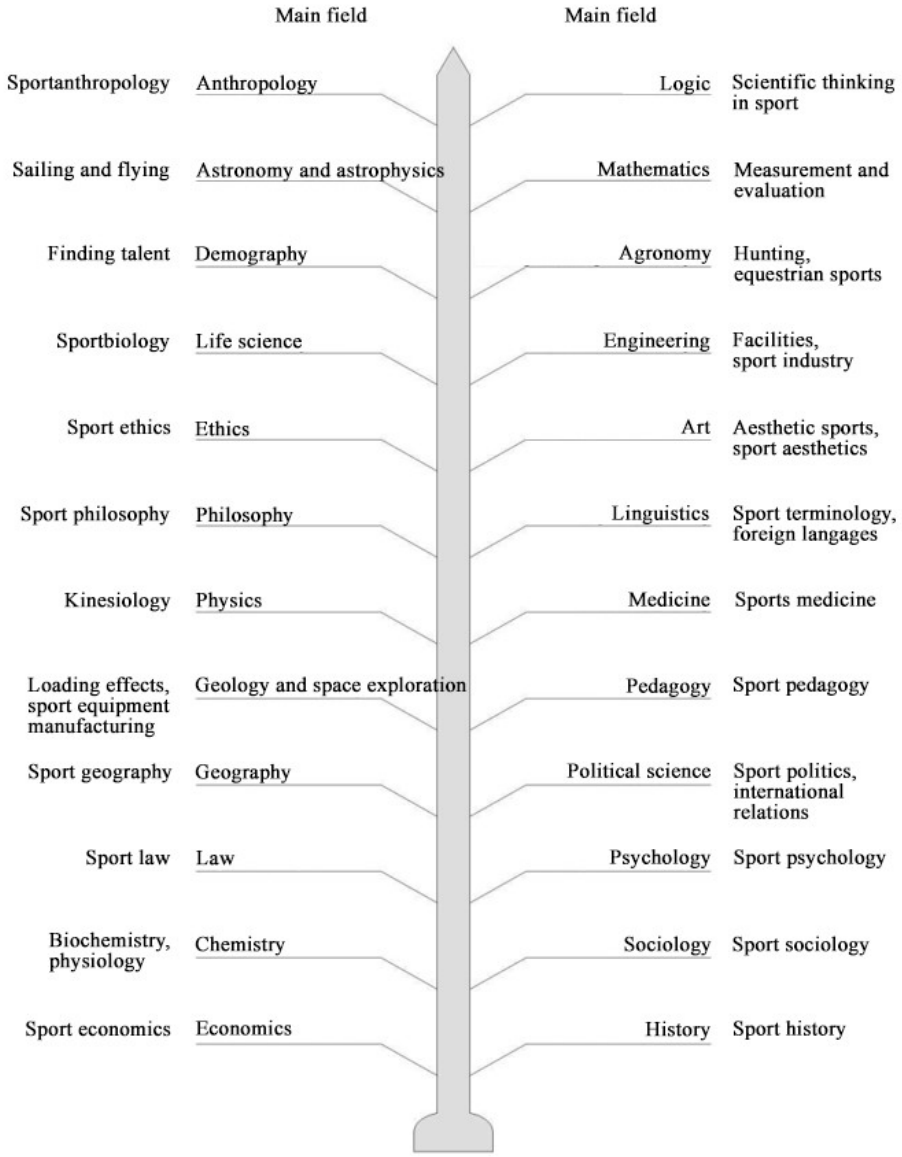


Figure 1/2. The relationship of sport science and the main scientific fields

Source: Zsolnai (1996), Vass, 2005; Horváth- Prisztóka, 2005

Disciplinary phase (1960-)

In this phase the practical issues of physical education and sport are being examined based on scientific knowledge. The examinations are carried out independently most of the time, while cooperation occurs in small groups, although such cooperations needed no special organisational tasks to fulfill. . This phase may also be called adaptational phase, as this is a time when appropriate methods of other fields are being adapted, at the same time terminology is being created. Eg. Ferenc Hepp (1962): *Sportwörterbuch in sieben Sprachen*; László Nádori (1985): *Sportlexikon* I-II [Sport Encyclopedia]

Differentiation phase: as a result of the accentuating of the processes that originate in the empirical phase, the subdisciplines of sport science also occur in handbooks:

- pedagogy → sportpedagogy
- psychology → sportpsychology
- sociology → sportsociology
- recreation → sportrecreation
- management → sportmanagement
- statistics → sportstatistics

Integrational phase: finding solutions for practical problems concerning the culture of the body, to be able to connect methods of various fields into an interconnected whole. E.g. Jenő Koltai and László Nádori (1973): *Sportképességek fejlesztése* [Development of Sport Skills]; József Czirják (1956): *Testneveléselmélet* ([heory of Physical Education]

Interdisciplinary phase (1990-

Scientific problems are being approached from several angles, with a multi-aspect character. This is the phase when well-organised team-work is born. The prerequisites of the interdisciplinary phase are:

a) Designating research strategy
b) Defining clear research concept and well-defined tasks.
c) Harmony of cooperating subdisciplines.
d) Harmony of participants.
e) High quality of the research results' presentation.
f) High standards and critical approach of users.

The result of the interdisciplinary team-work is the theory conception, which is a generalised system of reality. In sum, we may state that in the center of sport science we find the human being carrying out sports, the examination of which is carried out according to several aspects, from the point of view of ecological and social factors.

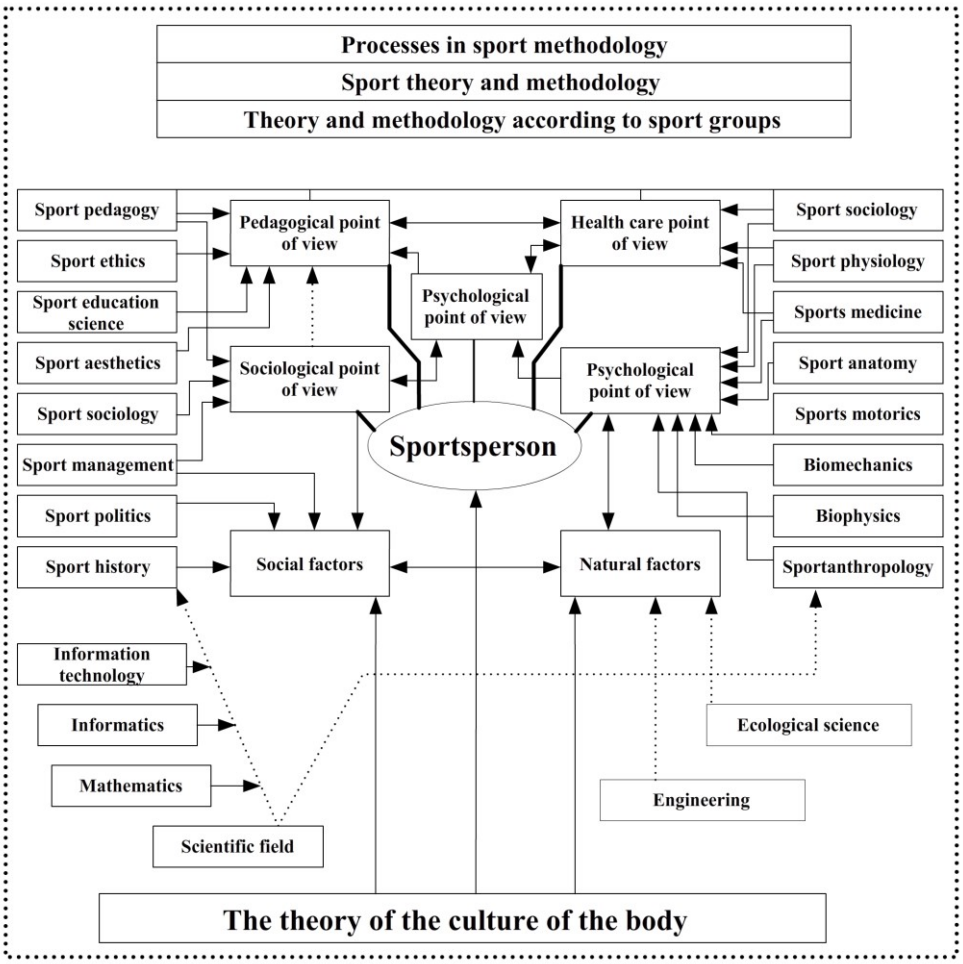


Figure 1/3. Complex definition of sport science

Source: Vass, 2005; Horváth-Prisztóka, 2005

1.2. Research in sport science

Sport science was introduced to Hungary through the works of Ferenc Hepp as among the courses he taught „Basics of scientific research" was also included at TF's predecessor institution in 1941.

„It is an important and particular step that the obligatory scientific study of physical education was introduced into the curriculum of the Hungarian University of Physical Education at the academic year of 1947-48. The course „Theory of physical education" in the 1st and 2nd year included the fundamental knowledge of scientific work, to be continued by the course „Scientific research" in the 3rd and 4th year. Within this latter course, students prepared their scientific thesis under the supervision of a mentor teacher, which was an obligatory part of the specialised final exam for teachers." (Nádori, 1971: 37). This is the date when the education on fundamental theories and methodology of sport science

has taken widespread recognition, inducing the establishment of several research programs at the field of sport science.

Research in sport sciences aims to explore principles and aquire and examine new knowledge, during which numerous methods are employed. Even today, experience should not be divided from goal-oriented and measurable scientific research. . Research in sport science must not be separated frompractical experience!

It is their specific features that differentiate science – and sport science – from other forms of aquiring knowledge of the world. These are:

- generalisability,
- repeatability,
- provability,
- coherence,
- analitical approach,
- simplicity (compactness, elegance)
- importance (usefulness),
- depth (new results can be related to several others) (Csermely, Gergely, Koltay, Tóth, 1999.).

We believe that the basis of research in sport science is observation, which is further examined by proven knowledge and methods.

Due to the nature of sport science research, we may differentiate three types of it, although they may not be strictly separated in all cases. For example, applicability (projects) is becoming an important aspect of basic researche, as well.

The three types of scientific research:

1. basic research (theoretical, acquiring knowledge),
2. applied (practical) research,
3. developmental research.

Basic research (theoretical, acquiring knowledge):

Basic reserach focuses on understanding certain phenomena or principles of the world. Generally speaking, they are theoretical, aim to aquire new knowledge, have no prerequisites of practical application, but form the basis of further scientific work. This type of research is rare in sport science, as sport science mostly carries out research of which immediate, practically applicable results are expected.

Applied (practical) research:

The applied (practical) research examines the possibility of practical usage of certain results produced by basic research, focusing on the practical application. If we look at changes in the circulatory system under stress (basic research), then we would be looking at the practical issues of for example speed development within an applied research. Apparently, we search for practical opportunities to use the results of the theoretical science.

Developmental research:

Most research activity is being carried out upon an assignment, where the client expects applicable (profitable) solutions (project plans, processes, products), thus the aim is to meet direct practical needs. We may state that developmental research further improves the results and principles produced by applied research.

Generally speaking, all three types of research include both quantitative and qualitative analyses, the two types do not separate strictly. *Qualitative analysis* is a non-numerical examination, with the aim of understanding and mapping the meaning of existing relationships – which is mainly characterizing sport history research. *Quantitative analysis* is a numerical examination, serving as a base to explain meanings.

1.2.1. Basic model of scientific research

Scientific activities can be divided in time into three distinct parts (Figure 4.): *theory, operationalisation and selecting the system of tools (observation)* (Babbie 2000).

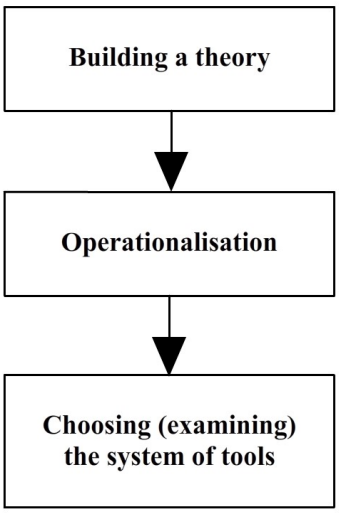


Figure 1/4. Basic model of science

Source: edited by the author, based on Babbie

As the figure shows, the first step is building a theory, which means that something raises the researcher's interest in the topic. There are two types of theory building: it can be *deductive* or *inductive.*

During *deductive theory building* we build assumptions and hypotheses using general basic principles. Our starting point is that in football, penalty kicks can only be performed badly – we may assume, that if the goalkeeper keeps the ball, than it was not kicked properly.

During *inductive theory building* we define general theories and principles through exact observations. In basketball, we looked at match records and visual tools to examine the scores of centers. Based on these, we can state that they are dangerous under the plank, but rarely try throwing the ball from a distance.

Falus (2000) explains three inductive research strategies that rely on each other:

- *descriptive*, when we look at the process only from the outside. This is usually applied when we do not know reality very well, or we do not know it at all;
- the strategy of *exploring relationships*, which is non-intrusive, too when we look at the relationship of processes with each other. Most often we already know the results of the descriptive startegy;
- *experimental*, when the researches intervenes with the process, changes the factors of the process on purpose, to be able to further explain results of the strategy of exploring relationships.

The next step after theory building is operationalisation, in which we plan the steps, processes and actions of the required variable's examinatinon. This could also be manifested as a research plan.

The final step is the choice of research tools (observation). We only supply a list of them here for now, as we will look at them in detail in the followings:

- observation,
- examination,
- experiment,
- interview,
- survey,
- sociometrics,
- etc.

2. STRUCTURE AND PROCESS OF RESEARCH IN SPORT SCIENCE, BASED ON THE RESEARCH PLAN

> *„Sucess doesn't only require energy and persistence, but also calmness, quick and solid orientation, observation and balance. All these are needed to aquire and then keep results, as whatever we reach in sports will quickly vanish should we not keep guard. Thus sport plants the seed of intellect and moral into the human soul."(Pierre de Coubertin)*

Planning the structure and procedure of the research work is a highly important task, as without the proper level of preparation, the result of the research becomes questionable. An indispensable and necessary step in the planning the particular actions is to define the time-frame. It is advisable to define the time-frame of each step during the preparation of the research plan. This depends on several variables (eg. sample size, data collection method, data analysis method, the persons conducting the interviews, etc.), but of course the time required for a given research also depends on whether it is a *cross-sectional* or *longitudinal* examination.

Cross-sectional studies are based on examinations that represent one particular point in time (a specific date). The researcher conducting a national research to examine the agressivity of professional male waterpolo players will probably look at only one cross-section in time.

Longitudinal research is the one that is being carried out in the course of a longer time span. Babbie (2000) distinguished between three important types of longitudinal studies:

- *Trend research* looks at processes in time within larger populations. E.g.: changes in the number of professional sportsmen at cencus surveys.
- *Cohort studies* look at smaller samples and detect changes of the same sample through time. In most cases cohorts are formulated according to age groups. E.g. 14-16 year old female tennis players in 2000.
- *Panel studies* mean longitudinal research gathering data at different times of the same sample (panel), from the same people. Thus from time to time it looks at the same sample taken from the whole society or population.

The most common type of research is the cross-sectional type, as they are cheaper and faster,however, longitudinal examination depict changes in time more precisely.

The suggested process of the research plan is the following (Falus, 2000):

1. Choice of the research topic.
2. Analysing the literature on the topic.
3. Formulating the main hypotheses.
4. Choice of research methods and tools applicable for the justification or rejection of the hypotheses.
5. Selecting the sample to be examined.
6. Completing the examination.
7. Data analysis and formulation of general findings.
8. Publication of research results.

2.1. Choice of the research topic

All scientific research starts with a question, the result of which is the development of interest. Curiousity is raised in people with interests, as they analyse things around them from various perspectives. One of the most common questions in sport science is „how can I develop my performance and push my limits?" Such questions originate from the interest, most often determined by the research environment. The reason for BSc/Msc students' thesis research is the necessity (the diploma can not be acquired without a thesis). The topic often depends on the environment of the person preparing the thesis. During topic choice it should be defined what and why will be examined. The expected result and (if there is) the practical applicability of the research also has to be mentioned.

„The most important things at topic choice are the theoretical and practical background and knowledge of the field, which will enable the student to navigate around important and insignificant phenomena, as well as to recognise real problems" Gyetvai-Kecskemétiné, 1997).

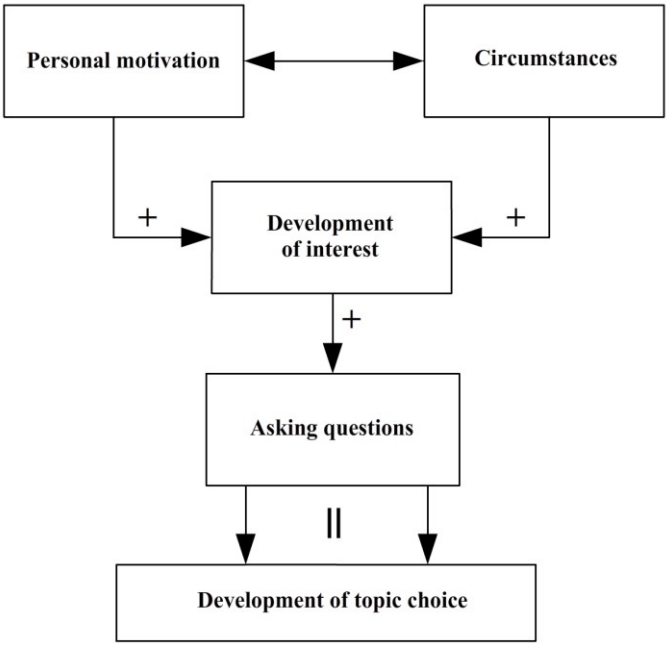


Figure 2/1. The process of development of topic choice

Source: edited by the authors

2.2. Analysing the literature on the topic.

As in case of all scientific research, the collection and understanding of already published literature is highly important in sport science research as well. Unfortunately we often encounter research and articles examining a topic well-examined by international and Hungarian scientists, still their results are not mentioned. It often happens that experts try to work on hypotheses that could be easily answered after a thorough literature review. Technically, they put a lot of effort into finding results that have already been explored and published by *valid*[8] and *reliable*[9] methods. Thus, to avoid such work, we must explore the relevant literature on the given topic and we have to understand the published and accepted results by similar researches.

Any kind of systematic data collection can be considered as a research. Generally we make the first step when there isa specific topic or the background of a topic that we wish to understand. One can remember only a limited amount of information about a particular topic, thus the first step of the research is to carry out a literature review. Although we

[8] Validity: shows how well our method or tool examines the penomena or concept that we were set out to examine.

[9] Reliability: a feature of our research method or tool that shows whether we get the same results if we repeat the examination.

might think that we have a unique and revolutionary idea, it is advisable to explore the field first, as it might be possible that the topic has already raised the interest of others as well. In most cases the literature review, the understanding of the research topic, and the review of the most important articles might take years, which is influenced by the aim of the research. Theses, articles, not to mention dissertations all have different criteria. It is important to select the relevant information from the large scope of available materials. The process of literature reviews have significantly changed in the age of information technology, as acquiring information is much faster now.

„*Literature review generally means that we perform high-level, critical analysis of the state of available knowledge on a well-defined topic , and we create synthesis concerning the topic.*" (Falus, 2004)

The following criteria should be considered during the literature review of a chosen topic:

1. **The scientific level of the resource must meet the level of interest.** The aim of the research, the reliability of expected results and the place of publication must be considered in all cases. A literature review must emphasise the works that meet scientific requirements, but other sources of information should also be looked at. Trainers, teachers and students with a practical problem often believe that popular magazines provide appropriate information for their question. However, it must be understood that while these resources usually do not meet the relevant scientific requirements, they may still contain important and useful information.
2. **The reliability of the source must be decided.** One should aim at interpreting the most recent scientific results, although it cannot mean that new research and results are necessarily better than older data. If there is an extensive scope of literature available on a given topic, we suggest to start the selection process with the abstracts of the articles – this will save a lot of time and effort.
3. **The appropriate key term should be defined, that connects the topic with the field of interest.** This means the mapping of the key words and expressions. Their exact definition is highly important, as they will again save us time and energy. The incorrect choice will have just the opposite effect, possibly leading to information overload.
4. **Similar views, problems and concepts should also be considered.** Young researchers and students often make the mistake of shifting the point of view when a new problem arises, or when they choose to apply a certain resource. Keeing the appropriate balance requires a lot of experience.

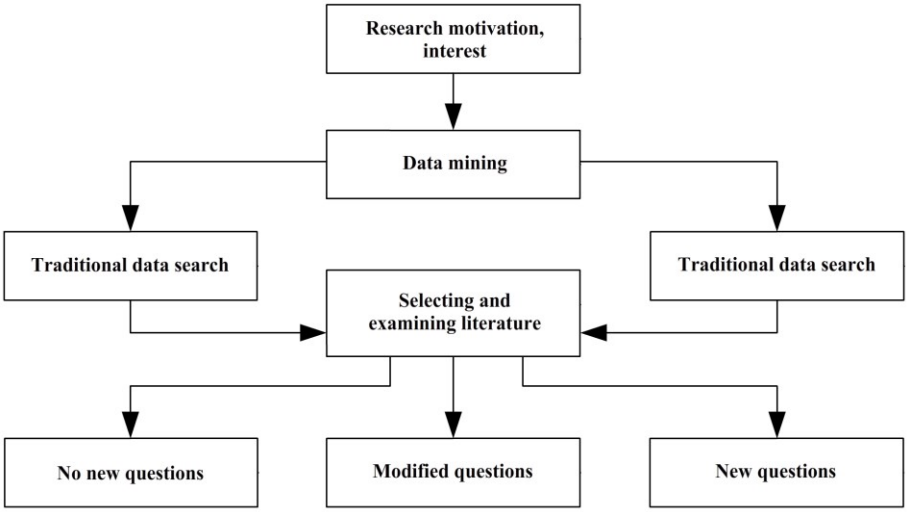


Figure 2/2. Process of literature review

Source: edited by the authors

We look at the process of literature reviews according to Figure 2/2. The first step is to define research motivation (interest), which is most often indicated by the environment. This may be grouped into **outer-** (e.g. it is obligatory or expected) and **inner incentives** (e.g. we are led by our own interest). Nowadays the data mining techniques have two clearly separable parts: *traditional (printed)* and *internet-based* searches. Both techniques belong to the secondary data search group, as their aim is to find and use knowledge that is already available and published. The most important locations for *traditional data mining* are libraries and archives. Archives are generally used during historical researches. The existence of particular libraries can also be considered a part of the independent institutional system of sport sciences (eg. the library of University of Pécs, Institute of Physical Education and Sport Science). These libraries keep the handbooks, course books, course materials, research reports, conference publications, journals, articles, theses, daily papers and other publications of sport sciences (e.g. publications in Hungarian: *Iskolai Testnevelés és Sport, Testnevelés, Testneveléstudomány, TF Tudományos Közlemények, Kalokagathia, Testnevelés- és Sportegészségügyi Szemle, Sportorvosi Szemle, A Sport és Testnevelés Időszerű Kérdései, Mesteredző, Sporttudomány, Magyar Edző, Asztalitenisz*, etc.).

Computers help literature reviews as they provide direct access to databases. There is no particular location to be mentioned at this data-mining technique, as the Internet can be accessed from anywhere. The following sources may be used to access Hungarian literature in sport science: www. sporttudomany.hu; www.magyaredzo.hu; www.threef.hu; www.nupi.hu; www.hupe.hu; www.leistungssport.net. Although there are a lot of

advantages to online data mining (eg. it is fast and cheap), we must emphasise that most of the knowledge available there is published without reviews, which makes the credibility of a lot of data questionable.

After completing literature review with the traditional (printed) data mining method, there is a need to compress data, which can be achieved by preparing notes. With this step we aim to downsize the number of data to minimal, while ensuring clear references to the original work throughout. Depending on the amount of literature we plan to process, we may distinguish between *keyword highlighting (extracting)* and the *card-based method.*

We deal with a fairly small amount of literature during *keyword higlighting*, and our aim is to prepare for a one-time occasion (e.g. a test or presentation). Here we higlight the most important ideas and key sentences, of which we make a list following the process.

We use the *card-based method* when we would like to preserve information from multiple sources. In this case we do not only note the keyword, but the specific idea itself. When we interpret longer ideas or quotes, we must precisely define the data of the source as well, so that we can reference it anytime, so it does not qualify as partial or complete plagiarism.

It should be understood that literature review may result in new questions and it may modify the previous ones, or in case of finding satisfactory answers, no further questions need to be raised.

2.2.1. Preparation of the reference list

The unambiguous aim of the bibliography and reference list is for resources to be:

- identifiable,
- and traceable.

The preparation of a bibliography and references is a necessary task which is often carried out inaccurately. We describe the process of preparing the reference list based on the manuscript of László Harsányi.[10]

All methods, thoughts, figures and tables are considered to be a reference that originate from another researcher. In the scientific world everybody aims to protect his own intellectual property, thus plagiarism is one of the most serious ethical offences.

Plagiarism: the partial or complete expropriation of intellectual property, by publishing it under one's own name. Nowadays there are several methods in use to point out plagiarism. There are large databases prepared (where articles, handbooks, theses and dissertations are

[10] Harsányi L. (2007)

uploaded), where new texts can be compared to already published ones. If there are matching parts and no references are given, then it provides ground for suspicion. For this reason, most higher educational institutions require theses to be submitted not only in a printed form but also in an electronic format (CD disk).

Similarly to many other products, intellectual property is also defined by law. In Hungary, the Act No. LXXVI of 1999 defines copyright issues of products of literature, science and art. Furthermore, this act disposes over free usage and other limits of copyright, which all must be taken into consideration.

ACT NO. LXXVI of 1999 ON COPYRIGHT

1. § (1) This Act shall provide protection for literary, scientific and art creations.

THE FREE USE OF THE WORK AND OTHER LIMITATIONS TO THE COPYRIGHT

General Provisions

33. § (1) Uses falling within the scope of the free use shall not be subject to the payment of remuneration and to any authorization of the author. Only works made available to the public may be used freely in accordance with the provisions of this Act.

33. § (4) For purposes of the provisions of this Chapter the use shall be taken to serve the purposes of illustration of teaching if it is implemented in accordance with the requirements of education and with the curriculum used in kindergarten, primary and secondary school, industrial school, vocational school education, the primary education of arts, as well as in higher education falling within the scope of the act on higher education.

The cases of free use

34. § (1) From a disclosed work any part may be cited by indication of the source and naming the author indicated as such. Such citation shall be true to the original and its scope shall be justified by the nature and purpose of the borrowing work.

35. § (4) In a manner and to the extent complying with the intended purpose as well as for internal use in an institution, if this is outside the scope of commercial activity, a copy may be made for own purposes if it is not designed for earning or increasing income even in an indirect way and

a) it is required for scientific research,
b) it is made from an own copy for the files to be used for scientific purpose or for the supply of a public library, or
c) it is made of a limited part of a published work or of an article in a newspaper or periodical.

35. § (5) Specific parts of a work published as a book as well as newspaper and periodical articles may be reproduced for educational purposes in a number corresponding to the number of pupils in a class or for purposes of exam in public and higher education in a number necessary for the said purpose.

35. § (6) The temporary reproduction of a work done with the exclusive purpose to permit the realization of the use of the work authorized by the author or permitted pursuant to the provisions of this Act shall be taken to fall within the scope of free use on the condition that the temporary reproduction is an integral part of the technological process aiming to achieve the said use and lacking any economic significance of its own.[11]

There is a different standard regulating different types of documents, which are nevertheless often mixed. The basic measurement of bibliographies and reference lists is the *bibliographic item.*

Bibliographic items in case of books are made up of the following data: full name of the authors (family- and first name), title of the volume (and subtitle if available), number of publication, place of publication, publisher, year of publication, number of pages. We may often seen a format where the year of publication comes after the name of the author.

Rétsági, Erzsébet: Kézikönyv a testnevelés tanításához (5-8. osztály). Budapest-Pécs, Dialóg Campus Kiadó, 2005. 352. o.

or

[11] 1999. LXXVI. tv.

Rétsági E. (2005): Kézikönyv a testnevelés tanításához (5-8. osztály). Budapest- Pécs, Dialóg Campus Kiadó, p. 352.

Hepp Ferenc: A mozgásérzékelés kísérleti vizsgálata sportolókon. Pszichológia a gyakorlatban 22. kötet. Budapest, Akadémiai Kiadó, 1973.

Eco, U. (1991): Hogyan írjunk szakdolgozatot? Budapest, Gondolat, p. 256.

Reference: The format of reference shall always be indicated at the end of the quotation. There are two acceptable formats:

1. The most popular is when the name of the author(s) and the year of publication is indicated. E.g. (Nyerges, 1981). No attention should be paid to the numbering of the reference list.
2. Just a number, e.g. (12), or the name of the author(s) and an ordinal number, e.g: Dubecz (12). In this case we should make sure that the number at the reference matches the number of the publication listed in the reference list.
3.

Examples:

- Berkes (2007)
- Ormai (1981, p. 126.)

When we insert quotations into the text, quotation marks are used only in case of word-for-word quotations.

Bibliographic items in case of articles are formulated according to the following list: author's name, title of the article (equation-mark is acceptable, but rare), title of journal, volume, year, month (or publication number), page number. The publication year often comes immediately after the name of the author in brackets here as well. If there are more than one articles referenced from the same author, then we use letters from the alphabet to distinguish.

Katics L.- Petrekanits M.- Derzsy B.- Gedő D.- Römer I. (1992): Szabad-, versenyaerobic és akrobatikus gyakorlatok hatásai. Mester-edző, 4. 12-20. o.

Pintér J. – Rappai G. (2001): A mintavételi tervek készítésének néhány gyakorlati megfontolása. Marketing & Menedzsment 2001/4. 4-11. o.

Harsányi L. (1992 a): A blokkszerű terhelés az éves felkészülésben. Testnevelés- és Sporttudomány 3. 109- 119. p.

Harsányi L. (1992 b): Die „Geheimnisse" der ungarischen Leistungen. Leistungsport, 6. 27-29. o.

Reference:

- Katics et al. (1992) or (Katics et al., 1992)
- Harsányi (1992 a)
- Harsányi (1992 b)

There were numerous early authors in Hungarian training studies (Vadas, 1927; Abád, 1962; Kutas, 1962; és Nádori, 1962, 1968, 1972).

Referening several works of the same author:
Preparation process in sport was divided by Nádori (1962, 1972, 1981), Harsányi (1992 ab) and Platonov (1987, 1999) as well.

Bibliographies may also be prepared in case of book chapters, published conference presentations or collections, and they may be used as references too, listing data in the following order: author's name, title, title of the publication, author(s) of the publication, number of the publication, place of publication, publisher, year of publication, page numbers of the referenced part.

Ács P. (2006): The analysis of the regional competitiveness of the Hungarian sport with multivariable statistical methods, In Hughes M. (editor): World Congress of Performance Analysis of Sports 7, Berzsenyi Dániel College, 299. – 309. o.
Reference

- Ács (2006)

In case of *weekly and daily papers*, we list data in the following order: name of the author, title, name of the paper, date of publishing, page number.

Kulcsár Gy. (2008):A kapitány visszaszúr. Nemzeti Sport, szeptember 24. 20. o.

Reference:

Kulcsár (2008)

We consider manuscripts, research reports, theses, dissertations, treatises, standards and patents as „other works". In these cases the order of data is the following: name of author, title, subtitle, type of work, name of higher education institution if availabe in brackets, place and year of publication. In case of standards: number of standard, year of introduction, title of standard. In case of patents: name of the patent's owner, title of invention, name of inventors, name of country, number of patent, year of introduction.

Rétsági E. (1996): Törekvések az iskolai testnevelés és a testnevelő tanárképzés megújítására. Kandidátusi értekezés. Magyar Tudományos Akadémia. Budapest. 186. p.

Mikroelektronikai Vállalat (Budapest). Kettős színhatású folyadékkristályos kijelző. Bence Gy- Seyfried É- Véghelyi T. HU 182-495. 1986. 06. 30

Ács P. (2007): A sportolói vándorlás, „Migráció – társadalmi összefüggések" című konferencia, Budapesti Corvinus Egyetem Szociológiai és Társadalompolitikai Intézete, valamint a Magyar Statisztikai Társaság Demográfiai szakosztálya. Budapest, 2007. október 19.

Reference:

- Rétsági (1996) wrote that…
- (SZAB)-HU 182-495 (1986)
- Ács (2007) mentioned in a presenetation that…

Computerised (online) data is also obligatory to be referenced nowadays, even if the authenticity and future searchability of the used data can not be guaranteed. That is why the date of download must be indicated after the name of author, year, title, and webpage's url.

Soós I. (2002): A sportpedagógia, mint prevenciós eszköz a fiatalok egészségnevelésében. Kalokagathia. 2002. 1-2, 130- 135. o. http:// http://www.hupe.hu/info/kg/cikkek/2002_1-2.pdf (2008. 07. 17)

Reference: same as in case of the other information carriers.

As a formal detail regarding the specifics of referencing, we must highlight that we never indicate the scientific degree with the authors' names. The reference list must contain all works that was referenced within the text, but must not contain any work that was not

dierctly referenced by the author in the given text. If a title was written by multiple authors, all their names should be indicated in the reference list, in the appropriate order. However, in the text an *et al* is enough after the first author's name. Foreign authors are also indicated by their family name, followed by the initial of the first name. The reference list must contain articles in the strict alphabetical order of the first authors' family names. Works from the same author(s) should be organised according to publication's year, and works published in the same year must also be distinguished. Some works lack place of publication (n.p. – no place), year of publication (n.y. – no year) or publisher (n.publ. – no publisher). There are several ways to indicate page numbers as well, as in Hungarian page numbers may be referenced using the English-format 'p' (pagina) as well. The indication „p. 146" or „146 p" means that the book has altogether 146 pages. The indication „45-49 p" or „p 45-49" means that the referenced part lasts from page 45 to page 49. When writing in Hungarian, we suggest using the Hungarian format ('o' for *oldal*), after the page number. The most important aim that we should keep in mind when preparing the reference list is the source's ability to be identified and searched in the future. It helps preparing the reference list if it is compiled continuounsly, and gets longer during the writing process. If the reference list is relatively long, the following grouping can be applied:

- *books in Hungarian*
- *articles in Hungarian*
- *other sources in Hungarian*
- *books in a foreign language*
- *articles in a foreign language*
- *online resources.*

Nowadays traditional literature reviews are outnumberedby online literature searches. Several sources are available for online literature reviews:

- *Catalogues of electronic libraries*
- *Electronic bibliographies*
- *Online databases*
- *Electronic books, journals and articles*
- *Internet-based search engines*

Other supplementary services can also be found in the field of sport sciences as well, which may help in literature searches:

- *Newspaper of the Hungarian Society of Sport Science (www.sporttudomány.hu)*
- *Website of the Hungarian Association fro University- and College Sport (www.mefs.hu)*
- *Newsletter of the Hungarian Association of Table Tennis (www.moatsz.hu)*

Detailed information about the use of electronic search engines and databases can be found in the book 'Data Analysis in Practice' (Ács, 2015). In this textbook we only describe the electronic 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 in 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 with 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 only or full text. Also, the user may define the time interval, collection or type of media to be searched (picture, text, audio or video) too (Figure 2/1).

Display of results

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

The first two forms show the relevant titles, 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 2/2.) (Source: Annamária Karamánné Pakai, András Oláh, 2015)

Access

The Digital Library of the University of Physical Education can be accessed via http://tf.hu/oktatas/konyvtar/tf-digitalis-konyvtar/digitalis-dokumentumok/

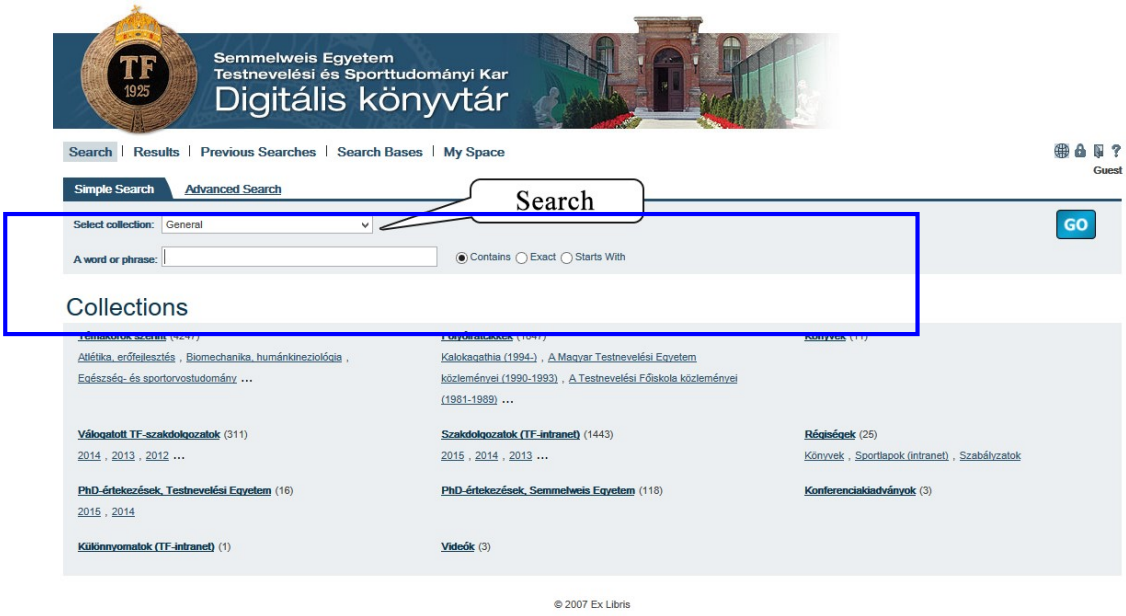


Figure 2/1. Search interface

Figure 2/2. Results list

2.3. Formulating the main research hypotheses

Generally speaking, we call hypothesis all assumptions on the nature of a given phenomenon originated in theory or practice (observation). Therefore, these are statements defining the major assumption regarding the variables and their relationships in the research. The main research hypotheses must include the expected final result of the research.

Most often the questions that come up during topic search will already lead us to formulate certain scientific assumptions. Literature review also supports this, as we must aim at collecting the most possible data already during the preparation (literature research) phase. Researchers believe that it is by means of scientific assumptions (hypotheses), that we are able to orient, interpret and predict events in the world. Basic assumptions and thougths are usually based on *reliable theories*, which *serve as guidelines*. Thus a major research assumption (hypothesis) should consist of *presumptions, definitions and arguments*, all interconnected on a logical basis, thus forming the methodological basis of a conceptual system. A reliable theory shall meet the following criteria:

- It must include logical statements, which make up a coherent unit and its events are provable.
- It should incorporate explanation and the concept, starting from which the research plan can be derived, examined and proved or rejected.

In light of all this we suggest to prepare the research plan after defining the main research hypotheses (Figure 2/3), as they will significantly influence the plan.

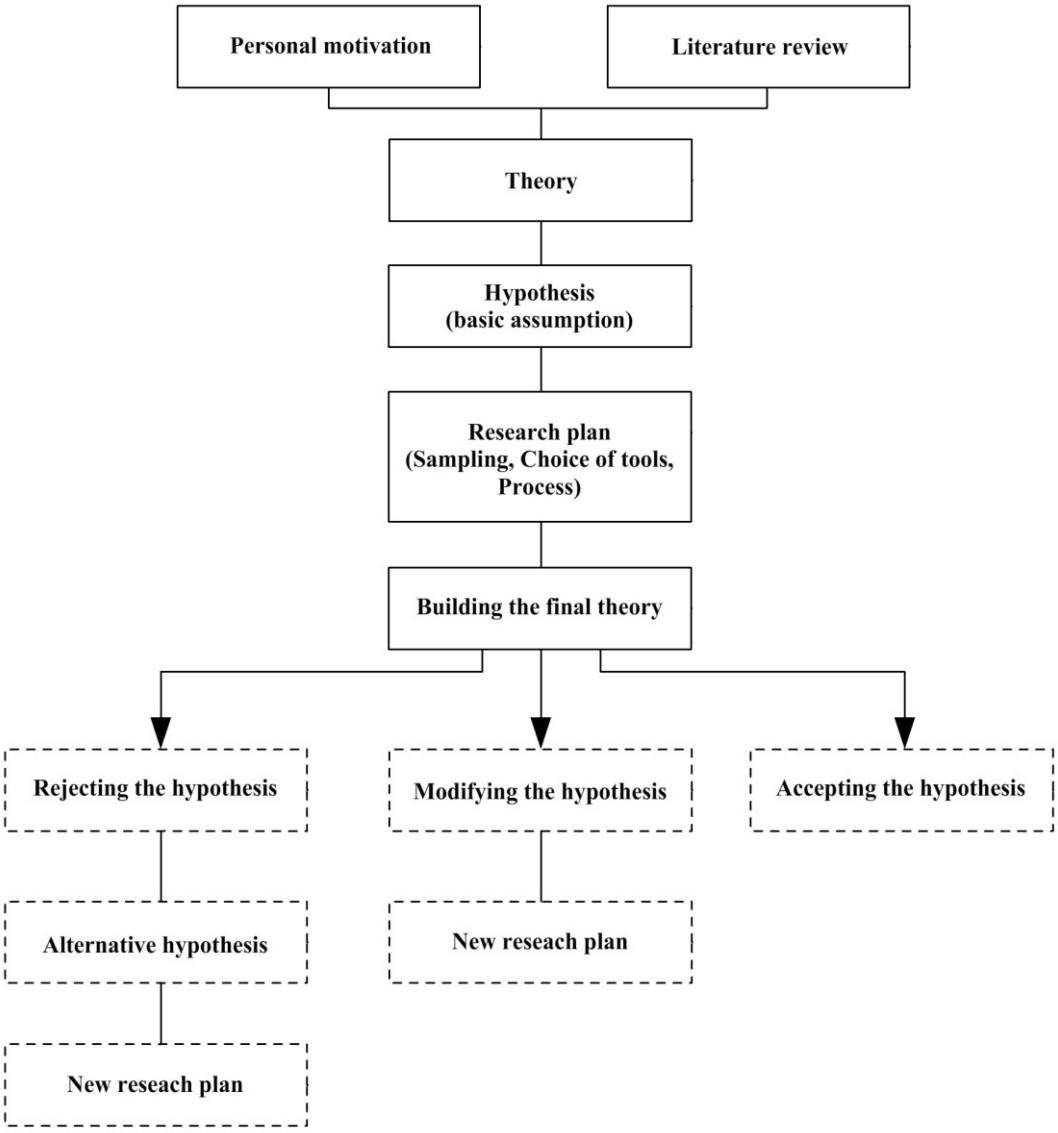


Figure 2/3. Schedule of defining research hypotheses

Source: edited by the authors based on G. Tenenbaum és M. Driscoll (2005)

The requirements of a good research hypothesis were defined by Falus (2000) according to the following aspects:

1. Hypotheses should be explanatory, which means that the suggested assumption and relationship must be imaginable for all.
2. A good hypothesis highlights the relationship between variables. E.g. in basketball, by the increase shooting preparation time, the scoring got better.
3. The hypothesis should be unambiguously proven or rejected. Continuing with the previous example, the hypothesis is supported by the match record of the given team.
4. The methods, techniques, processes, calculations and measurements used during the proof or rejection of the hypothees must prove to be possible to

carry out. The preparation time is measured by a stopwatch (in minutes), and the scoring is measured by the number of scores.

5. Hypothesis should be defined using unambiguous, clear, and operative terminology. We often see researches containing hypotheses with assumptions referring solely to professional sportsmen. E.g.: professional sportsmen smoke less. Such a formulation is incorrect as the term 'professional sportsmen' must be clearly defined.
6. The hypothesis should be based on existing knowledge on our part. We usually formulate our assumptions so that we end up with true answers. Yet this does not mean that with a solid theoretical basis we should not examine statements that others believe to be true. We are especially likely to encounter such occasions when examining a new sporting tool or technique.
7. The starting assumptions must be paraphrased in the most understandable, easy and compact manner. We may formulate subhypotheses if we face more complex problems.
8. The hypotheses altogether must provide an answer for the reearch problem that was selected during the topic choice.

The number of main assumptions (hypotheses) we examine during our research depends on the aim and circumstances of the research as well. It is definitely not useful if this number is too high. E.g.: it is not very beneficial for a thesis to examine more than five hypotheses, as it might prove difficult to test more, with respect to the standards of good quality and further requirements.

2.4. Choosing the research methods and tools to ensure the verification or rejection of the hypotheses

Scientific observations differ from everyday ones in being carried out goal-oriented, according to a plan, and on a regular basis.

Selecting the method of scientific research is always the privilege of the researcher. The scientific research method means the process that indicates how we carry out the collection, systematization, and storage of data, and what processes we use for data analysis.

Within research methods we differentiate between:

I. exploratory methods
II. processing methods

Exploratory methods usually apply qualitative techniques, and in the following we describe the most often used methods.

During **observation** we do not interfere with the events, but we examine the phenomena and their context within their natural environment. The success of the observation is determined by the preparation of the observer. Correct selection of the subject and person of observation is crucial. The effectiveness of the appropriate process is defined by awareness and planning. Researches in sport sciences often use tools to aid observatory methods, which should support perception and data collection. The most popular tools are: dictaphone, camera, photos, video, mirror, proceedings, etc. We often see sporstmen of individual sports record their own matches and movements on camera – this is what we call *introspection*. Usually the aim of this is to analyse the movements afterwards – this is what we call *retrospection.* The observation of others (objective observation) happens mostly before matches, to support strategy-building.

Advantages: it can present full value in itself, being a direct method ensures its flexibility, and it does not only show the result of education but also its process.

Disadvantages: it requires a lot of time and organisational effort, involves the possibility of subjective mistakes, empathy of the observer, possibility of observational mistakes due to perception.

In case of **examination** the researcher deliberately creates the prerequisites for the phenomenon he wants to observe and analyse (the advantage is that the phenomenon can be repeated and controlled). Probably the most reliable data in case of sport science researches are provided by experiments. We may distinguish between laboratory- (e.g. doping laboratory) and natural-environmentexaminations. We may distinguish between several methods: the most direct is the conversation (exploration), and the most standardised are the tests (e.g. the IQ-test), as most of them are standardised.

The aim of the **survey** method is to collect information within a research project by asking questions. Surveys can be used to explore knowledge, opinion, attitude, experience, patterns or lifestyle of individuals or groups. Basic types of the method are *oral* and *written* surveys.

During *oral surveys* there is a personal interaction between the interviewer and the interviewee. It may be *individual* or *group-based*, depending on the number of participants.

The most often used method during *individual surveys* are the *in-depth interviews*, the *narrative interviews* and the *exploration*.

In-depth interview: this variety of interviews provide a tool to explore the most intimate things. Generally it takes the form of a long conversation with famous sportsmen or leaders, where the interviewee may describe experiences which he never talked about before. This type of interview holds great freedom and responsibility alike, both for the interviewer and the interviewee.

Narrative interview: where the interviewee describes only those events that had happened to him and became his own experiences.

During *exploration, personal and guided survey* we apply a set of ready-made set of questions that were developed based on theoretical and practical knowledge, matched with specific questions relevant to the particular research. The leading researcher gets into direct contact with the research participant. Using the survey we may try to explore a particular relationship or prove a regularity. In this case it is a prerequisite to ensure representativity (as we are using a sample), reliability and validity.

Representativity means that the distribution of the sample's given feature matches that of the population, thus persons included in the survey represent the total sum of the group we wish to examine in terms of explorable features that are relevant for the research, which can be examined and demonstrated by statistical methods. This way it is facilitated that the results derived from the sample will provide similar features and results in terms of the population as well. From all sampling methods the probability methods enhance representativity.

Group interviews are useful to understand the „collective knowledge", explore the opinion of the group, and to demonstrate individual points within the group (team).

Naturally, group interviews may be challenging for the interviewer as a prerequisite is that he can find his place within the setting and find the appropriate comunicationalcommincational style. The most popular type of group interviews is the focus group interview.

Focus group researches are the most popular type of qualitative researches. Their aim is to explore feelings and attitudes of the invited subjects. During the examination there are 8-12 persons – selected by an appropriate test questionnaire – who talk about their previous experiences, share their opinion feelings and beliefes, all focusing on a given topic. The conversation is led by a moderator who is also the coordinator of the event. He has two tasks: to aid a conversation which enables balanced, free and open exchange of ideas

between the participants; and to make sure that they mention each important item of the research syllabus agreed by the client. All this must happen in an open atmosphere where respondents can become uninhibited. The conversation is recorded, thus not only the participating experts but also the client can receive the full material, which enables him to grasp personal insight on his target group's behaviour. The aim of the focus group research is to provide general background information on a given topic; examine potential problems during the launch of a new program, product or service; getting new, creative ideas; and to understand the main strengths and weaknesses of new ideas.

This method is usually applied prior to the launch of a new product or service, e.g. before the opening of a sport centre or a wellness hotel.

The process of the focus group research is depicted by Figure 2/4.

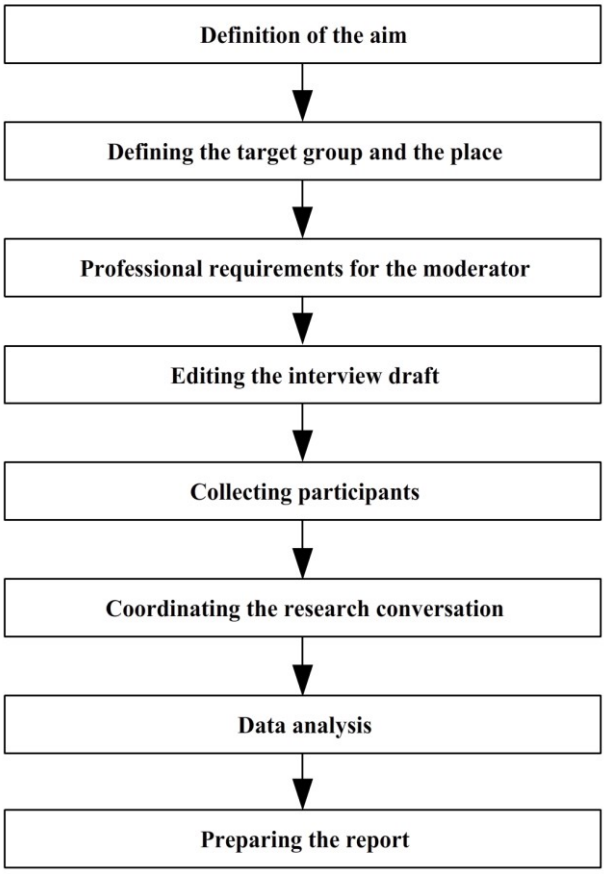


Figure 2/4. Process of the focus group research

Source: edited by the authors

Testing examinations are considered as basic examination nowadays at the field of sport sciences, as they are mostly examine the body (e.g. height, weight), skills (e.g. evaulation of forehand flip execution on a 1-to-5 scale), and competences (e.g. Cooper test). The

most important is the correct choice of tests and the criteria of qualitative measurement and evaluation.

The personal presence of the researcher is not necessary at the application of **written surveys**, as the aim is to gather information on tasks transmitted in written forms (e.g. surveys distributed through traditional letters or online). An important item of written surveys is that it has to be sent to the respondent with a formal explanatory letter, as this will substitute the interviewer and will have to motivate participants to fill out the questionnaire. It has to contain basic information, such as the name of the company that carries out the research, and it also has to contain a guideline for completing if the questionnaire is rather complex. The success of the survey – the return rate – highly depends on the accompanying letter. Small presents or vouchers are often used to motivate respondents to return the questinonaire.

Nowadays the most popular method is *surveys with questionnaires*, as it has several types, it is quick, economical, enables researchers to collect a large amount of data, and is also fairly objective. The method asks a sample with pre-set features, and uses a „tool" or person to carry out the data collection.

The questionnaire is a document containing a mix of questions aiming to gain information. The prerequisites of the method of using questionnaires are the following:

1. The topic must be suitable for a survey.
2. The aim of the survey must be set clearly. This method can only be used consciously if we know exactly what we want to ask and why, as this is the basis for the compilation of questions.
3. The competence of respondents must be ensured. The researcher must decide if respondents or participating groups are able to answer the questions.
4. The correct selection of the representative sample must be ensured, so that the composition of the sample matches that of the population, and consequently the answers given by members of the sample will be generalisable for the whole of the population.
5. The survey should be prepared appropriately, as this method requires a lot of time and effort. Thus a good plan for the smooth execution of the survey is a must. It is important to have the appropriate group of colleagues who will carry out interviews and data processing.
6. Results must be available for processing. Most surveys use large samples, with the possibility of numerical conversion, even in the case of open questions.
7. The required knowledge and software should be available for the data processing.

What we get with the survey are answers for pre-set questions that are available for processing, and that come from a sample selected based on various methods. There are several forms of surveys, thus it has to be chosen according to the research aim.
A survey can be completed *in writing, orally, through the phone, or online*. The topic, the target group, and the depth of the research vary from one examination to the other, thus questionnaires are accordingly different as well. There are some rules which must be obeyed while others are only suggestions – this develops gradually by each researcher through the years of experience.

Written questionnaires are a typical quantitative method. Their most common form is the questionnaire sent through post, where respondents fill out the survey without the presence, help or influence of the interviewer in their homes, and return the completed survey via post to a pre-defined address.
In these cases it is advised to attach an envelope for returning the questionnaire that is free to send and contains the address of the company responsible for the research. Nowadays a popular option is to attach questionnaires in newspapers or journals. It is cheap and fast but less effective, even if the completion of the survey is motivated by small presents. A proven method is to use short, easy questionnaires in shops or at exhibitions that contain questions in connection with the topic of the location.
Advantages: it can be properly used in case of a large sample. It is quick to distribute, enough time is available for filling, the respondent can be more honest because the interviewer does not influence the respondent, and it is low-cost.
Disadvantages: low return rate (8-25%), thus the representation is not satisfactory. Special reminding actions are required to make respondents remember. A relatively small amount of questions can be asked. A large number of the questionnaires come back with mistakes and misunderstandings are common (8-10% of the returned questionnaires), which is because the interviewer's absence.
The **„oral questionnaire"** is the most popular, as it is more in-depth, more precise and ensures appropriate answers according to the required representation. In this case the questionnaire and the professional guiding has an important role, but the most important factor in attaining good results is the interviewer.

Types:

1. traditional form (PP „paper and pencil"), which can be carried out at home or in a specifically designated place (eg. in a rented room of a store)
2. CAPI (Computer Assisted Personal Interviews), a new method worldwide, when interviewers work with laptops.

Advantages: 100% return rate, due to the presence of the interviewer. Controlling of the answers happen while the questions are asked, and misunderstandings can be clarified.
Disadvantages: Threats of the personal contact, as the personality of the interviewer may influence the respondent positively or negatively. The respondents may not confess to their ignorance about certain topics, or they may give untrue answers.
Remuneration of interviewers is given for each questionnaire, thus their reliability can only be ensured with frequent, unexpected checks.
Phone-based survey is a popular method in countries with good quality phone lines. The anwers are being recorded by the interviewer over the phone. In many cases these answers are recorded directly on computers, thus data processing is quick, and the results are available in only a few minutes.
Advantages: Quick, relatively cheap, it is similar to personal surveys, with an added technical tool. The voice and style of the interviewer is important.
Disadvantages: it can only be applied at set times by families, thus it is more popular in case of institutions and shops. It is often obtrusive. Not everybody has a phone.
Computer-based surveys are being carried out with the help of an online questionnaire, using a common network. Representativity is hard to ensure, as not everybody has access to the Internet. However, it is expected to be the most popular method in the future.
Advantage: fast, cheap, convenient, continuously spreads.
Disadvantage: respondents may not reply honestly.
During questionnaire-editing, the aim and expected answers shall be defined precisely once again. The **guidelines at questionnaire-editing** are the following:

1. Use simple questions that are easy to answer. A good idea is to apply closed-ended questions, which can be quickly replied orally, or can be underlined in writing.
2. Ask only a few questions. If the questionnaire is too long, and questions are getting more and more complex, then the answers are becoming increasingly less punctual.
3. Ask shortly, simply and unambiguously, and apply various types of questions. The applied words and terminology must be understandable for everyone. We can

only make an exception when we are asking a unique, homogenous group – in this case we might use professional terminology.

4. All possible answers must be indicated. If we are asking open-ended questions, we shall provide enough space for the respondent to write up his answer.
5. Questions must not influence the respondent.
6. We must emphasize – at the top of the questionnaire or the accompanying letter – that responding is voluntary. This should also be mentioned during personal surveys. Anonymity of the respondent must be respected.
7. Answers must be useful for numerical transformation. The answers must be summarised by themselves or combined with other questions – e.g. in case of direct computer recording – thus it is useful to apply coding squares to make numerisation easier.
8. Giving answers must be easy – by underlining, numbering or giving just one or two words.
9. The structure of the questionnaire must be logical. Not only the wording but also the order of questions is important. Transmitting parts are required between important and not so important sections. The main topic shall come at the 2/3 of the questionnaire. Objective and routine-like data and statistics shall come at the end, eg. questions regarding wage.
10. The questionnaire must raise interest. A well-edited accompanying letter is also important.

The content- and format-based editing are the next steps in editing, alongside with the order of question numbering and grouping according to aims, through which the questionnaire can become logically better structured.

The *introductory questions* can be found at the beginning of the questionnaire. They are simple as their aim is to prepare respondents for the topic. Then *transition questions* help shifting from one logical part to the next. *Checking questions* connect criteria that are logically related (creating factors). These questions aid future processing and analysis. The most important group is the *topic-related questions*, which can be checked by the checking questions as well. The latter is also available to make sure whether the respondent gave reliable answers or just wanted to get over with responding.

There are two main types of questions in surveys: closed-ended and open-ended questions. **Closed-ended questions** are the ones where all possible answers are already given. It may provide two or more outputs, may be an importance-scale, semantic differentiating scale, or Likert-scale.

In case of *two-output closed-ended* questions the number of options can only be two. These questions mostly include the ones that ask the gender of the respondent, or that provide simple yes/no options. It is a popular type of question as the coding (0,1) is rather simple.

e.g.: Do you like the football team FTC? Underline your answer!

Yes No

Multiple-output closed-ended questions are the same as two-output ones, except for providing more answer options. Its threat is that the options must be well-chosen – no possible options should be left out. A lot of attention shold be paid to this type of questions during testing. The advantage of this type lies in the simple coding.

eg.: Where do you live? (Mark with an X!)

County centre ☐
Town ☐
Village ☐

The *importance scale* includes the levels of importance of a chosen feature, and rankings can also be provided here. A disadvantage is that no reasoning is given for the ranking, and it may be more difficult to answer.

eg.: How important is the your sport equipment for you during matches?

Very important ☐
Important ☐
Not so important ☐
Not important at all ☐

Rank the following brands from 1 to 5! (1 is the worst and 5 is the best)

Puma ☐
Reebok ☐
Adidas ☐

Nike ☐

Asics ☐

Evaluation scales evaluate a given feature from poor to exceptional. Its advantage is that it provides a picture about the strength of feelings, and the disadvantage is if the scale is incorrectly given then the options may not always be understood.

eg.: How much do you like to swim? (1– hate to swim; 7– very much love to swim) ☐

Semantic differentiating scales can be used to measure consumer attitudes, as there are completely opposite words or sentences (statements) given at the two ends of the scale. Opposite-based scales can be given graphically, at the respondent may indicate the direction and intensity of his feelings. (The respondent may choose freely from the values represented by the scale.) The main idea was that features manifesting in opposites (such as beautiful-ugly, good-bad, old-young, cold-hot) are generally understandable. Semantic scales provide a comprehensive picture on sconsumers' (groups') opinion on a given product, advertisement or service. As given opposites are used to measure attitudes and opinions, it is higly important for them to be relevant regarding the examined product, service, etc. The most popular scale-type has seven levels.

e.g.: Please rate the "X" shoe you tested on the following scale:

cheap						expensive
1	2	3	4	5	6	7

poor quality						excellent quality
1	2	3	4	5	6	7

Likert-scale is basically a specific type of semantic differentiating scales. It is also known as "agreement scale" and is the most often used. In this case the respondent must evaluate a statement-list in connection with a given topic, and he should measure how much he agrees with the given statement. This scale-type is used most often for lifestyle researches, market segmentation, company layout evaluation, etc. The scale may have even or odd number of levels, although the most common are the 5- or 7-level scales. More information on these scales can be found in the study of Kehl and Rappai (2006).

eg.: Please mark the number that reflects your opinion! (1– don't agree at all, 5– completely agree)

It is an advantage for a sport shoe to be comfortable.

1	2	3	4	5

It is an advantage for a sport shoe to be light.

1	2	3	4	5

or

Mark the answer closest to your feelings.

	I strongly agree	*I agree*	*I'm neutral*	*I disagree*	*I strongly disagree*
My days are interesting	☐	☐	☐	☐	☐
I feel lonely	☐	☐	☐	☐	☐
I usually enjoy what I do	☐	☐	☐	☐	☐
I feel loved and needed	☐	☐	☐	☐	☐
I am able to relax	☐	☐	☐	☐	☐
I have plans and aims about my future	☐	☐	☐	☐	☐

Another group of questions in a survey are the *open-ended questions.* We can differentiate the following types: *completely open-ended questions, associative questions, sentence fill-in questions, picture-completion test questions, and thematic test questions.*

Completely open-ended questions: the respondent has complete freedom in formulating his answer.

e.g.: What is you opinion on doping in sport?

Associative questions: the respondent has to say out loud the word that comes into his mind immediately after a given word is said.

e.g.: What comes into your mind when you hear the word "doping"?

Sentence fill-in questions: The respondent have to continue a given sentence.
eg.: I would never go on dopes because…
Story completion question: The respondent have to continue a given story.
eg.: I went to a world championship and in the changing-room I saw that one sportsman used doping chemicals. This immediately made me think of…
Picture-completion test questions: The respondent is given a picture of two persons. One states something, and the respondent has to put the opinion of the other person at the empty space.
Thematic Apperception Test (TAT): The respondent is asked to describe what he can see in the picture he was given.

The following guidelines should be followed when collecting questions for a survey:

1. ask logical questions
2. ask specific questions
3. use full sentences
4. avoid abbreviations
5. avoid slang
6. avoid professional terminology
7. avoid stereotypical terms
8. do not ask double questions
9. do not ask negative questions

Questinonaires must always be tested before using, and they should be *modified* (if needed) then *finalised*, as there might be questions which are ambiguous or unclear. Supervision helps ensuring that the questionnaire can collect information on opinions, wills, motivations, or stereotypes. It should be checked according to the research plan whether the survey contains all questions that are relevant for the examination. The questionnaire can only be finalised after multiple checking, and the interviews can only start afterwards.

In summary we can state that the preparation of questionnaires contains three main steps:

1. Preparatory step: the specific aim and the expected results of the research has to be defined, and all relevant information for the editing should be collected.

2. The main part is the questionnaire's editing, in light of the order of the used questions:

- start with the general part to avoid confusing the respondent at the beginning with a difficult question
- compile the introduction in a thorough and polite manner that fits the topic, as it will determine further steps
- the questionnaire's structure must contain transitions from one part (topic) to the other
- statements should follow each other so that it fits the logic of the respondent and is clear
- the most problematic questions should come at the end
- personal data of the respondent (segmenting criteria) should be collected either at te beginning or the end – if at the end, and these answers are not given, then the rest of the data may be used.

3. **Closing part, pilot testing, finalising, copying**

The other group of methods is the **exploratory-evaluative methods**, where the information gathered by exploratory methods are evaluated according to two main aspects. The qualitative evaluation examines the content and features of the results through precise recording and categorization based on content. The quantitative evaluation describes quantitative features of the results through descriptive and inferential statistics.
During *qualitative evaluation* the results are rarely numerical and cannot be measured. Qualitative evaluation is useful when we would like to explore motives of different behaviours and their features.
Quantitative evaluations are based on the assumption that even human attitude and behaviour can be measured and presented numerically, and the retrived results may be analysed and evaluated by statistical methods.

2.5. Definition of the research sample (based on Pintér, Rappai, Herman, Rédei)

Individuals included in the scientific research can be counted and examined with statistical methods. If the examination is being carried out according to a predetermined aspect, then we make a complete observation (a classic complete observation is the census). If the observation includes only certain individuals of the population, then we carry out a partial observation. Representative observations have a crucial role in partial observations. In case of partial (and especially representative) observation we do not have information about all individuals of the whole population, thus we try to make approximative statements reflecting the population based on the sample[12]. We may easily understand that the inference will only be successful to the extent the chosen sample fits the population. The abovementioned theories require two further concepts to be explained: **success** and suggestivity.

The science of statistics measures the success of the inference with two categories (which are contradictory, as we will find out later):

- the "visible" sign of success is **reliability**, which means that the statement or value based on the sample proves to be correct in many cases;
- the other criteria for success is **accuracy**, which refers to the expectation that the information based on the sample is indeed informative, and the values that form the estimation's result vary within a small range.

This textbook does not describe in detail which features are required for a sample to be considered representative – those who are interested can find further information in the study of Pintér and Rappai (2001).

We may rightfully ask to examine the basic population (e.g. the professional sportsmen), as it would lead to the most precise results. The most common answer to this question is that it would be rather time-consuming and costly, and we would also face legal limitations. We must emphasise that such basic statistical data is highly incomplete in the field of sports (e.g. there is no data on the number of sportsmen), thus it is almost impossible to ensure representativity.

We must always calculate with statistical errors during sampling, which are added up of sampling- and nonsampling errors. Sampling errors are rooted in the fact that we examine

[12] Pintér- Ács (2006)

only a part of the population instead of the whole. Nonsampling errors are usually due to data collection mistakes (answering mistakes, executional mistakes, processing mistakes, etc).

Sampling methods may be grouped according to several aspects, but Babbie (2000) distinguishes between probability and non-probability sampling methods.

Probability sampling – as in calculating probability – is a general name for sampling with random choice methods. E.g. Simple random, stratified sampling.

In case of *non-probability sampling* we do not follow the rules of probability sampling. Eg.: quota sampling, snowball sampling.

There are several methods for representative sampling, differing in their execution or depending on the randomness of the sampling. The following figure shows one grouping of the sampling methods.

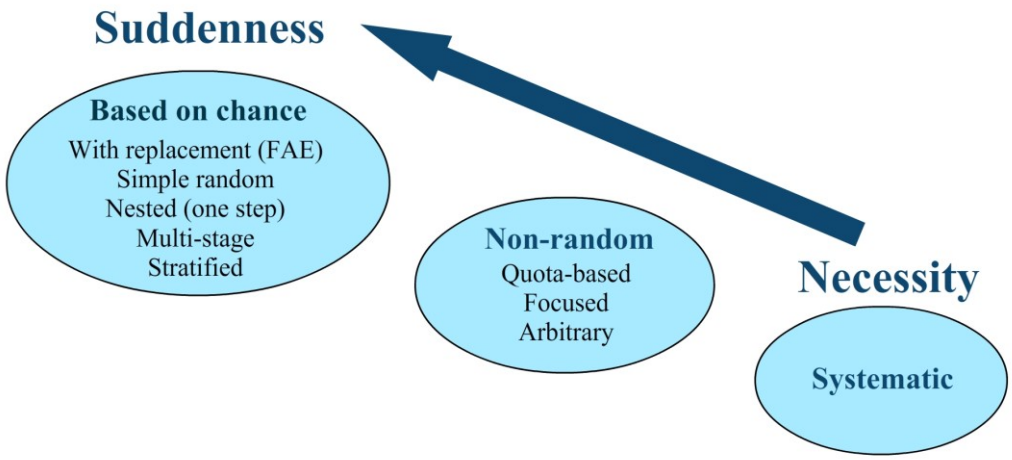


Figure 2/5. The most common representative sampling methods

Source: Pintér- Rappai: *Statisztika* (2007)

The figure above groups sampling methods according to the level of randomness. However, as we go 'down and right' through the figure, we feel like including certain individuals in the sample is more and more necessary. Let us look at the most commonly used sampling methods.

Simple random sampling is the most common method. In this case we choose pieces of the sample independently, by the same probability. Each member of the population has the same chance to be included in the sample. First of all, we need the list of the complete population, then we pick names randomly, or assemble the required size of sample using a random number table. One type of this sampling is the one *without replacement*, where one item can not be included in the sample twice. E.g. the lottery.

Sampling with replacement results in a sample where all items were returned to the pool before an item is selected. To execute the method first we need a list of all members of the population, from which we randomely choose the first piece of the sample. Examples for this method are hard to find in practice, but this is what happens during an oral exam when the lecturer returns each topic to the pool after each presentation, thus it may occur that one topic is discussed more than once.

In case of **systematic sampling** we pick for example every fifth member of the population to be included in the sample (this is the sampling interval: the distance between the items). To avoid certain bias in the sample we may choose the first number randomly, in this case from the first of 5 items, and then each fifth would be included in the sample. The sampling ratio is the ratio of the items included in the sample to the population; here: 1/5. The threat of the method occurs when the sampling frame consists of groups according to some sort of ranking, or is periodical. E.g.: basketball groups are being asked to participate, and the players are listed in the same order in each team according to their roles. If we choose every fifth placer, then our sample will include players of only one role, say centers.

Stratified sampling is a modification of the simple- and the systematic sampling, which may further enhance the representatitivity of the sample. It enables equal inclusion of items from the population from homogenous subgroups that we wish to examine.

The simplest – although not the only one – method of stratified sampling is when we choose a sample size based on a stratification feature from each group of the population, according to a predetermined ratio, with simple random sampling. The stratification depends on what information we have about the population. E.g. if we examine handballplayers, and the number of scores they make at each match, then we can make prior stratification based on roles, asroles may determine the number of scored goals.

Nested sampling aims to avoid the negative feature of simple random sampling when the listing of the population's members is problematic. Thus in this case we concur to practice as we do not need information about the actual members to be examined , only about certain important features (criteria or grouping) of them. Based on these clearly understood features we create **primary sampling units**, and pick the items of the sample from this pool. Then we examine each item of the thus – randomly – assembled sample (without having knowledge of their existence prior to the sampling). E.g. during a state-funded research to examine level of education among sportspeople we assemble a questionnaire that we plan to get filled out by 5% of all Hungarian sportspeople. We know that in 2007 there were 2 280 000 sportspersons in Hungary, although we do not know them by name.

All we know is that in the same year there were altogether 2780 registered sport teams. Thus we execute sampling using the nested method: in the first step, using simple random sampling we chose 139 sport teams, and we get the survey completed with all of their sportspeople. During the sampling we will only need to have a list of the primary sampling units (ie. the teams) and not about the sportspeople themselves.

Cluster sampling means that we sample the groups of items, and in the next step we sample again within these groups. Thus we prepare a list and apply the sampling multiple times, applying selection.

Quota sampling is quite similar to stratified sampling, as we apply auxiliary information in this case as well. The content of the domestic areas are defined according to residents and grouped by locations, and a quota (list) of required respondents is being assembled according to the most important features. The researcher who carries out the sampling – in the knowledge of the quota – must examine all individuals who meet the designated features as long as the predetermined quota is fulfilled.

Concentrated sampling further decreases the role of chance and enhances the responsibility of the researcher who carries out the sampling. In this case individuals are weighted, and only those are included who are considered to be opinion leaders.

In the case of **arbitrary (expert) sampling** the researcher can define which individuals he includes in the sample. The only limitation to influence the decision is the required number of the sample size. E.g. I would like to ask hardcore supporters about something, thus I tend to include respondents according to their physical features of being muscular and bald.

Sampling of easily available materials: the researcher examines only those items or individuals which are easily available. Representative data can rarely be produced this way, so this method is only useful if the focus of our interest actually coincides with those people whom we select. E.g. the opinion of 100 supporters queuing before a match, whom we meet at the cashier (non-representative sample regarding the whole scope of visitors).

Snowball sampling is useful in case of populations that are difficult to contact, with the sample growing through the network of participants.

2.6. Execution of the research

In the previous chapters we discussed in detail the execution and planning of the scientific research, thus here we would like to present another commonly used practical method, the SWOT-analysis.

The **SWOT-analysis** is one of the most popular methods applied at the field of sport sciences. It is used mostly for the evaluation or problem-solving in case of institutions or organisations, but could also be applied at several other domains to evaluate a starting position.

Strength
Weaknesses
Opportunities
Threats

These four words mark four aspects according to which all available information can be grouped. Applying this method, the researcher can evaluate the opinion of everybody involved in the operation; may evaluate the current situation; and may formulate statements about the opportunities or threats concerning the future.
Data-collection is usually carried out through a brainstorming session, but the method can also be used to sort out data acquired by surverys or interviews. It should be kept in mind that only those individuals should be included in the sample, who possess relevant information or experience in the research topic, and may provide useful information for the researcher. E.g. if we wish to examine the current state of Hungarian football, then we shall not carry out a SWOT-analysis based on information gathered form volleyball trainers.

Strengths and weaknesses also originate from the inner features of the research's subject, while the opportunities and threats originate in the environment.
Strengths: What do respondents consider good about the research's subject, what qualifies as its strength? The most common features included here are experience, knowledge, facility, tradition, good organisation, etc. What these features have in common is that they may be built upon in the future.
Weaknesses: These data formulate the domains where changes will have to be made. E.g. insufficient access, poor facilities, lacking sport equipments, etc.
Opportunities: competent respondents are necessary to define opportunities, as it requires proper understanding of the environment. E.g.: new gyms, need for new sport services, opening a new wellness center, etc. Mapping the opportunities may significantly influence future results.

Threats: similarly to opportunities this is an independent factor. If we stick to the previous example, e.g.: quick drop in the interest towards a new sport service.
It is useful to use a datasheet to organise information.

Strengths	**Weaknesses**
Opportunities	**Threats**

Table 2/1. Table of the SWOT-analysis

Completing the first two boxes of the SWOT-analysis is usually not very difficult. There are higher challenges in filling the boxes of opportunities and threats, as opportunities sometimes include features that should have been listed under strengths. The difference is whether we talk about an inner (e.g. organisational) or outer, environmental, but so far unused source. A similar issue may come up regarding threats and weaknesses. We should understand that weaknesses come from inner features.

Figure 2/6. Flowchart of the SWOT-anylsis
Source: edited by the authors

I. Preparatory phase:

a) Chosing the research domain
b) Deciding whose opinion we wish to explore with this method. E.g. trainers' opinion, official opinion of the team's management, etc.

c) The data-collection must be planned: time, location, number of participants, other circumstances
d) Participants must be invited, and informed about the aim of the data collection and its further usage
e) A specific time should be agreed
f) A place or room has to be set appropriately, and the required tools should be prepared

II. Information collection phase:

In the first step, the previously chosen participants – who meet all criteria – have to sit down individually or in groups. Next is that all participants are given a paper and a pen, and they may also receive the printed SWOT-table. Then the aim and method of the meeting is described verbally, alongside with the desctiption of how the table should be completed. After this the completion is being carried out gradually, first focusing on the easier aspects (strengths – weaknesses), followed by the opinions on the environment (opportunities and threats). The researcher should ask inspirational questions. The respondents should be asked to give honest answers and opinions and we should suggest to list a similar number of information in each box.

III. Evaluation phase

The evaluation phase starts with the summary of the data, which can be carried out immediately or at a later date. The latter is more time-consuming and participants will not be informed about the results.

Summarising on the spot may be carried out by listing al opinions on a board or other tool, in one big SWOT-table. It is a general rule that we shall aim for the most precise evaluative answers in case of all four boxes.

The easiest method is to weight the statements that come up more than once with their frequency, and the resulting suggestions may be prioritised accordingly.

2.7. Data analysis and formulating general statements

In this chapter we describe the basic statistical methods of processing and analysis of data collected during a sport science research. Nowadays no research data evaluation or analysis is being carried out without computers – and we must admit that it indeed helps the work – thus in this chapter, alongside some simple calculations, we discuss the most

common data analysis methods and output interpretations of the Excel and SPSS softwares.

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 textbook will discuss two subfields: estimations and testing hypotheses, which we will demonstrate both with the Excel and the SPSS software.

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.

2.7.1. Basic statistical concepts and scales

Statistical population is the sum of all the cases that formulate the subject of the statistical observation. The basic part of the population, the **element** is called *observatory item.* A concept that is strongly connected to statistical population is criterion.

Statistical criterion (variable) is the feature relevant for the elements of the statistical population. Possible outputs of the variables are the *variable versions*. Generally, variables can be *quantitative, qualitative, time-dependent* and *location-dependent.*

If a variable has only two versions then it is called **alternative** variable (e.g. gender has only two possible versions, male or female). **Quantitative variables** (measurement feature) describe variable versions by numbers, while the **qualitative variables** (qualificational feature) describe them by qualitative attirbutions, concepts or words. In most cases, research data is acquired through measurement. 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 for differentation. 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.

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 one football player is better than the other).**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 subject to interpretation. 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.

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.

For example, we might think of goal difference: everybody knows what a +3 goal difference means (the team scored 3 more goals than it received), and what a 0 means (the team scored 20 goals and received 20 as well). Also, a negative goal difference may exist as well (when the number of received goals is higher than the number of scored goals).

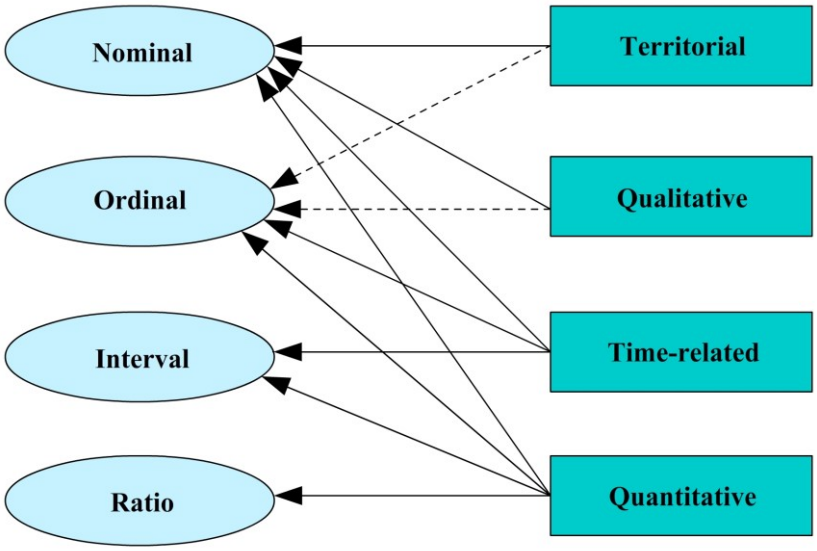


Figure 2/7. Relationship between variables and the measurement scales

Source: 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, p. 25.)

2.7.2. Descriptive statistical analysis

The most important domains of *descriptive statistics* are: collecting data, plotting data, grouping and clustering data, doing simple arithmetic operations with data, and displaying outputs. The aim is to describe the status of something.

In this chapter we look at the following concepts (univariate analysis):

- ratios
- information summary using measures of central tendency (arithmetic means, mode, median),
- dispersion and symmetry,
- tools of data presentation.

2.7.2.1. Ratios

Within this type of analysis we calculate the **ratio** of two related statistical data. The general definition is:

$$V = \frac{A}{B}$$

where: V – ratio

A – compared value

B – comparison basis

Ratios may fall into the following categories:

- *Partition coefficients* describe the relation between a given part and the whole, where we compare the sample size (frequency) of a partial population to the sample size of the full population (e.g. number of handball teams in Baranya county compared to the number of handball teams in Hungary).
- The *coordination ratio* compares two proportional data to each other (e.g. the number of junior players for each ten adult players in a team)
- *Intensity ratio* is the ratio of two different but related statistical data (e.g. gas consumption on 100 km)
- The *dynamic ratio* is a popular analytical tool for comparisons over time. The time-frame to which we compare is called *base period*, and the time-frame which we analyse is called *subject period.* Two types can be distinguished: *base-relative* and *chain-relative.*

In case of *base-relative*, the base period is a chosen, constant value, usually an important date or the first value of a time series.

If we consider the first date of the time series the base, then the formula of the base-relative is the following:

$$b_i = \frac{y_i}{y_0}$$

The base period of the *chain-relative* changes constantly as it is the value from the period before the subject period. The formula of the chain-relative is the following:

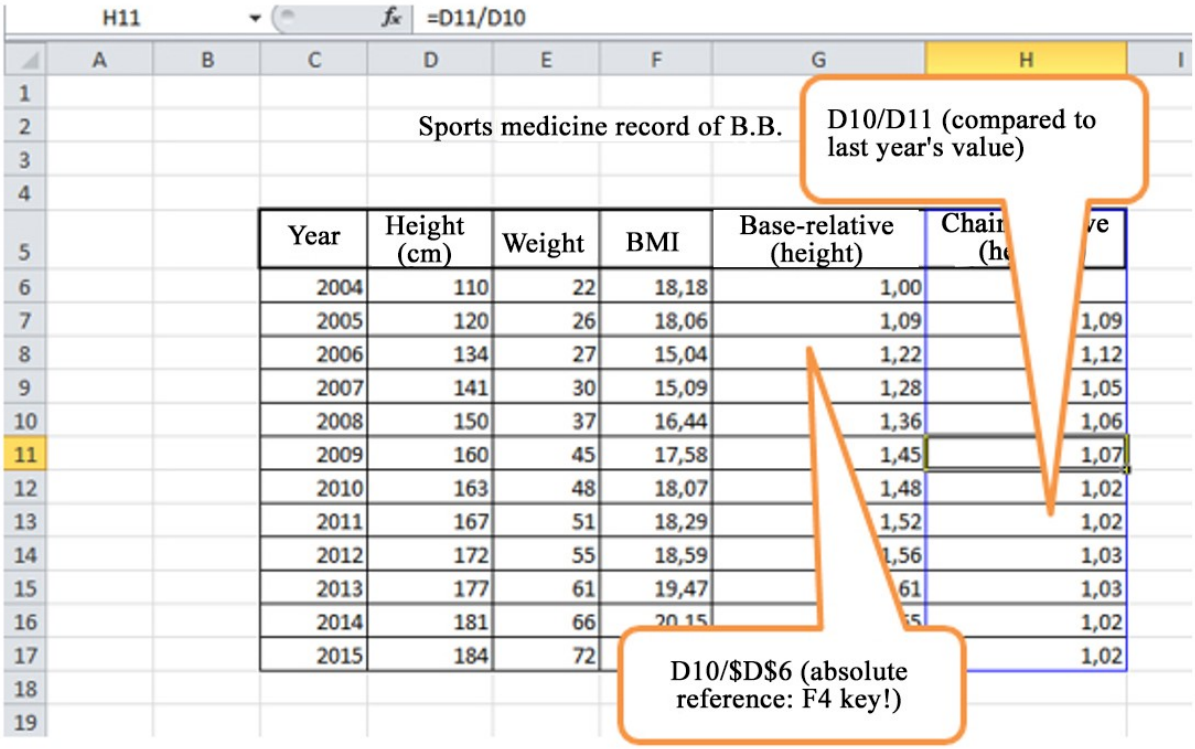
$$l_i = \frac{y_i}{y_{i-1}}$$

Using he following formula, we can count the chain-relative from the base-relative by a division, while the product of the chain-indices calculated by the end of any year gives the the concluding year's base-relative.

$$\frac{b_m}{b_{m-1}} = l_m$$

$$l_1 \times l_2 \times l_3 \times \ldots \times l_m = b_m$$

Calculate the base- and chain-relatives using the relationships. Here we represent the notable base- and chain-relatives using the Excel software.



Year	Height (cm)	Weight	BMI	Base-relative (height)	Chain (he	ve
2004	110	22	18,18	1,00		
2005	120	26	18,06	1,09		1,09
2006	134	27	15,04	1,22		1,12
2007	141	30	15,09	1,28		1,05
2008	150	37	16,44	1,36		1,06
2009	160	45	17,58	1,45		1,07
2010	163	48	18,07	1,48		1,02
2011	167	51	18,29	1,52		1,02
2012	172	55	18,59	1,56		1,03
2013	177	61	19,47	61		1,03
2014	181	66	20,15			1,02
2015	184	72				1,02

Figure 2/3. Base- and chain-relatives

Base-relatives (Figure 2/8) show the relative measurement of development, while chain-relatives (Figure 2/9) reflect the rhythm of the change (Forrás: ratios.xls; viszonyszámok.xls).

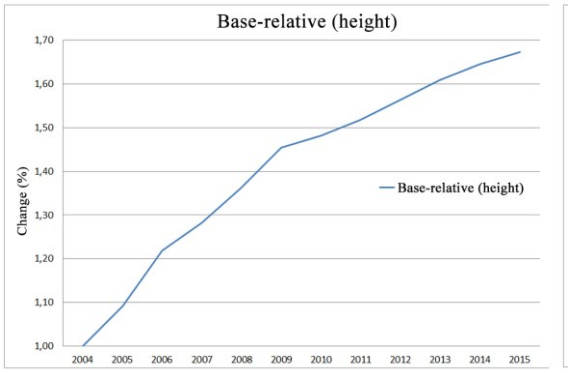


Figure 2/8. Base-relative

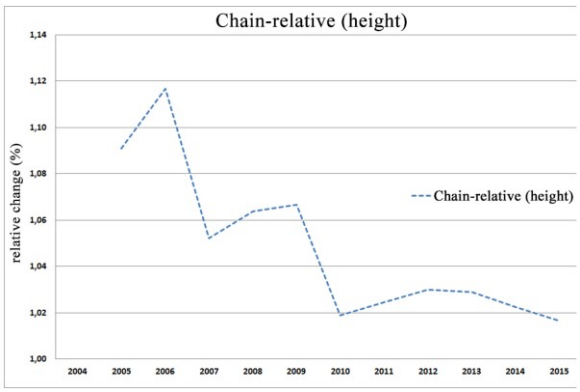


Figure 2/9. Chain-relative

2.7.2.2. Information summary using measures of central tendency (arithmetic mean, mode, median)

Overview and evaluation of a large amount of data may often be challenging. A primitive way of analysis may be the counting, ranking or summary of the data, but the demand may occur for major phenomena to be represented by a small amount of data. Thus there is a need for a numerical value that may be accepted as the common feature for all cases. The most important tool for this information summary is calculating the measures of central tendency. There are several values – which may not always be the same – that aim to explain common information about the cases. There are three statistical requirements of these indices: they should be *medial, robust,* and *typical.*

Two groups of mean values may be differentiated based on calculation (average, mean) and their place in the line of data (median, mode). Of all *calculated means (arithmetic-, geometric-, harmonic- and quadratic mean)* in this chapter we look only at the arithmetic mean – information on the rest may be found in the book of Jánosa (2005).

Furthermore, we discuss the *mode* and the *median* of the means that refer to their place in the line of data. First we illustrate these central tendency measures with simple frequency distributions, then with categorical frequency distribution.

We use our own primary database to present statistical methodologies (source: fitness 57_clarified data.xlsx; fittségi 57 tisztított adat.xlsx). The research included university students (N=57) and was carried out by the Hungarian School Sport Federation. It applied the NETFIT testing system (Kaj et al, *Kézikönyv a Nemzeti Egységes Tanulói Fittségi Teszt*

(NETFIT) alkalmazásához.[Handbook for the Application of the Hungarian National Student Fitness Test] 2014).

The database shows results of the following tests:

1. **Body composition and nutritional status**

We conducted body composition analysis, BMI-measurement (Body Mass Index – BMI) and examined body fat %, metabolism in rest and fisceral fat level using a bioimpedance analysis machine (OMRON BF 511).

2. **Paced curl-up test**

Test objective: abdominal muscle-activation measurement.

Equipent: test audiotape, audio player, gym map, measuring strip.

Starting position: The student lies in a supine position on the mat, knees bent at an angle of approximately 140°, feet flat on the floor, legs slightly apart, raising shoulderblades with arms straight and parallel to the trunk with palms of hands resting on the mat. The fingers are stretched out and the head is in contact with the mat. After partner A has assumed the correct position on the mat, partner B places a measuring strip on the mat under partner A's legs so that partner A's fingertips are just resting on the nearest edge of the measuring strip.

Task: Partner A has to perform as many correct curl-ups as possible in the rhythm set by the sound (1 curl in every 3 seconds) with bent knees, sliding fingers across the measuring strip until fingertips reach the other side, while the heels are continuously touching the ground. Partner A continues without pausing until he or she can no longer continue or has completed the maximum number of curl-ups, or mistakes the pace for the second time. Partner B counts the number of completed curl-ups and highlights mistakes.

Socoring: The score is the number of curl-ups performed.

3. **Trunklift test**

Test objective: Measurement of the power of muscles in the back.

Equipment: Yardstick with measure in cm, gym mat, marker point.

Number of trials: two

Starting position: The student being tested lies on the mat in a prone position (with face down). Toes are pointed and hands are placed under the thighs.

Task: The student lifts the upper body off the floor, in a very slow and controlled manner, focusing on a coin or marker right in front of her, thus the head is being held straight in line with the trunk. The position is held long enough to allow the tester to place the ruler

on the floor in front of the student and determine the distance from the floor to the student's chin. The exercise should be carried out in a controlled manner.

Scoring: The test should be repeated twice and the better result should be recorded, with precise cm data. (Distance above 30 cm should still be recorded as 30 cm.)

4. **Paced push-up test**

Test objective: Dynamic power measurement of the arm- and shoulder muscles.

Equipment: test audiotape, audio player

Starting position: The student being tested assumes a prone position on the mat with hands placed under or slightly wider than the shoulders, fingers stretched out, legs straight and slightly apart, and toes tucked under. (Knees may touch the ground until the audiotape starts).

Task: Student A performs the maximum number of push-ups he or she can in the pace given by the audiotape (1 push-up/3 secons), keeping the body straight and keeping the elbows bent in 90° at each movement. Partner A continues without pausing until he or she can no longer continue or has completed the maximum number of push-ups (86), or mistakes the pace for the second time. Partner B counts the number of completed push-ups and highlights mistakes.

Scoring: The score is the number of push-ups performed. The test may be administered with the help of a summary table.

5. **Grip strength measurement**

Test objective: Maximum strength measurement of lower arm muscles

Equipment: customisable hand dynamometer

Number of trials: two + two

Starting position: The student holds the dynamometer with the palm of his better hand, and lowers his arm so that the hand and the lower arm is at the same level.

Task: The student pushes the hand dynamometer with his better arm using maximum power and holds position for 2 seconds. The movement should be carried out with a straight wrist, and a well-paced, confident movement, without quick and pulling movements. Also, during the test the arm shall not be lifted and/or the measurement tool shall not be pulled close to the body. The test should be carried out twice with the better arm, with a small break between the two movements; and then it should be carried out again with the weaker arm.

Scoring: The test should be repeated twice and the better result should be recorded, with precise kg data.

6. **Standing long jump test**

Test objective: Measurement of the dynamic power of the leg.

Equipment: yardstick

Number of trials: two

Starting position: The student stands behind the marked line, knees are bent, arms are in front of the body, parallel to the ground, and toes touch the marked jumping line.

Task: The student pulls arms back and then forth to gain swing, then jumps, aiming to cover the farthest possible distance.

Scoring: The test should be repeated twice and the better result should be recorded, with precise cm data. The shortest distance of the arrival point of the knee closer to the marked line should be recorded.

7. **Flexibility test**

Test objective: Movement range examination of the joints and flexibility measurement of the popliteus muscle.

Equipment: measuring scale

Number of trials: two + two

Starting position: The student sits across the test apparatus with one knee bent keeping the foot on the floor, and keeping the other foot flat against the measuring apparatus (box).

Task: After three forward bends, the student reaches directly forward over the measurement tool for the maximum level, keeping the given posture. The test should be repeated on the other side of the body as well.

Scoring: Results of each side should be recorded with 0,5 cm precision.

The test differs from regular "sit and reach" tests as only one side of the student is being measured at a time, thus the difference between the two sides may be measured more easily.

8. **Endurance run test (20 meters)**

Test objective: Measurement of aerob capacity

Equipment: audiotape of the test, audioplayer, marker cones

Starting position: students stand behind their designated starting points, their partners sit behind them and look at the performance of their pair.

Task: Students should run as long as possible with continuous movement back and forth across a 20-meter space at a pace specified by the audiotape. The test has progressive intensity: the beginning is easy and it is becoming harder with the progression of time.

There are 21 different levels in the audiotape: the first level provides 9 seconds to complete the 20 meter distance, this time limit decreases by 1,5 meters by every level. Switch from one level to the other is marked by a triple beep sound, indicating that a faster pace is coming next.

A lap is completed when the student reached the 20-meter mark line with at least one leg at least at the time of the beep, or passed the line by that time. If the student passes the line earlier, he has to wait for the beep, and shall continue running back afterwards. The test finishes after the second mistake of the student: when he does not reach the line by the beep or is unable to continue running.

Scoring: The test is evaluated according to the completed laps. Using a table we state which level the student reached and give the completed distance in meters.

The most commonly used mean value is the **arithmetic average (mean)**, which is a calculated mean value. Technically it is the number that makes the sum if all values are replaced by this number. We can calculate I if we devide the sum of the data by the number of the data (source: fitness 57_clarified data.xlsx; fittségi 57 tisztított adat.xlsx).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} x_i}{n}$$

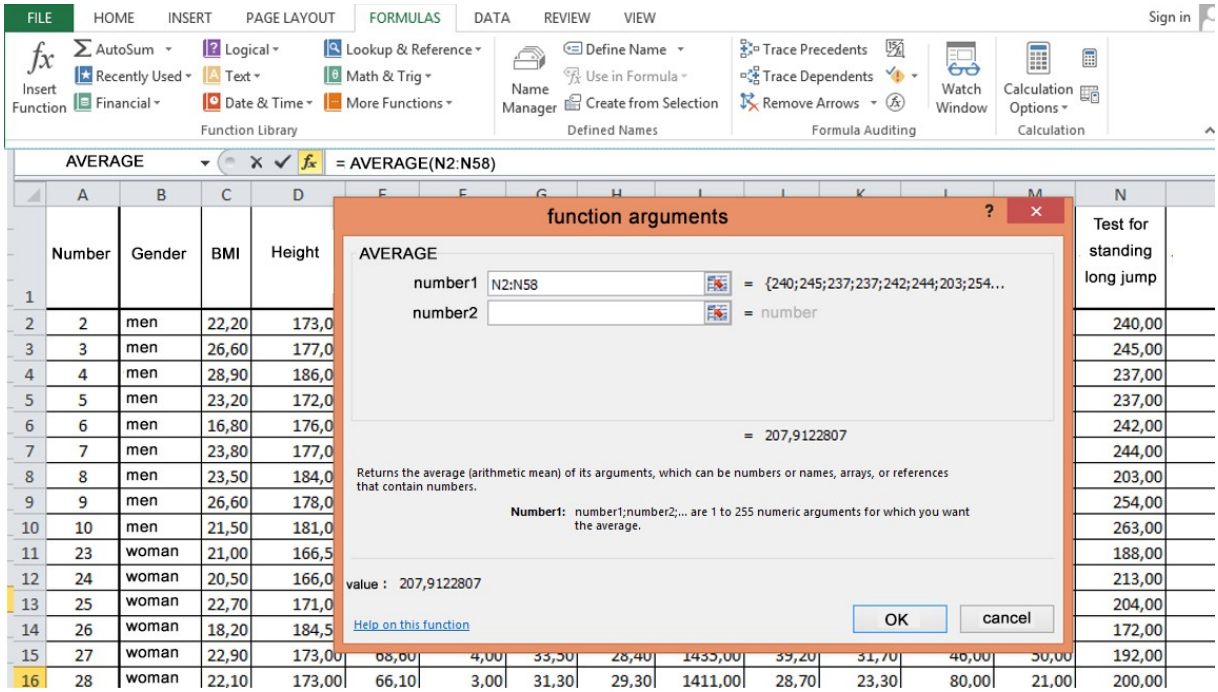


Figure 2/4. Calculation of the arithmetic mean

The Excel software offers several options to calculate the arithmetic mean. One of the most popular methods – as it can be used for other functions as well – is the function wizard. The function wizard can be found by clicking “Formulas” and then “Insert”.

We often have to calculate the arithmetic mean of a frequency list. In this case certain values may occur multiple times, as indicated by the frequencies. To count the arithmetic mean, these frequencies have to be used, which may be carried out by the formula of the ***weighted arithmetic mean.***

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

In practice we first calculate the total values (which is the sum of the product of variables and frequencies), which has its own meaning (total jumped distance in cm). After we calculated this value, it should be divided by the case number. We should calculate like this in cases when our datalist groups cases and frequencies of the values are given. We carry out the abovementioned calculation in two steps. First we calculate the summarised value (list summary 11851 cm), which we then divide by the number of cases (57).

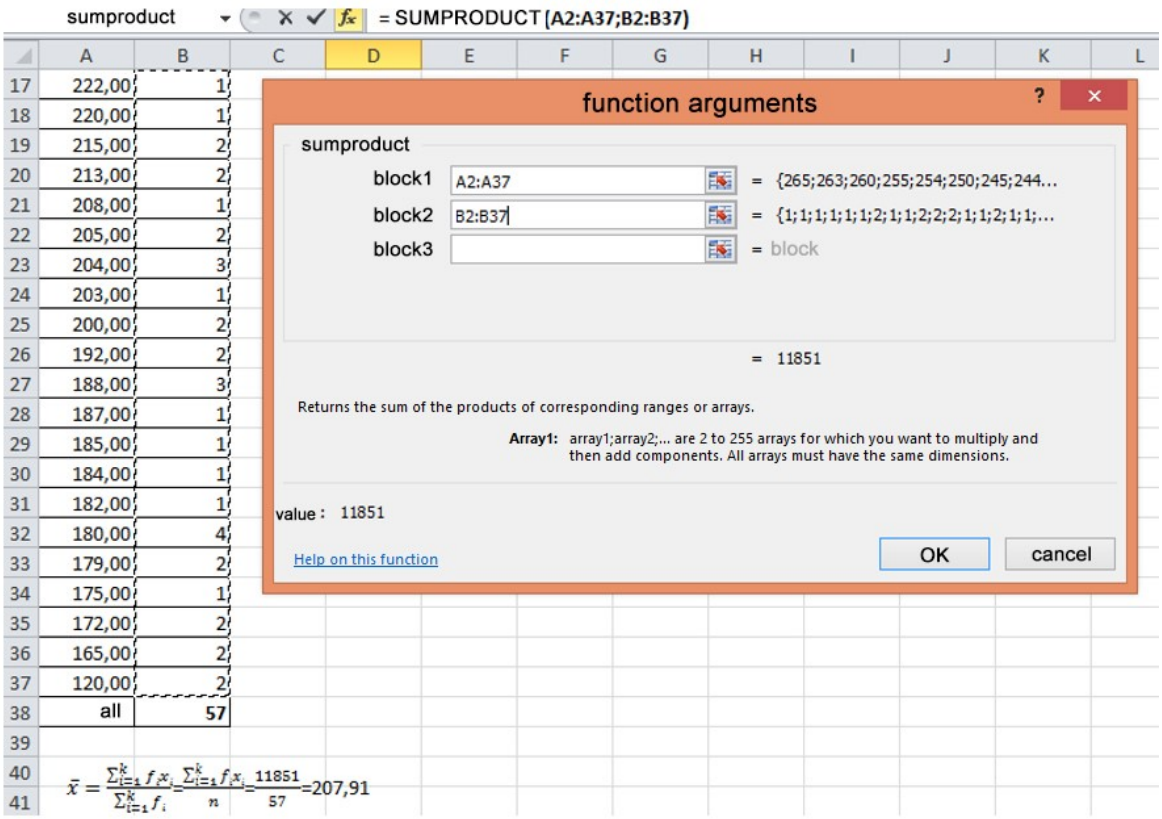


Figure 2/5. Calculation of the summarised product

We can see that the result is the same by both methods of calculating the mean. The average distance the examined 57 students could achieve in standing long jump was 207.91 cm. The chosen method is defined by what data is available at the beginning. Interpretation of the results may often be challenging (eg. 24.11 pieces of sit-ups), and unfortunately the formula of the arithmetic mean is rather sensitive for extreme values (eg. mistyped values), thus it is advisable to examine the applicability of another measure of central tendency.

The **median** denotes or relates to a value or quantity laying at the midpoint of a frequency distribution of observed values or quantities, presenting 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: $Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$.

Excel offers the function of the median as well, which is either available from the function wizard or we can simply type it in as well (=median). To demonstrate the two steps, let's start with the more complex one. To define the median of the sit-up test first we rank our data – we can perform this task by clicking "Data" at the menu bar, and then clicking "Sort". The menu that occurs is rather simple (starting page/sorting and pooling/customised ranking), thus we do not specify each step, only give the result. However, we suggest that the "customised ranking" is the most popular module in sorting as it can be set easily.

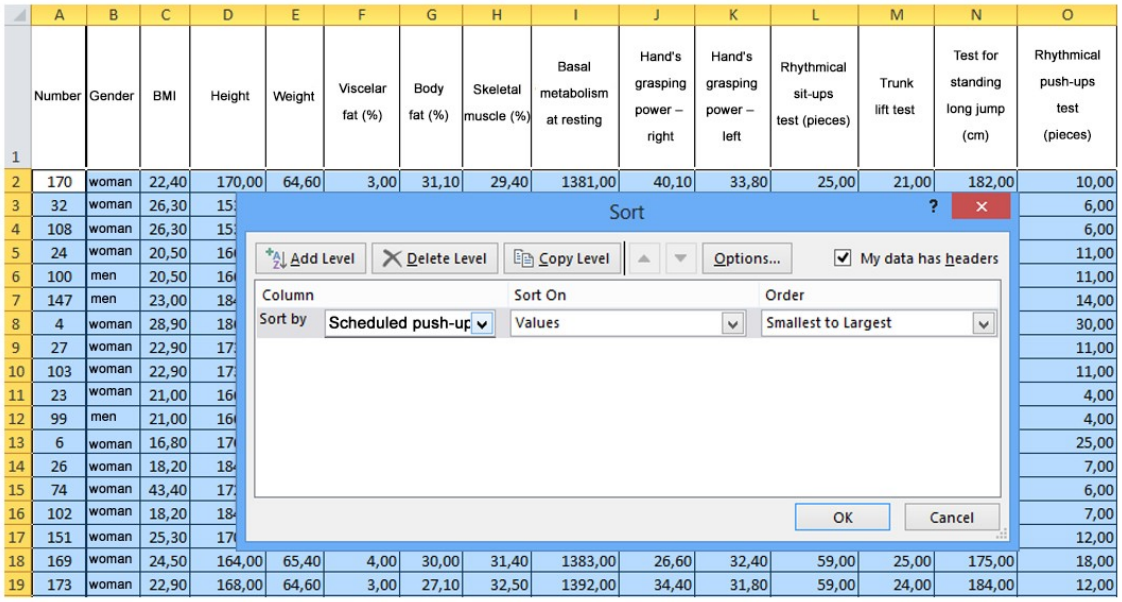


Figure 2/6. Process of sorting

We use the abovementioned formula to define the median from the resulting list. We have altogether 57 observations, thus the arithmetic mean of the rank's (57+1)/2 items will be the median: Me=26. This means that half of the students completed 26 sit-ups, and half of them completed more. This result is rather easy to interpret, but we often end up with fractions, that may be hard to interpret.

Mode is the most typical, most characteristic 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 denser, i.e. at the maximum of the frequency curve (see later).

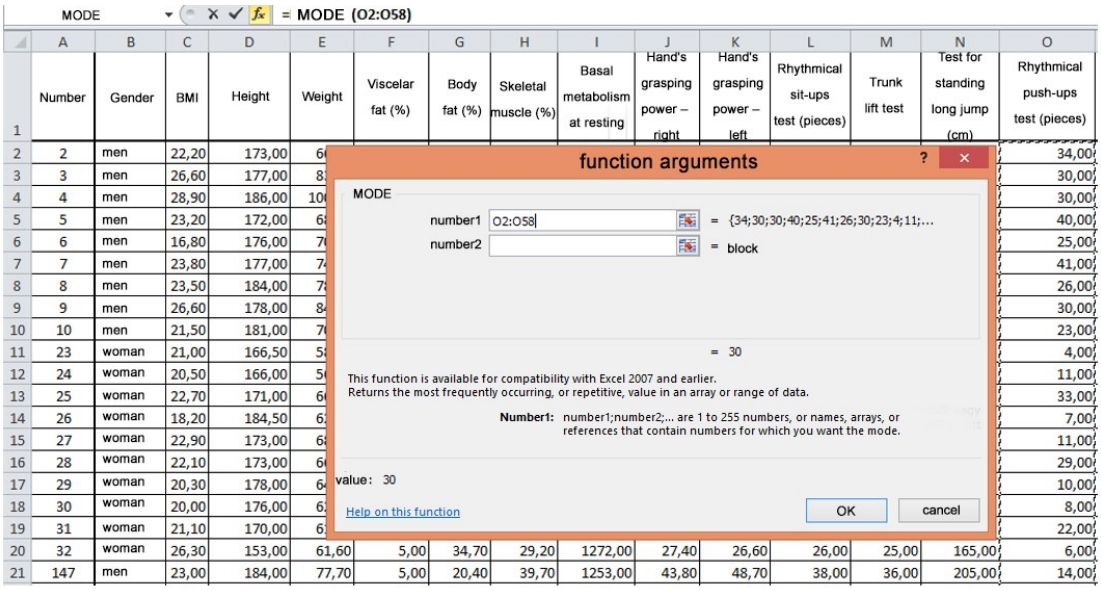


Figure 2/7. Calculation of the mode

Using the function wizard we can define the mode as well, which is 30 in this case. Thus the most frequent result among the participating students was 30 sit-ups. We should be careful, because the mode may not be the typical measure of central tendency in all cases. Not all measures of central tendency meet all criteria, i.e. there is no measure of central tendency that meets all criteria completely. We shall use the one that is in line with the research aim and can be interpreted logically.

Defining **measures of central tendency based on categorical frequency distributions** is a more complex task. Categorical frequency distributions are worth creating when the decrease in the number of values is required, as we have to work with a large amount of data. Thus we produce intervals (groups) from the variable values. Naturally we get the most precise results if we use all the individual data, and computer softwares can deal with a large number of data relatively well, too. Categorical frequency distributions are most often used when the basic data is already given in this format, or when the secondary database is already organised like this.

We would like to highlight that if the original data is available then there is no need to create categorical frequency distributions, because it may lead to less precise results. Thus the next example only demonstrates the issue, as we use the given data to create categorical frequency distribution – this will decrease precision but the process will demonstrate the method in question.

We organise the height (cm) data of the 57 participating students into categorical frequency distributions. The basic data is show by the table below (source: categorical frequency distributions.xlsx; osztályközös gyakoriság. xlsx).

The first step is to create categories. There are several statistical methods available, now we choose one that is quick and easy to apply.

The number of categories can be defined based on prior information on a formula:

$$r = \sqrt{\sum f_j} + 1 \text{ . or } r = 1 + (3{,}3 \times \lg n)$$

Length of categories: $$h = \frac{x_{\max} - x_{\min}}{r}$$

Thus if we would like to define the average height of the students, first we should organise our data into categorical frequency distributions.

$$r = x_{\max} - x_{\min} = 188{,}5 - 153 = 35{,}5$$

Here we may use min- and max functions, which define the smallest and biggest values in the distribution. Range is the difference between the maximum and the minimum values, which is a measure of standard deviation.

$r = \sqrt{57} + 1 = 8{,}55$, in this case we shall round upwards, so we group values into 9 categories. The range of the intervals is: $h = \frac{188{,}5 - 153}{9} = 3{,}94$; rounded up to 4

Then we prepare the categorical frequency distributions:

Table 2/2. Categorical frequency distributions in 9 categories

lower limit of the category	upper limit of the category	frequency (participants)
153	157,00	3
157,00	161,00	0
161,00	165,00	3
165,00	169,00	6
169,00	173,00	17
173,00	177,00	8
177,00	181,00	7
181,00	185,00	8
185,00	189,00	5

Source: calculated by the authors

Looking at the prepared categorical frequency distribution we can see that there are four categories without or with very few respondents. Thus we assume that applying five categories instead of nine would better serve our purposes. The original data is available to one decimal digit, so we shall round the interval range up to one decimal digit (function wizard: round up function).

Range length: $h = \frac{188{,}5 - 153}{5} = 7{,}1$; rounding up strongly: **8** (the aim is to minimise categories)

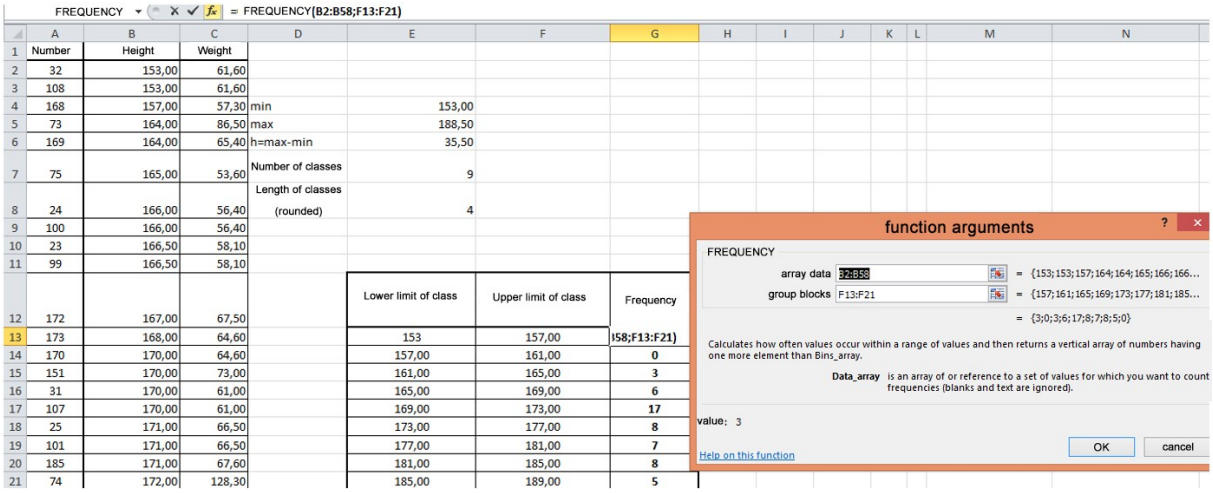


Figure 2/8. Defining frequency

Data array includes the basic data and bins array include the upper limits of the categories. Then we expand the function to the rest of the categories, to calculate the results for other intervals as well. Then we press F2 and then simultaneously the shift-ctrl-enter keys.

Looking at the prepared categorical frequency distribution we can see that there are four categories without or with very few respondents. Thus we assume that applying five categories instead of nine would better serve our purposes more. The original data is available to one decimal digit, so we shall round the interval range up to one decimal digit (function wizard: round up function).

Range length: $h = \frac{188{,}5 - 153}{5} = 7{,}1$; rounding up strongly: **8** (the aim is to minimise categories)

This way we can get the new categorical frequency distribution. To evaluate this correctly we can look at the relative and cumulative frequency distribution. When preparing the categorical frequency distributions we should also pay attention to defining limits of the given range. A general rule is that limits shall always enable unambiguous categorisation. A given dividing value shall clearly belong to a particular category[13]. This may be particularly challenging when we use continuous variable values, thus the researches must be very careful. Rounding continuous values may help a lot, which can finally make the categorisation unambiguous.

[13] In these cases we may provide more precise category limits at the observations as well, which we treat then as „technical numbers".

Tbale 2/3. Working table for categorical frequency distributions in 95categories

lower limit of the category	upper limit of the category	frequency (participants)	Relative frequency (gi)	Cumulative relative frequency (gi')
–	161	3	5,26%	5,26%
161	169	9	15,79%	21,05%
169	177	25	43,86%	64,91%
177	185	15	26,32%	91,23%
185	–	5	8,77%	100,00%
Total		**57**	**100,00%**	

Source: calculated by the authors

Apart from frequencies the table above also shows the *relative frequencies*, which can be created and interpreted as a ratio distribution. It means that 5.26% of participants are not as tall as 161 cm. The (upper) *cumulative relative frequency* shows that 64.91% of participating students are not as tall as 177 cm. If there are no particularly deviant values in the relative frequency list's two starting categories, then we can accept the number of categories as satisfactory.

The lower and upper intervals in the frequency distribution depicted by the table above are both **open-ended intervals**. This solution is applied when there are extreme values in the data set. This is not the case in our example, but in case of a repeated test (examining several thousands of cases, thus extreme values cannot be closed out) the above grouping provides a solid ground for comparison. During further calculations we take these open-ended intervals as if they were closed-ended – we assume that the first interval is as big as the next; and the last interval is as big as the one before.

In the next step we use the categorical frequency distribution to estimate the average height of the examined students, for which we have to define the class marks. **Class marks (x_i)** – which are going to have an important role in our further calculations as well – can be calculated by averaging the lower and upper limits of the intervals. During the calculation we do not take into consideration the technical lower- and upper limits that were distinguished solely for unambiguous categorisation. The dataset that we get by multiplying class marks and frequency distributions is the **summary distribution** (s_i). After this we devide the sum of the summary distribution by the number of cases.

Table 2/4. Working table of defining arithmetical mean

lower limit of the category	upper limit of the category	frequency (participants)	class mark (x_i)	sum ($s_i=f_i*x_i$)
–	161	3	157	471
161	169	9	165	1485
169	177	25	173	4325
177	185	15	181	2715
185	–	5	189	945
Total		57		9941

Source: calculated by the authors

$$\bar{x} = \frac{\sum_{i=1}^{5} f_i x_i}{n} = \frac{9941}{57} = 174{,}4$$

The result means that the average height of the 57 respondents is 174.4 cm. We used class mark during this calculation thus the result is an estimation. The calculation based on actual data also resulted in 174.3 cm, which means that there was no significant difference between the results of the two methods.

Definition of the median using categorical frequency distribution aslo takes place by an approximation procedure (using the cumulative frequency distribution), according to the following formula:

:

$$\mathrm{Me} = x_{me,a} + \frac{\mathrm{s} - \mathrm{f}'_{\mathrm{me-1}}}{\mathrm{f}_{\mathrm{me}}} \times \mathrm{h}$$

where : $x_{me,a}$ - lower (non - technical) limit of ther range including the median

s - n/2 rank of the median

$\mathrm{f}'_{\mathrm{me-1}}$ - cumulative frequency of the range before the median

f_{me} - frequency of the range including the median

h - length of the range including the median.

Table 2/5. Working table of calculating the median

lower limit of the category	**upper limit of the category**	**frequency (participants)**	**cumulative frequency (f_i')**
–	161	3	3
161	169	9	12
169	177	25	37
177	185	15	52
185	–	5	57
Total		**57**	

Source: calculated by the authors

$$Me = 169 + \frac{28{,}5 - 12}{25} \times 8 = 174{,}28\,cm.$$

This means that half of the students are taller than 174.28 cm and half of them are shorter. If we do not calculate the median based on the categorical frequency distribution, then the value will be 173 cm. Using the formula of the median we can easily define the specialised quantiles, as the median is such itself. If we divide the sorted population to 2, 3, 4, ...,k **equal parts**, then the variable values at the dividing points are the **quantiles.** In other words, the quantiles are the variable values of which the $\frac{1}{k}, \frac{2}{k}, ..., \frac{k-1}{k}$, i.*e.* $\frac{\mathrm{j}}{\mathrm{k}} (\mathrm{j} = 1,2,..., \mathrm{k}\text{ -}1)$ part of all values are smaller, and 1-(j/k) part of all values are bigger.

Certain specialised quantiles have their own name:

Table 2/6. A few specialised quantiles

Q	Name	Sign
2	Median	M_e
3	Tertiles	T_j
4	Quartiles	Q_j
5	Quintiles	K_j
10	Deciles	D_j
100	Percentiles	P_j

An often-used piece of quantiles are the quartiles (Q_j), and usually we mean *upper quartiles* (Q_3) and *lower quartiles* (Q_1) by them, as Q_2 is the median. Accordingly, we may look at quartiles as the medians of the values lower and upper of the median. Quarter of all values are lower than the lower quartile and three-quarter are upper; while three-quarter of all values are lower than the upper quartile and quarter are lower. Looking at our example, this means that first we have to calculate the appropriate j/k ratio of the n=57 sample size. Half of the sample size is at 28.5, quarter is at 14.25 and three-quarter is at 42.75. Looking at the cumulative frequency distribution we can see that 7.5 overreaches the 13.1 category-limit, thus – as the median – the lower quartile can be found in the 13.1-14.6 category.

$$Q_1 = 169 + \frac{14{,}25 - 12}{25} \times 8 = 169{,}72.$$

$$Q_3 = 177 + \frac{42{,}5 - 37}{15} \times 8 = 179{,}93\,.$$

Consequently, quarter of the students are shorter than 169.72 cm but quarter of the sudents are taller than 179.93 cm. In other words, three-quarter of the students are shorter than 179.93 cm.

Calculating the mode from a simple frequency distribution shall not be challenging as we have to choose the value with the highest frequency. In case of categorical frequency distribution the approximative definition of the mode starts with the calculation of the **modal class interval**. The modal class interval – which includes the mode – is the interval with the highest frequency. Obviously, the immediate definition of the modal class interval is only possible when we have intervals with equal ranges. If this is not the case, then we should carry out a correction: we must recalculate the frequencies for equal-sized intervals. To determine modal class intervals and also during further calculations we are going to use the following corrected frequencies:

We use the following formula to define the mode:

$$Mo = x_{mo,a} + \frac{k_1}{k_1 + k_2} \times h$$

Where : $x_{mo,a}$ – is the lower limit of the modal class,

k_1 – is the subtraction of the modal class interval and the frequency of the previous class interval

k_2 – is the subtraction of the modal class interval and the frequency of the following class interval

h – is the length of the modal class interval

Frequencies also aid the definition of modal class intervals, as the frequency is the highest by the category of the modal class intervals. It is possible that two equally high frequencies belong to two neighbouring classes – in which case a common class limit should be selected. The explanation of this can be found in the book of Pintér and Rappai (2007, p. 131).

$$Mo = 169 + \frac{25-9}{(25-9)+(25-15)} \times 8 = 181{,}8cm.$$

This means that the most common height within students is 181.8 cm.

2.7.2.3. Variability and symmetry

As it can be seen, none of the mean values meet all the requirements and therefore a more sophisticated analysis is required. It is possible that mean values are equal but variable values (population units) differ significantly. The more homogenous the population is, the less the dispersion from the examined variable's point of view.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.

The most widely used measures of dispersion include:

- range (R),
- interquartile range (TQ),
- standard deviation (σ), variance (σ^2)
- relative standard deviation (V)

Range is the difference between the greatest and the lowest value:

$$R = x_{\max} - x_{\min}$$

From the point of view of sportsmen's training, difference from the best own result (world record) is an informative data. It is an important measure to be considered when determining the record-breaking form since its decrease shows improvement.

The **interquartile range** is the interval, where the medial 50% of all the values is found. It can be calculated as:

$$TQ = Q_3 - Q_1$$

Continuing the example above: TQ=179.93-169.72= 10.21 cm means that 50% of students' height is in a 10.21 cm range.

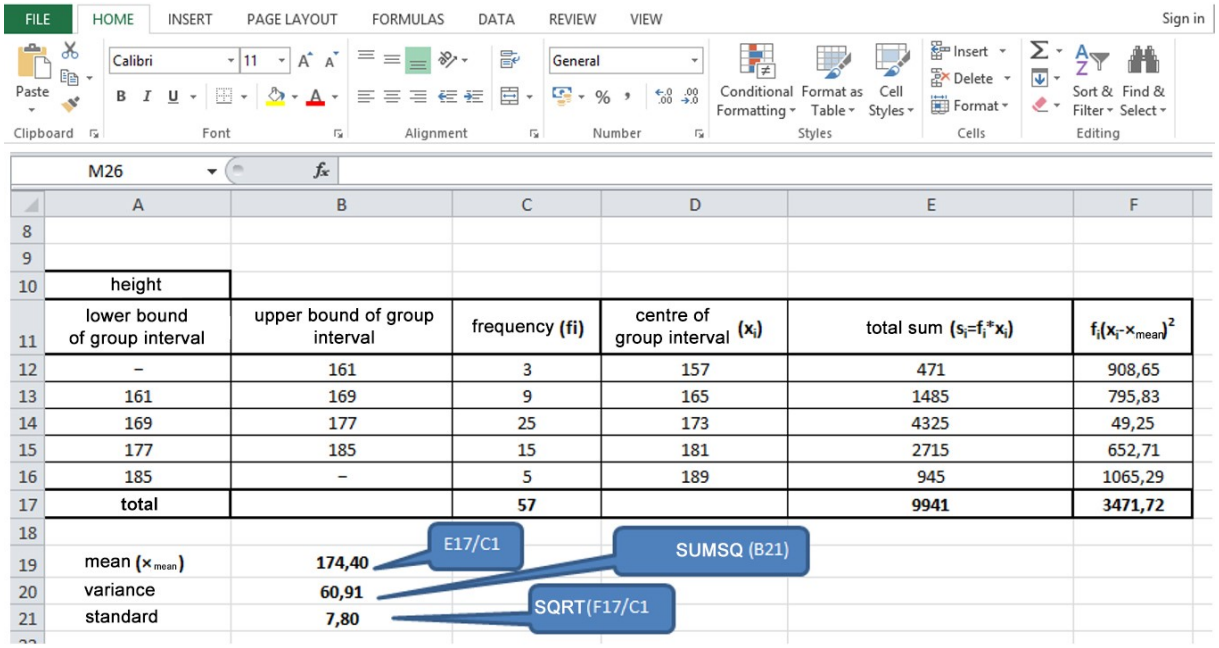
The most commonly applied measure of dispersion is **standard deviation** which is the square root of the average square deviation from the mean.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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** (σ^2). It has no special meaning but is a key index in several statistical methods. Standard deviation and variance is calculated from a frequency distribution table for the example above.



	A	B	C	D	E	F
8						
9						
10	height					
11	lower bound of group interval	upper bound of group interval	frequency (fi)	centre of group interval (x_i)	total sum ($s_i=f_i*x_i$)	$f_i(x_i-x_{mean})^2$
12	–	161	3	157	471	908,65
13	161	169	9	165	1485	795,83
14	169	177	25	173	4325	49,25
15	177	185	15	181	2715	652,71
16	185	–	5	189	945	1065,29
17	**total**		**57**		**9941**	**3471,72**
18						
19	mean (x_{mean})	**174,40**				
20	variance	**60,91**				
21	standard	**7,80**				

Screen view 2/9. Calculating standard deviation

Means of heights differ from the mean (174.4) by 7.8 cm on average. 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.

$$V = \frac{\sigma}{\bar{x}} = \frac{7.8}{174.4} = 4.47\%$$

Relative standard deviation is small, since it makes 4.47% of average results. (The ratio of mean and standard deviation was defined by Müller as performance consistence index)[14].
As we have previously seen when examining the frequency tables, mean values characterize the place of the distribution and dispersion measures refer to the range and the "density" of the population. Additional information on pseudo randomness and nature of the population can be gathered when examining frequency distributions. If variables are not equally "condensing" between the highest and the lowest value, then graphic illustrations show which ranges include more of them. **Distribution** shows the frequency of variable values at a given range, i.e. if values rather belong to the lower or the higher value range. Distributions can be clustered based on this (Pintér- Ács, 2006, page 95).

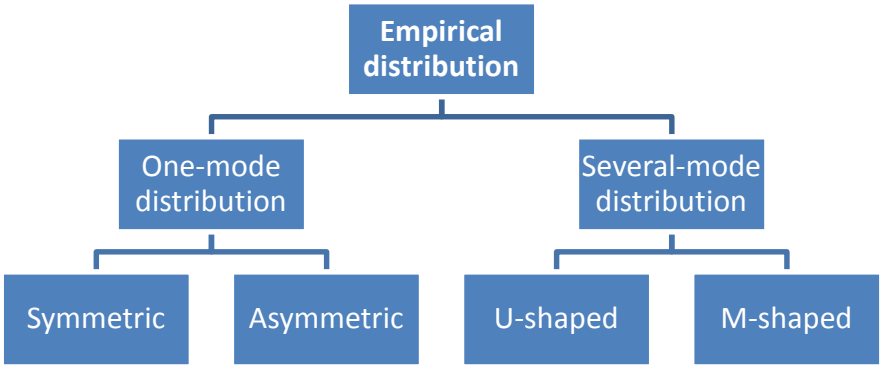


Figure 2/10. Types of empirical distributions

One type of one- and several-mode distribution curves is illustrated below.

[14] Müller A.(2004), page 103

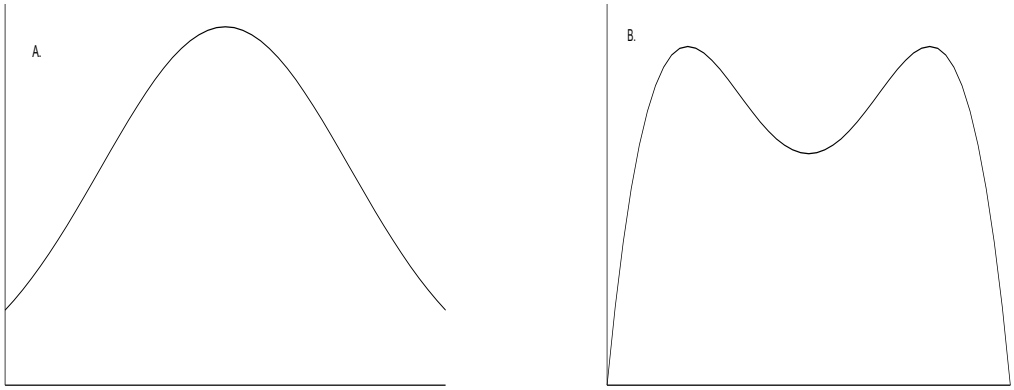


Figure 2/11. One- and several-mode distribution

Part A int he figure above is a one-mode, **symmetric** distribution. Several-mode distributions (part B) used to refer to heterogeneous[15] population. Local maximums of the frequency curve occur at modes of sub-populations. Distribution of the salary of a handball team's members can be a typical example for an M-shaped distribution (part B of the figure) since salaries can have different local maximums (here: two maximums) according to recognition (number of front-rank matches), position, etc.[16] U-shaped distribution is relatively rare in practice since it means that the two modes are the two extreme values as well. Teachers often hear after university exams that the examiner only gave 1 and 5 as result mark. If these comments by students would be valid, grades would shape a U-distribution.

Measures of central tendency (arithmetic average, median, and mode) have the same numeric measure $Mo = Me = \bar{x}$ (in practice, only approximately the same) for symmetric distributions (Fig. A). The most common item is the central value of the population.

This equality is not true for asymmetric distributions. There are left-skewed and right-skewed distributions which are not presented the same way in statisctics. This book handles it based on the statistical alma mater in Pécs"[17].

Left-skewed:

$$\mathrm{Mo} > \mathrm{Me} > \bar{\mathrm{x}}$$

Right-skewed:

[15] The population is considered to be heterogeneous if it can be separated into more homogeneous parts based on a variable.

[16] The most frequently used example of M-shaped distribution is the height of students (male and female) in the same year.

[17] The Anglosaxon terminology is used in Pécs – as applied by most statistical softwares.

$$Mo < Me < \bar{x}$$

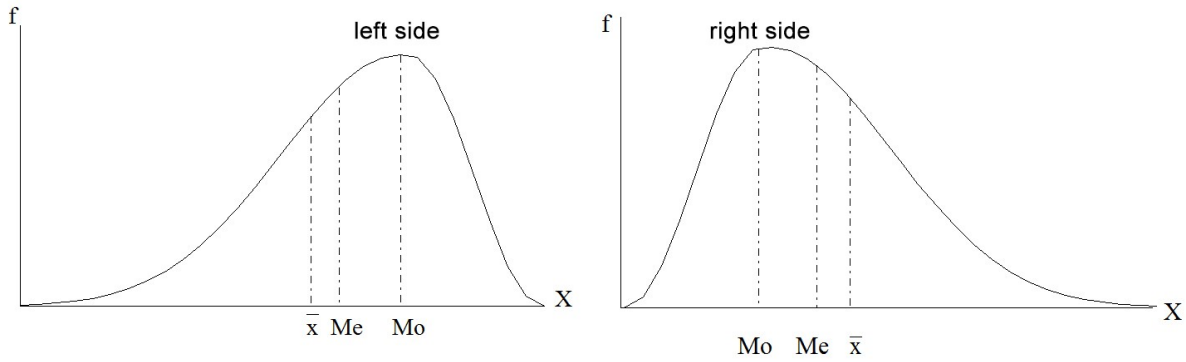


Figure 2/12. Left- and right-skewed graphs

Two types of asymmetric frequency curves

In sport, right-skewness is more frequent since not many can reach outstanding results; most sportsmen are able to perform less "only".

In the case of left-skewedness, higher values are more frequent, e.g. the improper training (planned training levels) will be overperformed by most sportsmen. Consider the analysis of an exercise with the result of left-skewed curve. This means that the training is too easy, so some alteration (in range, intensity) is needed in order to improve performance. If training is suitable, the shapes of curves of the analyses have to be close to normal distribution since a sample (of finite units) can only approach the normal distribution[18] (Gaussian curve).

Pearson's A is often applied to calculate asymmetry, the formula of which is:

$$A = \frac{\bar{x} - Mo}{\sigma}$$

The measure is zero for symmetric distributions. Positive sign refers to right-skewness, and negative one means left-skewness. The measure has no upper limit (in absolute value) but an absolute value greater than one refers to significant asymmetry.

Left-skewness can be detected in the example above: $A = \frac{174.4 - 181.8}{7.8} = -0.9$

The value of F is the other important measure: $F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}$

[18] In normal distribution, 68% of data is situated in one standard deviation (upper and lower) difference from the mean. One sixth is outside the standard deviation, approx. 2/3 is more than two std. deviations apart, and 0.1% is even further away than three standard deviations.

The measure value is negative for right-skewness, positive for left-skewness, and zero for symmetric distribution. It has definite lower and upper limits: $-1 \le F \le 1$

$$F = \frac{(179.93 - 174.28) - (174.28 - 169.72)}{(179.93 - 174.28) + (174.28 - 169.72)} = 0.11$$

The F measure detects again *right-skewness*. Difference comes from the different types of central measures.

Excel applies corrected S' (skewness) for individual data which can be interpreted the same way as the Pearson A.

$$S' = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^3$$

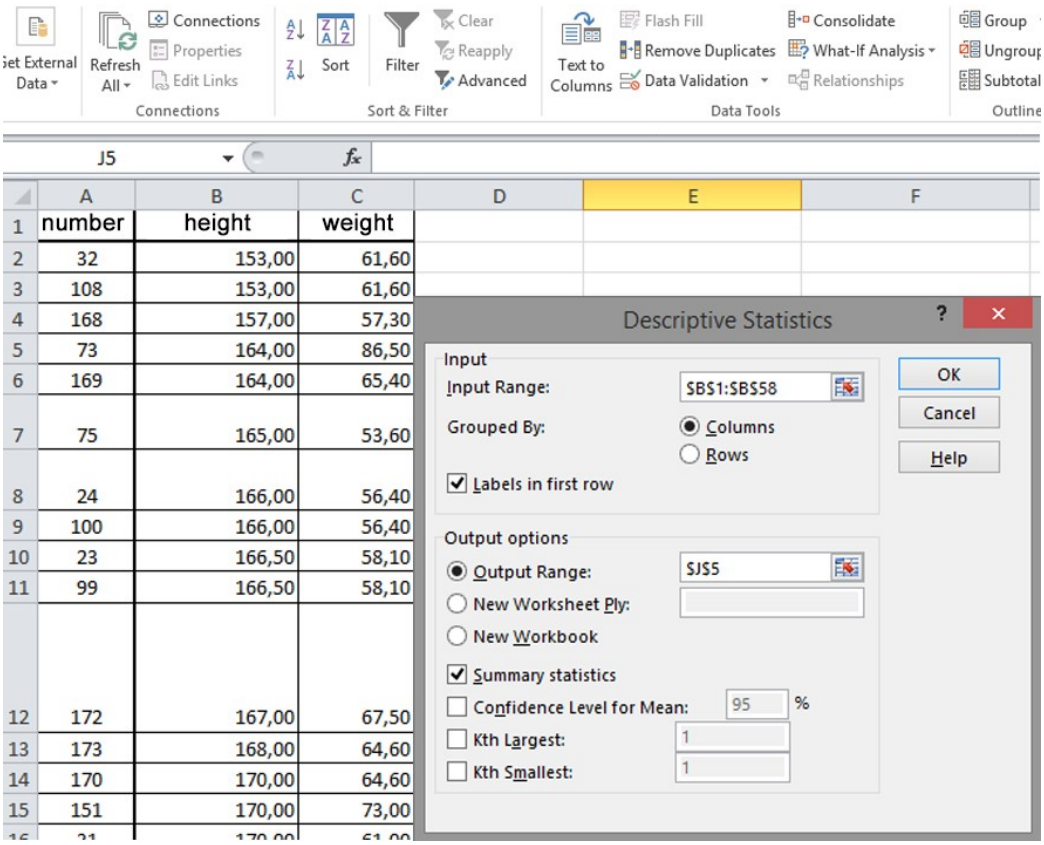
Besides symmetry and skewness, kurtosis can be examined as well. The value of the K-measure is 3 for normal distribution. If it is smaller, then the distribution is flat (close to uniform distribution), otherwise it is peaky (dense around the mean).

$$K = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^4$$

The built-in kurtosis measure K' calculates with small sample correction. That is why the positive values refer to peak and negative ones refer to flat distribution.

Descriptive statistical analysis can be carried out step by step (e.g. function wizard) or in Data/Data analysis/Descriptives. This is not a default option so go to Add-In (File/ Excel Options/Add-Ins) and select Analysis ToolPak. (Source: osztályközös gyakoriság.xlsx)

Request descriptives for summary.



Screen view 2/10. Descriptives settings

It is reasonable to display the variable names (in the first row) in order to know what the descriptives are about. Results are the same as the ones above. Slights differences stem from the fact that our data were "estimated" from frequency distribution table while Excel calculated with the actual frequencies. The software considers data to stem from sample as default so it applies correction factor in calculations. E.g. corrected standard deviation which means that the denominator n-1.

Table 2/7. Descriptives in Excel

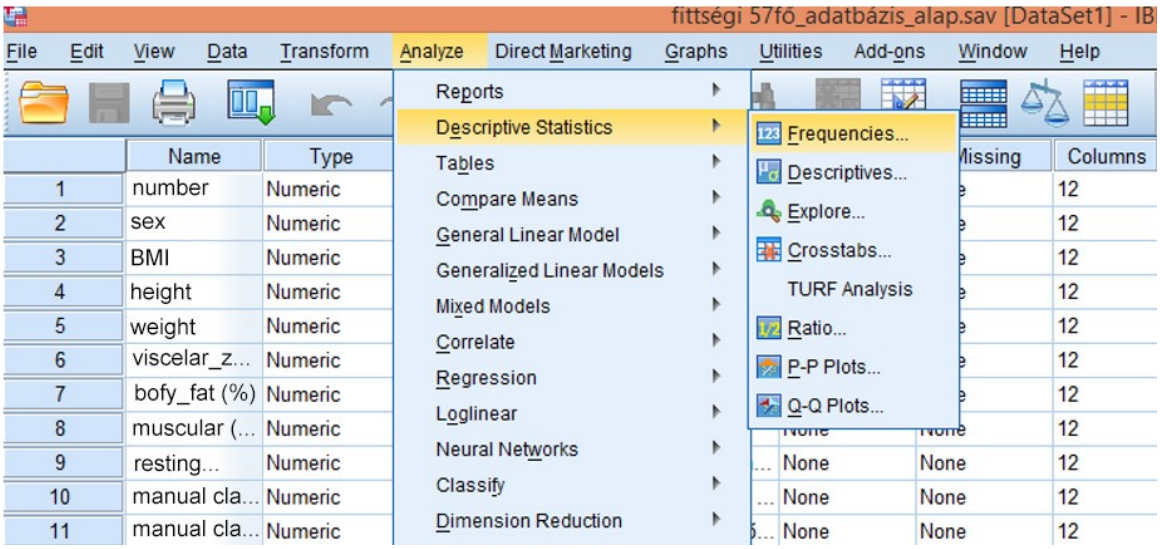
	height
expected value	174,27
standard error	1,07
median	173,00
mode	173,00
standard deviation	8,05
sample variance	64,80
kurtosis	0,36
skewness	-0,47
range	35,50
minimum	153,00
maximum	188,50
sum	9933,50
sample size	57,00

The first value is the numerical average, referred to as expected value by the program. Range means the range of standard deviation.

There are three ways in **SPSS** to access descriptive statistics: ANALYSE/DESCRIPTIVE STATISTICS/DESCRIPTIVE or ANALYSE/DESCRIPTIVE STATISTICS/FREQUENCIES or ANALYSE/DESCRIPTIVE STATISTICS/EXPLORE.

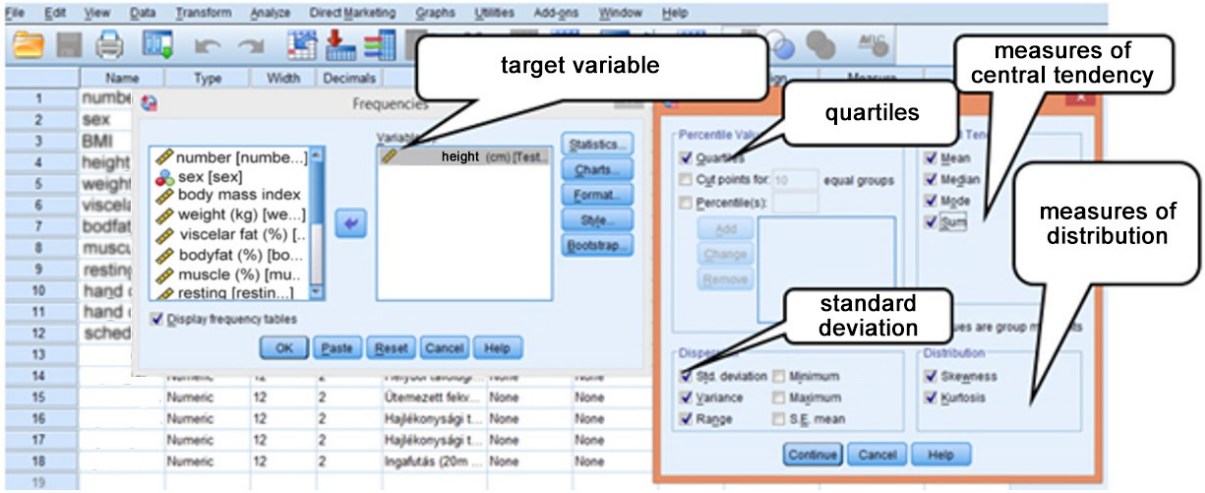
There is no binding rule to decide which method should be used from descriptive statistics. The module labelled DESCRIPTIVE is mostly used for interval or ratio scale (in SPSS. scale), provided there is no need for a frequency table. The module FREQUENCIES is rather used for variables of nominal and 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, descriptive indices can also be calculated, where the sample can be separated into groups. More details are available in the book titled Data analysis (Ács, 2015).

Consider the example above but in SPSS (forrás: fittségi 57fő_adatbázis_alap.sav). First, go to Frequencies to start the analysis. (Analyze/Deccriptive Statistics/Frequencies)



Screen view 2/11. Start with Frequencies

Select the appropriate variable (height), and then press Statistics to select the proper measures.



Screen view 2/12. Settings of Frequencies in SPSS

Press CONTINUE and OK to get the following results in Output View:

Table 2/8. The output of Frequencies

Statistics

height (cm)

N	Valid	57
	Missing	0
Mean		174,2719
Median		173,0000
Mode		173,00
Std. Deviation		8,04955
Variance		64,795
Skewness		-,469
Std. Error of Skewness		,316
Kurtosis		,364
Std. Error of Kurtosis		,623
Range		35,50
Sum		9933,50
Percentiles	25	170,0000
	50	173,0000
	75	180,0000

Results are the same as Descriptives in Excel. Frequency table appears as default.

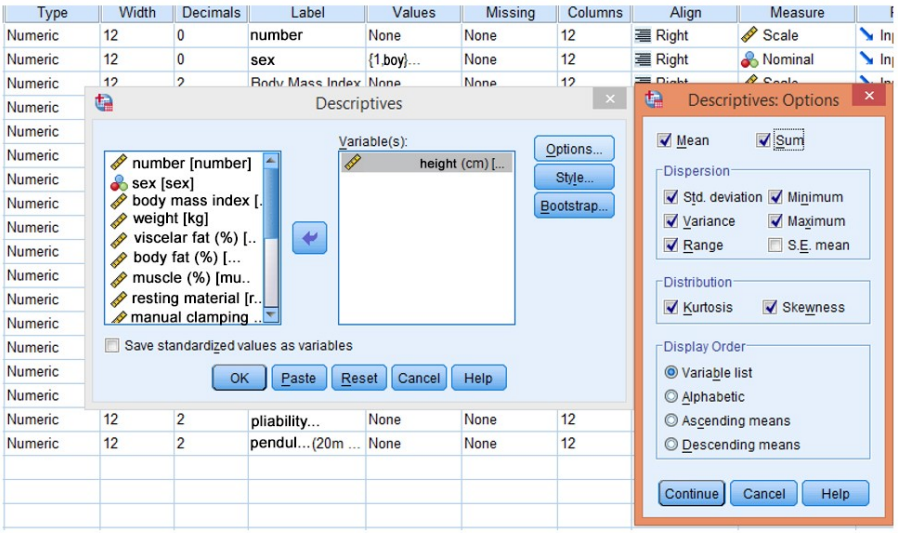
Table 2/9. Frequency table of height (cm)

height (cm)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	153,00	2	3,5	3,5	3,5
	157,00	1	1,8	1,8	5,3
	164,00	2	3,5	3,5	8,8
	165,00	1	1,8	1,8	10,5
	166,00	2	3,5	3,5	14,0
	166,50	2	3,5	3,5	17,5
	167,00	1	1,8	1,8	19,3
	168,00	1	1,8	1,8	21,1
	170,00	4	7,0	7,0	28,1
	171,00	3	5,3	5,3	33,3
	172,00	3	5,3	5,3	38,6
	172,50	1	1,8	1,8	40,4
	173,00	6	10,5	10,5	50,9
	174,00	2	3,5	3,5	54,4
	176,00	4	7,0	7,0	61,4
	177,00	2	3,5	3,5	64,9
	178,00	5	8,8	8,8	73,7
	179,00	1	1,8	1,8	75,4
	181,00	1	1,8	1,8	77,2
	182,00	2	3,5	3,5	80,7
	183,00	1	1,8	1,8	82,5
	183,50	1	1,8	1,8	84,2
	184,00	2	3,5	3,5	87,7
	184,50	2	3,5	3,5	91,2

One of the main advantages here is that a direct option of graphic illustration (charts) is provided (see later).

Results are similar under Analyze/Descreptive Statisics/Descriptives. Of course, the variable has to be selected first, and then relevant methods can be chosen in Options. (Source: fittségi 57fő_adatbázis_alap.sav)



Screen view 2/13. Descriptives: Options

Press Continue and Ok to get the results:

Table 2/10. Descriptive Statistics

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
height (cm)	57	35,50	153,00	188,50	9933,50	174,2719	8,04955	64,795	-,469	,316	,364	,623
Valid N (listwise)	57											

Of course, the results are the same as the ones we already had.

2.7.2.4. Tools for data visualisation

If there number of data exceeds the level which is simple and easy to be dealt with, it is useful to summarize them for illustration and transparency. Effective and well-know data visualisation tools include *graphic illustration* and *statistical tables*.

Graphic illustration has the important role to visually display the main characteristics, ratios, tendencies and connections of the examined phenomena. Its aim can range from simple data statement to exploration of more complex relations.

Statistical graphs can be *simple* or *complex*. Simple ones include point (xy), bar, pie and line charts. Complex graphs are the result of mathematical or statistical calculations, and are applied to analyse frequency tables, e.g. polygon, histogram, ogiva, Box-plot, Lorenz-curve, dendogram. The basis of graphic illustration is the coordinate system, the essence of

which is taught in primary school already. The system consists of a “y” and an “x” axis, crossing each other in the point 0. Situation of one point can be determined by its distance from the two axes (this is the principle of the point chart).

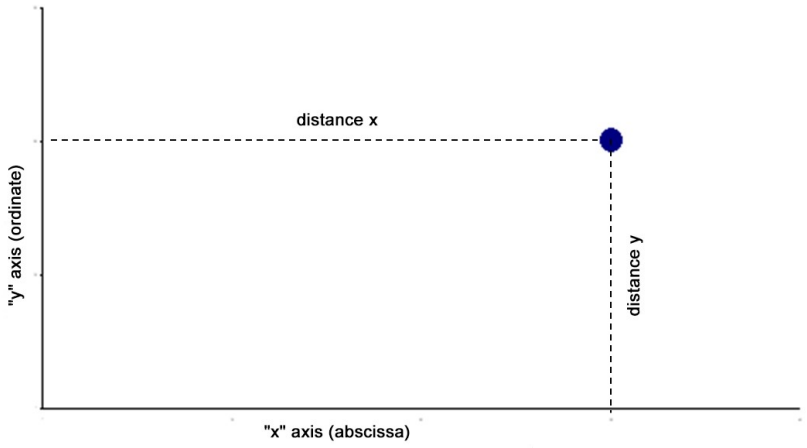
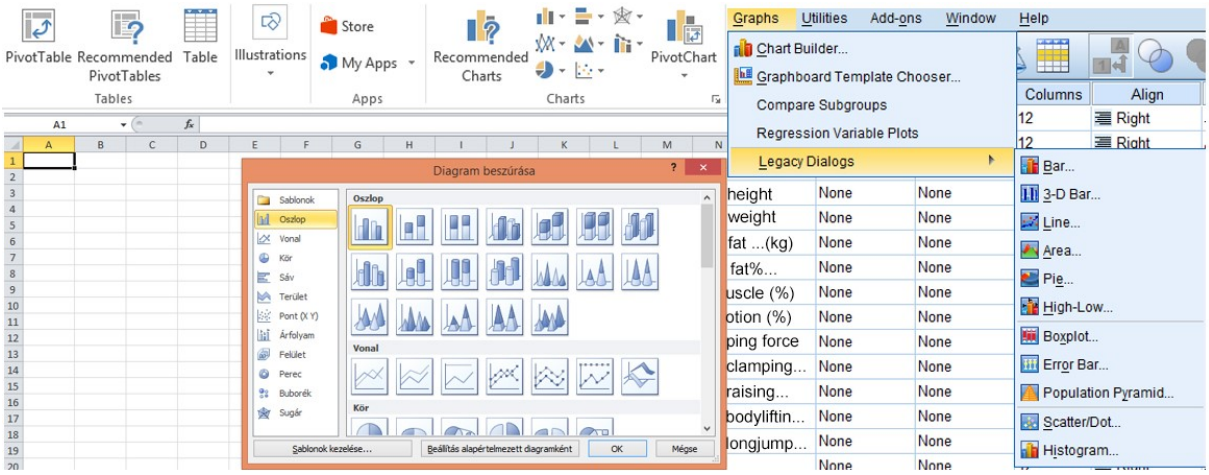


Figure 2/13. The basis of graphic illustration (coordinate system)

It is of most importance that the “simple figure” types have to be applied for different types of data. This does not mean that some figures could not be used for other types of data but there are some that are not valid in particular cases. That is why the following examples present graphic illustrations as a suggestion only. The main principle is to illustrate the steps of creating them. Graphs are meant to illustrate the research data in a valid way, and the values have to be easy to identify and interpret. For details on graphic illustration consult HUNYADI 2001.

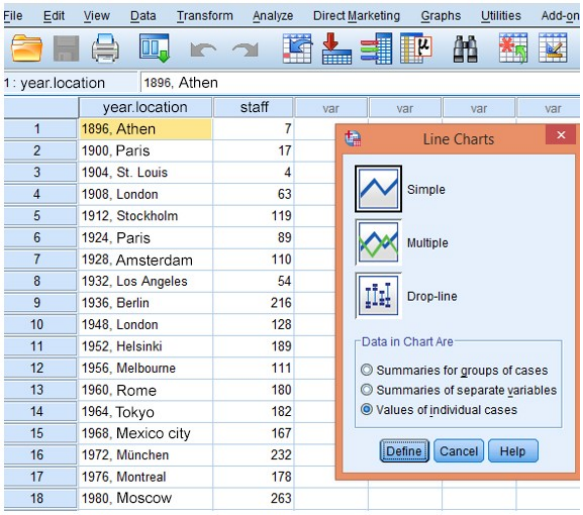
Of course, both Excel (Figure a.) and SPSS (Figure b.) offer the opportunity to create graphs.



Screen view 2/14. Moduls of graphic illustration

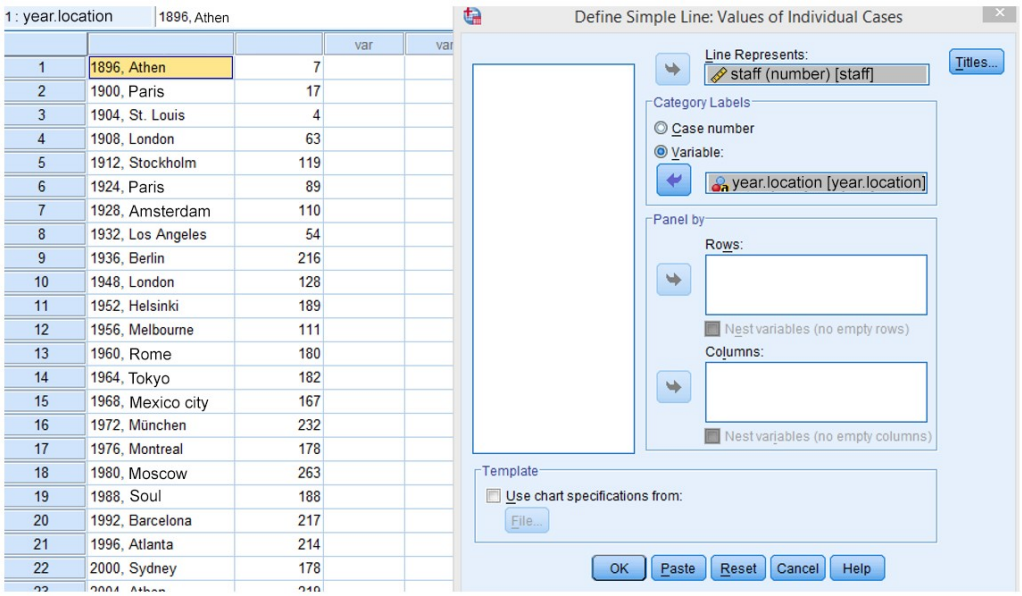
The left-hand side of the figure shows the diagram module of Excel, available under Insert/Diagram. The right-hand side shows the options of Graphs in SPSS. All graphic illustrations are available from these modules.

Line chart will be presented first which is applicable for the analysis of time series and dynamic partitions. Data include the Olympics participation of Hungarian sportsmen (Source: vonaldiagram.adat.sav). Select Line in the Graph module of SPSS to get the following screen:



Screen view 2/15. Line Charts in SPSS

Select Simple and Values of induvidual cases, then press Define to continue.



Screen view 2/16. Line Chart settings in SPSS

Variables have to be put in this window. Membership data comes to the top since these will form the “diagram line”. The values of the cathegorical variable (here: x axis) are the year and place of the Olympics.

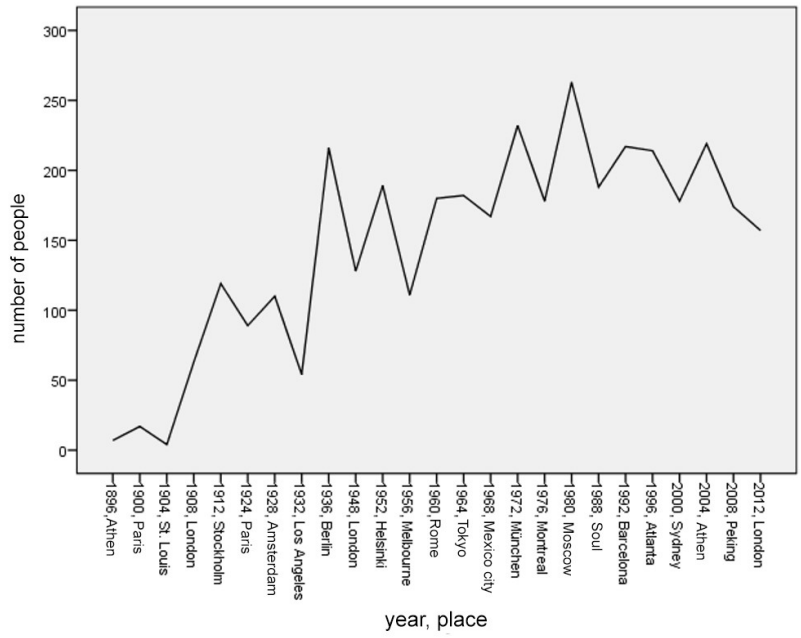
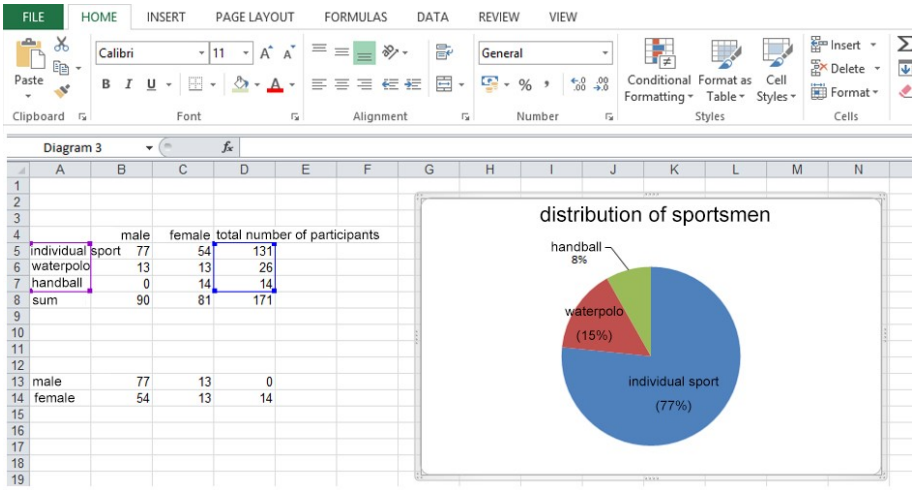


Figure 2/14. Graphic display of the line chart

The line chart shows that the number of Hungarian participants of the Olypmics is growing in tendency, even though in the last some Olympics it went a bit down.

Pie charts make comparison possible since it separates parts of the round into “slices”. It is the most applicable for partition coefficients and data ordered by qualitative variables. Pie charts in SPSS are the same but the access path is Graph/Pie.

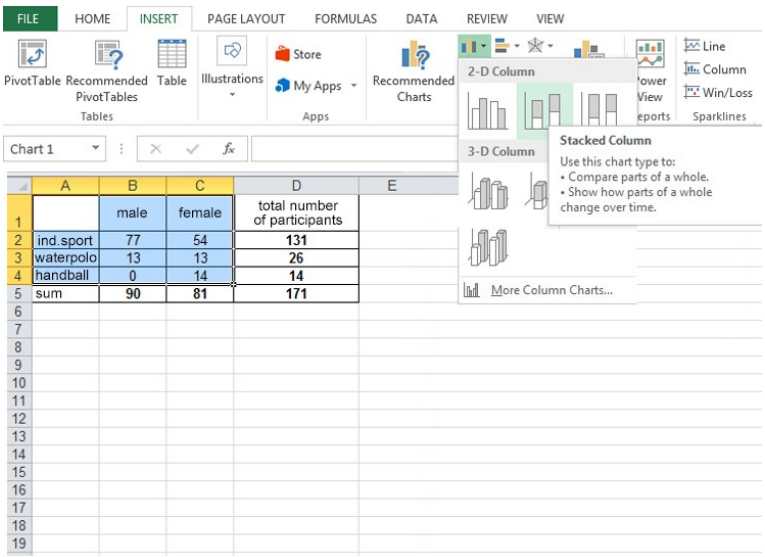
Let us examine the sports (individual, waterpolo, handball) participating at the 2008 Olympics the basic data of which are available at: http://www.mob.hu (26 July 2008). Source: grafikus ábrázolás alapadatok.xls.



Screen view 2/17. Pie chart settings in Excel

Results include that the ratio of individual sportsmen is the highest (77%), followed by waterpolo (15%) and handball (8%). From which it follows that participation of individuals (77%) exceeds the ratio of teams (23%).

Column diagram is probably the most commonly used form of graphic illustration because it can be applied for time series[19], divisions (e.g. territories)[20], and comparisons as well. Let us examine the ratio of men and women according to types of sports. Use the diagram wizard to create a column diagram, and select the second option from the types. As a next step, select that data are ordered in rows. (Source: grafikus ábrázolás alapadatok.xls.)

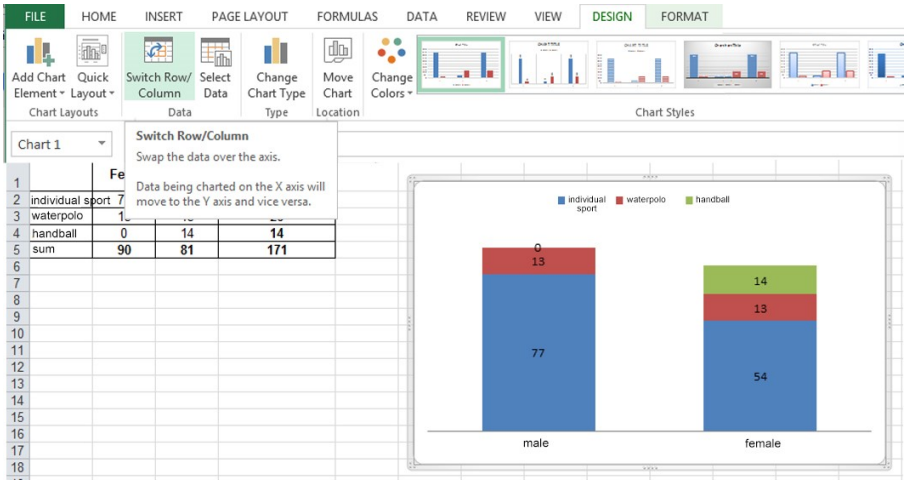


Screen view 2/18. Column diagram settings in Excel

[19] Rather for time intervals when data refer to duration

[20] Horizontal bar charts are often used for territorial comparisons.

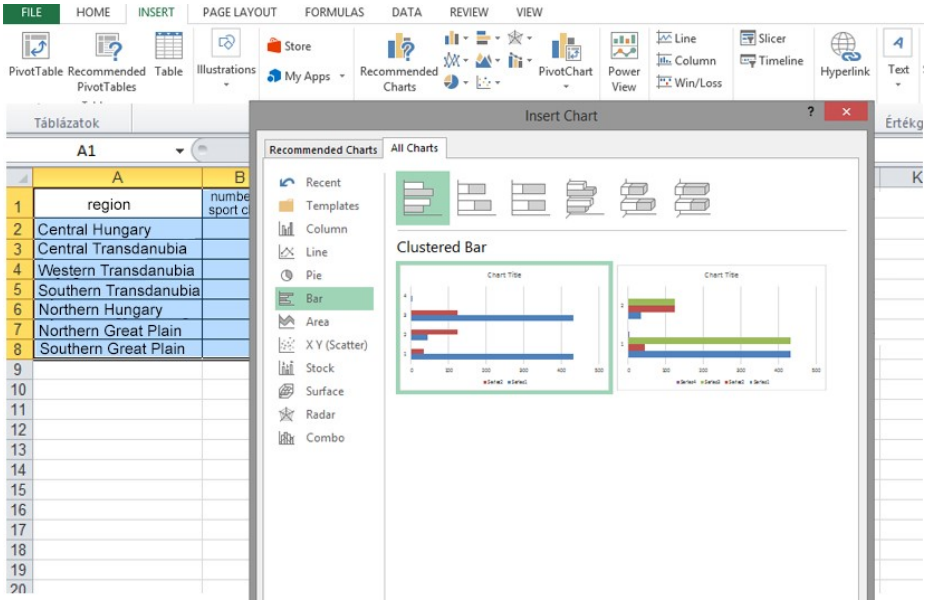
Next, request display of values, and click on Finish.



Screen view 2/19. The column diagram generated

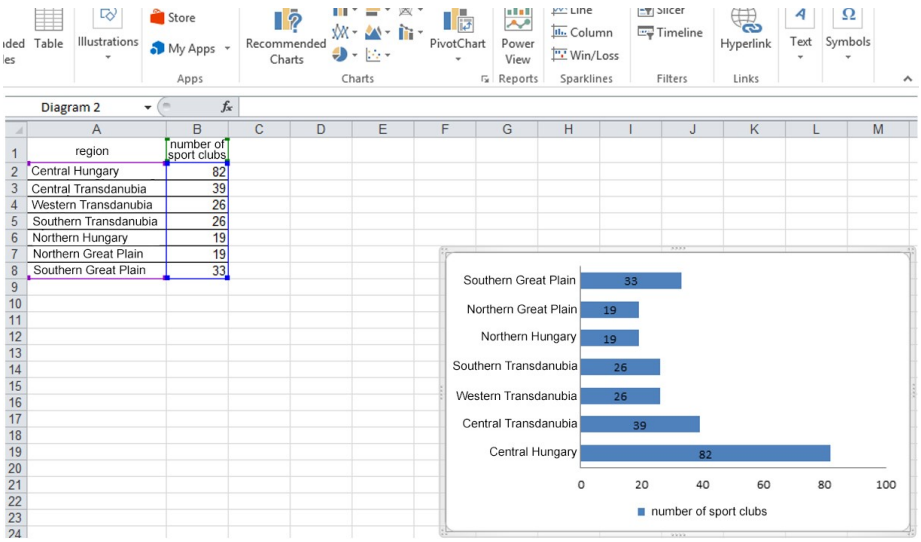
Comparison is possible now.

The next bar chart shows the geographic situation of Olympic athletes' sport clubs according to region. (Source: graphic display basic data; grafikus ábrázolás alapadatok.xlsx.). Select bar chart in the diagram wizard.



Screen view 2/20. Bar chart settings in Excel

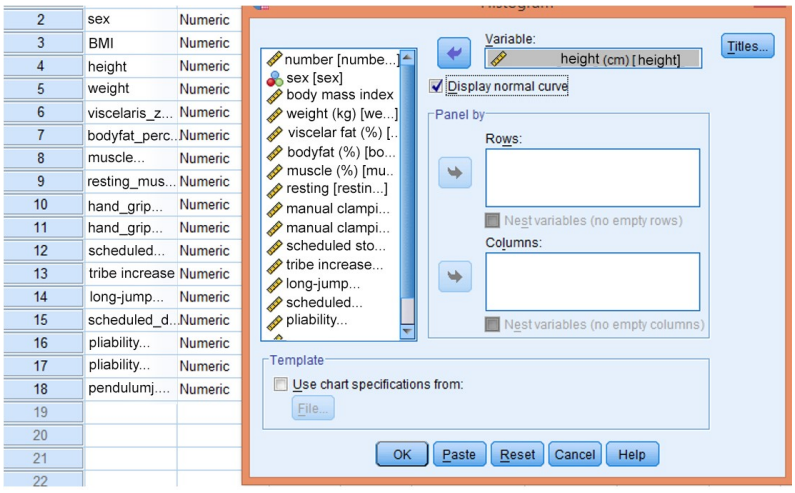
Follow the wizard options to get the following result.



Screen view 2/21. Graphic display of bar chart

The most sport clubs with Olympic athletes are in Central Hungary, while the least come from Northern Great Plain and Northern Hungary.

The bar chart without gaps is called histogram, and a special attention is paid to it. Histogram is the basic method of graphic illustration, the area of its columns refers to frequencies. If the group intervals are the same, then height of the columns represent frequencies. Access path in SPSS is Decriptive Statiscics/Frequencies/Charts but is also available under graphs. The program offers to display the curve of normal distribution on the histogram. The next example is about the histogram of high jumps. (Source: fittségi 57fő_adatbázis_alap.sav)



Screen view 2/22. Edit panel for histogram in SPSS

The result is displayed on the following figure.

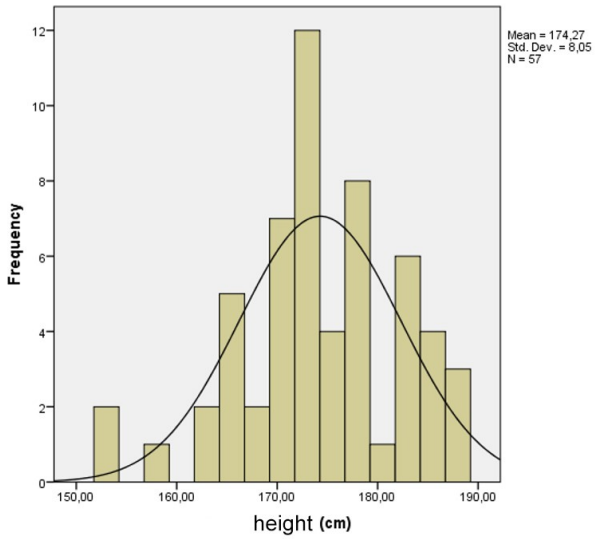
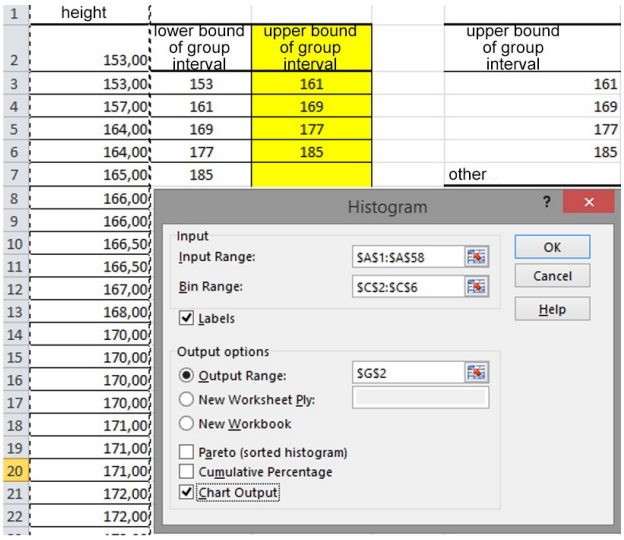


Figure 2/15. Graphic display of the histogram

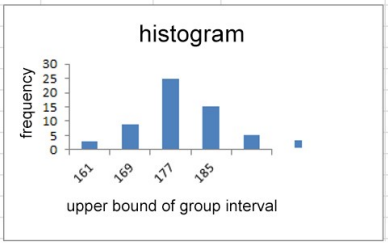
Histogram is available in the Data analysis option of Excel. It makes quick creation of frequency distribution table possible since both he input and the bin range can be selected to determine class intervals. (Source: osztályközös gyakoriság. xlsx)



Screen view 2/23. Histogram in Excel

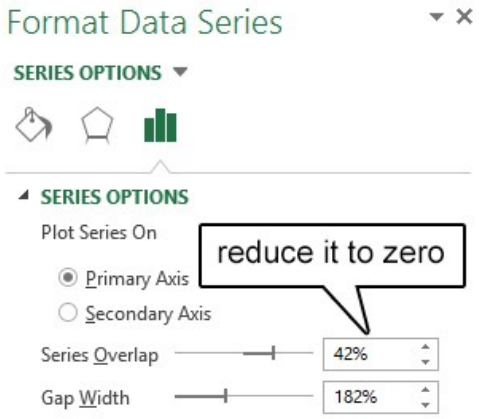
Select the diagram output and put the calculated **upper bound of group interval** in the next row (bin range). The results are shown by the next figure.

	A	B	C	D	E	F
1	height					
2	153,00	lower bound of group interval	upper bound of group interval		upper bound of group interval	frequency
3	153,00	153	161		161	3
4	157,00	161	169		169	9
5	164,00	169	177		177	25
6	164,00	177	185		185	15
7	165,00	185			other	5
8	166,00					
9	166,00					
10	166,50					
11	166,50					
12	167,00					
13	168,00					
14	170,00					
15	170,00					
16	170,00					
17	170,00					
18	171,00					
19	171,00					



Screen view 2/24. The graphic view of histogram in Excel

As shown by the screen view, frequencies of the class intervals (bins) are the same as above but the graphic illustration includes a column diagram which is not a histogram because there are gaps between the columns. To remove the space between them, right click a bar, select Format Data Series and change the Gap Width to 0%.



Screen view 2/25. Formatting data

The requested diagram is the following histogram.

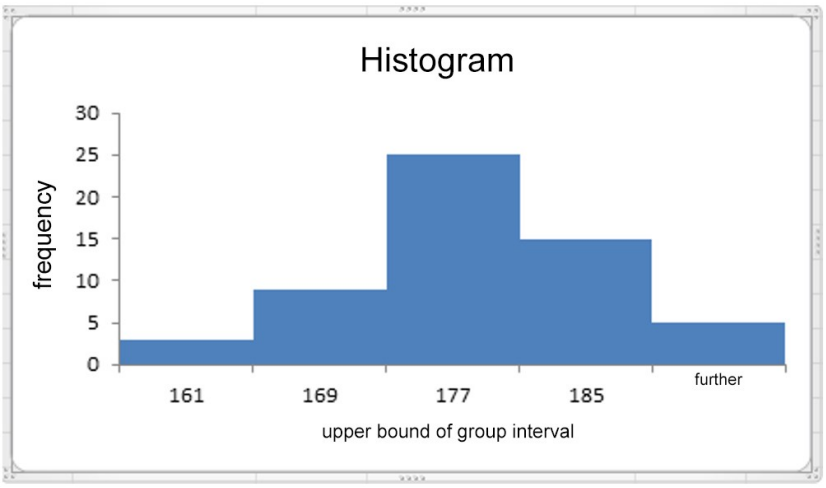
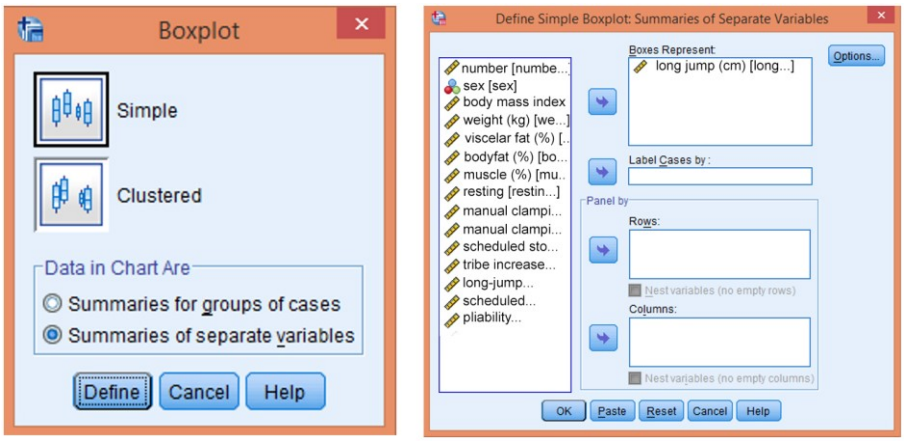


Figure 2/16. The histogram generated

"Box-plot" is a popular form of complex graphs because it displays the most important measures (mean, quartiles, range, outliers) of the quantitative data. It is only available in SPSS (Graph/ Boxplot). See the Boxplot of standing long-jump for illustration. (forrás: fittségi 57fő_adatbázis_alap.sav)



Screen view 2/26-27. Boxplot settings

Select "Simple" as type and "Summaries of separate variables" since data are not grouped. Press Define and add standing long-jump ("Helyből távolugrás") as target variable. Press OK to get the following screen.

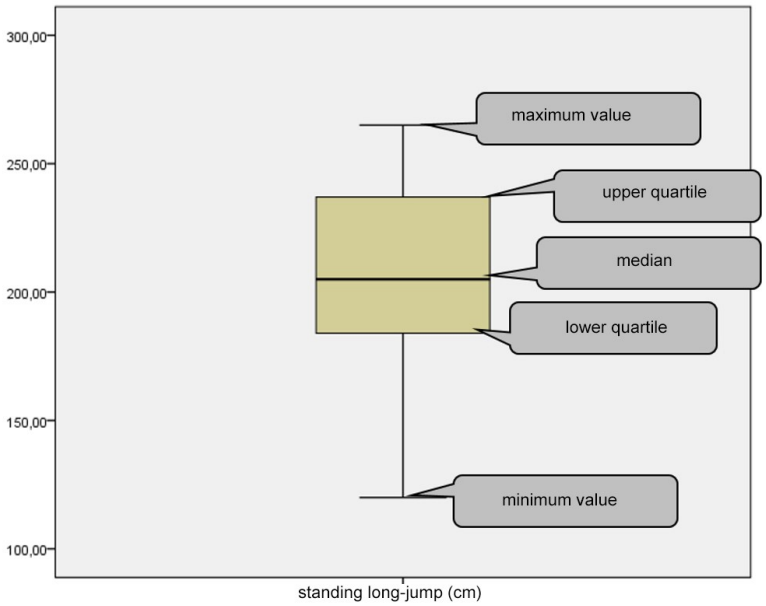


Figure 2/17. Box- plot

The Lorenz curve is another complex method of illustration which is often applied when analyzing concentration. In territorial concentration, the Lorenz curve is one of the most commonly used method. It is actually the graphical form of the concentration table.

It is a figure placed in a unit-length square, displaying the accumulated relative total value (z_i') as a function of accumulated relative frequencies (g_i'). It is not reasonable to display only one Lorenz curve for illustration and comparison since the extent of territorial disparity can not be determined with certainty.

The method is often applied since the phenomenon displayed in different times provides information on changes in territorial disproportion (concentration).

Before drawing a Lorenz curve, put data (in this case: regions) in increasing or decreasing order according to a given relative variable. If data are in increasing order, then the curve will appear under the diagonal, and it will be found over the line in case of decreasing ordering. Ács (2007) carried out a research on the territorial concentration of talented domestic athlets. He drew up the hypothesis that older athletes (Olympic athletes) show greater territorial concentration than young talents (Heracles athletes) which is mostly due to financial (salary) reasons.

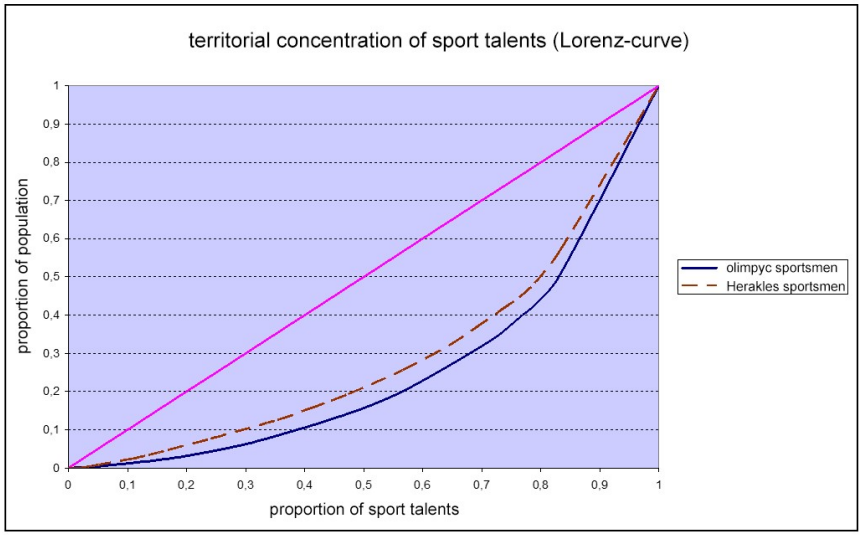


Figure 2/18. Lorenz curve

Source: Ács P. (2007)

Interpretation of the curve: if there exists a region which possesses of a high proportion of the examined variable's total sum, i.e. relative frequencies and total sums differ significantly, then the curve is far from the diagonal, and in case of total concentration, it is equal to the sides of the square.[21] If units possess of the same amount of the total sum, then the accumulated relative frequencies and accumulated relative total sums are equal ($g_i' = z_i'$). The curve is identical to the line in this case, showing the lack of concentration, i.e. there is absolute equality.

As shown by the figure, geographic concentration of Olympic athletes is greater. Concentration is illustrated well but no numeric data is provided on its measure. Statistical table is also a common tool for data presentation since it aims to order and summarize data.

Statistical table is a system of statistical data in which data are listed according to one or more variables. Statistical tables include time series, territorial, qualitative or quantitative variable. Tables used to be typified according to two aspects. According to the *number of dimensions* there can be two- or three-dimensional tables. Decision on this can be made based on the number of variables in the table. The other clustering is based on the aim of listing the variables since it can be done for comparison or grouping. Based on t*ype*, the following tables can be mentioned:

[21] Hajdu (1997)

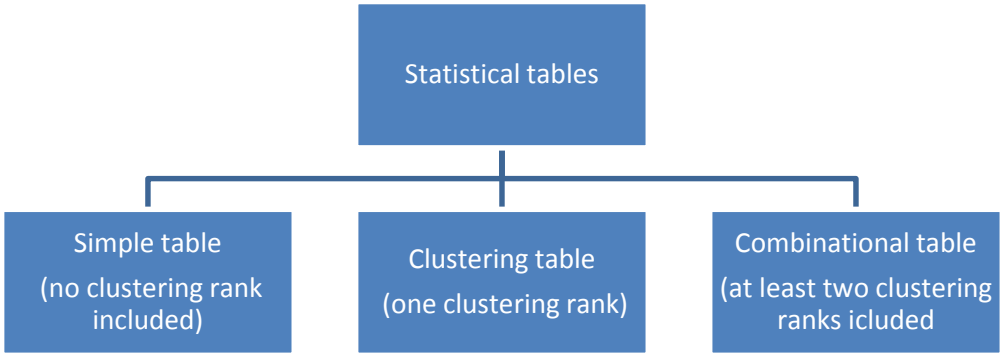


Figure 2/19. Clustering statistical tables

Most statistical tables are combinational ones. If the table contains frequencies, it is a **contingency table**. The next table is a three-dimensional contingeny table.

Table 2/11. Some team sports according to regions and sex (2005/2006)

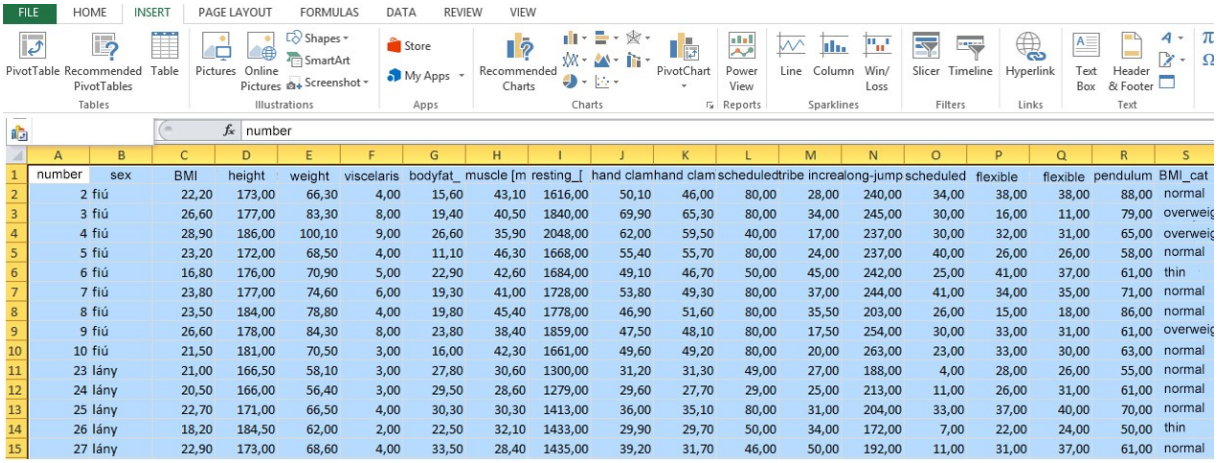
Region	Handball			Basketball			Volleyball			Total		
	m	fm	sum	m	fm	sum	m	fm	sum	m	fm	sum
Central Hungary	2	4	6	1	2	3	0	4	4	3	10	13
Central Transdanubia	3	2	5	1	0	1	2	0	2	6	2	8
Western Transdanubia	1	1	2	4	4	8	0	0	0	5	5	10
Southern Transdanubia	1	0	1	4	2	6	2	0	2	7	2	9
Northern Hungary	1	0	1	0	1	1	1	1	2	2	2	4
Northern Great Plain	2	1	3	3	1	4	2	2	4	7	4	12
Southern Great Plain	2	3	5	1	1	2	1	1	2	4	5	9
Total	**12**	**11**	**23**	**14**	**11**	**25**	**8**	**8**	**16**	**34**	**30**	**64**

Source: Pintér- Ács (2006)

There are crucial formal expectations related to statistical tables, the lack of which can reduce the niveau of the research (dissertation, thesis). These include: title, source, explanatory text. Expectations about the content: all units have to have an exclusive place where it can be ordered to, according to concerning data.

Statistical tables can be established in both programs. First, we present the option in Excel, creating a table from the fitness ("fittségi") database. (Source: fittségi 57fő_adatbázis_alap_bmikat. xlsx).

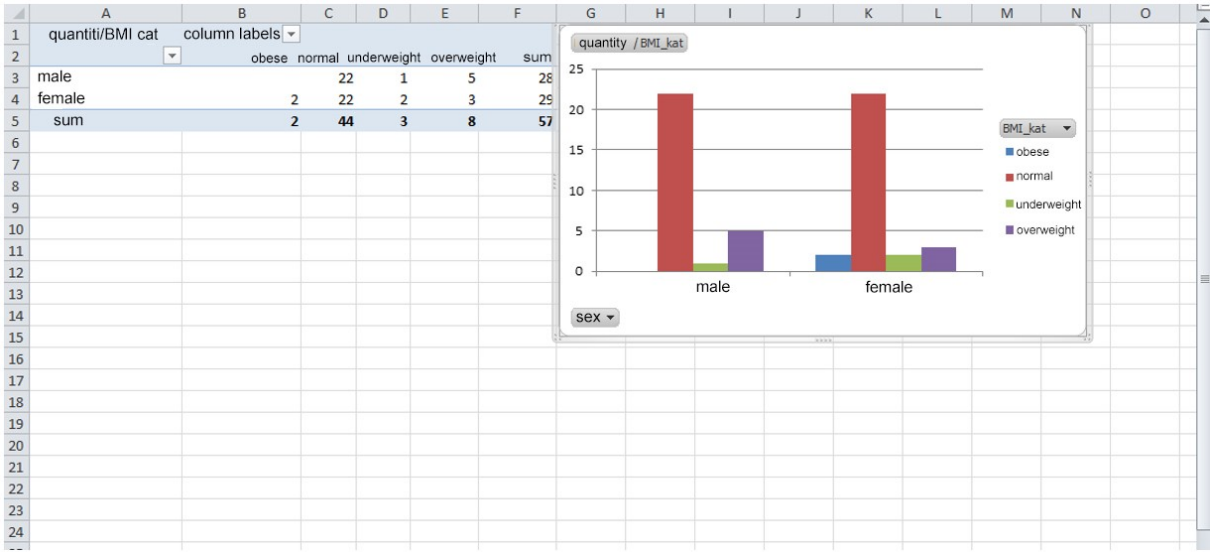
Let us insert a statistical table with the sex and BMI categories (underweight, normal, overweight, obese) of students. There are two options to create a pivot table from secondary data. Go to Insert to find the two options: PivotTable and PivotChart.



number	sex	BMI	height	weight	viscelaris	bodyfat_	muscle [m	resting_[	hand clam	hand clam	scheduledtribe	increalong-jump	scheduled	flexible	flexible	pendulum	BMI_cat
2	fiú	22,20	173,00	66,30	4,00	15,60	43,10	1616,00	50,10	46,00	80,00	28,00	240,00	34,00	38,00	38,00	88,00	normal
3	fiú	26,60	177,00	83,30	8,00	19,40	40,50	1840,00	69,90	65,30	80,00	34,00	245,00	30,00	16,00	11,00	79,00	overweig
4	fiú	28,90	186,00	100,10	9,00	26,60	35,90	2048,00	62,00	59,50	40,00	17,00	237,00	30,00	32,00	31,00	65,00	overweig
5	fiú	23,20	172,00	68,50	4,00	11,10	46,30	1668,00	55,40	55,70	80,00	24,00	237,00	40,00	26,00	26,00	58,00	normal
6	fiú	16,80	176,00	70,90	5,00	22,90	42,60	1684,00	49,10	46,70	50,00	45,00	242,00	25,00	41,00	37,00	61,00	thin
7	fiú	23,80	177,00	74,60	6,00	19,30	41,00	1728,00	53,80	49,30	80,00	37,00	244,00	41,00	34,00	35,00	71,00	normal
8	fiú	23,50	184,00	78,80	4,00	19,80	45,40	1778,00	46,90	51,60	80,00	35,50	203,00	26,00	15,00	18,00	86,00	normal
9	fiú	26,60	178,00	84,30	8,00	23,80	38,40	1859,00	47,50	48,10	80,00	17,50	254,00	30,00	33,00	31,00	61,00	overweig
10	fiú	21,50	181,00	70,50	3,00	16,00	42,30	1661,00	49,60	49,20	80,00	20,00	263,00	23,00	33,00	30,00	63,00	normal
23	lány	21,00	166,50	58,10	3,00	27,80	30,60	1300,00	31,20	31,30	49,00	27,00	188,00	4,00	28,00	26,00	55,00	normal
24	lány	20,50	166,00	56,40	3,00	29,50	28,60	1279,00	29,60	27,70	29,00	25,00	213,00	11,00	26,00	31,00	61,00	normal
25	lány	22,70	171,00	66,50	4,00	30,30	30,30	1413,00	36,00	35,10	80,00	31,00	204,00	33,00	37,00	40,00	70,00	normal
26	lány	18,20	184,50	62,00	2,00	22,50	32,10	1433,00	29,90	29,70	50,00	34,00	172,00	7,00	22,00	24,00	50,00	thin
27	lány	22,90	173,00	68,60	4,00	33,50	28,40	1435,00	39,20	31,70	46,00	50,00	192,00	11,00	31,00	37,00	61,00	normal

Screen view 2/28. Menu of statistical table (pivot)

Select “PivotChart” to get a diagram besides the table. The program asks if one would like to export data from Excel or an external source, and data range also has to be added. Leave the default settings unchanged and press OK.



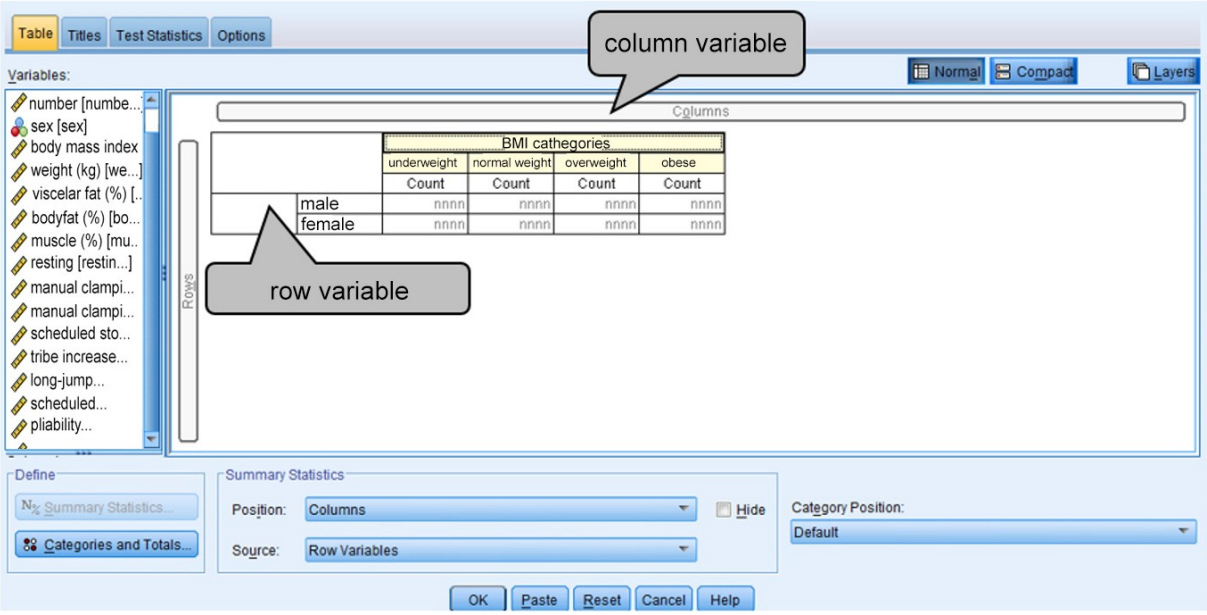
quantiti/BMI cat	column labels				
	obese	normal	underweight	overweight	sum
male		22	1	5	28
female	2	22	2	3	29
sum	2	44	3	8	57

Screen view 2/29. Pivot settings

First, add row labels and colimn labels by clicking on the proper variables in the four squares on the right-bottom corner. Add BMI categories (“BMI_kategóriák”) to Column Labels, Sex (“Nem”) to Row Labels, and BMI quantities (“Mennyiség”) to Values.

A table as well as a column diagram will be established, showing the number (value) of units in BMI categories according to sex.

The same option is available in SPSS under Analyze/Custom Tables. The first step is to move the variables to rows and columns. In continuation, we seethe following screen.



Screen view 2/30. Creating tables in SPSS

Select Table-margin totals to get totals of columns and rows. Select zero option in Format to put zeros instead of the missing values. Press Continue and Ok to get the requested statistical table.

Table 2/12. The statistical table created

		BMI cathegory			
		underweight	normal weight	overweight	obese
		Count	Count	Count	Count
sex	male	1	22	5	0
	female	2	22	3	2

Of course, the table can be formatted further by clicking twice with the left-hand side mouse button and once on the right-hand sided one. The number of dimensions and categories can be extended, too. One has to practice how to compute tables since only the basic settings have been presented here. In order to learn more about creating tables in SPSS, consult Ács: Data Analysis.

Basic table can be easily and quickly generated in both programs.

2.7.3. Analyzing two-variable 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. It can also make it easier for us to understand phenomena in the field of sport if we do not only look at them as separate entities, but we also examine their relationship with other phenomena.

We may examine the efficiency of a team – its scores achieved in the league – by the measures introduced before (e.g. mean, standard deviation) but it provides a more complex picture if we extend our examination with other influencing factors (variables). Adding the venue of the game (home field, away) to the analysis will probably result in a more precise summary since it is known that teams used to apply more offensive tactics at their home premises.

Phenomena and processes can be classified as follows:

- variables are ***independent*** from one another if there can be no consequences made based on one with respect 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.***
- In case of ***mixed association***[22], the cause is a qualitative, while the effect is a quantitative variable.
- If all (both) variables are quantitative then it is called ***correlation***.

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 of the index number is: $[-1;1]$.

[22] This expression does not exist in the English literature; it was translated from the Hungarian term "vegyes kapcsolat". – translator's note

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 2/13. Interpretation of association/correlation coefficient

$T = 0$	no association/correlation
$0 < T < 0.3$	weak association/correlation
$0.3 \leq T \geq 0.7$	moderate association/correlation
$0.7 < T < 1$	strong association/correlation
$T = 1$	deterministic (perfect association/correlation)

2.7.3.1. Association 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. If the variables are dichotomous or there are only two possible answers (alternatives) 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 2/14. General form of a two-dimensional contingency table

Versions of variable A	Versions of variable B		**Sum:**
	B_1	B_2	
A_1	f_{11}	f_{12}	S_1
A_2	f_{21}	f_{22}	S_2
Sum:	O_1	O_2	n

Source: author

n – the number of elements,

f_{11} – frequency of first group of variables A and B (the frequencies of the other cells can be interpreted similarly!),

S_1 – the first row is the sum of frequencies (belonging to the first group of variable A),

O_1 – 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 a completed survey[23] carried out by the Instite of Sport Sciences and Physical Education at the University of Pécs as commissioned by the State Secretariat of Sport of the Ministry of Local Government and Regional Development in 2008.

Data of the example come from the answers to the following two questions:

SEX (Please write an x in the proper box!)

Male ☐ Female ☐

HAVE YOU EVER BEEN ON A SLIMMING DIET? (Please write an x in the proper box!)

Yes ☐ No ☐

The results of the answers are summarized in the contingency table:

Table 2/15. Combinational table

Count

		Have you ever made a slimming diet?		total
		yes	no	
sex	male	53	261	314
	female	117	165	282
	total	170	426	596

Calculating the coefficient, we get the following result:

$$Y = \frac{53 \times 165 - 261 \times 117}{53 \times 165 + 261 \times 117} = -0.55$$

[23] The title of the survey, representative for Southern Transdanubia: **Phisycal activity and quality of life of pubescents.** Research team members: Dr. Erzsébet Rétsági, Zsuzsanna Pótó, Dr. Pongrác Ács.

The absolute value is between 0.3 and 0.7, so we found a moderate association between the two variables so as the number of respondents who had slimming diet is growing, the number of male respondents is shrinking. 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). 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 elements 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 * :

$$\frac{S_1 \times O_1}{n} = f_{11}^* \qquad \frac{S_1 \times O_2}{n} = f_{12}^*$$
$$\frac{S_2 \times O_1}{n} = f_{21}^* \qquad \frac{S_2 \times O_2}{n} = f_{22}^*$$

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 frequencies 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 χ^2 (chi square) value.

$$\chi^2 = \sum\sum \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}$$

χ^2 itself does not meet the requirements of measures of stochastic relationships. Its lower limit is zero but its upper limit may exceed 1 to a high degree. This dilemma is solved by the Cramer coefficient (Cramer's V), which can be calculated as:

$$C = \sqrt{\frac{\chi^2}{n \times (s-1)}}$$

where **s** means the minimum of the versions of variables (the number of less versions).

In the following, the calculation of the Cramer coefficient will be illustrated by the example introduced when preparing tables. (Source: fittségi 57fő_adatbázis_alap_bmikat.xlsx)

Table 2/16. Basic data of the contingency table:

sex	BMI cathegories				
	obese	normal	under-weight	over-weight	total
male	0	22	1	5	28
female	2	22	2	3	29
total	**2**	**44**	**3**	**8**	**57**

Source: author's own survey

The frequencies under the assumption of independence with marginal totals:

$$\frac{28 \times 2}{57} = 0.98 \quad \frac{28 \times 44}{57} = 21.61 \quad \text{etc.}$$

Frequencies under the assumption of independence are listed in the following table:

Table 2/17 Frequencies under the assumptions of independence

sex	BMI cathegories				
	obese	normal	under-weight	over-weight	total
male	0,98	21,61	1,47	3,93	28
female	1,02	22,39	1,53	4,07	29
total	**2**	**44**	**3**	**8**	**57**

First, the relative frequencies need to be calculated in every cell:

$$\frac{(0-0.98)^2}{0.98} = 0.98 \text{ etc.}$$

Here is the newly generated contingency table:

Table 2/18. Calculation of χ^2 values

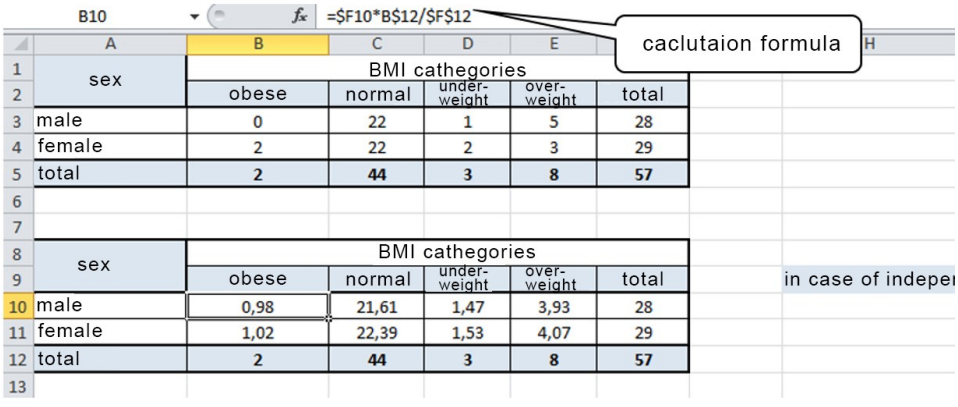
sex	BMI cathegories				
	obese	normal	under-weight	over-weight	total
male	0,98	0,01	0,15	0,29	28
female	0,95	0,01	0,15	0,28	29
total	**2**	**44**	**3**	**8**	**57**

Cramer's V in our example can be calculated as follows:

$$C = \sqrt{\frac{2.82}{57 \times (2-1)}} = 0.22$$

Based on the coefficient, the stochastic relationship between sex and BMI categories is weak. The measure C^2 can also be interpreted. It shows that the type of sex determine BMI categories to an extent of 4.94%.

The Cramer coefficient can be calculated – from the contingency table – in Excel as follows.. (Source: fittségi 57fő_adatbázis_alap_bmikat.xlsx)



Screen view 2/31. Calculating the Cramer coefficient in Excel

Applying the formulas in the calculations before, the fictitious frequency table (assuming independency) wil be generated from the table of observed frequencies. Only the first frequency will be calculated by the formula, for the next ones, we will copy the formula. This solution also provides an excellent opportunity to practice absolute and relative references. Fictitious frequencies will be gained through marginal frequencies which will

show the types of variables in the phenomenon. Marginal frequencies are to be found in the "total" rows that is why the references "column-absolute" and "row-absolute" have to appear in the formula. Although, the sample number appears in the formula with a reference to "absolute-absolute". The formula is to be found in the editor.

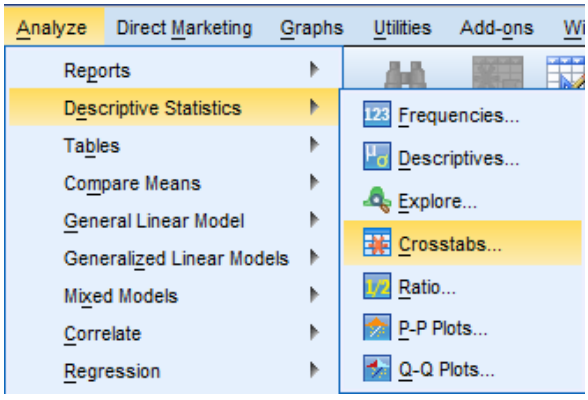
As shown, the fictitious frequencies calculated like this are not equal (but similar in measures) to the observed (valid) ones (neither have we found independency nor deterministic relationship) so there must be a stochastic relationship. Finally, we compare the valid and fictitious frequencies as a result of which we get the squared contingency measure making it easy to calculate the Cramer coefficient.

sex	BMI cathegories				
	obese	normal-weight	under-weight	over-weight	total
male	0,98	0,01	0,15	0,29	28
female	0,95	0,01	0,15	0,28	29
total	2	44	3	8	57
chi square					
Cramer	0,22				
C^2	4,94%				

x^2 squared contingency measure, relative

Screen view 2/32. Calculating squared contingency measure in Excel

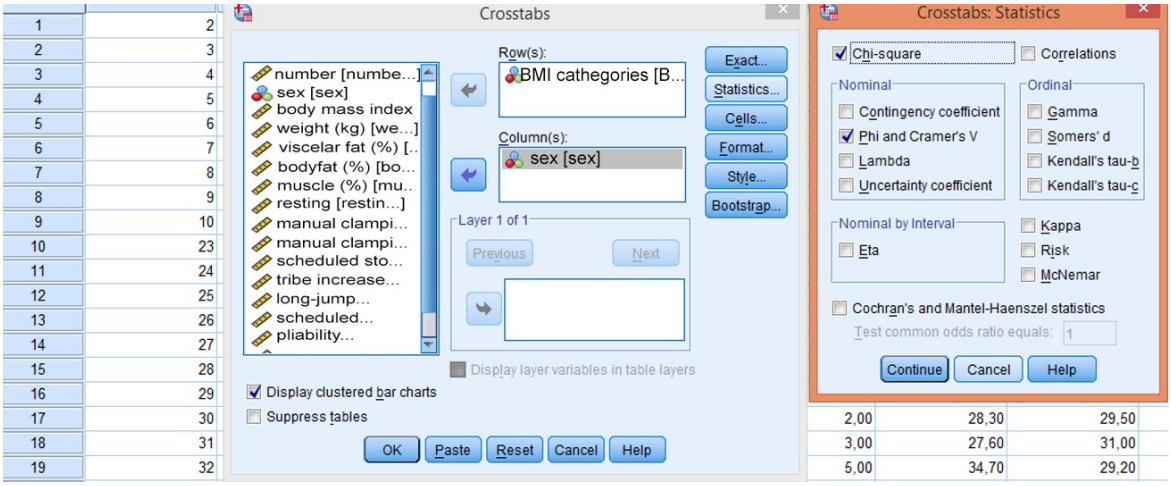
Association can be analysed in module CROSSTABS of SPSS, after choosing ANALYSE / DECRIPTIVE STATISTICS. (Source: fittségi 57fő_adatbázis_alap_bmikat.sav)



Screen view 2/33. 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. We may suggest that 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).



Screen view 2/34. Associtaion (crosstab) settings in SPSS

First, the row and column variables have to be added, and then we get a lot of options in Statistics since measures of different scale types appear. In our case, it will be enough to select Chi-square and the Phi and Cramer's V coefficients. We give a short summary on the additional measures in order to help users.

- 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 (Φ): used in the special case of crosstabs, in case of 2x2 tables, since it is easily 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 2/19. Measures of association in case of ordinal scales

Measure (for ordinal scales)	**Types of tables**
Gamma	For every type of tables and size (easy to interpret)
Sommers' d	For every type of tables and size (not easy to interpret)
Kendall tau-b	In case of symmetric tables
Kendall tau-c	In case of asymmetric tables

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[24]**:** 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 (γ>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.
- Kappa (KAPPA)[25] is a measure of agreement that measures the agreement of values (raters). Base on Landis-Koch (1977), it can be interpreted as:

[24] Is is often referred to as Goodman and Kruskal's Gamma.

[25] It is often referred to as Cohen's Kappa.

Table 2/20. Interpretation of Kappa

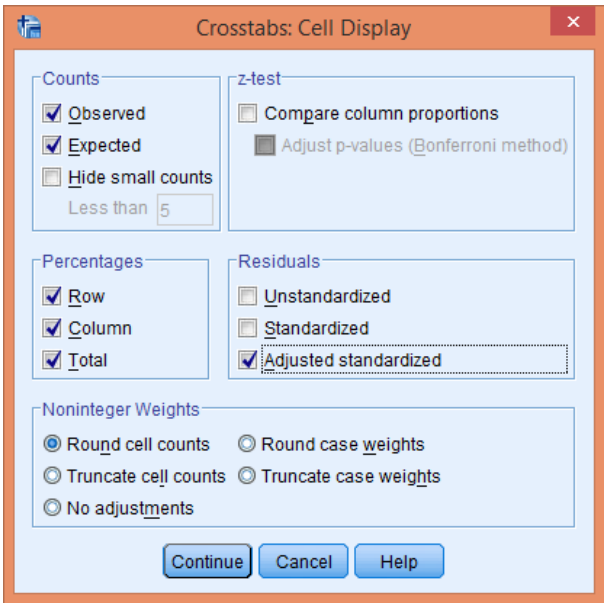
κ	**Interpretation**
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Source: Jánosa, 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 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)* is 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. In sum, 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.
- 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.

After changing the settings, press OK and select Cells.



Screen view 2/35. Association (crosstab) settings in SPSS (Cells)

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.

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 2/21. Case processing summary

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
sex * BMI categories	57	100,0%	0	0,0%	57	100,0%

The sample size is 57, 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 Nem)
- Percentage of total sample (% OF TOTAL)
- Standardized adjusted residual (ADJUSTED RESIDUAL)

The next table contains the Pearson Chi-Square and a measure equal to the square contingency we calculated before.

Table 2/22. Calculated chi-square values

BMI cathegories * Nem Crosstabulation

			sex		
			male	female	Total
BMI cathegories	underweight	Count	1	2	3
		Expected Count	1,5	1,5	
		% within BMI kategóriák	33,3%	66,7%	
		% within Nem	3,6%	6,9%	
		% of Total	1,8%	3,5%	
		Adjusted Residual	-,6		
	normal weight	Count	22	22	44
		Expected Count	21,6	22,4	44,0
		% within BMI kategóriák	50,0%	50,0%	100,0%
		% within Nem	78,6%	75,9%	77,2%
		% of Total	38,6%	38,6%	77,2%
		Adjusted Residual	,2	-,2	
		Count	5	3	8
		Expected Count	3,9	4,1	8,0
		% within BMI kategóriák	62,5%	37,5%	100,0%
		% within Nem	17,9%	10,3%	14,0%
		% of Total	8,8%	5,3%	14,0%
		Adjusted Residual	,8	-,8	
	obese	Count	0	2	2
		Expected Count	1,0	1,0	2,0
		% within BMI kategóriák	0,0%	100,0%	100,0%
		% within Nem	0,0%	6,9%	3,5%
		% of Total	0,0%	3,5%	3,5%
		Adjusted Residual	-1,4	1,4	
Total		Count	28	29	57
		Expected Count	28,0	29,0	57,0
		% within BMI kategóriák	49,1%	50,9%	100,0%
		% within Nem	100,0%	100,0%	100,0%
		% of Total	49,1%	50,9%	100,0%

From all with normal BMI, 50% (22/44) are male

78.6% (22/28) of male belong to the normal BMI category

38.6% (22/57) of the sample is male and belongs to the normal BMI category

If the measure is above +2, then there is a significant association for sure

Table 2/23. Calculated chi-square values

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,817[a]	3	,421
Likelihood Ratio	3,600	3	,308
Linear-by-Linear Association	,040	1	,842
N of Valid Cases	57		

a. 6 cells (75,0%) have expected count less than 5. The minimum expected count is ,98.

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 (χ^2), and it has to be compared to the theoretical one in order to decide if the null hypothesis should be accepted or rejected.[26] The table of χ^2 distribution provides a basis for comparison for a given degree of freedom (3) and error probability (0.05), its value in our case is 7.81 (in Excel: =inverz.khi (0.05;3)). As the

[26] For more details see following chapters.

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 sex of students has no connection with the BMI categories so the sex does not determine the BMI categories. The same consequences can be made based on the tables about significance (*DIRECTIONAL MEASURES, SYMMETRIC MEASURES*) since the value is higher than the 5% we picked.

Table 2/24. Calculated measures of correlation

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,222	,421
	Cramer's V	,222	,421
N of Valid Cases		57	

The Cramer coefficient suggests that there is a weak relationship between the two variables.

The Phi coefficient cans generally reasonably applied for alternative variables since the upper limit will be 1 in this case only – otherwise there is no limit, and interpretation is problematic. The value of the Phi coefficient is the chi-square value modified by the sample size (N): $\phi = \sqrt{\frac{\chi^2}{N}}$

When generating crosstabs, the programme offers the opportunity to display charts (*DISPLAY CLUSTERED BAR CHARTS*) as well. As default, the association is illustrated on a bar chart.

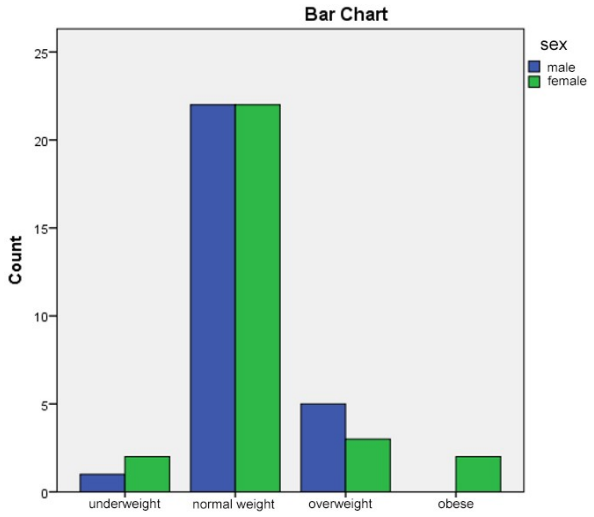


Figure 2/20. Graphic illustration of the BMI categories

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.

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 allow easy interpretation. Categories similar to one other will be located close to one another also in the graphic 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).

Let us find out if there is an association between women's progressive shuttle categories and BMI categories, and display the coherent categories in a graph. (Source: fittségi 57fő_adatbázis_alap_bmikat.sav). Categories have been calibrated based on NEFTIT.

For illustration, let us decide if there is a relationship between the BMI categories and the level of qualification.

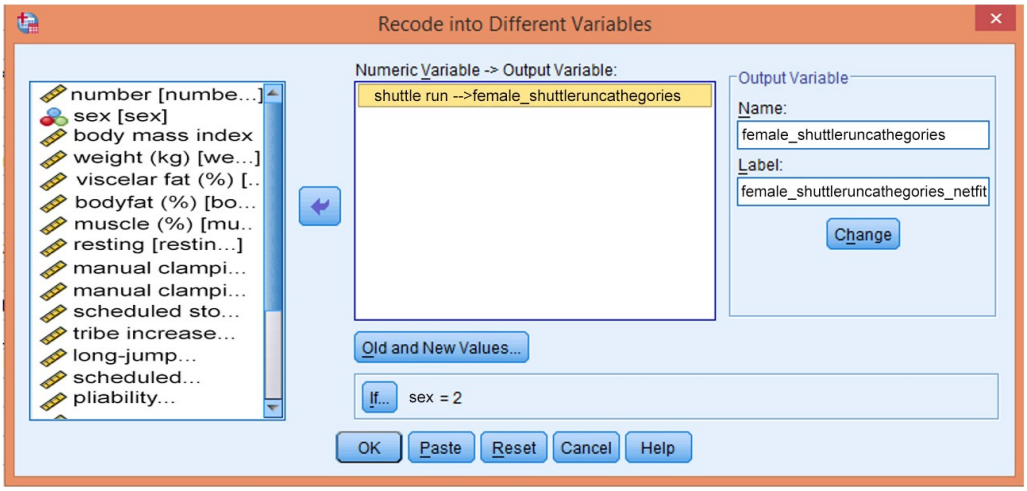
The access path to this method is the following: *DATA REDUCTION / CORRESPONDENCE ANALYSIS*.

Table 2/25. BMI and progressive shuttle run categories of adult women

BMI					20 meters progressive shuttle run			
	Women					Women		
Age	*under-weight*	*healthy*	*developement needed*	*intensive development needed*	*Age*	*healthy*	*developement needed*	*intensive developme nt needed*
18-	<18.5	18.6-24.9	25.0-29.9	30.0<	**18-**	38<	29-37	<28

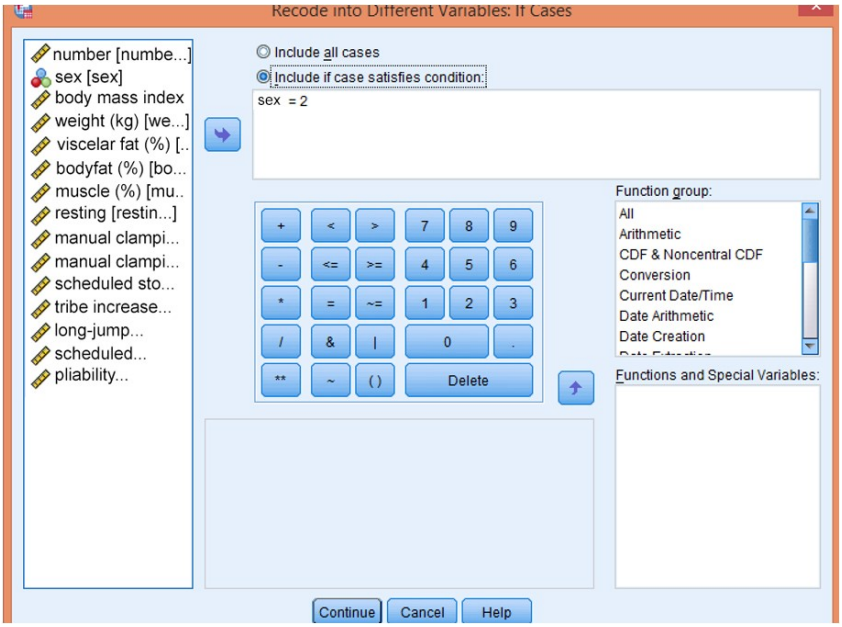
In order to be able to carry out the correspondence analysis, continuous variables have to be converted into categorical ones. As our database contains the BMI categories, it is enough to determine the categories from the 20 m progressive shuttle run results of women above the age 18. This can be done under TRANSFORM / RECODE INTO DIFFERENT VARIABLES. Next, a new variable (women_shuttlecategories, lány_ingakategóriák) has to be created with an already existing one (progressive shuttle run, Ingafutás). To do so, the existing variable has to be moved from the left box to the Numeric Variable box with the arrow. Name the new variable (lány_ingakategóriák) and label it (lány_ingafutás Netfit

categories). After naming the new variable, press CHANGE to see the new variable name next to the old one in the box in the middle.



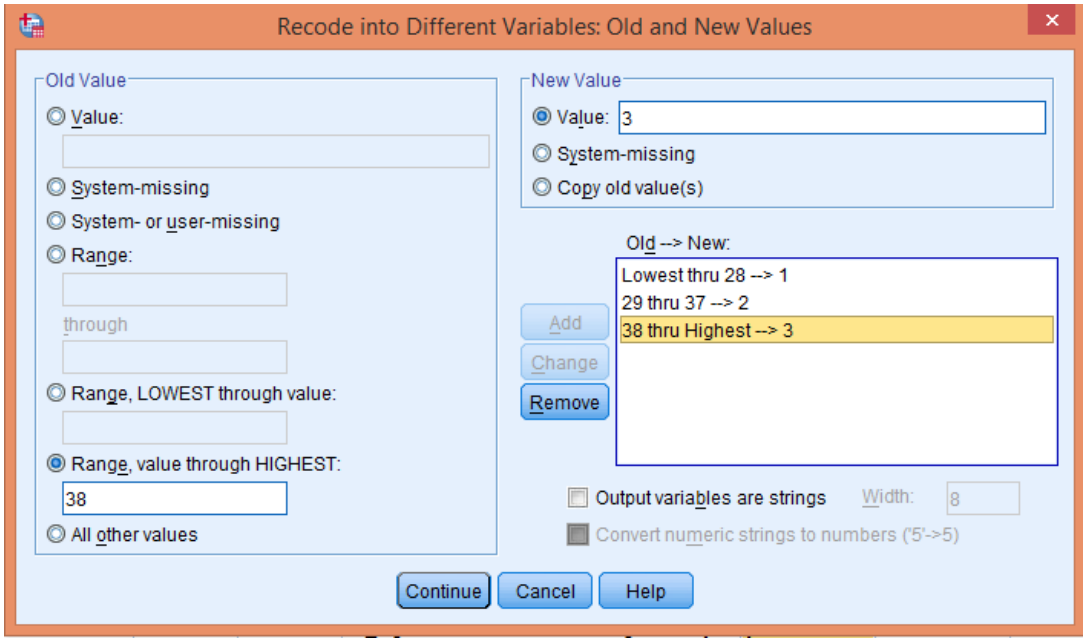
Screen view 2/35. Decoding the variable progressive shuttle run

Use filter to make sure that the categorization will be valid for women only. Filtering can be done by pressing IF.



Screen view 2/36. Selecting filters

Select “Include if case satisfies condition” and move the variable Sex (“Nem”) to the box in the middle. Set that from the variable Sex only women (2) will be taken into account. Press continue and select value ranges with the Old and New Values option.



Screen view 2/37. Adding range

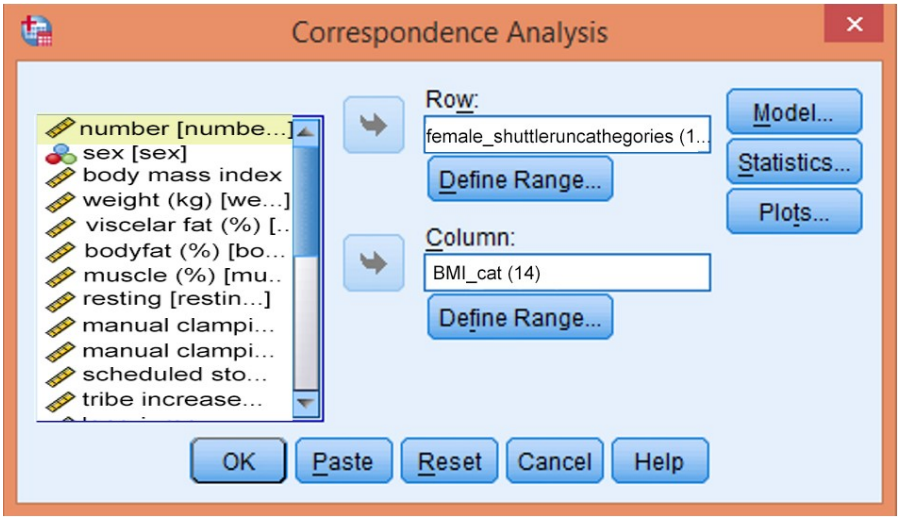
The pop-up window can be separated into two parts because original values are listed on the left hand side, while new ones are on the right one. Selecting the option VALUE, one can add the old values one by one. The options SYSTEM-MISSING or SYSTEM -OR USER- MISSING can exclude items that do not meet 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 29 and 37). 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 38, items get code 3). 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 three categories will become valid after clicking on CONTINUE and OK.

As the three categories only contain one numerical value, the categories can be named in variable view, at VALUE LABELS.

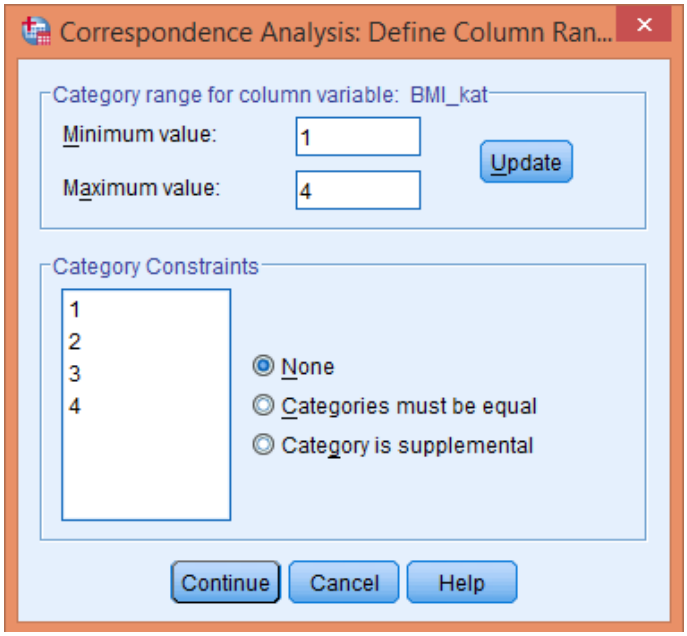
Screen view 2/38. Naming and labelling new groups

When modifying data, it is often not necessary to generate a new variable but it suffices to 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. Correspondence analysis can be carried out after these data transformations. This method is available under DATA REDUCTION/ CORRESPONDENCE ANALYSIS. (Source: fittségi 57fő_adatbázis_alap_bmikat.sav).



Screen view 2/39. Correspondence Analysis settings in SPSS

First, select the row and column variables, then define all variables with the number of their versions.



Screen view 2/40. Setting the number of variable versions

In this case, BMI categories have been defined from 1 to 4, depending on the versions of variables. After defining both variables, press OK. Results are listed in the next table.

The first table (Correspondence Table) contains the actual frequencies; the second one lists the results. The association is significant (p=0.00) and the two new dimensions can be displayed graphically since 100% of variance is explained. The next two tables contain coordinates of the versions of variables according to the two default variables. Graphic illustration (Biplot) makes coherent values visible in two dimensions.

Table 2/25. Correspondence Table

Correspondence Table

female_shuttlerun_netfit cathegories	BMI cathegories				
	underweight	normal weight	overweight	obese	Active Margin
increased development needed	0	1	1	2	4
development needed	0	2	2	0	4
healthy zone	2	19	0	0	21
Active Margin	2	22	3	2	29

Table 2/26. Summary of Chi-Square

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	,778	,605			,712	,712	,124	,090
2	,495	,245			,288	1,000	,231	
Total		,850	24,653	,000[a]	1,000	1,000		

a. 6 degrees of freedom

Table 2/27. Progressive shuttle run results for two dimensions

Overview Row Points[a]

female_shuttlerun_netfit cathegories	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
					1	2	1	2	Total
increased development needed	,138	1,961	-,804	,457	,682	,180	,903	,097	1,000
development needed	,138	,681	1,673	,241	,082	,780	,207	,793	1,000
healthy zone	,724	-,503	-,166	,152	,236	,040	,936	,064	1,000
Active Total	1,000			,850	1,000	1,000			

a. Symmetrical normalization

Table 2/28. Result of BMI categories in case of two dimensions

Overview Column Points[a]

BMI cathegories	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
					1	2	1	2	Total
underweight	,069	-,647	-,334	,026	,037	,016	,855	,145	1,000
normal	,759	-,365	-,055	,080	,130	,005	,986	,014	1,000
overweight	,103	1,424	1,712	,313	,270	,612	,521	,479	1,000
obese	,069	2,521	-1,624	,431	,564	,367	,791	,209	1,000
Active Total	1,000			,850	1,000	1,000			

a. Symmetrical normalization

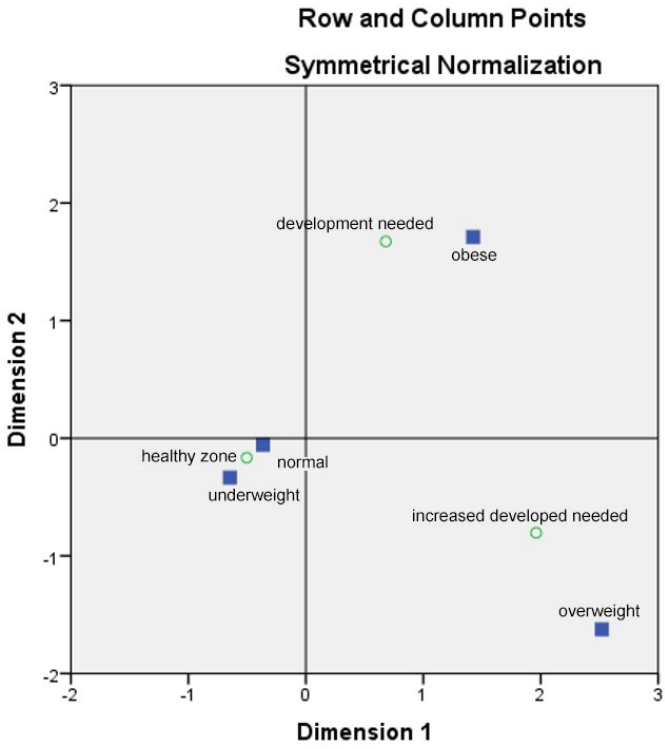


Figure 2/21. Two-dimensional graphic illustration of the correspondence analysis

From the graph it is easy to see which values belong together. Healthy zone rather includes students with normal and underweigh BMI while increased development of standing power is suggested for units from the BMI category of obese.

Naming dimensions is of high importance since it aims to help understanding which is the task of the researcher. Naming of the dimensions in this example is left to the reader' choice.

2.7.3.2. Mixed association

As stated before, in mixed associations, 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 connection between squares of standard deviations is the partition of the square. 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. It can be seen easily that 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 with 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:

$$H = \frac{\sigma_K}{\sigma} = \sqrt{\frac{\sigma_K^{\ 2}}{\sigma^2}} = \sqrt{1 - \frac{\sigma^2{}_B}{\sigma^2}}$$

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:

$$H^2 = \frac{SS_K}{SS}$$

Let us consider the following example, based on data from the webpage www.nemzetsport.hu (2nd August 2008). We examine to what extent the type of sport determines the number of members (how many sportsmen are needed), i.e. if there is an

association between the type of sport and number of members. (Source: vegyes kapcsolat.xls)

Table 2/29. Basic data of the mixed association

Team	Sport	Members
Siófok	Soccer	26
Budapest Honvéd	Soccer	22
DVSC	Soccer	35
Paks Fc	Soccer	24
MTK	Soccer	22
Kaposvár	Soccer	21
Pick- Szeged	Handball	18
MKB Veszprém Kc	Handball	16
Komlói KBSK	Handball	17
Dunaferr SE	Handball	18
Albacomp	Basketball	12
Atomerőmű SE	Basketball	12
Falco KC	Basketball	13
PVSK Expo Center	Basketball	14
BVSC	Waterpolo	18
Domino BHSE	Waterpolo	15
PVSK- Füszért	Waterpolo	16

Data arranged in groups can be found in following table:

Table 2/30. Grouped data

Sport	Number of teams examined (f_i)	Means of members	Std. deviation of members
Soccer	6	25,00	4,76
Handball	4	17,25	0,83
Basketball	4	12,75	0,83
Waterpolo	3	16,33	1,25
total	17		

The main average of the 17 observables can (also) be calculated with the group means (weighted average):

$$\bar{x} = \frac{6*25 + 4*17.25 + 4*12.75 + 3*16.33}{17} = 18.76 \text{ ppl}$$

Internal variance and standard deviation can be calculated as follows (using groups' standard deviation):

$$\sigma_B^2 = \frac{6\times 4.76^2 + 4\times 0.83^2 + 4\times 0.82^2 + 3\times 1.25^2}{17} = 8.60$$

$$\sigma_B = \sqrt{8.60} = 2.93$$

External variance and standard deviation:

$$\sigma_K^2 = \frac{6\times(25-18.76)^2 + 4\times(17.25-18.76)^2 + 4\times(12.75-18.76)^2 + 3\times(16.33-18.76)^2}{17} = 23.82$$

$$\sigma_K = \sqrt{23.82} = 4.88$$

Variance and standard deviation of the whole population (using the additive function):

$$\sigma^2 = 8.6 + 23.82 = 32.42$$

$$\sigma = \sqrt{32.42} = 5.69$$

Total standard deviation of members: 5.69 people. The internal standard deviation shows that the standard deviation of members differs by 2.93 on average from the mean of its group (the type of sport). The external standard deviation shows that the average member data by sports differ by 4.88 people on average from the average of all the teams involved in the survey (main average).

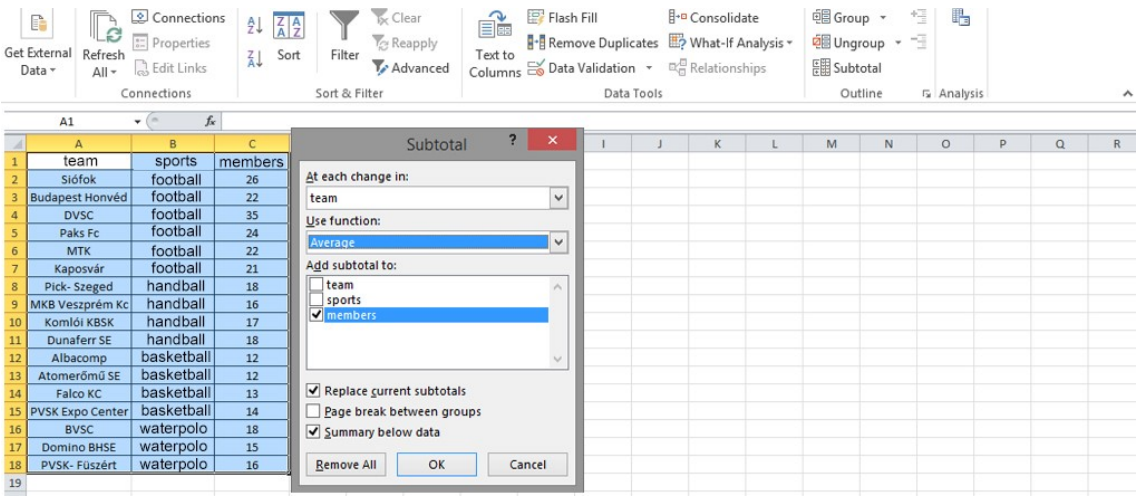
Based on the data above, the strength of the mixed association can be calculated with the standard deviation ratio.

$$H^2 = \frac{23.82}{32.42} = 0.73$$

$$H = \frac{4.88}{5.69} = \sqrt{\frac{23.82}{32.42}} = \sqrt{1 - \frac{8.6}{32.42}} = 0.86$$

The standard deviation ratio shows that there is a strong association between the type of sport and the number of team members. According to the variance ratio, the type of sport explains 73% of the number of members.

Strength of mixed association can be calculated in Excel as follows. First, sub-means have to be calculated. The easiest way to achieve this is by selecting the followings in Data/ Sub-totals:



	A	B	C
1	team	sports	members
2	Siófok	football	26
3	Budapest Honvéd	football	22
4	DVSC	football	35
5	Paks Fc	football	24
6	MTK	football	22
7	Kaposvár	football	21
8	Pick- Szeged	handball	18
9	MKB Veszprém Kc	handball	16
10	Komlói KBSK	handball	17
11	Dunaferr SE	handball	18
12	Albacomp	basketball	12
13	Atomerőmű SE	basketball	12
14	Falco KC	basketball	13
15	PVSK Expo Center	basketball	14
16	BVSC	waterpolo	18
17	Domino BHSE	waterpolo	15
18	PVSK- Füszért	waterpolo	16

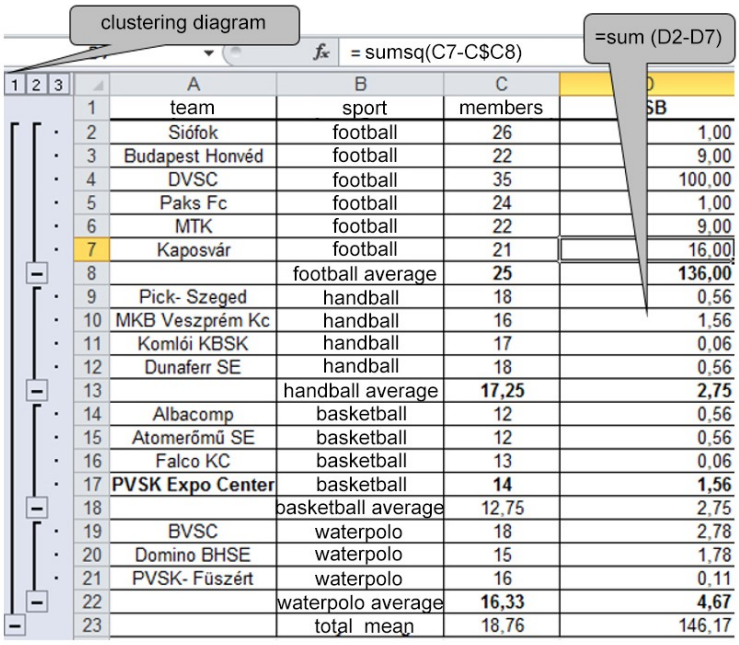
Screen view 2/41. Calculating mixed association in Excel

The name of the sport is the grouping variable in this case. The characteristic value is the average to be calculated for membership data. As a result, the program calculates the average of all groups which will be displayed after all groups.

	A	B	C	D
1	team	sport	members	
2	Siófok	football	26	
3	Budapest Honvéd	football	22	
4	DVSC	football	35	
5	Paks Fc	football	24	
6	MTK	football	22	
7	Kaposvár	football	21	
8		football average	25	
9	Pick- Szeged	handball	18	
10	MKB Veszprém Kc	handball	16	
11	Komlói KBSK	handball	17	
12	Dunaferr SE	handball	18	
13		handball average	17,25	
14	Albacomp	basketball	12	
15	Atomerőmű SE	basketball	12	
16	Falco KC	basketball	13	
17	PVSK Expo Center	basketball	14	
18		basketball average	12,75	
19	BVSC	waterpolo	18	
20	Domino BHSE	waterpolo	15	
21	PVSK- Füszért	waterpolo	16	
22		waterpolo average	16,33333333	
23		total average	18,76470588	
24				
25				

Screen view 2/42. Mixed association calculation in Excel

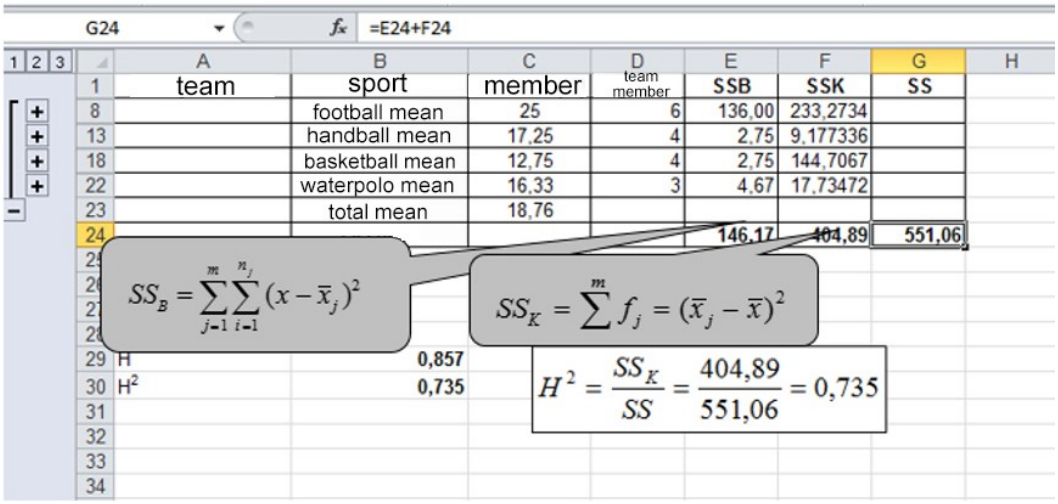
Next, the internal and external sum of squares and their sum have to be calculated. Create a column containing the number of members – you can do this with the function “count”. Display the results in the row of the means of sports. The next columns contain sums of squares, calculated with the SUMSQ function.



	A	B	C	D
1	team	sport	members	SB
2	Siófok	football	26	1,00
3	Budapest Honvéd	football	22	9,00
4	DVSC	football	35	100,00
5	Paks Fc	football	24	1,00
6	MTK	football	22	9,00
7	Kaposvár	football	21	16,00
8		football average	**25**	**136,00**
9	Pick- Szeged	handball	18	0,56
10	MKB Veszprém Kc	handball	16	1,56
11	Komlói KBSK	handball	17	0,06
12	Dunaferr SE	handball	18	0,56
13		handball average	**17,25**	**2,75**
14	Albacomp	basketball	12	0,56
15	Atomerőmű SE	basketball	12	0,56
16	Falco KC	basketball	13	0,06
17	**PVSK Expo Center**	basketball	**14**	**1,56**
18		basketball average	12,75	2,75
19	BVSC	waterpolo	18	2,78
20	Domino BHSE	waterpolo	15	1,78
21	PVSK- Füszért	waterpolo	16	0,11
22		waterpolo average	**16,33**	**4,67**
23		total mean	18,76	146,17

Screen view 2/43. Calculating the internal sum of squares

When copying data, make sure that you use absolute-absolute references for sum of squares functions. Next, sum all internal sum of squares one by one. In order to get sub-data, press “grouping diagram”.

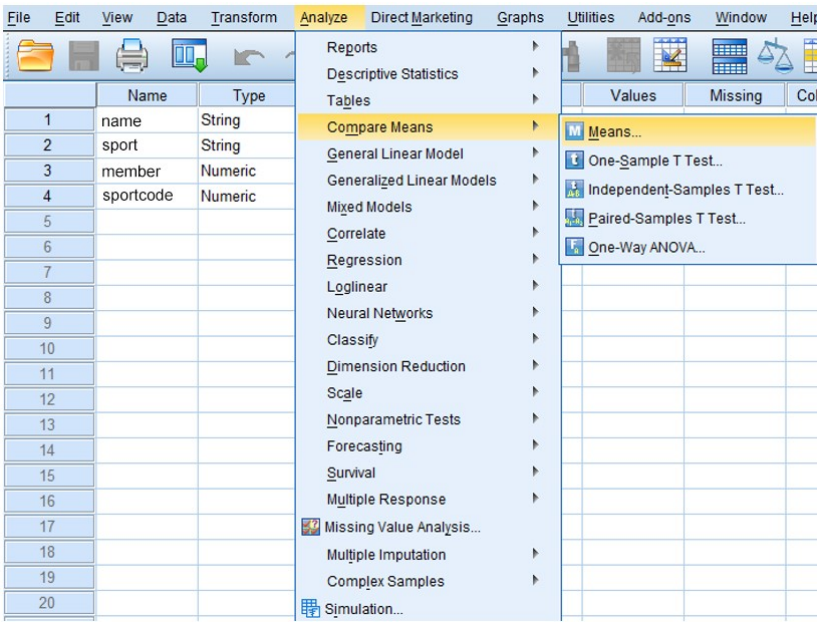


	A	B	C	D	E	F	G
1	team	sport	member	team member	SSB	SSK	SS
8		football mean	25	6	136,00	233,2734	
13		handball mean	17,25	4	2,75	9,177336	
18		basketball mean	12,75	4	2,75	144,7067	
22		waterpolo mean	16,33	3	4,67	17,73472	
23		total mean	18,76				
24					146,17	404,89	551,06
29	H		0,857				
30	H^2		0,735				

Screen view 2/44. Mixed association results in Excel

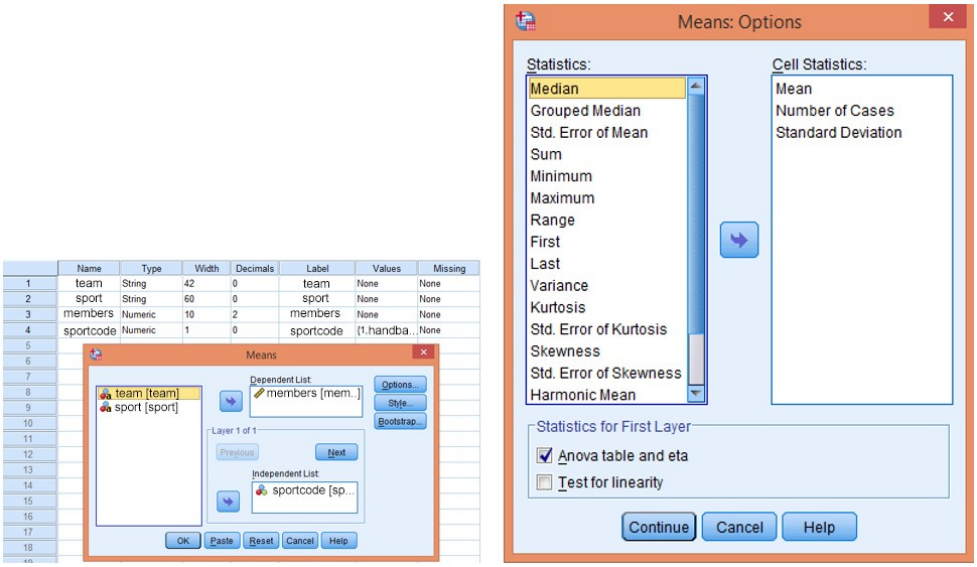
The variance ratio calculated by the program (0.73) slightly differs from the results above because of rounding.

Using SPSS makes the researcher’s task even easier since settings leading to the results can be done quickly. (Source: vegyes kapcsolat.sav)



Screen view 2/45. Calculating mixed association

Go to ANALYSE / COMPARE MEANS / MEANS, add dependent and independent variables, then click options to select statistics to be calculated.



Figures 2/46. and 2/47. Mixed association settings in SPSS

The most important thing is to select ANOVA TABLE AND ETA. Eta (η) can be calculated the following way (as we already know):

$$\eta = \sqrt{\frac{\sigma^2{}_K}{\sigma^2}}$$

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 2/31. Basic statistics of the types of sports (mean, number of elements, standard deviation)

Report

members

sport	Mean	N	Std. Deviation
handball	17,2500	4	,95743
basketball	12,7500	4	,95743
football	25,0000	6	5,21536
waterpolo	16,3333	3	1,52753
Total	18,7647	17	5,86866

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 2/32. ANOVA Table

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
members *sport	Between Groups	(Combined)	404,892	3	134,964	12,004	,000
	Within Groups		146,167	13	11,244		
	Total		551,059				

internal sum of squares

external sum of squares

As follows we do also get the measures of association.

Table 2/33. Measures of Association

Measures of Association

	Eta	Eta Squared
number * sport	,857	,735

H

H^2

2.7.3.3. Correlation analysis

If both the causes and the effects are quantitative variables, then the relationship is called **correlation**. In the followings we will introduce 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 see 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:

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}, where\ \mathrm{cov}ariance\ is\ C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)(y_i - y) = \frac{\sum d_x d_y}{n} = \frac{\sum xy}{n} - xy$$

where σ_X and σ_Y are the standard deviations of the variables.

Now, we examine the strength of correlation in an example of BMI measures and the 20 m progressive shuttle run. First, we plot the results in a scatter chart. (Source: fittségi 57fő_adatbázis_alap_bmikat.xlsx)

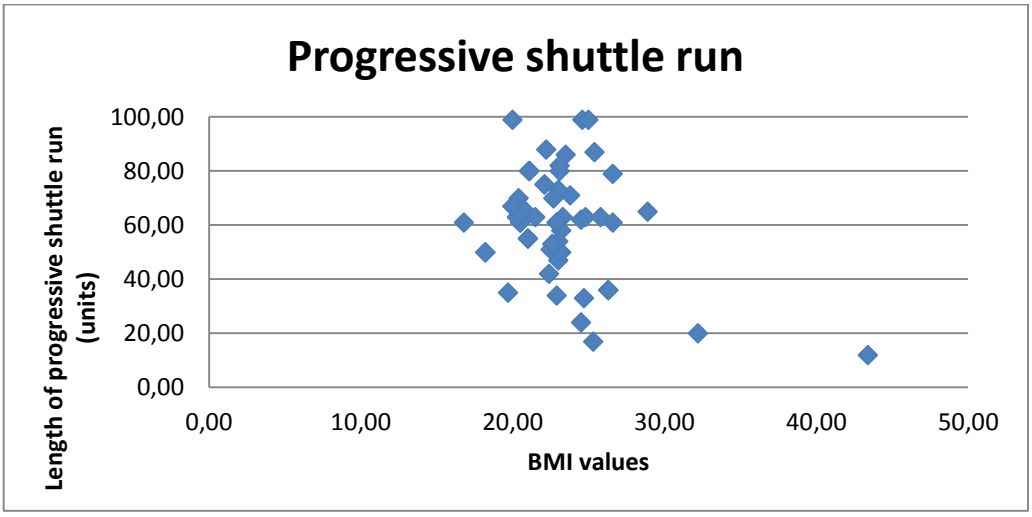


Figure 2/22. Scatter chart

Based on the chart, a negative linear correlation can be expected. As the BMI value is increasing the number of runned lengths is decreasing. The following working table contains data and calculations needed to determine the strength of correlation:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	number	BMI	shuttle run	dx_i	dy_i	$dx_i dy_i$							
2		22,20	88,00	-0,97	27,02	-26,16							
3		26,60	79,00	3,43	18,02	61,83		number	BMI	shuttle run	dx_i	dy_i	$dx_i dy_i$
4		28,90	65,00	5,73	4,02	23,03		Összeg	1320,6	3476			-1547,93
5		23,20	58,00	0,03	-2,98	-0,09		Átlag	23,16842	60,98246			-27,1567
6		16,80	61,00	-6,37	0,02	-0,11		Szórás	3,757386	19,53359			0
7		23,80	71,00	0,63	10,02	6,33							
8		23,50	86,00	0,33	25,02	8,30							
9		26,60	61,00	3,43	0,02	0,06							
10		21,50	63,00	-1,67	2,02	-3,37							
11		21,00	55,00	-2,17	-5,98	12,97							
12		20,50	61,00	-2,67	0,02	-0,05							
13		22,70	70,00	-0,47	9,02	-4,22							
14		18,20	50,00	-4,97	-10,98	54,57							
15		22,90	61,00	-0,27	0,02	0,00							
16		22,10	75,00	-1,07	14,02	-14,98							
17		20,30	63,00	-2,87	2,02	-5,79							
18		20,00	67,00	-3,17	6,02	-19,07							
19		21,10	80,00	-2,07	19,02	-39,34							
20		26,30	36,00	3,13	-24,98	-78,23							
21		23,00	47,00	-0,17	-13,98	2,35							
22		23,00	73,00	-0,17	12,02	-2,02							
23		23,00	54,00	-0,17	-6,98	1,18							
24		24,70	33,00	1,53	-27,98	-42,86							

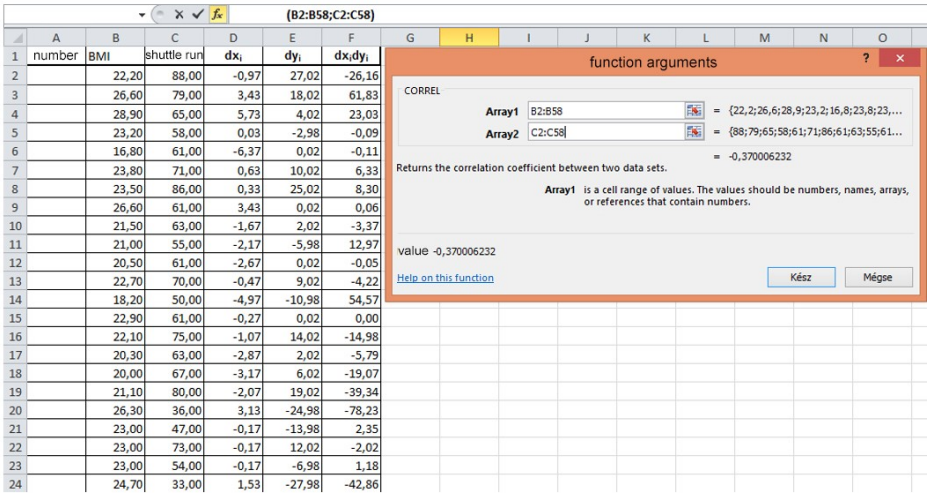
Screen view 2/48. Working table for correlation

Based on data in the working table, the linear correlation coefficient can be calculated as:

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{-27.16}{3.75 \times 19.53} = -0{,}37, \text{so the coefficient of determination}: r^2_{xy} = 0.1369 \text{azaz } 13.69\%.$$

The result shows that there is moderate negative correlation between flat race and high jump. Knowing the coefficient of determination, it can be stated that BMI values determine progressive shuttle run results to an extent of 13.69%. The remaining part is probably determined by other factors (e.g. toughness, etc.). The negative correlation may be a bit confusing but it can be interpreted like this: as the BMI values are increasing (unfavourable), the performance in progressive shuttle run (number of lengths run) is decreasing.

The strength of correlation can be calculated in Excel with the help of an available function. In Excel, the Korrel and the Pearson functions give the same results.



	A	B	C	D	E	F
1	number	BMI	shuttle run	dx_i	dy_i	$dx_i dy_i$
2		22,20	88,00	-0,97	27,02	-26,16
3		26,60	79,00	3,43	18,02	61,83
4		28,90	65,00	5,73	4,02	23,03
5		23,20	58,00	0,03	-2,98	-0,09
6		16,80	61,00	-6,37	0,02	-0,11
7		23,80	71,00	0,63	10,02	6,33
8		23,50	86,00	0,33	25,02	8,30
9		26,60	61,00	3,43	0,02	0,06
10		21,50	63,00	-1,67	2,02	-3,37
11		21,00	55,00	-2,17	-5,98	12,97
12		20,50	61,00	-2,67	0,02	-0,05
13		22,70	70,00	-0,47	9,02	-4,22
14		18,20	50,00	-4,97	-10,98	54,57
15		22,90	61,00	-0,27	0,02	0,00
16		22,10	75,00	-1,07	14,02	-14,98
17		20,30	63,00	-2,87	2,02	-5,79
18		20,00	67,00	-3,17	6,02	-19,07
19		21,10	80,00	-2,07	19,02	-39,34
20		26,30	36,00	3,13	-24,98	-78,23
21		23,00	47,00	-0,17	-13,98	2,35
22		23,00	73,00	-0,17	12,02	-2,02
23		23,00	54,00	-0,17	-6,98	1,18
24		24,70	33,00	1,53	-27,98	-42,86

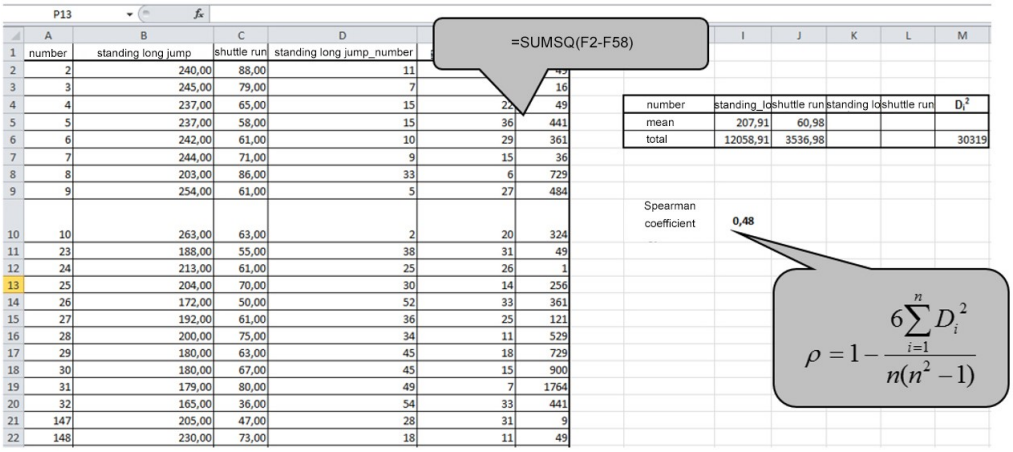
Screen view 2/49. Calculating correlation coefficient in Excel

As a result, we got the values calculated before. The method is a bit different if data are ordered (ranked). In the case of monotonic relationship[27] 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:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}\left[R(y_i) - R(x_i)\right]^2}{n(n^2-1)} = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}.$$

Ranked data will be used to examine the strength of the relationship between length of progressive shuttle runs and standing long-jump. (Source: fittségi 57fő_adatbázis_alap_bmikat.xlsx)

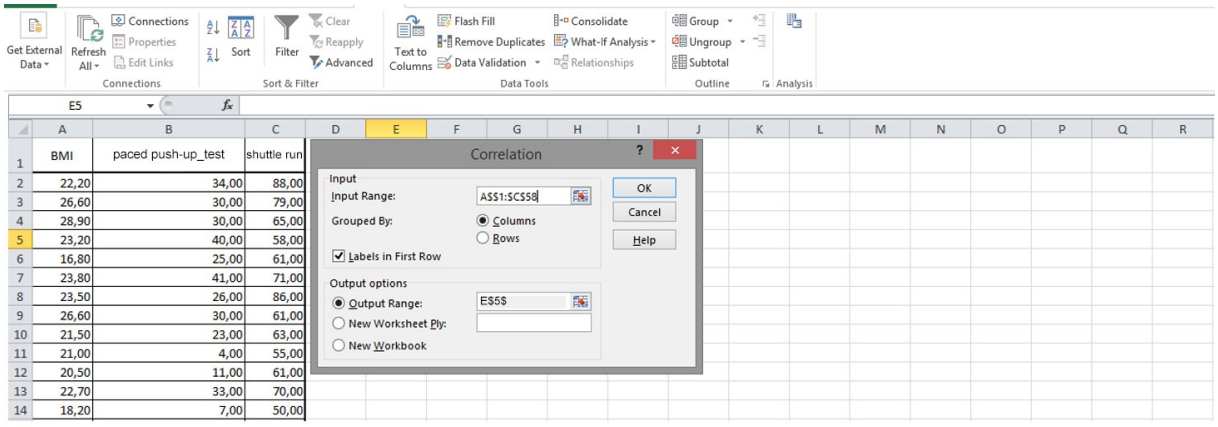
[27] In case of monotonic relationship, the measure of the change in Y is not constant for a unit change in X.



Screen view 2/50. Working table of Spearman's rank correlation

There is a moderate positive correlation between progressive shuttle run orders and standing long-jump orders. The shuttle run results determine long-jump results by 22.69% (determination coefficient).

In order to examine more correlations, go to Tools/Data Analysis/Correlations analysis. In this case, the database has been extended by the paced push-up results. (Source: fittségi 57fő_adatbázis_alap_bmikat.xlsx)



Screen view 2/51. Correlation analysis in Excel

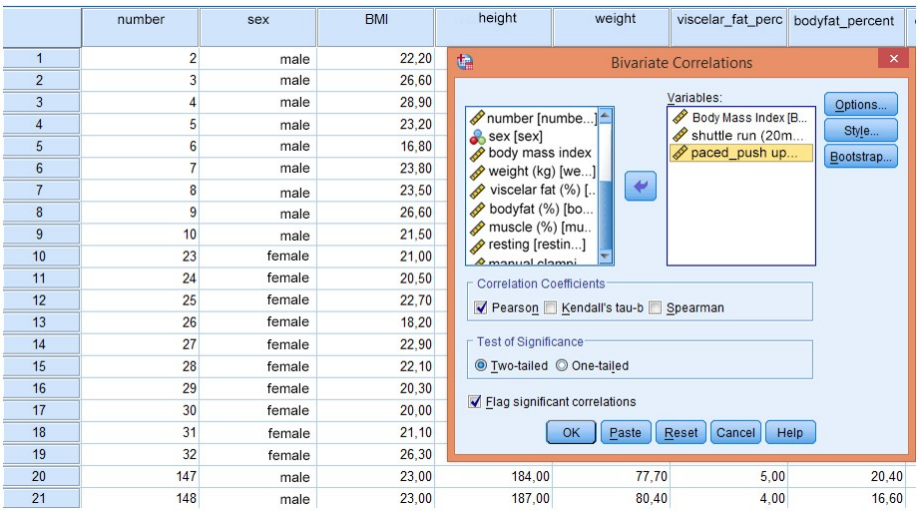
Select input data in the proper box. Add variable names and select the option "labels in the first row". Press OK to get the correlations matrix as the result. It contains total and two-variable correlation coefficients.

	A	B	C	D	E	F	G	H
1	BMI	paced push-up_test	shuttle run					
2	22,20	34,00	88,00					
3	26,60	30,00	79,00					
4	28,90	30,00	65,00					
5	23,20	40,00	58,00			*BMI*	paced_push up_test	shuttle run
6	16,80	25,00	61,00		BMI	1		
7	23,80	41,00	71,00		paced_pu	-0,00336	1	
8	23,50	26,00	86,00		shuttle run	-0,37001	0,415345302	1
9	26,60	30,00	61,00					
10	21,50	23,00	63,00					
11	21,00	4,00	55,00					
12	20,50	11,00	61,00					
13	22,70	33,00	70,00					
14	18,20	7,00	50,00					
15	22,90	11,00	61,00					

Screen view 2/52. Correlations matrix

All values in the diagonal have the value one since the variables have deterministic correlation with themselves. It is a new result that BMI values and paced push-ups are slightly negatively correlated (-0.01) while push-up and shuttle run are moderately correlated (0.42).

Correlation analysis in SPSS is available under Analyze / Correlate / Bivariate. Move the three variables (BMI, shuttle run, push-up) to the variables box. (Source: fittségi 57fő_adatbázis_alap_bmikat.sav)



	number	sex	BMI	height	weight	viscelar_fat_perc	bodyfat_percent
1	2	male	22,20				
2	3	male	26,60				
3	4	male	28,90				
4	5	male	23,20				
5	6	male	16,80				
6	7	male	23,80				
7	8	male	23,50				
8	9	male	26,60				
9	10	male	21,50				
10	23	female	21,00				
11	24	female	20,50				
12	25	female	22,70				
13	26	female	18,20				
14	27	female	22,90				
15	28	female	22,10				
16	29	female	20,30				
17	30	female	20,00				
18	31	female	21,10				
19	32	female	26,30				
20	147	male	23,00	184,00	77,70	5,00	20,40
21	148	male	23,00	187,00	80,40	4,00	16,60

Screen view 2/53. Calculating correlation in SPSS

The default is Pearson's correlation coefficient but Spearman's rank correlation is also available for calculation. The final result is a correlations matrix which equals to the values calculated in Excel.

Table 2/34. Correlations Matrix

Correlations

		Body Mass Index	shuttle run (20, number of lengths)	paced push-ups
Body Mass Index	Pearson Correlation	1	-,370**	-,003
	Sig. (2-tailed)		,005	,980
	N	57	57	57
shuttle run (20, number of lengths)	Pearson Correlation	-,370**	1	,415**
	Sig. (2-tailed)	,005		,001
	N	57	57	57
paced push-ups	Pearson Correlation	-,003	,415**	1
	Sig. (2-tailed)	,980	,001	
	N	57	57	57

**. Correlation is significant at the 0.01 level (2-tailed).

Stars after values mean significance, their interpretation will be discussed later. It is shown that significant correlation is only lacking between paced push-up and BMI index.

2.7.3.4. 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. It is worth noting that 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 can be defined:

$$\hat{Y} = b_0 + b_1 X$$

An estimation of the constant parameters of the linear can be carried out by the method referred to as the ordinary least squares (OLS)[28].

The two parameters can be calculated in the following way:

[28]This book does not present a description of this estimation method only the formulas obtained via the application of the method.

$$b_1 = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

$$b_0 = \bar{Y} - b_1\bar{X} = \frac{\sum Y}{n} - \frac{b_1 \sum X}{n}$$

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.

In the followings the two-variable linear regression will be introduced in a practical example.

Let us examine the relationship between standing long-jump and BMI, and calculate the extent of change in standing long-jump based on one unit change in the BMI.

The next table contains values of the standing long-jump and the BMI. (Source: fittségi 57fő_adatbázis_alap_bmikat.xlsx)

Table 2/35. Basic data (part of the table)

	A	B	C	D	E	F
1	number	BMI (X)	standing long jump_test (Y)	XY	x^2	y^2
2	2	22,20	240,00	5328	492,84	57600
3	3	26,60	245,00	6517	707,56	60025
4	4	28,90	237,00	6849,3	835,21	56169
5	5	23,20	237,00	5498,4	538,24	56169
6	6	16,80	242,00	4065,6	282,24	58564
7	7	23,80	244,00	5807,2	566,44	59536
8	8	23,50	203,00	4770,5	552,25	41209
9	9	26,60	254,00	6756,4	707,56	64516
10	10	21,50	263,00	5654,5	462,25	69169
11	23	21,00	188,00	3948	441	35344
12	24	20,50	213,00	4366,5	420,25	45369
13	25	22,70	204,00	4630,8	515,29	41616
14	26	18,20	172,00	3130,4	331,24	29584
15	27	22,90	192,00	4396,8	524,41	36864
16	28	22,10	200,00	4420	488,41	40000
17	29	20,30	180,00	3654	412,09	32400
18	30	20,00	180,00	3600	400	32400
19	31	21,10	179,00	3776,9	445,21	32041

It is advisable to plot basic data in a scatter chart first since it shows the type of relationship.

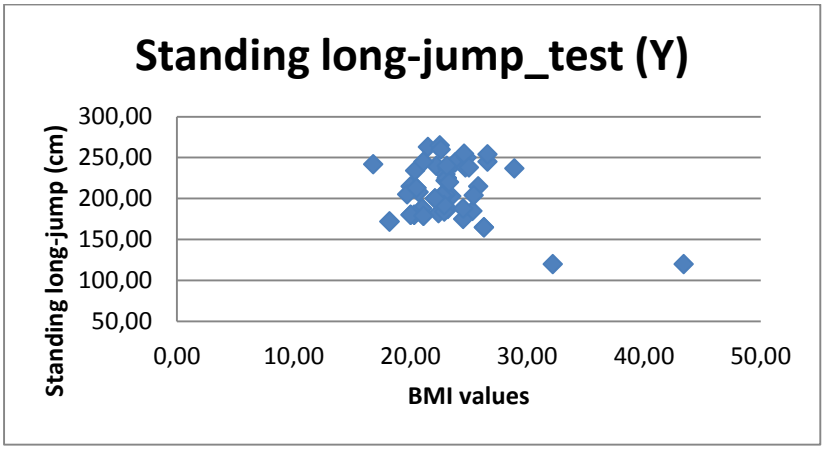


Figure 2/23. The relationship between the variables examined (long-jump, BMI)

Calculation of estimated parameters is shown by the following working table.

	A	B	C	D	E	F	G	H	I	J	K	L
1	number	BMI (X)	standing long jump_test (Y)	XY	x^2	y^2	number	BMI	standing long jump_test (Y)	XY	x^2	y^2
2	2	22,20	240,00	5328	492,84	57600	total	1320,60	11851,00	272679,40	31400,94	2523819,00
3	3	26,60	245,00	6517	707,56	60025	mean	23,17	207,91	4783,85	550,89	44277,53
4	4	28,90	237,00	6849,3	835,21	56169	standard deviation	3,76	32,40			
5	5	23,20	237,00	5498,4	538,24	56169						
6	6	16,80	242,00	4065,6	282,24	58564						
7	7	23,80	244,00	5807,2	566,44	59536						
8	8	23,50	203,00	4770,5	552,25	41209						
9	9	26,60	254,00	6756,4	707,56	64516						

Screen view 2/54. Working table of regression (part of the table)

Applying the formulas above the parameters will be the following:

$$b_1 = \frac{57 \times 272679.4 - 1320.6 \times 11851}{57 \times 31400.94 - 1320{,}6^2} = -2.35$$

$$b_0 = 207.91 - (-2.35) \times 23.17 = 262.31$$

The regression function: $\hat{Y} = 262.31 - 2.35X$

Based on the regression coefficient one can state that one unit increase in BMI may be expected to decrease standing long-jump by 2.35 cm.

Expected standing long-jump can be estimated for a student with a BMI value of 21 like this: $\hat{Y} = 262.31 - 2.35 \times 21 = 212.96$

Having identified the measure of the regression coefficient, it is 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:

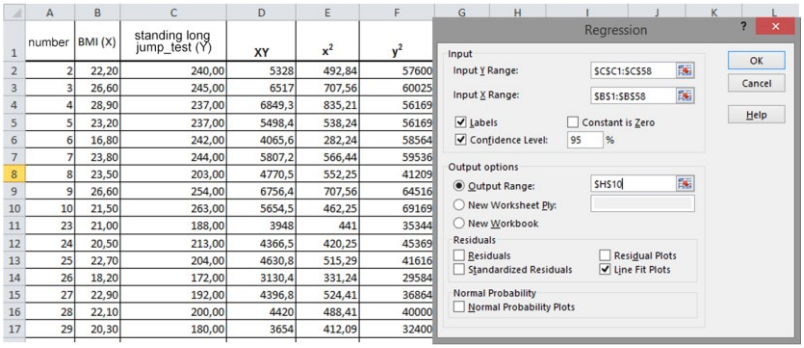
$$El = b_1 \frac{\overline{x}}{\overline{y}}$$

The average elasticity in the example: $El = -2.35 \frac{23.17}{207.91} = -0.26\ \%$

This means that the standing long-jump value reacts to BMI changes by less extent.

There is a lot of literature available in the topic of regression analysis, see for example PINTÉR-RAPPAI (2007), MUNDRUCZÓ (1981) or RAMANATHAN (2003).

Related analysis can be carried out in Excel under Data/Data analysis/Regression.



	A	B	C	D	E	F
1	number	BMI (X)	standing long jump_test (Y)	XY	x^2	y^2
2	2	22,20	240,00	5328	492,84	57600
3	3	26,60	245,00	6517	707,56	60025
4	4	28,90	237,00	6849,3	835,21	56169
5	5	23,20	237,00	5498,4	538,24	56169
6	6	16,80	242,00	4065,6	282,24	58564
7	7	23,80	244,00	5807,2	566,44	59536
8	8	23,50	203,00	4770,5	552,25	41209
9	9	26,60	254,00	6756,4	707,56	64516
10	10	21,50	263,00	5654,5	462,25	69169
11	23	21,00	188,00	3948	441	35344
12	24	20,50	213,00	4366,5	420,25	45369
13	25	22,70	204,00	4630,8	515,29	41616
14	26	18,20	172,00	3130,4	331,24	29584
15	27	22,90	192,00	4396,8	524,41	36864
16	28	22,10	200,00	4420	488,41	40000
17	29	20,30	180,00	3654	412,09	32400

Screen view 2/55. Calculating regression in Excel

Add standing long-jump (cm) to input Y range and BMI data to X range. Select titles, confidence level, and request graphic illustration. Results include:

regression summary

regression statistics	
R value	0,272271911
R square	0,074131993
adjusted	0,05729803
standard	31,74148991
observab	57

variance analysis

	df	*SS*	*MS*	*F*	significance
regression	1	4436,841	4436,841	4,4037159	0,040467
remainder	55	55413,72	1007,522		
total	56	59850,56			

	coefficients	standard	t-value	p-value	lower 95%	upper 95%	lower 95%	upper 95%
ordinate	262,3136888	26,26261	9,988103	5,824E-14	209,6822	314,9451	209,6822	314,9451
BMI	-2,348084402	1,118933	-2,098503	0,0404667	-4,59048	-0,10569	-4,59048	-0,10569

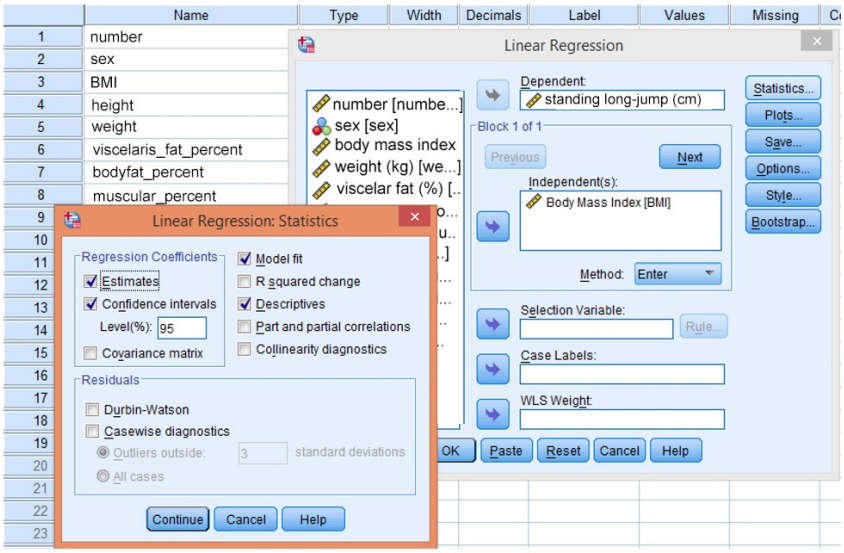
Screen view 2/56. Final results of regression

The table contains data calculated above in details. R value shows that there is a weak association between the two variables (BMI and standing long-jump). BMI determines

standing long-jump values by 7.41%. High value of R^2 may mean that the linear fits the point set well.

Estimated parameters (b_0; b_1) are displayed in the column called Coefficient. Parameter b_0 will be named ordinate.[29] At the end of the parameter estimation, the interpretation of interval estimation is possible, too (see later). Multiple linear regression can be calculated in a similar way.

We can get the same results just as quickly in SPSS by the access path ANALYSE / REGRESSION / LINEAR. (Source: fittségi 57fő_adatbázis_alap_bmikat.sav)



Screen view 2/57. Regression calculation in SPSS

Standing long-jump has to be defined as the dependent variable, and BMI as the independent one, then by pressing STATISTICS, one has to choose the options CONFIDENCE INTERVALS, MODEL FIT and DESCRIPTIVES. Press CONTINUE and OK, and see the results of calculation in the OUTPUT VIEW.

Table 2/36. Descriptive Statistics

Descriptive Statistics

	Mean	Std. Deviation	N
standing long-jump (cm)	207,9123	32,69190	57
Body Mass Index	23,1684	3,79079	57

[29] T-tests test if the parameters are equal to zero.

Table 2/37.Correlations matrix

Correlations

		standing long-jump (cm)	Body Mass Index
Pearson Correlation	standing long-jump (cm)	1,000	-,272
	Body Mass Index	-,272	1,000
Sig. (1-tailed)	standing long-jump (cm)	.	,020
	Body Mass Index	,020	.
N	standing long-jump (cm)	57	57
	Body Mass Index	57	57

The first table contains descriptive which is followed by the correlations matrix. Regression model summary is provided by the following tables:

Table 2/38. Measures of correlation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,272[a]	,074	,057	31,74149

a. Predictors: (Constant), Body Mass Index

Table 2/39. ANOVA table

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4436,841	1	4436,841	4,404	,040[b]
	Residual	55413,720	55	1007,522		
	Total	59850,561	56			

a. Dependent Variable: Helyből távolugrás (cm)

b. Predictors: (Constant), Body Mass Index

The first "Model Summary" table contains Pearson's correlation coefficient which refers to a weak correlation (R=0.272). The determination coefficient (R^2 =0.074) expresses the strength of the relationship. It shows that the independent variable can explain 7.4% of the total standard deviation, i.e. BMI values only influence changes in standing long-jump by 7%. The higher the R^2, the better the line fits the point set. Next, the standard error follows which is the proxy of the precision of estimation (if it is high, then the model is not capable of good estimation). The ANOVA table contains the F-value and the significance value ($p<0.05$) – the latter is proving the existence of the relationship. Based on the table, it can be decided that the correlation between the two variables exists, and this is not a matter of a random event. It is followed by the table, containing the regression parameters.

Table 2/40.Regression parameters

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	262,314	26,263		9,988	,000	209,682	314,945
	Body Mass Index	-2,348	1,119	-,272	-2,099	,040	-4,590	-,106

a. Dependent Variable: Helyből távolugrás (cm)

Before interpreting the model, see that t-values and significance values ($p<0.05$) of both variables prove its existence. At the end of the parameter estimation, the interval estimation can be interpreted, too.

The regression equation can be drawn up with the help of the non-standardized coefficients:

$$b_1 = -2.35$$
$$b_0 = 262.31$$

The regression function:

$$\hat{Y} = 262.31 - 2.35X$$

2.7.4. Inferential statistical methods

In sport as well as in other areas of 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. Two main groups of inferential statistical methods, estimation and hypothesis testing will be presented. Mathematical basis of inferential statistics is provided by probability theory that is discussed in details by Hunyadi (2001).

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 (μ) and the variance, i.e. the square of standard deviation (σ^2). Normal distribution can easily 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, and performance.

Depending on the subject of analysis, expected values and standard deviations can have a lot of different values, which can cause 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 2/76).** 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 relatively quickly and easily. Although, in the age of computers, similar tables are gradually forgotten.

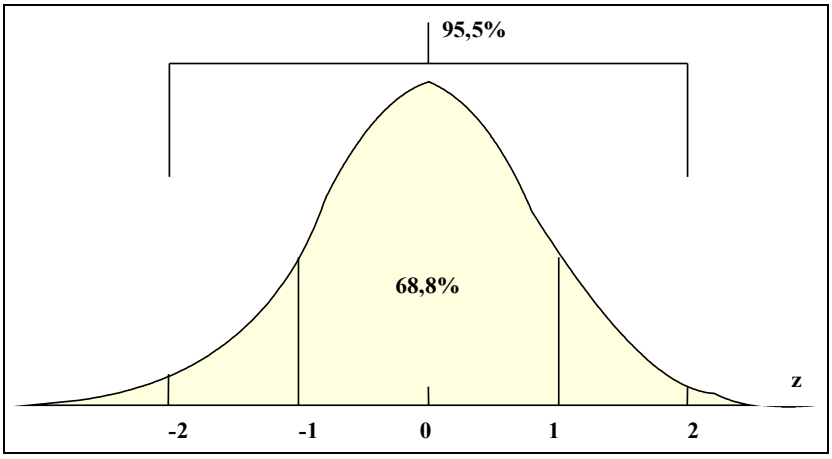


Figure 2/24. Most important probability values depending on z

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 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[30] 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.

---

[30] Sampling without replacement (e.g. simple random sampling) is quite popular in practice since its application causes no waste of information.

Elements of the random sample are random variables, which is why any of their transformations – similarly to their arithmetic average – will be random variables, too. If the distribution of the population is normal, then the sample mean does also follow normal distribution, independently from the sample size. Keeping this in mind, sample means can be standardized with the $z = \frac{x - \mu}{\sigma}$ formula[31].

2.7.4.1. Statistical estimations

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 direct 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:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

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.

---

[31] Pintér- Ács (2006)

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 2/41. Estimators of the most relevant population parameters

Population parameter	Unbiased estimatos	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

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 the most important:

Table 2/42. Often applied critical values

α	1-α	$Z_{(\alpha/2)}$	$Z_{(1-\alpha/2)}$
0.01	0.99	-2.576	2.576
0.05	0.95	-1.96	1.96
0.1	0.9	-1.645	1.645

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)[32].

Let us have a look at an example based on the above:

Based on our survey, let us consider the height of 57 university students and generate a sample of 30 with simple random sampling. The task is to estimate the height (cm) of the students with 95% confidence interval (Source: fittségi 57fő_adatbázis_alap_bmikat.xlsx).

Table 2/43. Height and weight parameters in the random sample

Number	Height	Weight	Number	Height	Weight	Number	Height	Weight
2	173,00	66,30	24	166,00	56,40	148	187,00	80,40
3	177,00	83,30	25	171,00	66,50	149	187,00	80,40
4	186,00	100,10	26	184,50	62,00	150	178,00	73,80
5	172,00	68,50	27	173,00	68,60	151	170,00	73,00
6	176,00	70,90	28	173,00	66,10	168	157,00	57,30
7	177,00	74,60	29	178,00	64,40	169	164,00	65,40
8	184,00	78,80	30	176,00	62,00	170	170,00	64,60
9	178,00	84,30	31	170,00	61,00	172	167,00	67,50
10	181,00	70,50	32	153,00	61,60	173	168,00	64,60
23	166,50	58,10	147	184,00	77,70	180	176,00	63,30

Calculating the (sample) mean: $\overline{\mu} = \frac{\sum x}{n} = 174.1$ and the standard deviation (corrected standard deviation because we have data from a sample): $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = 8.34$ 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): $s_{\overline{x}} = \frac{s}{\sqrt{n}} = 1.52$. As the sample is small (n<50) the critical value from the t-distribution has to be identified which is z=2.05. ($t^{df,\ confidence}$)

The value of error limit is ($z \times \sigma_{\overline{x}}$): 2.99, from which the following result can be obtained: 174.1±2.99

[32] Pintér – Rappai (2001)

E	F	G	H	I	J	K	L	M	N	O	P	Q	R
number	height	weight	number	height	weight	number	height	weight					
2	173,00	66,30	24	166,00	56,40	148	187,00	80,40		mean	174,10		
3	177,00	83,30	25	171,00	66,50	149	187,00	80,40		standard deviati	8,34		
4	186,00	100,10	26	184,50	62,00	150	178,00	73,80		sample size (n)	30,00		
5	172,00	68,50	27	173,00	68,60	151	170,00	73,00		standard error	1,52		
6	176,00	70,90	28	173,00	66,10	168	157,00	57,30		critical value	2,05		
7	177,00	74,60	29	178,00	64,40	169	164,00	65,40		error limit	2,99		
8	184,00	78,80	30	176,00	62,00	170	170,00	64,60		lower bound	171,11		
9	178,00	84,30	31	170,00	61,00	172	167,00	67,50		upper bound	177,09		
10	181,00	70,50	32	153,00	61,60	173	168,00	64,60					
23	166,50	58,10	147	184,00	77,70	180	176,00	63,30					

=P3/SQRT(P4)

=1,96*P5

=TINV (0,05;29)

Screen view 2/58 Working table of the estimation

We can state with 95% confidence that the average height of students is between 171.11 and 177.09 cm.

Critical values can be calculated in Excel by the functions inverz.stnorm(pobability)[33], and inverz.t[34](probability, degree of freedom).

The following formulas can be used for estimating total values: $x' = N \times \bar{x}$,*and* $\sigma_{x'} = N \times \sigma_{\bar{x}}$. These cannot be interpreted in our example.

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:

$$p \pm z \times \sigma_p$$

[33] Inverz.stnorm: calculates critical value from the standard normal. Inverz.stnorm(α/2) provides the value for (1-α) confidence.

[34] Inverz.t(probability, degree of freedom): provides critical value by calculating the half of the probability value given from the t-distribution.

Let us use the 30 elements-sample above to determine by 95.5% confidence the proportion of students being taller than 180 cm.

$$p = \frac{k}{n} = \frac{7}{30} = 0.233$$

$$0.233 \pm 2 \times \sqrt{\frac{0.267 \times 0.767}{30}}$$

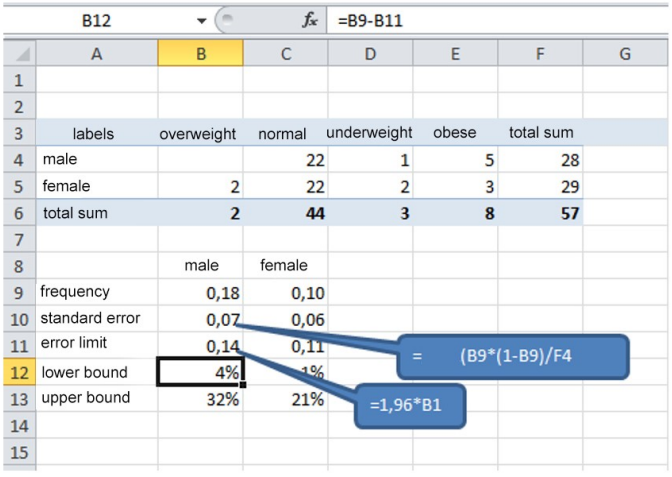
$$0.233 \pm 0.165$$

6.8% 39.8%

We can state with 95.5% confidence that minimum 6.8% of students are taller, and 39.8% of them are smaller than 180 cm.

Ratio estimation slightly differs since the estimator is the relative frequency in the sample – it is not determined in advance – but if a new variable will be generated, then it traces back to mean estimation. Let the value of the new variable be 1 if the height is greater than 180 cm and zero otherwise. This can be easily done by the "if" function (if B2>180;1;0). Upcoming calculations are the same as the ones introduced above.

As a practice excercise, let us determine the ratio of overweight men and women with 95% confidence interval.



Screen view 2/59 Working table of ratio estimation

Besides these, Excel provides options to quick estimations in Data/Data Analysis/Descriptive Statistics. The only alteration in the known module has to be the confidence level of the expected value. After changing these settings, the following results will be displayed:

Table 2/44. Statistical estimation results with Data analysis in Excel

height	
expected value	174,10
standard error	1,52
median	174,50
mode	173,00
standard deviation	8,34
sample variance	69,59
kurtosis	0,31
skewness	-0,48
range	34,00
minimum	153,00
maximum	187,00
sum	5223,00
sample size	30,00
confidence level (95%)	3,12

The slight differences stem from the fact that the computer always calculates the critical value from the t-distribution with the corresponding degree of freedom[35]. This is not an error because as the t-distribution fits the standard normal distribution with as the sample size is growing.

As a practice exercise, let us display the graphic illustration of the expected value of height with the confidence interval.

- Use Ctrl to select the names of columns and their expected values.
- Create diagram with the Bar chart option of Insert.
- The (positive and negative) error limits can be given in Diagram tools, Error zones, further error zone settings.

[35] Inverz.t(0.05;29)

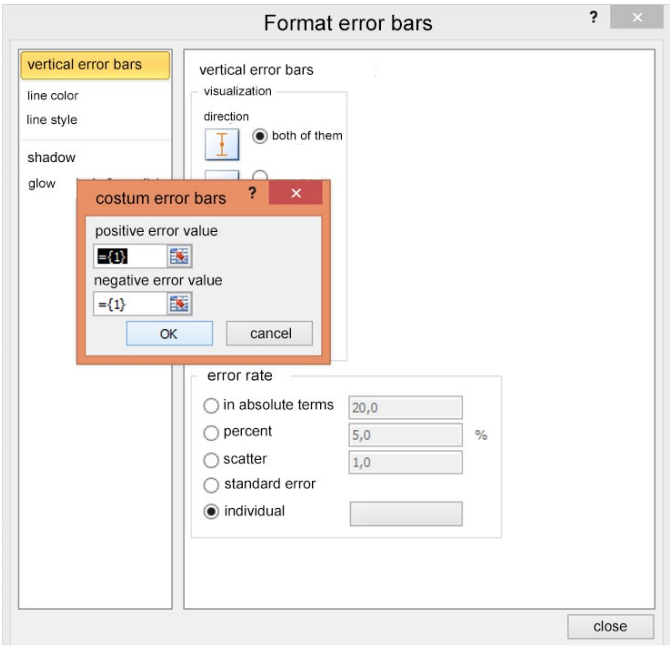


Figure 2/60. Error limit settings

After setting the error limits (2.99 both for the negative and the positive value) the propram puts the confidence interval on the bar chart.

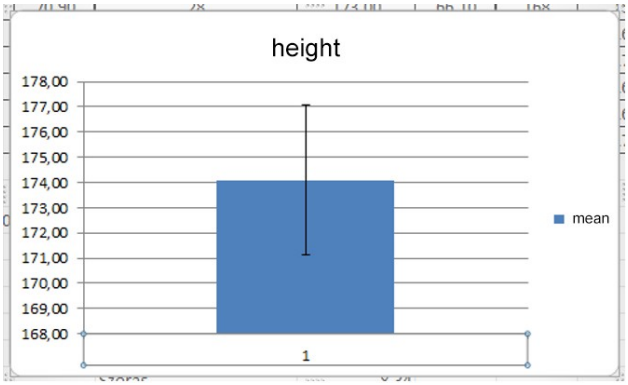
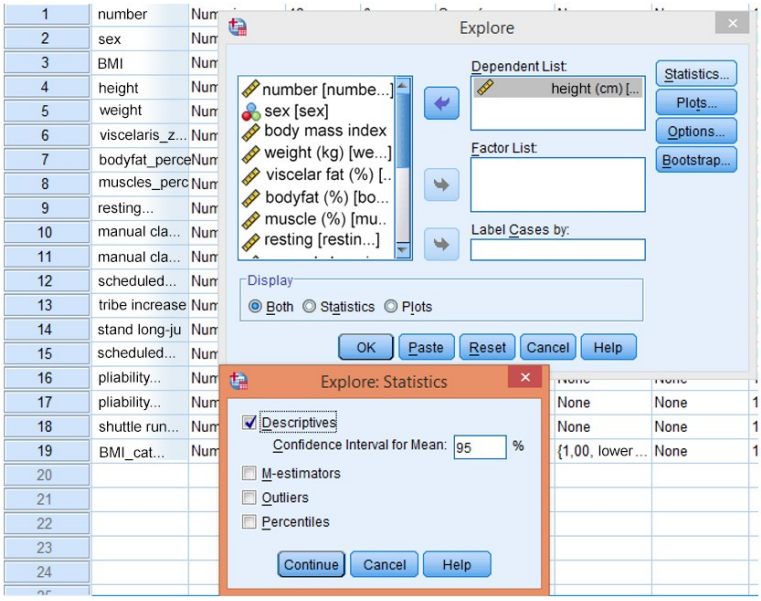


Figure 2/25. Graphic illustration of the height estimation (with 95% confidence interval)

Simple estimations are also carried out relatively quickly by SPSS under Analyze/ Descriptive Statistics/Explore. (Source: fittségi 57fő_adatbázis_alap_bmikat.sav).
As a practice exercise, let us use the height data of the 57 students to estimate the expected value of height with 95% confidence interval.



Screen view 2/61. Statistical estimation in SPSS

Choose height to be the independent variable and write the confidence interval for mean in Statistics/Descriptives. The output table:

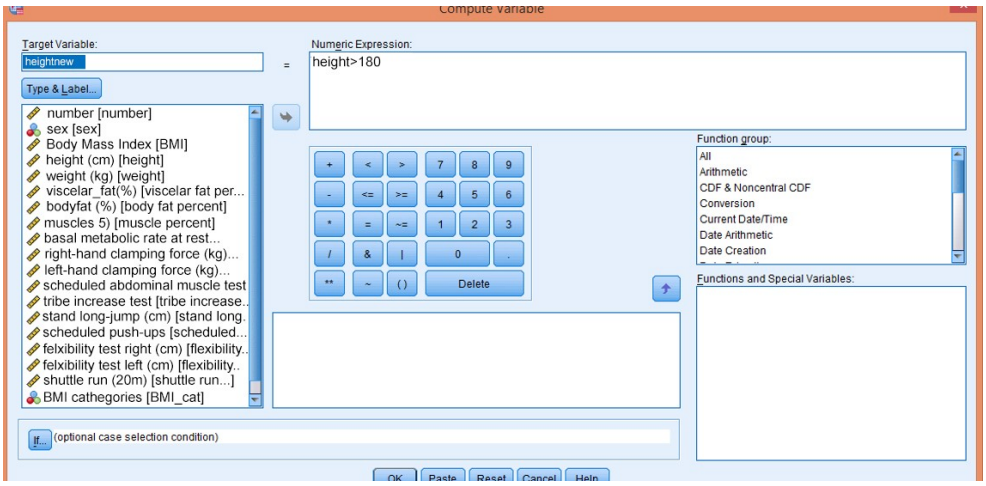
Table 2/45. Descriptives

Descriptives

			Statistic	Std. Error
height (cm)	Mean		174,2719	1,06619
	95% Confidence Interval for Mean	Lower Bound	172,1361	
		Upper Bound	176,4078	
	5% Trimmed Mean		174,6511	
	Median		173,0000	
	Variance		64,795	
	Std. Deviation		8,04955	
	Minimum		153,00	
	Maximum		188,50	
	Range		35,50	
	Interquartile Range		10,00	
	Skewness		-,469	,316
	Kurtosis		,364	,623

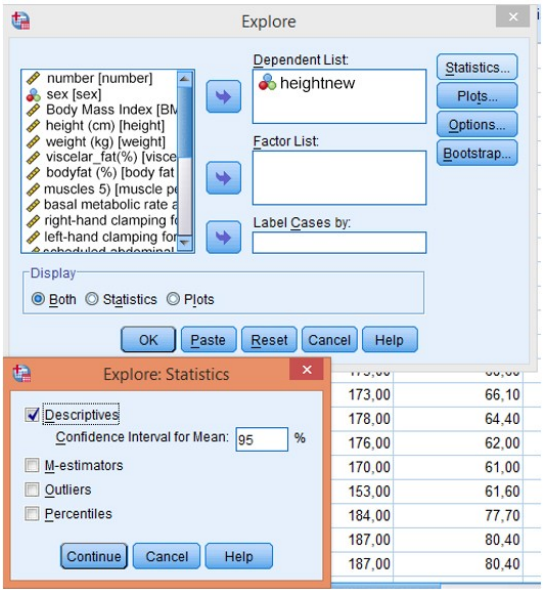
The results include all the calculations (even the upper and lower bound) we have done above.

Ratio estimation requires the creation of a new variable (magasságúj; heightnew) in Transform/ Compute Variable.



Screen view 2/62. Establishing a new variable in SPSS

A new variable will be computed from an already existing one. The new one (magasságúj) can be named in the box Target Variable, and constraints can be given in the box Numeric Expression (magasságúj>180). In the new variable, observables with higher values than the given one (180 cm) get the new value 1, while others become 0. Estimation from here is similar to the exercise before.



Screen view 2/63. Ratio estimation settings

Add the new variable as the dependent one, and select 95% confidence interval to get the following results:

Table 2/46. Descriptives

Descriptives

			Statistic	Std. Error
heightnew	Mean		,2456	,05752
	95% Confidence Interval for Mean	Lower Bound	,1304	
		Upper Bound	,3608	
	5% Trimmed Mean		,2173	
	Median		,0000	
	Variance		,189	
	Std. Deviation		,43428	
	Minimum		,00	
	Maximum		1,00	
	Range		1,00	
	Interquartile Range		,50	
	Skewness		1,214	,316
	Kurtosis		-,546	,623

2.7.4.2. Hypothesis testing

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 Θ (theta) and the expected value by Θ_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. For example, two groups of swimmers are generated so that one group will be trained by the traditional and the other by the "Széchy method". 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:

$$H_1: \Theta \langle \Theta_0$$

vagy

$$H_1: \Theta \rangle \Theta_0$$

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 population, 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 α, then the probability to be in the acceptance region is: 1-α.

Let us assume that the hypothesis is about the equality of the expected value (μ) 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.

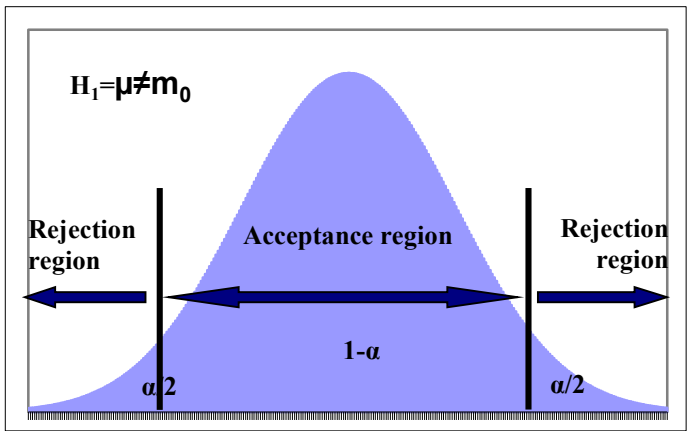


Figure 2/26. Acceptance and rejection region in case of a two-tailed hypothesis

The probability that the test statistic is within the rejection region is α. As the rejection region consists of two parts of the same size, both represent the probability of α/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.

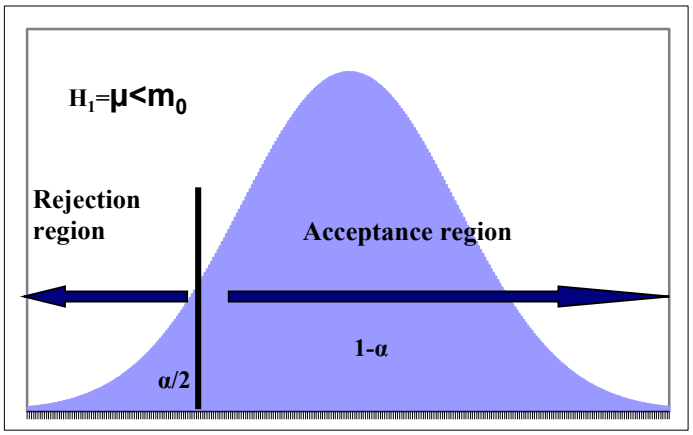


Figure 2/27. Acceptance and rejection region in case of a left-tailed alternative hypothesis

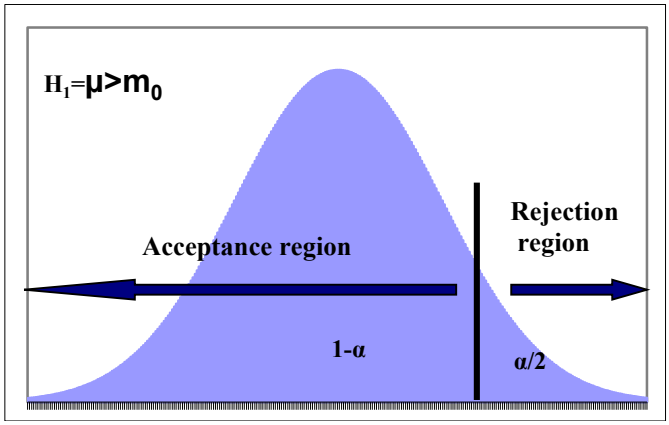


Figure 2/28. Acceptance and rejection region in case of a right-tailed alternative hypothesis

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:

Table 2/47. Statistical tests

Null hypothesis	Large sample (100≤n)	Small sample (n<100)
$H_0 : \mu = \mu_0$	$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \mid H_0 \sim N(0;1)$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \mid H_0 \sim {}_{n-1}t$
$H_0 : P = P_0$	$z = \frac{P - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \mid H_0 \sim N(0;1)$	
$H_0 : \sigma^2 = \sigma_0^2$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \mid H_0 \sim {}_{n-1}\chi^2$	

Let us now consider a test in practice:

When consulting literature for his/her thesis, a student reads in a foreign scientific article[36] (Bai et al, 2013) that the average of BMI index of Chinese university students from Tibet was 21.35. The average BMI of Hungarian university students in our database of 57 elements is 23.17 and the corrected standard deviation is 3.79. Let us test if there is a difference in this case between Tibetan and Hungarian. Can we accept the statement that the BMI value of Hungarian students cannot exceed that of Tibetan ones?

H_0: μ=21.35

H_1: μ>21.35

[36] Bai Jingya és társai: Quantitative Analysis and Comparison of BMI among Han, Tibetan, and Uygur University Students in Northwest China. The Scientific World Journal Volume 2013 (2013), Article ID 180863, 6 pages. http://www.hindawi.com/journals/tswj/2013/180863/

The null hypothesis states that the expected mean value of Hungarian students' BMI is equal to the same of the Chinese dtudents (difference from the Chinese mean is random). The alternative hypothesis says that this value is greater than 21.35, which means that there is a systematic reason for the difference[37] (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 χ^2 distribution it is one, in case of F distribution d.f. is two).

$$t = \frac{23.17 - 21.35}{\frac{3.79}{\sqrt{57}}} = 3.62$$

The critical value of the t-distribution in case of 5% significance level and 56 (n-1) degrees of freedom is 1.67 as written in the table of t-distribution.

As the calculated value is greater than the one in the table, the null hypothesis has to be rejected (there is no reason to accept it), i.e. the alternative hypothesis is accepcted (the rejection region of the chart is valid). This means that the difference between the sample value and the value expected is probably caused by a systematic factor (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 Tibetan sample, then a two-tailed alternative hypothesis has to be drawn up:

$$H_1: \mu \neq 23.17$$

[37] This assumption is based on our professional opinion, the literature review of which is not included in this book.

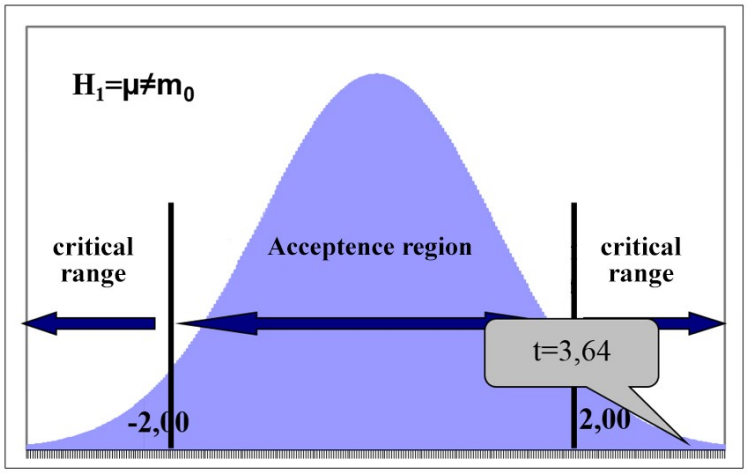


Figure 2/29. Illustration supporting the decision in case of two-tailed alternative hypothesis

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: ±2.00. Comparing the t-value to these, the null hypothesis has to be rejected, once again.

The method is similar if the assumption does not refer to the population mean but the ratio. The standard normal distribution is again only applicable in case of a large sample.

$$z = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

In the session of association we ordered the students into a contingency table, based on BMI categories.

Table 2/48. BMI categories according to sex

Sex	BMI categories				
	obese	normal weight	underweight	overweight	Total
male	0	22	1	5	28
female	2	22	2	3	29
Total	**2**	**44**	**3**	**8**	**57**

Altogether, 10 students belong to the categories obese and overweight.

We examine if a proportion of 15% overweighted and obese students can be expected.

H_0: P=0.15

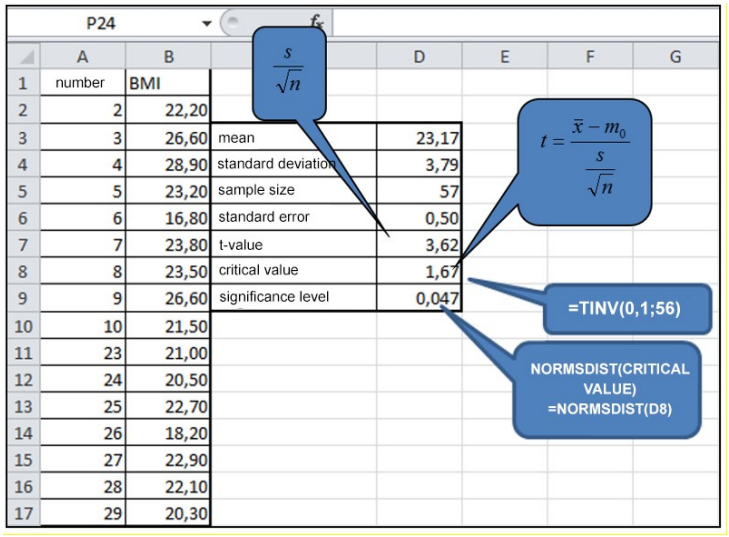
H_1: P<0.15

In the alternative hypothesis it is drawn up that this proportion will be smaller than 15%.

$$p = \frac{10}{57} = 0.176$$

$$z = \frac{0.176 - 0.15}{\sqrt{\frac{0.15 \times (1 - 0.15)}{57}}} = 0.55$$

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 overweight and obese university students is above 15%.

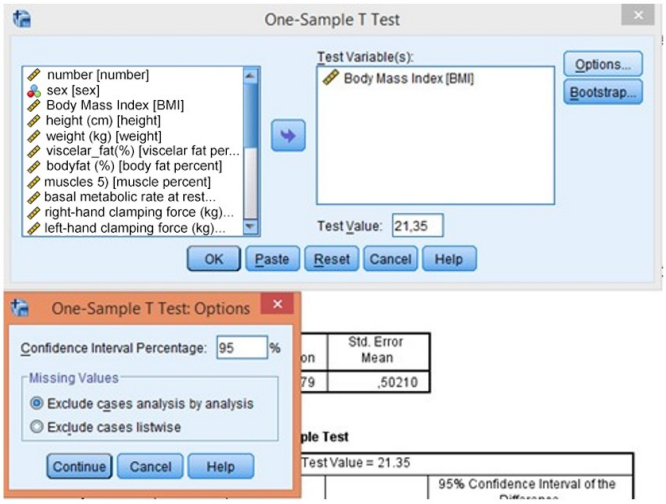


	A	B		D	E	F	G
1	number	BMI					
2	2	22,20					
3	3	26,60	mean	23,17			
4	4	28,90	standard deviation	3,79			
5	5	23,20	sample size	57			
6	6	16,80	standard error	0,50			
7	7	23,80	t-value	3,62			
8	8	23,50	critical value	1,67			
9	9	26,60	significance level	0,047			
10	10	21,50					
11	23	21,00					
12	24	20,50					
13	25	22,70					
14	26	18,20					
15	27	22,90					
16	28	22,10					
17	29	20,30					

Screen view 2/64. The calculation in Excel

As we have a small sample, we have to determine the critical value of the t-distribution but Excel does always calculate as if the critical region was two-tailed, i.e. it halves the region. The critical region will be calculated by the function inverz.t, and if we need a significance level of 5%, we have to add 10% instead.

SPSS provides the results more easily. Choose Analyze/Compare Means/ One-Sample T Test to get the results. (Source: Egymintás t-próba alapadatai.sav)



Screen view 2/65. One-Sample t-test in SPSS (BMI)

First, the user has to select the variable BMI as test variable, and then write the value in the null hypothesis (21.35) into the Test Value box. This constant will be the basis of comparison. Add significance level under Options. Press OK and go to Output View:
The results consist of two tables. The first contains descriptive statistical data we did also gain in Excel.

Table 2/49. Descriptive Statistics

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Body Mass Index	57	23,1684	3,79079	,50210

The second table supports the decision if the hypothesis has to be accepted or rejected.

Table 2/50. Results of the One-Sample Test

One-Sample Test

	Test Value = 21.35					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Body Mass Index	3,622	56	,001	1,81842	,8126	2,8243

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) which is our case, too. The confidence interval shows limits between which 95% of the difference is situated.
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:

$H_1: \mu_1 - \mu_2 < \delta$; $H_1: \mu_1 < \mu_2$ (left-tailed)

$H_1: \mu_1 - \mu_2 > \delta$; $H_1: \mu_1 > \mu_2$ (right-tailed)

$H_1: \mu_1 - \mu_2 \neq \delta$; $H_1: \mu_1 \neq \mu_2$ (two-tailed)

Now, we consider the most common **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[38]):
The test statistic to be applied:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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 has to test 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.

$$F = \frac{s_1^2}{s_2^2} \Big| H_0 \approx F_{n_1 - 1; n_2 - 1}$$

Differences of the two expected values are interpreted in the following example. Let us examine if there is a difference between height of boys and girls at 5% significance level. Average of the height of (the 29) girls is 169.62 cm with a standard deviation of 7.49. Same data of (the 28) boys are 179.04 cm and 5.37.

Since it is a two-sample t-test, we first have to examine if standard deviations can be considered identical:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$F = \frac{7.49^2}{5.37^2} = 1.94$$

[38]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.

Following the above logical assumptions, the upper critical value is to be found at 2.5% since we would like to consider 5%, and then lower critical value has to be calculated based on this. Calculations are a bit more difficult in the case of F-distribution because it is not symmetric and it will be only interpreted in the positive region. The lower critical value will be calculated with the following formula:

$$ {}_{n;m}F_{1-\alpha} = \frac{1}{{}_{m;n}F_{\alpha}} $$

The two critical values in our example (the first comes from the table; the second one has been calculated):

$$ {}_{28;27}F_{0.025} = 2.15 $$

$$ {}_{27;28}F_{0.975} = \frac{1}{{}_{11;12}F_{0.025}} = \frac{1}{2.18} = 0.46 $$

To sum up, identity of variances has to be accepted at 5% significance level since the value we got belongs to the acceptance region (it is situated between the two critical values). Then the two-sample t-test follows:

$$ H_0: \mu^1 = \mu^2 $$

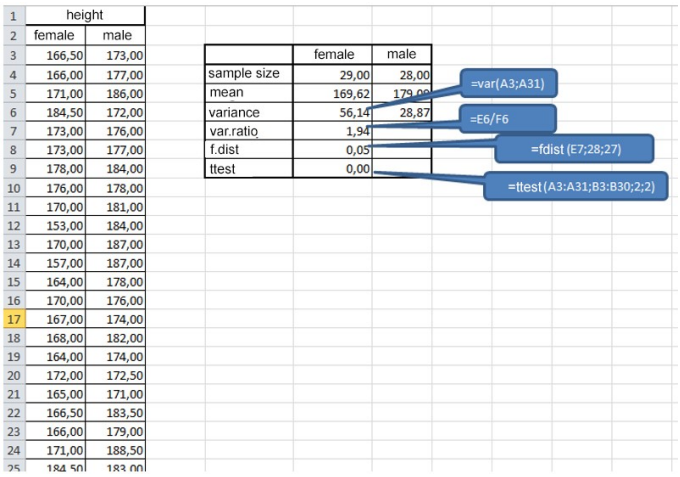
$$ H_1: \mu^1 \neq \mu^2 $$

$$ s_p^2 = \frac{(29-1)\times 7.49^2 + (28-1)\times 5.37^2}{29+28-2} = 42.75 $$

$$ s_p = \sqrt{42.75} = 6.54 $$

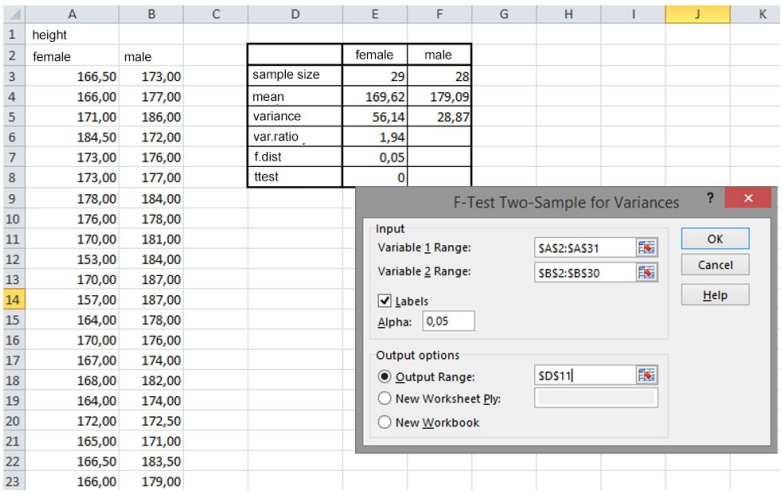
$$ t = \frac{169.62 - 174.04}{6.54 \times \sqrt{\frac{1}{29} + \frac{1}{28}}} = -5.5 $$

The table value of the t-distribution with 55 degrees of freedom is 2.00 since the hypothesis is two-tailed. The calculated value is in the rejection region so at 5% significance level there is a significant difference in the height of boys and girls ($p<0.05$).



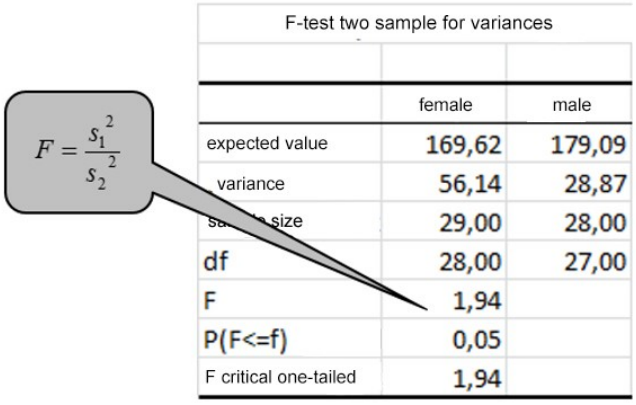
Screen view 2/66. Calculating the two-sample t-test

The two-sample t-test will be carried out in two steps in Excel (Tools/Data analysis) since the prerequisits have to be tested, too. (Two-samples F-test for variances)



Screen view 2/67. Fit test (two-sample F-test)

Add group data to variable ranges (with label), select the labels box, and determine the output range to get the results:



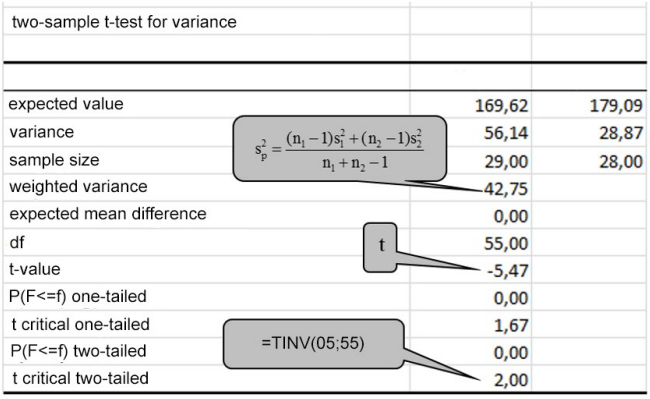
F-test two sample for variances

	female	male
expected value	169,62	179,09
variance	56,14	28,87
sample size	29,00	28,00
df	28,00	27,00
F	1,94	
P(F<=f)	0,05	
F critical one-tailed	1,94	

Screen view 2/68. Fit-test results

It may seem surprising that Excel carries out a one-tailed test because we referred to two-tailed test (unequality as counterhypothesis) but the program is "economical" since it takes advantage of the fact that one of the critical values in the F-test is greater while the other is smaller than 1. Based on the scale of the empirical test statistic, it can decide if the lower or the upper critical value is relevant.

Decisions can be made as follows: if the calculated F-value is between the critical value calculated by Excel and 1, then the null hypothesis has to be accepted, otherwise (if F-value is too small or too big) is has to be rejected[39]. In this case, the calculated values are around the limits but considering the variances to be equal, the two-sample t-test can be carried out. (If the variances were not equal, the unequal variances modul of the two-sample t-test had to be selected in Data analysis). Settings are the same as the ones for F-test.



two-sample t-test for variance

expected value	169,62	179,09
variance	56,14	28,87
sample size	29,00	28,00
weighted variance	42,75	
expected mean difference	0,00	
df	55,00	
t-value	-5,47	
P(F<=f) one-tailed	0,00	
t critical one-tailed	1,67	
P(F<=f) two-tailed	0,00	
t critical two-tailed	2,00	

Screen view 2/69. Results of the two-tailed t-test in Excel

[39] Pintér- Rappai 2007, page 385.

The results are equal so the null hypothesis has to be rejected. Making a decision based on the p-value (significance value), the maximal significance level until the null hypothesis had to be accepted was 0.00, so the rejection decision is clear.

The same test can be carried out in SPSS. First, the user has to check if the BMI follows a normal distribution. The normality test is applied when comparing the distributions of two random variables or 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.

Graphic and numerical tests are to be found under ANALYZE/DESCRIPTIVE STATISTICS/ EXPOLRE. The access path of the Kolmogorov-Smirnov test is ANALYSE / LEGACY DIALOGS / 1-SAMPLE K-S where height has to be selected as TEST VARIABLE. In OPTIONS, descriptive statistics can be requested (DESCRIPTIVE). Finally, CONTINUE and OK will have to be pressed. Let us examine quantitative and graphic settings (ANALYZE/DESCRIPTIVE STATISTICS/EXPLORE) and results of the normality test

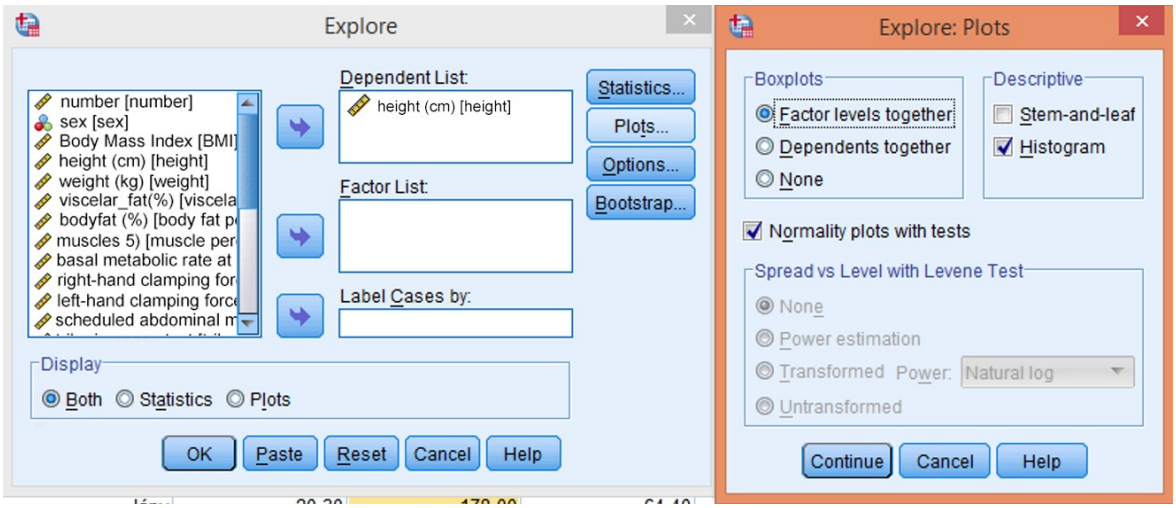


Figure 2/70. Settings of normality test

The variable height (Testmagasság (cm)) 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 view contains more tables and figure of results. We introduce only the ones that are relevant from our point of view.

Table 2/51. Numerical results of the Test of Normality

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
height (cm)	,087	57	,200*	,964	57	,087

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

a. Lilliefors Significance Correction

Significance for both tests (Kolmogorov-Smirov and Shapiro-Wilk) are higher than 0.05, thus the normality requirement is complied with.

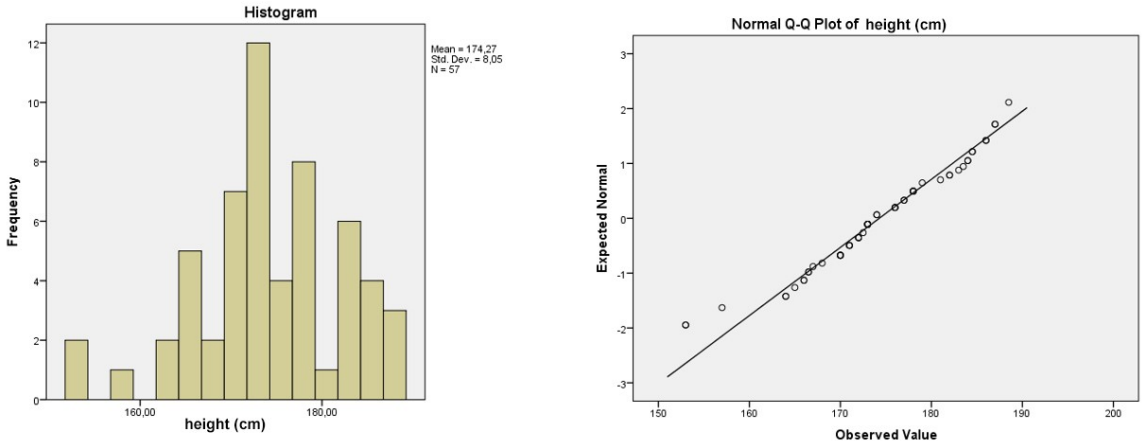
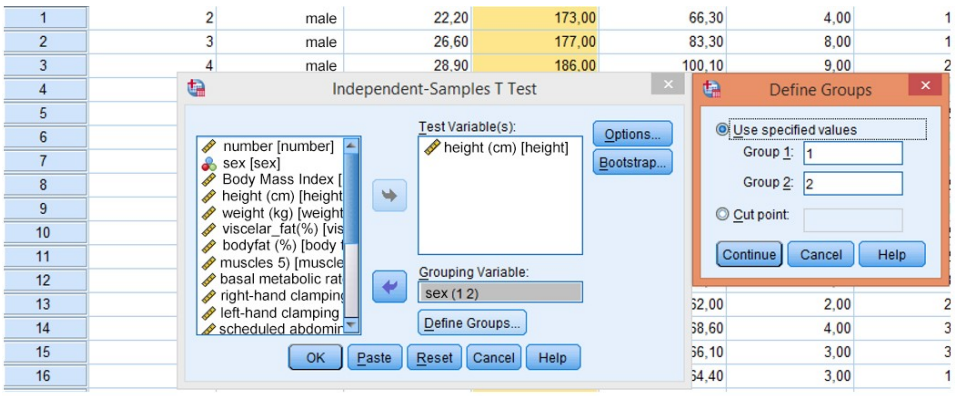


Figure 2/30. Figures of normality test of age (histogram, Q-Q plot)

The figures also state that the distribution is normal, since the histogram fits a Gauss curve, and the Q-Q plot fits the line, too. To sum up, the variable height follows normal distribution ($p<0.05$) so parametric analyses can be carried out that is why the two-sample t-test can be applied to decide on the hípothesis.

The proper method can be found in ANALYSE / COMPARE MEANS / INDEPENDENT-SAMPLES T TEST. Height (Testmagasság (cm)) has to be moved to the box of test variable(s), and where sex (nem) should be indicated as the grouping variable. Here, the codes (1 and 2) of the two sexes will need to be specified, too.



Screen view 2/71. Settings of the two-variable t-test

The first table of the results contains descriptive statistics:

Table 2/52. Descriptive statistics

Group Statistics

	sex	N	Mean	Std. Deviation	Std. Error Mean
height (cm)	male	28	179,0893	5,37321	1,01544
	female	29	169,6207	7,49244	1,39131

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 2/53. Results of the two-samples t-test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
height (cm)	Equal variances assumed	,631	,430	5,466	55	,000	9,46860	1,73236	5,99688	12,94031
	Equal variances not assumed			5,497	50,821	,000	9,46860	1,72246	6,01032	12,92688

Further results are the same as the ones above, and thus the null hypothesis will also have to be rejected. The positive sign before the t-value suggests that the average of the first variance category is the greater one. This means that men's height is significantly higher than women's, and this phenomenon is not random.

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.

In the followings we present the method of variance analysis which is applicable in case of more than one population. This method tries to measure subsample differences based on quantitative variables where the subsamples were generated based on one or more qualitative variables. The aim of the variance analysis (**An**alysis **O**f **Va**riance=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:

1. testing a hypothesis if the expected values of more than two (sub)populations are equal;
2. a homogeneity test;
3. 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 (μ), the group effect of class j (τ_j) and the random effect ε_{ji}. The following hypothesis system is tested:

$$H_0 : \mu_1 = \mu_2 = ...\mu_m = \mu$$
$$H_1 : \mu_j \neq \mu$$

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 2/54. Sum of squares formulas (total, external, internal)

Sum of squares	Type of sum of squares
$SS = \sum_{j=1}^{m} \sum_{i=1}^{nj} (x_{ji} - \bar{x})^2$	Total
$SS_K = \sum_{j=1}^{m} (\bar{x}_j - \bar{x})^2$	External (between groups)
$SS_B = \sum_{j=1}^{m} \sum_{i=1}^{nj} (x_{ji} - \bar{x}_j)^2$	Internal (within groups)

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 alternative 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 2/55. Summary of factors and formulas

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	$\frac{SS}{n-1}$	

Let us figure out if the long-jump values differ in the different BMI categories. To do so, let us consider the following random sample:

D	E	F	G
underweight	normal	overweight	obese
242,00	240,00	245,00	120,00
172,00	237,00	237,00	120,00
172,00	244,00	254,00	
	203,00	165,00	
	263,00	185,00	
	188,00	215,00	
	213,00	204,00	
	204,00	165,00	
	192,00		
	200,00		
	180,00		
	180,00		
	179,00		
	205,00		
	230,00		
	222,00		
	238,00		
	187,00		
	175,00		
	182,00		
	188,00		
	184,00		

Screen view 2/72. Long-jump data (part of the database)

Let us examine if the long-jump data in differet BMI categories can be considered to be equal, i.e. if the BMI category and the long-jump are independent, and standing long jump results of students are homogeneous.

Assuming that standing long-jump values follow normal distribution, and standard deviations are equal in all BMI categories, variance analysis can be applied. (Source:fittségi 57fő_adatbázis_alap_bmikat. xlsx). Partial results to be used later are listed below:

Table 2/56. Partial results

		Standing long-jump (cm)		
		Mean	Std. Dev.	Number
BMI categories	underw.	195.33	40.41	3
	normal	212.61	26.88	44
	overw.	208.75	35.08	8
	obese	120.00	0.00	2
	Total	**207.91**	**32.69**	**57**

First, determine the main average:

$$\bar{x} = \frac{3 \times 195.33 + 44 \times 212.61 + 8 \times 208.75 + 2 \times 120}{57} = 207.91$$

Sum of squares:

$$SS_K = 3(195.33 - 207.91)^2 + 44(212.61 - 207.91)^2 + 8(208.75 - 207.91)^2 + 2(120 - 207.91)^2 = 16{,}909.96$$

$$\mathrm{SS_B} = 3 \times (40.41)^2 + 44 \times (26.88)^2 + 8 \times (35.08)^2 + 2 \times (0.00)^2 = 46{,}526.77$$

$$F = \frac{\frac{SS_K}{m-1}}{\frac{SS_B}{n-m}} = \frac{\frac{16{,}909.96}{4-1}}{\frac{46{,}526.77}{57-4}} = 6.42$$

Consider the F critical value with the degree of freedom (3;57): ${}_{3;53}F_{0.05}=2.78$

The null hypothesis has to be rejected, weight of sportsmen differ in types of sports. The calculation of the standard deviation ratio:

$$H = \sqrt{\frac{SS_K}{SS}} = \sqrt{\frac{SS_K}{SS_K + SS_B}} = \sqrt{1 - \frac{SS_B}{SS}} = \sqrt{\frac{16{,}909.96}{63{,}436.73}} = 0.51$$

A moderate association between standing long-jump results and BMI categories has been detected for university students. BMI category determines standing long-jump results by 26.66%, i.e. 26.66% of standard deviation (H^2).

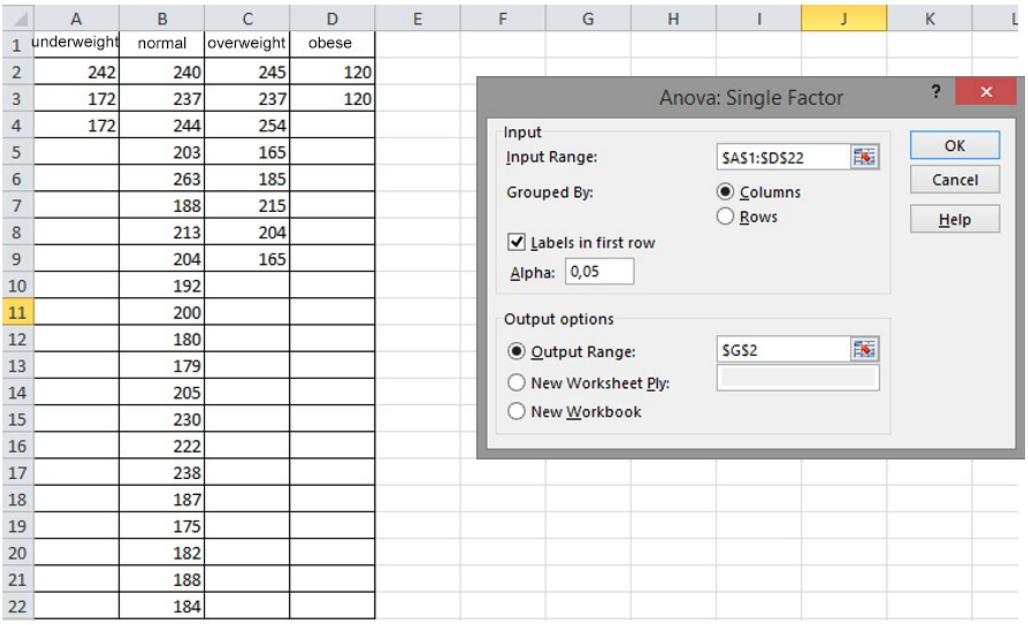
One-way variance analysis can be calculated quickly in Excel since there is a Data analysis module for this purpose. Data have to form a correlate range, and the subsamples have to be ordered according to row and column as well. It is assumed that weights follow a normal distribution.

The hypothesis system remains the same:

$$H_0 : \mu_1 = \mu_2 = ...\mu_m = \mu$$

$$H_1 : \mu_j \neq \mu$$

Let data ordered in column be the input. As BMI category names are included, the option "labels in the first row" has to be selected. Default settings of the alpha parameter (significance level) can remain 0.05, and then the output range has to be given where we would like to put the results.



Screen view 2/73. One-way variance analysis in Excel

Press OK to get the following results:

one-way variance analysis

summary

groups	size	sum	mean	variance
underweight	3	586	195,3333	1633,333
normal weight	44	9355	212,6136	722,3356
overweight	8	1670	208,75	1230,5
obese	2	240	120	0

variance analysis

factors	*SS*	*df*	*MS*	*F*	p-value	F crit.
between groups	16909,96292	3	5636,654	6,957115	0,000493	2,779114
within groups	42940,59848	53	810,2			
total	59850,5614	56				

Screen view 2/74. Results of the variance analysis

The first part of the results contains basic statistics of BMI categories, showing that the mean long-jump result of the three underweigh students is 195.33 cm with a variance of 1633.33. Note that the default interpretation of the program – like in the most statistical softwares – is the corrected standard deviation.[40] The corrected standard deviation slightly

[40] $s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{n-1}}$

exceeds the standard deviation but the difference is small, especially in the case of big samples. That is why the internal standard deviation is not equal to the value calculated before but the difference is not significant.
Slight differences are caused by standard deviation data so the test function value is 6.96 which is higher than the critical value 2.78 which means that the null hypothesis has to be rejected; weight of sportsmen is heterogeneous according to types of sport. Significance-level gives the same result since if we reject the null hypothesis, the probability of the error is very small (0.1%).

Let us present a one-way variance analysis in SPSS.

First, the distribution of the continuous variable has to be tested since normality is a requirement for a variance analysis. The 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 the standing long-jump 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 57 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 2/57. Results of the Kolmogorov- Smirnov test

Descriptive Statistics

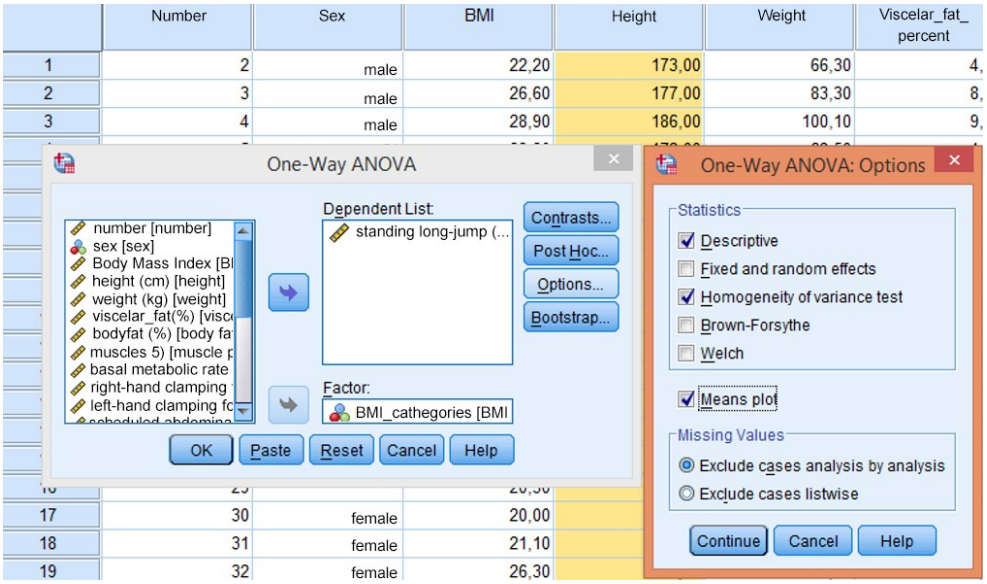
	N	Mean	Std. Deviation	Minimum	Maximum
standing long-jump (cm)	57	207,9123	32,69190	120,00	265,00

One-Sample Kolmogorov-Smirnov Test

		Helyből távolugrás (cm)
N		57
Normal Parameters[a,b]	Mean	207,9123
	Std. Deviation	32,69190
Most Extreme Differences	Absolute	,094
	Positive	,080
	Negative	-,094
Test Statistic		,094
Asymp. Sig. (2-tailed)		,200[c,d]

Based on the z value of the Kolmogorov-Smirnov test and the significance level (p=0.20), 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 Ács Pongrác: Data Analysis). 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 standing long-jump as the dependent variable (DEPENDENT LIST) and the BMI category as the independent (grouping) variable (FACTOR).



Screen view 2/75. Variance analysis settings 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 (sample size, mean, standard deviation, standard error, the lower and upper bound of confidence interval, and the minimum and maximum values):

Table 2/58. Descriptive Statistics

Descriptives

standing long-jump (cm)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
underweight	3	195,3333	40,41452	23,33333	94,9381	295,7286	172,00	242,00
normal	44	212,6136	26,87630	4,05176	204,4425	220,7848	175,00	265,00
overweight	8	208,7500	35,07848	12,40212	179,4237	238,0763	165,00	254,00
obese	2	120,0000	,00000	,00000	120,0000	120,0000	120,00	120,00
Total	57	207,9123	32,69190	4,33015	199,2380	216,5866	120,00	265,00

According to the table, there were 3 people in the category of underweight ("sovány"). Mean of the standing long-jump in the category of overweight ("túlsúlyos") was 208.75 cm with a standard deviation of 35.08 cm. It is also clear that 95% confidence interval for standing long-jump was between 165 and 254 cm in the category.

The next table tests the homogeneity of variances by Levene's test.

Table 2/59. Homogenity of Standard Deviation

Test of Homogeneity of Variances

standing long-jump (cm)

Levene Statistic	df1	df2	Sig.
2,720	3	53	,054

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=0.054), the variances can be considered to be equal. The next table is about variance analysis.

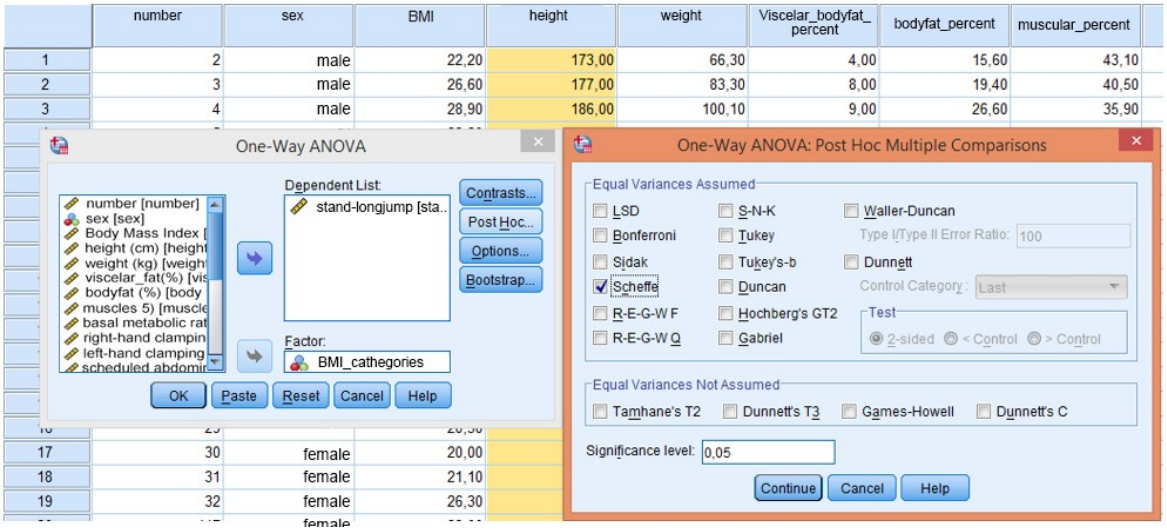
Table 2/60. Anova table

ANOVA

standing long-jump (cm)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	16909,963	3	5636,654	6,957	,000
Within Groups	42940,598	53	810,200		
Total	59850,561	56			

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=6.96) by comparing the mean squares between and within groups. Significance is lower than 0.05 so the null hypothesis will be rejected which means that the long-jump results will differ according to the 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.



Screen view 2/76. Post Hoc Test settings

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 α-error according to the number of independent comparisons. Test statistics of the Bonferroni test:

$$L = t(Table)\sqrt{S_p^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

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:

$$L = \sqrt{s_p^2(k-1)F_{(Table)}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

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:

$$\bar{x}_i - \bar{x}_o \pm |d| s_p \sqrt{\frac{2}{n}}$$

$$\bar{x}_o = control\ group$$

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 2/61. Post Hoc Tests of variance analysis

Post Hoc Tests

Multiple Comparisons

Dependent Variable: standing long-jump (cm)

Scheffe

(I) BMI cathegories	(J) BMI cathegories	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
underweight	normal	-17,28030	16,98471	,793	-66,3227	31,7621
	overweight	-13,41667	19,27023	,922	-69,0584	42,2251
	obese	75,33333*	25,98397	,049	,3060	150,3606
normal	underweight	17,28030	16,98471	,793	-31,7621	66,3227
	overweight	3,86364	10,94023	,989	-27,7257	35,4530
	obese	92,61364*	20,57945	,001	33,1916	152,0357
overweight	underweight	13,41667	19,27023	,922	-42,2251	69,0584
	normal	-3,86364	10,94023	,989	-35,4530	27,7257
	obese	88,75000*	22,50278	,003	23,7745	153,7255
obese	underweight	-75,33333*	25,98397	,049	-150,3606	-,3060
	normal	-92,61364*	20,57945	,001	-152,0357	-33,1916
	overweight	-88,75000*	22,50278	,003	-153,7255	-23,7745

*. The mean difference is significant at the 0.05 level.

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 ("elhízott") ($p<0.049$). 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:

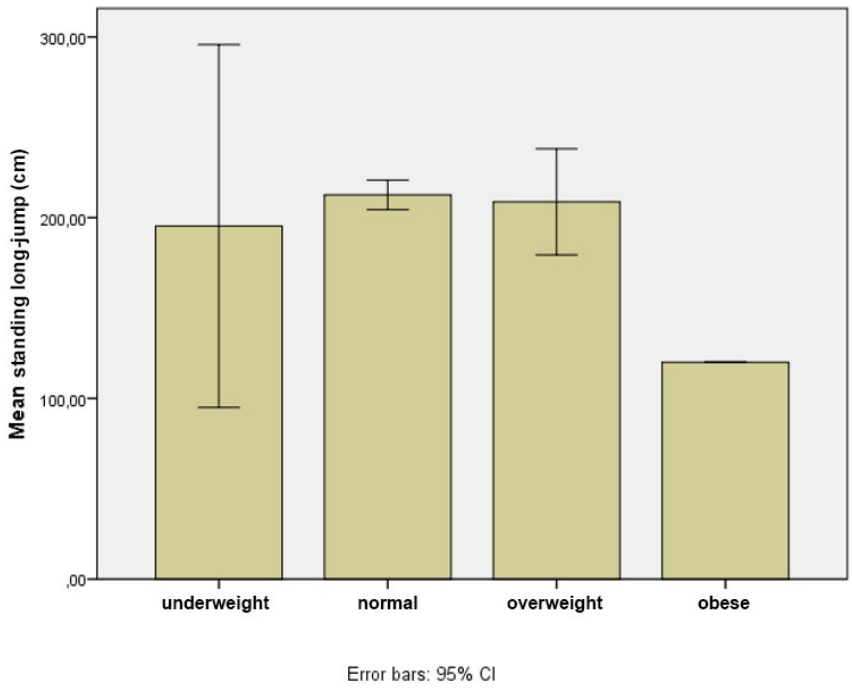


Figure 2/31. Standing long-jump according to BMI categories

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. Unfortunately, post hoc test cannot be carried out in the case of non-parametric tests (Ács, 2015).

In practice, multiple-way variance analysis is usually applied which differs from the one-way one as it measures the effect of more independent variables at once. For further details in the topic consult Sajtos László- Mitev Ariel (2007) or Székelyi Mária- Barna Ildikó (2005).

3. MULTIPLE VARIABLE METHODS

Multiple variable methods can also be applied in individual research. Databases generated like these are mostly analyzed with the help of the SPSS software, since there are more than 250,000 users from the fields of business, academia and government relying on SPSS technology. SPSS is a helpful tool when one has to handle a lot of data, test hypotheses, and look for consequences, because it provides useful results very quickly.

We would like to present the most commonly used methods with the help of an additional database (Source: motor.sav). The database was prepared by Dániel Kehl, associate professor, at the Faculty of Business and Economics (Pécsi Tudományegyetem Közgazdaságtudományi Kar), the University of Pécs, with the aim to make the fundamental methods of multivariate statistics relatively easy to understand. 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 ("gyártó"): most manufacturers have several products in the database,
2. type ("típus"): type mark applied by the manufacturers,
3. origin ("származás"): nationality of the manufacturer,
4. engine displacement ("lökettér", cm^3)
5. performance ("telj_kw", KW)
6. performance (horse-power, "telj_le")
7. weight ("tömeg", kg)
8. consumption ("fogyasztás", l/100km)
9. acceleration ("gyorsulás", 0-100 km/h (s))
10. terminal velocity ("végsebesség", km/h)
11. price ("ár", Ft).

The original database is to be seen on the following screen:

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	manufacturer	String	45	0		None	None	8	Left	Nominal	Input
2	type	String	78	0		None	None	11	Left	Nominal	Input
3	code of origin	String	24	0	place of origin	None	None	7	Left	Nominal	Input
4	displacement	Numeric	11	0	displacement (cm	None	None	4	Right	Scale	Input
5	perf. (kW)	Numeric	11	0	perf. (kW)	None	None	4	Right	Scale	Input
6	perf. (LE)	Numeric	11	0	perf. (LE)	None	None	5	Right	Scale	Input
7	moment	Numeric	11	0	momen (Nm)	None	None	4	Right	Scale	Input
8	weight	Numeric	11	0	weight (Kg)	None	None	4	Right	Scale	Input
9	consumption	Numeric	11	1	consumption (l/10	None	None	3	Right	Scale	Input
10	acceleration	Numeric	11	1	acceleration (0-10	None	None	4	Right	Scale	Input
11	terminal veloci	Numeric	15	0	terminal velocity (	None	None	5	Right	Scale	Input
12	price	Numeric	15	0	price (Ft)	None	None	7	Right	Scale	Input
13											
14											

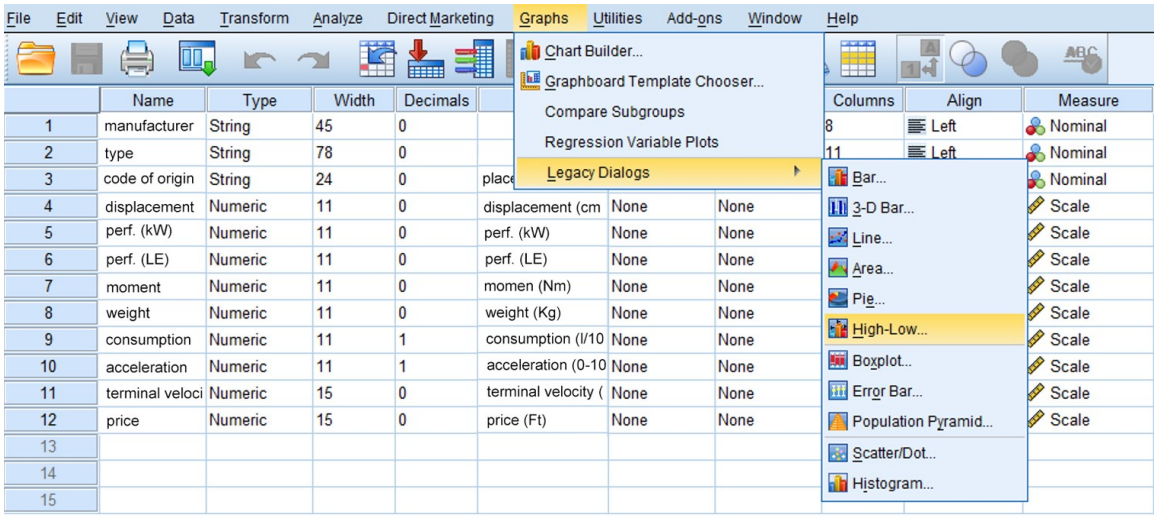
Screen view 3/1. The original database

Before analysing the continuous variables, it is reasonable to check the "unusual" values of the variables that may arise. This method will be shortly presented here.

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.

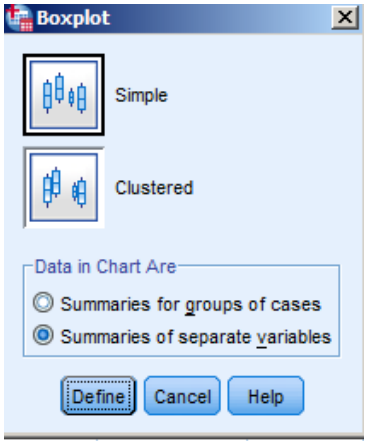
Let us examine the consumption of the motors in our database (motor.sav) with the help of a Boxplot.

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



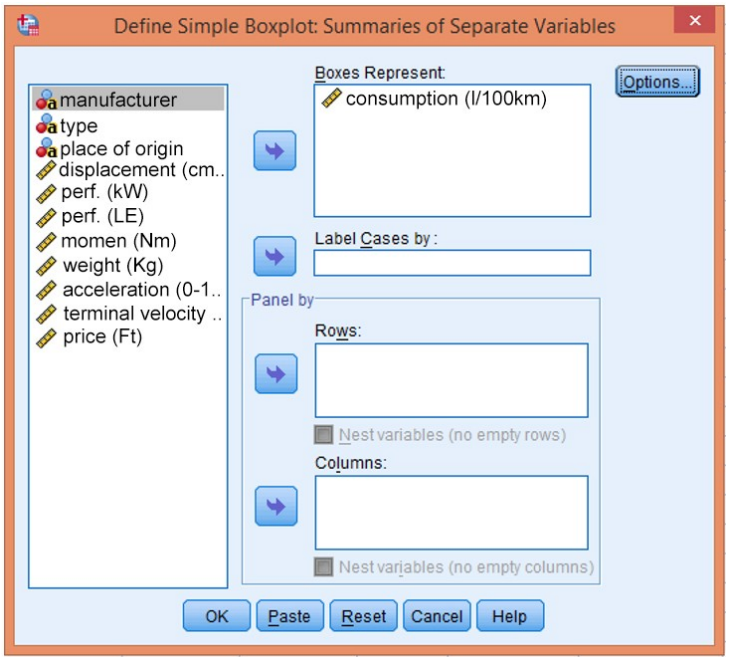
Screen view 3/2. Access path to the option Boxplot

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. terminal velocity) can be plotted depending on categories of another variable (e.g. origin). If we choose summaries of separate variables, the option *CLUSTERED* displays at least two continuous variables (e.g. performance, terminal velocity) according to the categories of another variable (e.g. origin).



Screen view 3/3. Calibrating the type of Boxplot

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



Screen view 3/4. Selecting the variable(s)

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 *.

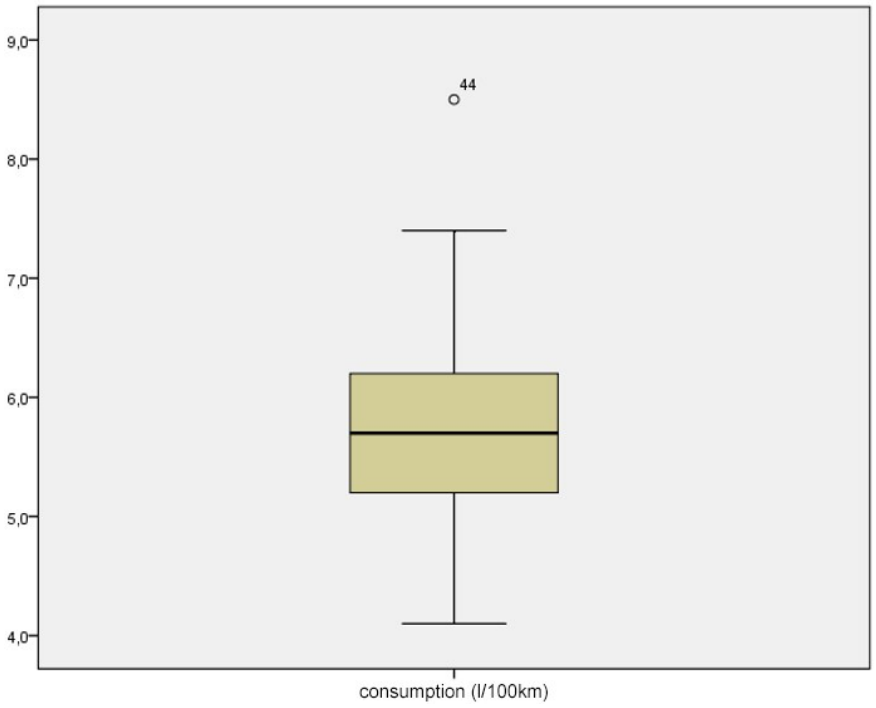


Figure 3/1. The Box-plot of variable consumption

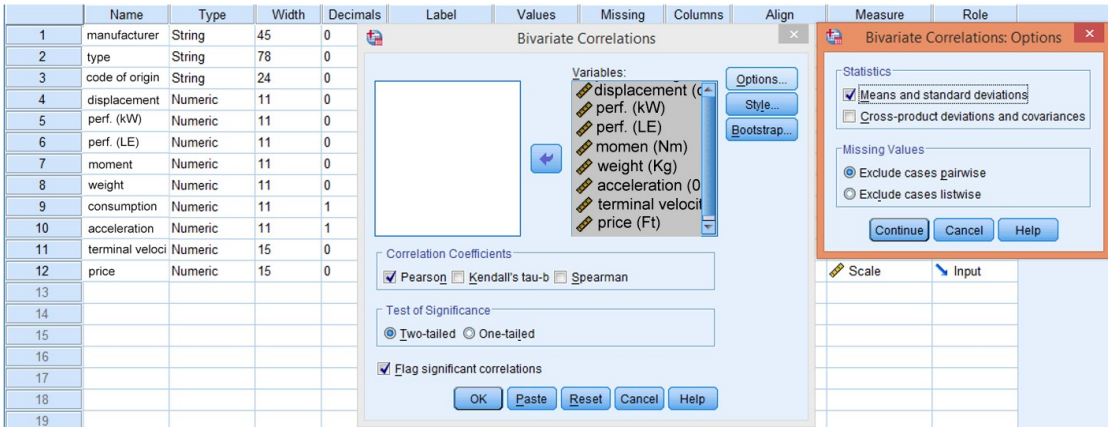
As shown by the figure, one outlier value appears for consumption (motor #44) but there is no extreme value. The data of motorcycle #44 can be checked in data view.

	manufac	type	code of	disp	perf_	perf_L	mome	weigh	con	accel	termina	price
26	Honda	Fireblade	japanese	954	110	150	104	199	5,3	3,0	277	3149000
27	Honda	VTR 1000 SP-2	japanese	999	99	135	102	218	5,7	3,1	278	3348000
28	Honda	X-11 CBS	japanese	1137	100	136	113	254	6,4	2,9	251	2590000
29	Honda	CBR 1100 XX	japanese	1137	112	152	119	254	6,5	3,0	290	2890000
30	Honda	GL 1800 Gold ...	japanese	1832	87	118	167	399	6,8	4,1	200	6490000
31	Kawasaki	ZZR 600	japanese	599	72	98	66	221	6,1	3,8	236	2190000
32	Kawasaki	Ninja ZX-6R	japanese	636	87	118	67	188	6,1	3,1	263	2550000
33	Kawasaki	Ninja ZX-9R	japanese	899	105	143	100	211	5,3	2,9	276	3250000
34	Kawasaki	ZZR 1200	japanese	1164	112	152	124	276	6,7	2,7	275	3350000
35	Kawasaki	Ninja ZX-12R	japanese	1199	131	178	134	249	6,8	2,7	298	3630000
36	Kawasaki	VN 1500 Mean ...	japanese	1471	53	72	114	314	6,2	4,8	182	3400000
37	Moto Guzzi	V11	italian	1064	67	91	94	246	6,2	3,9	214	2699000
38	MV Agusta	F4 S EVO3	italian	749	101	137	80	205	6,1	3,1	274	.
39	Suzuki	GSX-R 600	japanese	600	85	116	68	188	5,5	3,1	254	2398000
40	Suzuki	GSX 750 F	japanese	750	68	92	67	235	4,8	3,6	223	1798000
41	Suzuki	DL 1000 V-Storm	japanese	996	72	98	100	235	5,1	3,4	200	2498000
42	Suzuki	GSF 1200 SsB...	japanese	1157	72	98	91	244	4,7	3,2	235	1998000
43	Suzuki	GSX 1300 R Ha...	japanese	1299	129	175	138	251	6,6	2,7	295	3198000
44	Suzuki	VL 1500 Intrude...	japanese	1462	49	67	112	314	8,5	5,5	160	2748000
45	Triumph	Speed Triple	english	955	88	120	100	210	5,0	3,0	240	2998000
46	Triumph	Daytona 955i	english	955	108	147	100	212	5,3	3,1	260	3499000

Screen view 3/5. The consumption of motor #44 in data view

One can see that consumption of motorcycle #44 is 8.5 l/100 km which - based on engine displacement (1462) - seems to be a realistic result. That is why it is not a consequence of typing mistakes so the data does not have to be excluded. Should there be a typing mistake detected, the data had to be excluded. In these cases a *missing* value interval may be defined, or a data filter may be applied (*DATA* / *SELECT CASES*).

The first step is to find out the relations of the quantitative variables (engine displacement, performance (kW), performance (LE), moment ("nyomaték"), weight, consumption, acceleration, terminal velocity, price) The analysis method is calculating correlation under ANALYZE/CORRELATE/BIVARITE



Screen view 3/6. Settings of calculating correlation

First, one has to select the variables, then after pressing the OPTIONS button, choose to display means and standard deviations. Press CONTINUE and OK to get the following results:

Table 3/1. Descriptive statistics

Descriptive Statistics

	Mean	Std. Deviation	N
displacement (cm3)	1044,57	277,637	53
performance (kW)	78,45	24,664	53
performance (LE)	106,72	33,458	53
momen (Nm)	98,13	22,349	53
weight (Kg)	248,36	51,095	53
consumption (l/100km)	5,779	,8092	53
acceleration (0-100 k	3,825	1,1437	53
terminal velocity (km/h)	225,94	42,198	53
price (Ft)	3440618,00	1343598,867	50

The first table contains descriptive statistics (mean, standard deviation, number of items considered in the analysis). Data size of price (N=50) differs from that of the other

variables (N=53), which is possible due to the fact that the seller does only provide information on price of the three types in case of direct request.

The next table shows the correlation analysis of the variables. Besides the existence of the correlation (Sig. 2-tailed), the strength of correlation is easy to be found. Significant correlations are denoted by one or two stars.

Table 3/2. Correlations matrix

Correlations

		displacement (cm3)	perf. (kW)	perf. (LE)	moment (Nm)	weight (Kg)	consumption (l/100km)	acceleration (0-100 km/h (s))	terminal velocity (km/h)	price (Ft)
displacement (cm3)	Pearson Correlation	1	-,087	-,086	,850**	,820**	,408**	,404**	-,335*	,607**
	Sig. (2-tailed)		,537	,539	,000	,000	,002	,003	,014	,000
	N	53	53	53	53	53	53	53	53	50
performance (kW)	Pearson Correlation	-,087	1	1,000**	,401**	-,327*	,118	-,823**	,937**	-,004
	Sig. (2-tailed)	,537		,000	,003	,017	,400	,000	,000	,977
	N	53	53	53	53	53	53	53	53	50
performance (LE)	Pearson Correlation	-,086	1,000**	1	,401**	-,327*	,119	-,823**	,937**	-,004
	Sig. (2-tailed)	,539	,000		,003	,017	,395	,000	,000	,979
	N	53	53	53	53	53	53	53	53	50
moment (Nm)	Pearson Correlation	,850**	,401**	,401**	1	,589**	,415**	-,044	,129	,537**
	Sig. (2-tailed)	,000	,003	,003		,000	,002	,757	,357	,000
	N	53	53	53	53	53	53	53	53	50
weight (Kg)	Pearson Correlation	,820**	-,327*	-,327*	,589**	1	,351**	,616**	-,542**	,658**
	Sig. (2-tailed)	,000	,017	,017	,000		,010	,000	,000	,000
	N	53	53	53	53	53	53	53	53	50
consumption (l/100km)	Pearson Correlation	,408**	,118	,119	,415**	,351**	1	,104	,008	,221
	Sig. (2-tailed)	,002	,400	,395	,002	,010		,458	,954	,124
	N	53	53	53	53	53	53	53	53	50
acceleration (0-100 km/h (s))	Pearson Correlation	,404**	-,823**	-,823**	-,044	,616**	,104	1	-,883**	,305*
	Sig. (2-tailed)	,003	,000	,000	,757	,000	,458		,000	,031
	N	53	53	53	53	53	53	53	53	50
terminal velocity (km/h)	Pearson Correlation	-,335*	,937**	,937**	,129	-,542**	,008	-,883**	1	-,191
	Sig. (2-tailed)	,014	,000	,000	,357	,000	,954	,000		,184
	N	53	53	53	53	53	53	53	53	50
price (Ft)	Pearson Correlation	,607**	-,004	-,004	,537**	,658**	,221	,305*	-,191	1
	Sig. (2-tailed)	,000	,977	,979	,000	,000	,124	,031	,184	
	N	50	50	50	50	50	50	50	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Based on the given results, some typical features can be found:

- engine displacement is strongly correlated to moment which is a consequence of the principle of the internal-combustion motor: bigger cylinder capacity means bigger combustion,
- engine displacement does actually not correlate with performance which is not exactly the fact in reality: if only sport motorcycles had been considered, this correlation would be stronger,
- weight and terminal velocity have a moderate negative correlation which does not have to be explained in more details,
- the strong correlation between acceleration and terminal velocity means that the motors with the higher terminal velocity reach 100 km/h earlier (the reason behind this is that both values are determined by the motor performance),
- price is not correlated to other variables stronger than moderate; it is interesting that it has the strongest correlation with weight,
- consumption has no correlation stronger than moderate; moreover, according to the database, it is absolutely independent of the terminal velocity

- performance in kW and horsepower (LE) have a deterministic correlation because 1 LE ≈ 0,735 kW.

In the next subsections the most commonly used multivariate statistical methods (factor analysis, cluster analysis, discriminate analysis) are shortly presented.

3.1. Factor analysis

Applying factor analysis in multivariate practical analysis has become more usual since it can compress data and explore relations. The method makes it possible to summarize a high number of variables in factor variables that can not be directly observed. A small number of factor variables are seeked for instead of a lot of stochastically connecting variables. This makes data interpretation and analysis simpler since the number of variables shrinks.

The new factors do not correlate at all. Practical application is due to the common usage of survey research methodology, since surveys tend to overanalyze some topics (habits, characteristics, lifestyles, etc.) which can make data processing very complex. In cases like this, researchers tend to apply this method since it attempts to explore relations between features by reducing the number of variables. Factor analysis is a structure exploring method which means that dependent and independent variables are not predetermined so it attempts to explore relations. (Sajtos L.- Mitev A. ,2007)

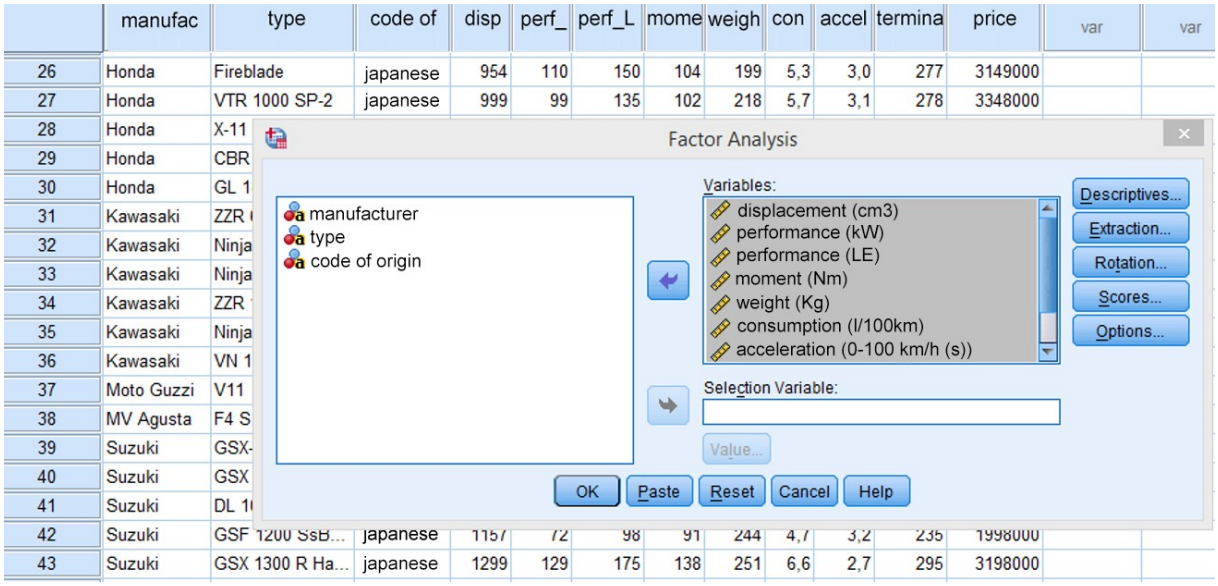
An additional advantage of factor analysis is that the factors created can be used in further multivariate analyses.

Most common questions arising in factor analysis:

- How can information be explained by several variables expressed by a small number of uncorrelated factors?
- To what extent do the new factors explain the original variables?
- Which variables are included in the same factors?
- What do the factors mean, how can they be named?

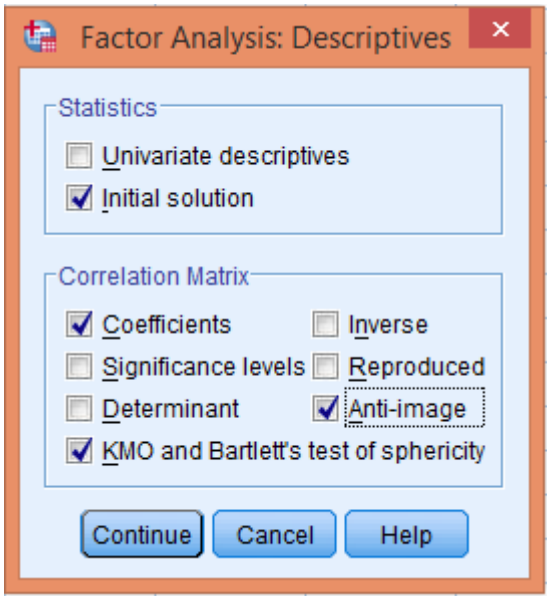
(Source: Ketskeméty-Izsó, 2005)

Factor analysis can be carried out under ANALYZE/DATA REDUCTION/FACTOR where variables have to be selected and moved to the VARIABLES window with the arrow. (Source: motor.sav)



Screen view 3/7. Factor analysis settings

Next, in the box DESCRIPTIVES one can test if the variables considered above are suitable for factor analysis. Besides the basic settings in STATISTICS one can also select univariate decriptives (UNIVARIATE DECREPTIVES), resulting in the table (mean, standard deviation, sample size) introduced above.



Screen view 3/8. Requirements settings

Correlations matrix can be created here as well. It is an important prerequisite of the analysis since correlation of variables is a basic requirement of factor analysis. Strong correlation between variables indicates the variables being suitable for factor analysis. Selecting the box COEFFICIENT displays the correlation values (coefficients) of the correlations matrix.

Table 3/3. Correlations matrix

Correlation Matrix

		displacement (cm3)	perf. (kW)	perf. (LE)	moment (Nm)	weight (Kg)	consumption (l/100km)	acceleration (0-100 km/h (s))	terminal velocity (km/h)	price (Ft)
Correlation	displacement (cm3)	1,000	-,069	-,069	,850	,821	,429	,396	-,321	,607
	performance (kW)	-,069	1,000	1,000	,421	-,319	,111	-,826	,937	-,004
	performance (LE)	-,069	1,000	1,000	,421	-,319	,112	-,825	,937	-,004
	moment (Nm)	,850	,421	,421	1,000	,593	,424	-,052	,149	,537
	weight (Kg)	,821	-,319	-,319	,593	1,000	,385	,608	-,542	,658
	consumption (l/100km)	,429	,111	,112	,424	,385	1,000	,122	,000	,221
	acceleration (0-100 km	,396	-,826	-,825	-,052	,608	,122	1,000	-,890	,305
	terminal velocity (km/h)	-,321	,937	,937	,149	-,542	,000	-,890	1,000	-,191
	price (Ft)	,607	-,004	-,004	,537	,658	,221	,305	-,191	1,000

This table is the same as the one above. In the box DESCREPTIVE the test of the other important prerequisite, the ANTI- IMAGE has been selected. It assumes that the square of the variables' standard deviation can be separated into explained and unexplained parts which are shown by the anti-image covariance and variance matrices. The diagonal values of the anti-image correlations matrix are the MSA values. These can range from 0 to 1, and values in the diagonal show how string the given variable is correlated to the others considered. If the MSA value is high, then the variable does fit the factor structure well. If it is low (under 0.5) then it is very likely that the variable will have to be excluded from the analysis.

Table 3/4. Table Anti- image

Anti-image Matrices

		displacement (cm3)	perf. (kW)	perf. (LE)	moment (Nm)	weight (Kg)	consumption (l/100km)	acceleration (0-100 km/h (s))	terminal velocity (km/h)	price (Ft)
Anti-image Covariance	displacement (cm3)	,050	7,482E-5	-3,695E-5	-,039	-,008	-,019	-,014	-,008	-,015
	performance (kW)	7,482E-5	6,422E-5	-6,448E-5	-9,820E-5	,000	,001	,001	,000	-2,067E-5
	performance (LE)	-3,695E-5	-6,448E-5	6,487E-5	5,116E-5	,000	-,001	-,001	4,589E-5	-1,933E-5
	moment (Nm)	-,039	-9,820E-5	5,116E-5	,038	-,013	,006	,012	,014	,010
	weight (Kg)	-,008	,000	,000	-,013	,177	-,070	-,030	,015	-,112
	consumption (l/100km)	-,019	,001	-,001	,006	-,070	,723	-,045	-,018	,099
	acceleration (0-100 km	-,014	,001	-,001	,012	-,030	-,045	,156	,020	-,042
	terminal velocity (km/h)	-,008	,000	4,589E-5	,014	,015	-,018	,020	,043	,002
	price (Ft)	-,015	-2,067E-5	-1,933E-5	,010	-,112	,099	-,042	,002	,490
Anti-image Correlation	displacement (cm3)	,714[a]	,042	-,021	-,894	-,085	-,100	-,162	-,182	-,094
	performance (kW)	,042	,729[a]	-,999	-,063	-,085	,112	,185	-,065	-,004
	performance (LE)	-,021	-,999	,730[a]	,032	,085	-,115	-,182	,028	-,003
	moment (Nm)	-,894	-,063	,032	,662[a]	-,161	,034	,160	,346	,075
	weight (Kg)	-,085	-,085	,085	-,161	,894[a]	-,197	-,183	,172	-,380
	consumption (l/100km)	-,100	,112	-,115	,034	-,197	,820[a]	-,134	-,103	,167
	acceleration (0-100 km	-,162	,185	-,182	,160	-,183	-,134	,917[a]	,239	-,153
	terminal velocity (km/h)	-,182	-,065	,028	,346	,172	-,103	,239	,922[a]	,016
	price (Ft)	-,094	-,004	-,003	,075	-,380	,167	-,153	,016	,858[a]

a. Measures of Sampling Adequacy(MSA)

MSA values range here from 0.66 to 0.92. The thing to be tested in case of almost every factor analysis: the KMO (Kaiser- Meyer- Olkin) criterion and the Bartlett test. Applying the KMO criterion is the best and easiest way to decide to what extent the variables are valid for factor analysis. The KMO value is the average of the MSA values which tests all variables at once. From the point of view of factor analysis, KMO values can be categorized as:

0.9 ≤KMO≤1 excellent

0.8 ≤KMO≤0.9 good

0.7 ≤KMO≤0.8 medium

0.6 ≤KMO≤0.7 mediocre

0.5 ≤KMO≤0.6 poor

KMO≤0.5 unacceptable

The null hypothesis of the Bartlett test is that there is no correlation between the original variables, i.e. they are uncorrelated. For us, it would be favourable to be able to reject the null hypothesis, meaning that the variables were correlated.

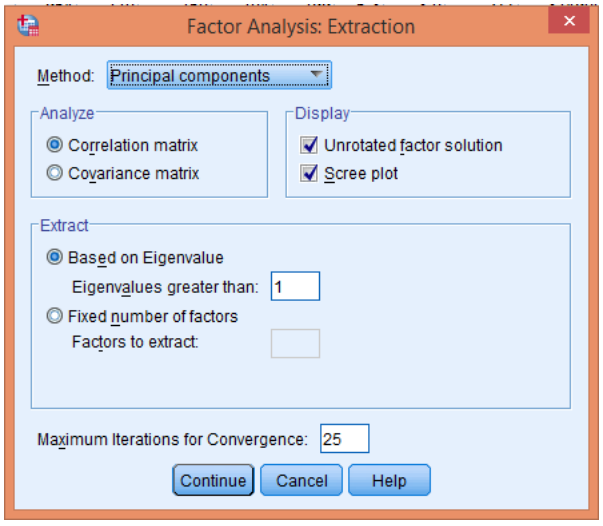
Table 3/5. Values of the KMO Test

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,796
Bartlett's Test of Sphericity	Approx. Chi-Square	901,966
	df	36
	Sig.	,000

The results show that the significance value of the Bartlett test is less than 0.05, i.e. the variables are correlated, so the factor analysis can be carried out.The KMO value (0.796) gives similar results so tha variables are valid for factor analysis.

In the dialog box of the factor analysis, the window EXTRACTION makes it possible to select methods as factor analysis can refer to different methods.

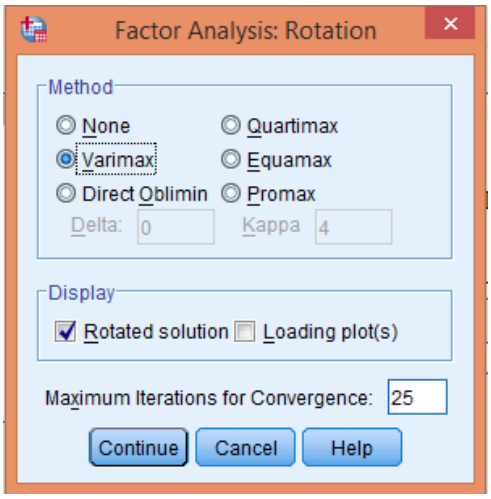


Screen view 3/9. Selecting the method

Let us choose the method PRINCIPAL COMPONENTS since it reduces the number of variables in a way that makes us loose the least information about the population. In the box EXTRACT, one can set the number of factors. If the researcher has an idea about the number of factors, he/she can set them after selecting NUMBER OF FACTORS (the maximum number of factors can not exceed the number of variables). As default, the program uses the KAISER criterion (eigenvalue) which only considers factors with an eigenvalue of minimum 1, since a lower value would mean that the factor contains less information than a variable.

The graphics SCREE PLOT (scree test) helps us to define the number of factors, too. This is the so-called elbow method that says that the number of factors has to be determined where the slope is decreasing and turns to become straight. This means that there can be factors that are important but have an eigenvalue under 1. Compared to the Kaiser criterion, this rule is usually less strict, and permits 1-3 factors more. The decision has to be made by the researcher.

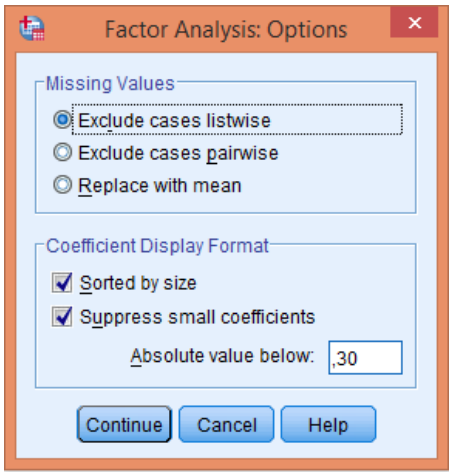
Press CONTINUE and set factor rotation under ROTATION. This means that in the favour of simpler and easier interpretation, the axes of the factors will be rotated. Rotation of factors makes variance explained by factors more proportional. Let us choose the method VARIMAX which is the most commonly used factor analysis method. Its advantage comes from the fact that it does separate factors better, making interpretation easier.



Screen view 3/10. Rotation Settings

After selecting the method, select only ROTATED SOLUTIONS in the DISPLAY box – now, the graphics of componenets (LOADING PLOT) will not be displayed in the rotated

space. The next step is setting OPTIONS where one can ease the interpretation of future factors in advance. By selecting the option SORTED BY SIZE makes the weights be displayed in decreasing order in the rotated factor weight matrix, making interpretation easier.



Screen view 3/11. Settings of rotated factor weight matrix

It can also be set here (SUPPRESS ABSOLUTE VALUES LESS THAN) to display values greater than given factor weights. Let us make it display values greater than 0.3, making interpreting and naming factors quicker. After gaining appropriate factors we can save them in SCORES / SAVE AS VARIABLES so that we can use them in further multivariate analyses (e.g. cluster analysis).

Let us now run the analysis. The output window contains the following results from which the first three tables have already been mentioned.

The fouth table shows the communality test of the variables. Here, one has to accept the "rule of thumb" that the final communality has to exceed 0.3 - otherwise the variables have no explanatory power.

Table 3/6. Communality values

Communalities

	Initial	Extraction
displacement (cm3)	1,000	,894
performance (kW)	1,000	,983
performance (LE)	1,000	,982
moment (Nm)	1,000	,908
weight (Kg)	1,000	,890
consumption (l/100km)	1,000	,331
acceleration (0-100 km	1,000	,894
terminal velocity (km/h)	1,000	,963
price (Ft)	1,000	,574

Extraction Method: Principal Component Analysis.

The Initial value in the table is always 1 while the column Extraction contains the communalities after factor analysis. Based on this, variables do not have to be excluded since all values exceed 0.3.

The next table contains the variance explained by the factors. The three parts of the table include the initial values (Initial), the values after factor analysis (Extraction Sums of Squared Loadings), and the values after rotation (Rotation Sums of Squared Loadings).

Table 3/7. Variance explained by the factors

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,255	47,281	47,281	4,255	47,281	47,281	4,035	44,834	44,834
2	3,162	35,136	82,417	3,162	35,136	82,417	3,382	37,583	82,417
3	,793	8,808	91,225						
4	,457	5,081	96,305						
5	,141	1,571	97,877						
6	,132	1,466	99,343						
7	,042	,464	99,807						
8	,017	,192	100,000						
9	3,229E-5	,000	100,000						

Extraction Method: Principal Component Analysis.

Values resulting from factor analysis and rotation are important for us since their eigenvalues exceed 1 as we requested. The first factor displayed is the factor with the highest eigenvalue (4.255/47.281). It is of highest importance that the two factors coming into existence have a higher cumulative variance (Comulative %) than the 60% criterion since the 82.417 percent means that only 17.583 % of the information has been lost. Values after rotation show that the cumulative variance has remained the same but the distribution has become more uniform. The next figure is the Scree Plot where the slope is decreasing after the third facor, and after that it starts to be flatter.

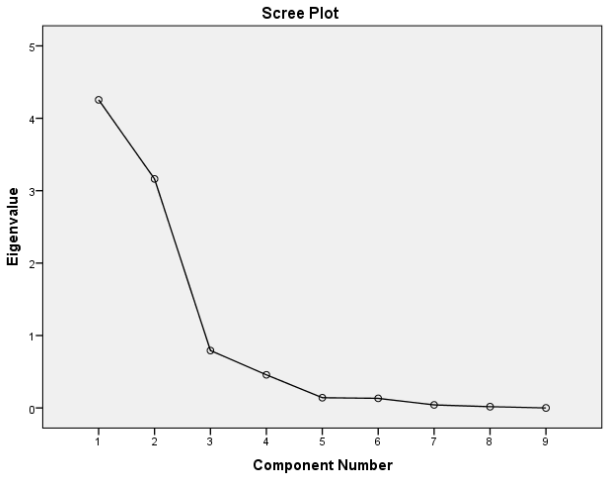


Figure 3/2. The graphics supporting the decision of the number of factors

In pursuing of the elbow method, we maximize the number of factors when the curve starts to get flat. In this case, three factors had to be created so also the third factor may be important despite the fact that its eigenvalue is less than one. Next we gain the matrices containing the factor weights before (Component Matrix) and after rotation. For us, the Rotated Component Matrix is more interesting.

Table 3/8. Rotated Component Matrix

Rotated Component Matrix[a]

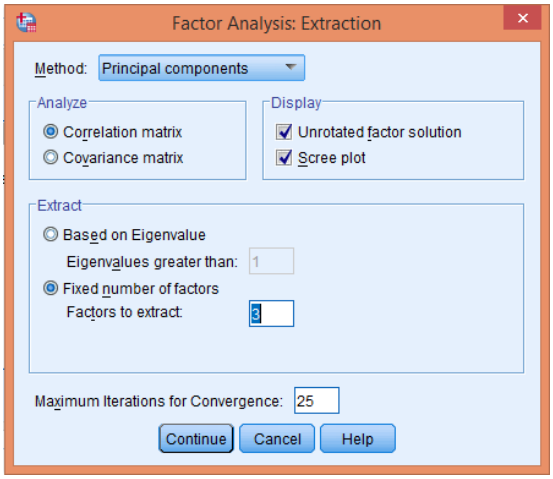
	Component	
	1	2
performance (kW)	,984	
performance (LE)	,984	
terminal velocity (km	,970	
acceleration (0-100	-,903	
displacement (cm3)		,930
moment (Nm)	,323	,896
weight (Kg)	-,438	,835
price (Ft)		,749
consumption (l/100		,571

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

The rotated matrix does only contain values exceeding the factor weights 0.3 as we requested. The higher the absolute value of the factor weight, the more important role is has in the factor. Based on this, the variables in the first factor are: performance (kW), performance (LE), terminal velocity, acceleration. All other variables are included in the second factor.

Let us now see the analysis in case of three factors. To do so, we only have to change one of the settings: let number of factors to extract be 3.



Screen view 3/12. Determining the method and the number of factors

Running the analysis shows that the three factors explain 91.225 percent of the total variance so we will loose a minimal amount of information in case of having three factors.

Table 3/9. Variance explained by the factors

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,255	47,281	47,281	4,255	47,281	47,281	4,000	44,448	44,448
2	3,162	35,136	82,417	3,162	35,136	82,417	3,069	34,098	78,546
3	,793	8,808	91,225	,793	8,808	91,225	1,141	12,679	91,225
4	,457	5,081	96,305						
5	,141	1,571	97,877						
6	,132	1,466	99,343						
7	,042	,464	99,807						
8	,017	,192	100,000						
9	3,229E-5	,000	100,000						

Extraction Method: Principal Component Analysis.

Finally, let us name the three gained factors based on the matrix containing the factor weights after rotation.

Table 3/10. Rotated Component Matrix

Rotated Component Matrix[a]

	Component		
	1	2	3
performance (kW)	,987		
performance (LE)	,987		
terminal velocity (km/h	,964		
acceleration (0-100 k	-,897		
displacement (cm3)		,887	
price (Ft)		,846	
moment (Nm)	,346	,845	
weight (Kg)	-,415	,817	
consumption (l/100km			,947

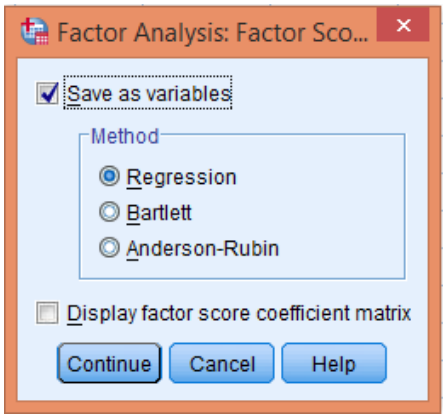
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

- the first component is strongly related to the performance measures, the terminal velocity, and the acceleration. As seen at the descriptive statistics, these variables are strongly correlated that is why they could become members of the same group. The name of the group could be motorcycle power. Acceleration in this component has a negative value, having an opposite meaning: not the high but the low number of seconds is favourable. It is better if the shorter time (sec) is needed to reach 100 km/h.
- the second component is related to engine displacement, price, moment and weight. We may name this component motorical feature.

- the third component is related to consumption. This variable is alone in the group which is not surprising based on the correlation analysis since it is not strongly correlated to any variables.

Finding this solution acceptable, we can save the gained values.



Screen view 3/12. Saving factors

After saving, it is reasonable to name the new factors under LABEL in the VARIABLE VIEW.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	manufacturer	String	45	0		None	None	8	Left	Nominal	Input
2	type	String	78	0		None	None	11	Left	Nominal	Input
3	code of origin	String	24	0	place of origin	None	None	7	Left	Nominal	Input
4	displacement	Numeric	11	0	displacement (cm3)	None	None	4	Right	Scale	Input
5	perf. (kW)	Numeric	11	0	performance (kW)	None	None	4	Right	Scale	Input
6	perf. (LE)	Numeric	11	0	performance (LE)	None	None	5	Right	Scale	Input
7	moment	Numeric	11	0	moment (Nm)	None	None	4	Right	Scale	Input
8	weight	Numeric	11	0	weight (Kg)	None	None	4	Right	Scale	Input
9	consumption	Numeric	11	1	consumption (l/100km	None	None	3	Right	Scale	Input
10	acceleration	Numeric	11	1	acceleration (0-100 k	None	None	4	Right	Scale	Input
11	terminal veloci	Numeric	15	0	terminal velocity (km/	None	None	5	Right	Scale	Input
12	price	Numeric	15	0	price (Ft)	None	None	7	Right	Scale	Input
13	FAC1_1	Numeric	11	5	engine performance	None	None	13	Right	Scale	Input
14	FAC2_1	Numeric	11	5	motor features	None	None	13	Right	Scale	Input
15	FAC3_1	Numeric	11	5	consumption	None	None	13	Right	Scale	Input
16											
17											

Screen view 3/13. Naming factors

3.2. Cluster analysis

Cluster analysis is a method for grouping (clustering) variables and reducing dimensions. Its aim is to reduce the number of observables (factor analysis reduces the number of variables), order them into coherent groups based on the variable included in the analysis. The analysis is successful if the ones belonging to the same group are close to each other but are far from the other groups, clusters.

There are two big methodical groups of cluster analysis. These include the hierarchical (tree-like structure) and the non-hierarchical (K-means) methods. In case of hierarchical methods, contracting (merging) cluster analysis (simple linkage method, complete linkage

method, average linkage method, ward method, centroid method) is applied the most commonly. It means that the items being separated (clusters) at the beginning of the process will be merged into bigger clusters, and eventually into one cluster. This method will be applied when the researcher does not know the number of clusters in advance. The non-hierarchical K-means method is reasonable to be used in case of bigger samples since it can be interpreted easier than hierarchical methods. The number of clusters to be generated has to be fixed in advance.

It is hard to decide which method should be used. The decision depends on the researcher's experstise and his/her surveys carried out so far. That is why in most cases both methods are applied. First, the number of clusters is determined by the hierarchical method, and then the analysis and the clustering of variables is carried out by the non-hierarchical method. In this case we apply the non-hierarchical method since we have information on the number of clusters. That is why we are going to order the types into three clusters. Let us note that if the variables in the analysis had been measured by different scales, then we had to standardize[41] the values first so that the comparison of different-scale data could be carried out. The easiest access path for standardization is ANALYZE/ DESCRIPTIVE STATISTICS /DESCRIPTIVES since the standardized values of variables can be saved in the box SAVE STANDARDIZED VALUES AS VARIABLES.

Cluster analysis can be carried out in ANALYZE/CLASSIFY/K-MEANS CLUSTER.

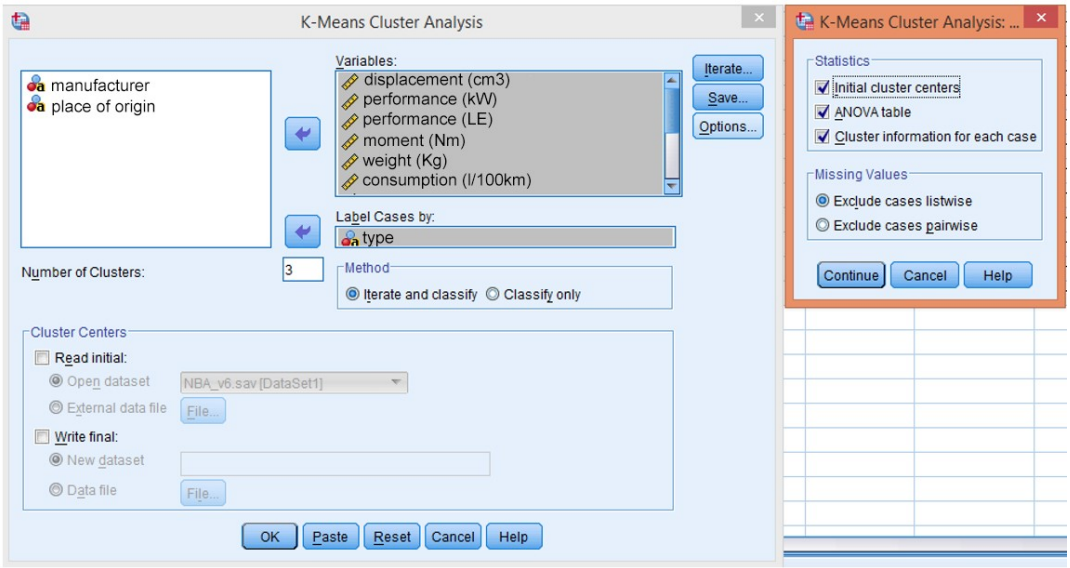


Figure 3/14. Cluster analysis settings

[41] The mean will be substracted from all the values and devided by the standard deviation. As a result, we get zero means and 1 standard deviation. In SPSS this option is to be found in Analyze/Classify/Hierarchial Cluster/Method/Transform Values/Standardize: Z Scores/ By Variable.

First, the variables to be used (engine displacement, performances, moment, weight, consumption, acceleration, terminal velocity, price) have to be moved to the box VARIABLES with the help of the arrow in the middle. The types has to be in the LABEL CASES box since we would like to label based on this. Afterwards, the ANOVA table and cluster information for each case (CLUSTER INFORRMATION FOR EACH CASE) have to be selected in OPTIONS. We do not deal with the ITERATE[42] box now, let us leave the default setting stay unaltered. Press CONTINUE and OK to get the following results:

Table 3/11. Initial Cluster Centers

Initial Cluster Centers

	Cluster		
	1	2	3
displacement (cm3)	750	1298	1449
performance (kW)	68	106	50
performance (LE)	92	144	68
moment (Nm)	67	134	110
weight (Kg)	235	263	385
consumption (l/100km)	4,8	4,9	5,4
acceleration (0-100 km/h	3,6	2,9	6,5
terminal velocity (km/h)	223	245	158
price (Ft)	1798000	3750000	7309000

The first table above shows what original centers the program has used. As we requested three clusters, it generated three centers, based on the number of variables we added to the analysis.

Four iterations have been carried out in the following table.

Table 3/12. Iteration History

Iteration History

	Change in Cluster Centers		
Iteration	1	2	3
1	521368,4	86888,712	764600,0
2	78631,594	51211,558	340828,6
3	50000,000	64621,056	,000
4	,000	,000	,000

[42] Iteration means that the program does reuse cluster centers until a new case is added to the cluster. This lasts until the centroids stop to change, i.e. we get a stable structure.

The Cluster Membership table shows which cluster the different types have been added to. The part of the table displayed contains the cluster number and the distance from the center. Based on this, the motorcycle type Aprilia RST 1000 Futura belongs to cluster #1.

Table 3/13. Cluster Members

Case Number	type	Cluster	Distance
1	RST 1000 Futura	1	401000,0
2	RSV mille R	2	176621,1
3	Tornado 900 i.e.	2	722479,0
4	F 650 GS	1	266000,2
5	R 1100 S	2	378521,1
6	R 1150 RT	2	412479,0

The table of the eventual cluster centers contains very important pieces of information since the clusters ca be characterized and named based on them.

Table 3/14. Final Cluster Centers

Final Cluster Centers

	Cluster		
	1	2	3
displacement (cm3)	931	1071	1418
performance (kW)	70	94	62
performance (LE)	95	128	85
moment (Nm)	86	107	117
weight (Kg)	236	234	345
consumption (l/100km)	5,7	5,7	6,1
acceleration (0-100 km/h (s)	3,9	3,3	5,3
terminal velocity (km/h)	217	252	181
price (Ft)	2448000	3676521	6203571

Consequently, the following clusters can be differentiated:

Cluster #1 (**„street motorcycles”**): this group contains relatively cheap, small and average performance motorcycles. Mostly their engine displacement is low (600-1000 cm^3). Their acceleration and terminal velocity is average.

Cluster #2 („**sport - touring motorcycles**”): vehicles with high engine displacement, high performance, high terminal velocity and moment belong to this cluster. They are generally chosen by customers who are sporty but like touring as well.

Cluster #3 („**cruiser motorcycles**”): heavy but slow-moment and worse-acceleration motorcycles with enormous engine displacement and high prices belong to this cluster. Typical cruiser owners have a special “feeling and lifestyle”.

Table 3/15. Distances between Final Cluster Centers

Distances between Final Cluster Centers

Cluster	1	2	3
1		1228521	3755571
2	1228521		2527050
3	3755571	2527050	

The table Distances between Final Cluster Centers demonstrates that the generated clusters got far from each other. The table shows distances between clusters.

The next table is similar to the already known Anova table but the usual Sum of Squares and Total columns are missing. The explanatory text under the table also points out that this is not a traditional significance test.

Table 3/16. ANOVA table

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
displacement (cm3)	49108,695	2	55645,613	47	11,665	,000
performance (kW)	4284,065	2	478,847	47	8,947	,001
performance (LE)	7858,366	2	882,425	47	8,905	,001
moment (Nm)	3888,055	2	378,853	47	10,263	,000
weight (Kg)	36552,543	2	1238,199	47	29,521	,000
consumption (l/100	,642	2	,683	47	,940	,398
acceleration (0-100	10,136	2	,991	47	10,226	,000
terminal velocity (k	14510,333	2	1299,967	47	11,162	,000
price (Ft)	,907E+013	2	,195E+011	47	178,009	,000

The F tests should be used only for descriptive purposes because the clusters have maximize the differences among cases in different clusters. The observed significa corrected for this and thus cannot be interpreted as tests of the hypothesis that the

The low values of Sig. show that the cluster centers based on all three cluster generators differ significantly. Based on data in the table we see that significant differences can be found in every variable, except the consumption. That is why we will run the analysis once again, excluding the variable above (consumption). The F-values in the table can signal along which variable clusters could be separated the best. The higher the F-value, the better

the cluster meaning that the variable has a more important in the clustering process. That is why the price is the strongest cluster generating variable.

Based on this knowledge, let us run the analysis again, without the consumption variable. Interpretation of the tables explained before is the same. The last new table has not been introduced yet which shows the number of cases in the clusters.

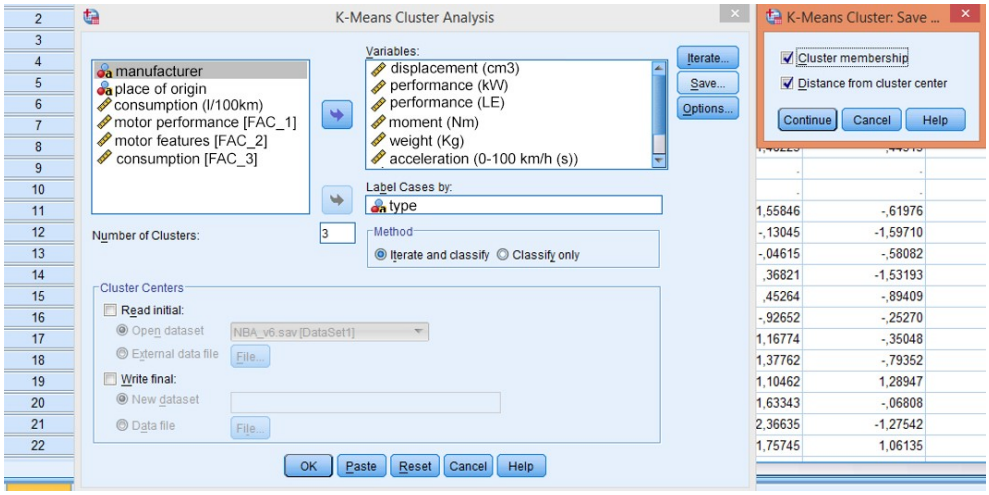
Table 3/17. Number of Cases in each Cluster

Number of Cases in each Cluster

Cluster	1	24,000
	2	19,000
	3	7,000
Valid		50,000
Missing		3,000

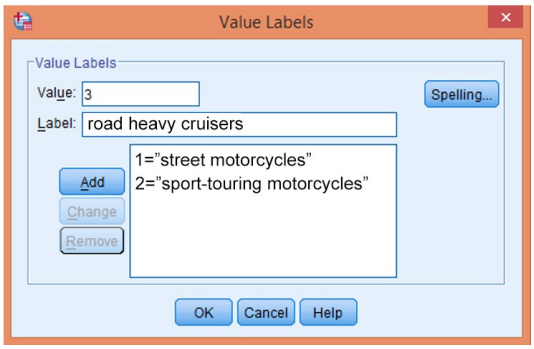
The program has separated fifty motorcycles into three clusters. It could not classify three cases because their price data is missing. The first cluster (street motorcycles) includes 24, the second (sport-touring motorcycles) includes 19, and the third (cruising motorcycles) includes 7 cases.

This analysis does also refer to the strategy of the bigger producers: five BMW products are included in the database from which there is one “street motorcycle”, one “cruiser motorcycle” and the others are “sport-touring motorcycles” as we expected. The Italian Ducati has only motorcycles in cluster #1 while the five from the six motorcycles of the American Harley-Davidson belongs to cluster #3. Let us note that cluster #3 consists of seven cases only. There is one “Harley-imitator” among the nine Honda models (at least its parameters are like that) but all others belong to the other two groups just like all the types of Kawasaki. Suzuki offers almost cluster #1 motorcycles only, and so does Yamaha as well, (Of course, these are only the consequences of our database). It is important to save and name the three new clusters. Select SAVE / CLUSTER MEMBERSHIP / DISTANCE FROM CLUSTER CENTER which means that the cluster distances from the cluster centers and the number of clusters will be saved.



Screen view 3/15. Saving new clusters

After saving, we can save the new factors under LABEL in the VARIABLE VIEW.



Screen view 3/16. Naming clusters

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	manufacturer	String	45	0		None	None	8	Left	Nominal	Input
2	type	String	78	0		None	None	11	Left	Nominal	Input
3	code of origin	String	24	0	place of origin	None	None	7	Left	Nominal	Input
4	displacement (	Numeric	11	0	displacement (cm3)	None	None	4	Right	Scale	Input
5	perf. (kW)	Numeric	11	0	performance (kW)	None	None	4	Right	Scale	Input
6	perf. (LE)	Numeric	11	0	performance (LE)	None	None	5	Right	Scale	Input
7	moment	Numeric	11	0	moment (Nm)	None	None	4	Right	Scale	Input
8	weight	Numeric	11	0	weight (Kg)	None	None	4	Right	Scale	Input
9	consumption	Numeric	11	1	consumption (l/100km)	None	None	3	Right	Scale	Input
10	acceleration	Numeric	11	1	acceleration (0-100 km/h (s	None	None	4	Right	Scale	Input
11	terminal veloci	Numeric	15	0	terminal velocity (km/h)	None	None	5	Right	Scale	Input
12	price	Numeric	15	0	price (Ft)	None	None	7	Right	Scale	Input
13	FAC1_1	Numeric	11	5	motor performance	None	None	13	Right	Scale	Input
14	FAC2_1	Numeric	11	5	motor features	None	None	13	Right	Scale	Input
15	FAC3_1	Numeric	11	5	consumption	None	None	13	Right	Scale	Input
16	QCL_1	Numeric	8	0	clusters	{1, street mot	None	10	Right	Nominal	Input
17	QCL_2	Numeric	20	5	Distance of Case from its...	None	None	22	Right	Scale	Input
18											

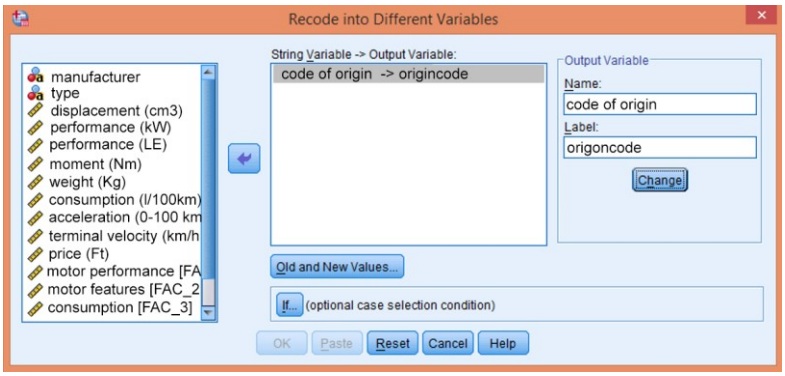
Screen view 3/17. Labeling clusters

3.3. Correspondence analysis

The method has shortly been described in the section about association so this can be considered as a new practice exercise. The question is if there is a connection between the country of production and the clusters, i.e. to what extent the producer's nationality influences the results of the former cluster analysis. To put it another way, the question is if the motorcycle producers have segmentated the market.

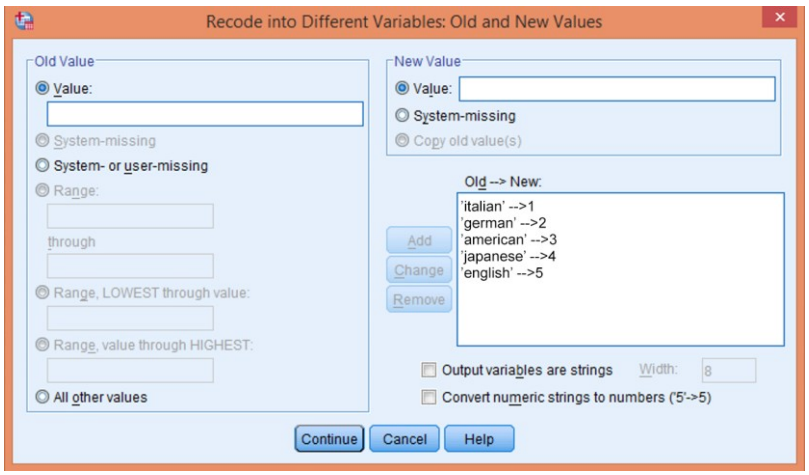
To give an anwer, the generated clusters - we have already named - have to be used. First, the countries in the database have to be coded into a different variable (code of origin). Codes of origins have been generated by the following method:

Countries of origin can be coded into a different variable in ANALYZE / INTO DIFFERENT VARIABLE



Screen view 3/18. Selecting coding

Let us select the variable - based on which the new one will be generated - with the help of the arrow. In the OUTPUT VARIABLE box, let us name the new variable (country of origin, "szarmazaskod") keeping in mind that no accents are permitted. In the LABEL box the name of the label can be given (accents are allowed here). Actual coding can be done with the help of the option Old and New Values.



Screen view 3/19. Coding origins

The name of the original variable has to be written in the OLD VALUE box (spelling has to be taken care of) and the new code in the NEW VALUE box. Press ADD and continue the process until all variables get a new code. New variable are created by pressing CONTINUE and OK. After that, let us check if there is a relation existing between the country of origin and the generated clusters. This question can be answered with the help of the table Symmetric Measures of the crosstab analysis.

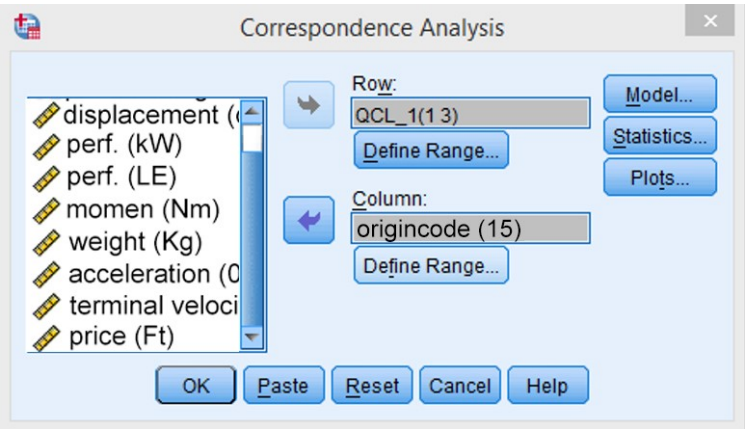
Table 3/18. Symmetric Measures

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,803	,012
	Cramer's V	,568	,012
N of Valid Cases		20	

One can see that the correlation does exist and its strength is moderate (0.57). Now, we can carry out the correspondence analysis.

The analysis is available in DIMENSION REDUCTION / CORRESPONDENCE ANALYSIS where first we have to select the cluster code as the row variable, and the new code of origin (szarmazaskod) as the column variable.



Screen view 3/20. Correspondence Analysis Settings

Both variables have to be defined afterwards, with the help of the number of versions included. The row variable (code of clusters) will be defined from 1 to 3 while the column variable (country of origin) ranges from 1 to 5 as shown before. Let us run the analyses, leaving all other settings stay the same. Examining the graphic illustration of the gained results, the values belonging together can clearly be seen.

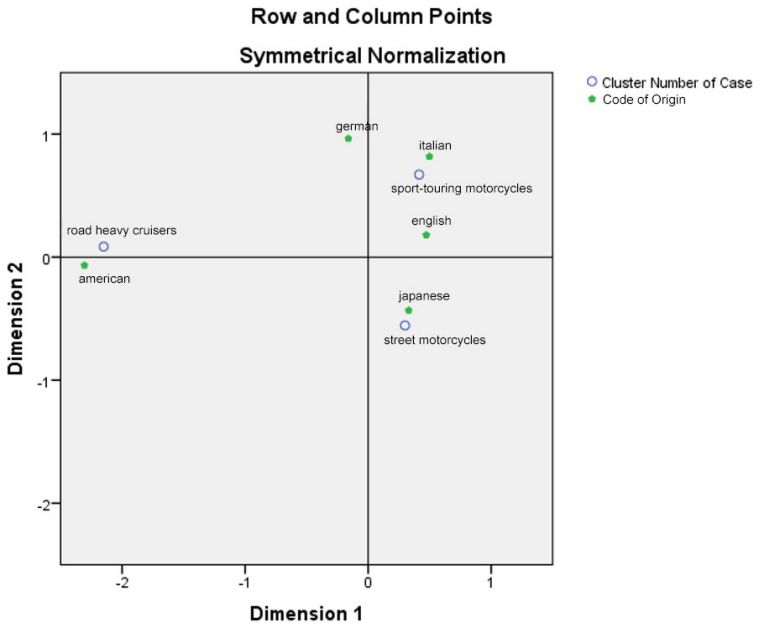
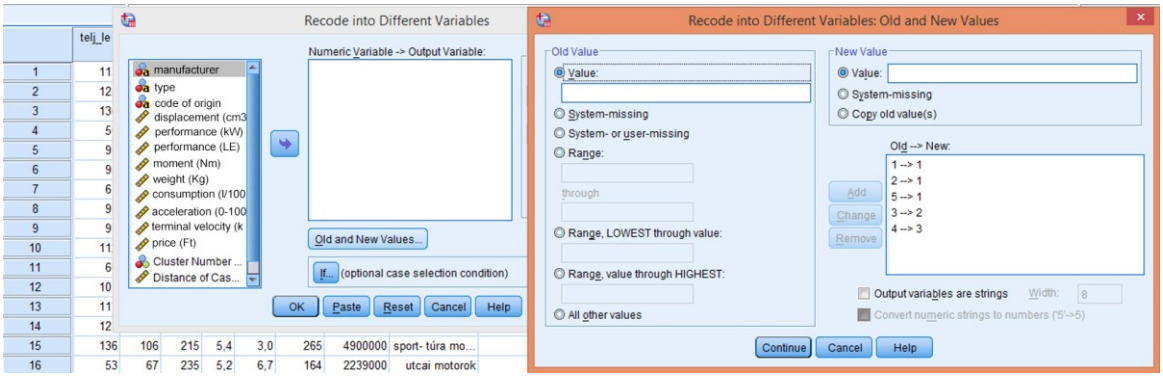


Figure 3/3. Graphic illustration of the Correspondence Analysis

Independent of the name of the dimensions, it is clear that the "sport-touring" motorcycles mostly origin from Europe, while the "street" motorcycles come from Japan and the "cruisers" are American. This makes the demand arise to illustrate the phenomenon based on the continents. First, let us again generate a new variable with the continent codes (name: származásföldrész).



Screen view 3/21. Coding based on continents

Regarding the coding, the numbers 1, 2, and 5 get the new code 1 (Europe), the original 3 becomes #2 (American), and let 4 become 3 (Asian). Let us label this variable as written in the brackets. The graphic illustration of the correspondence analysis will be the following:

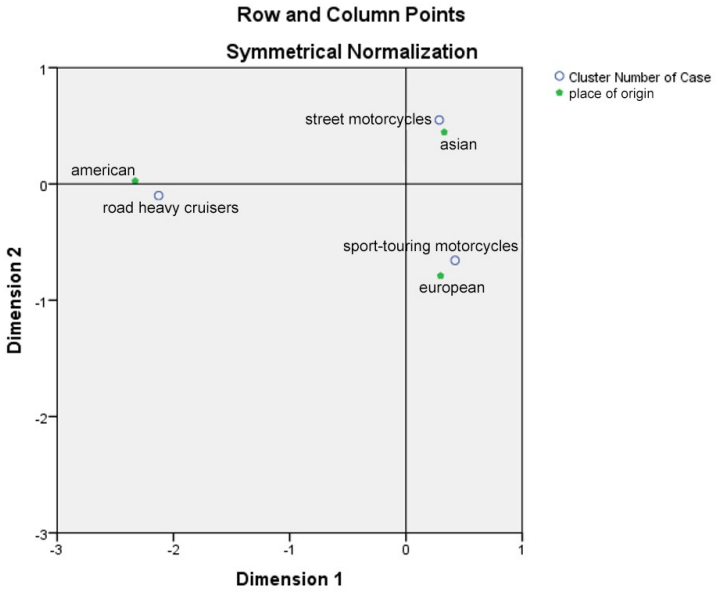


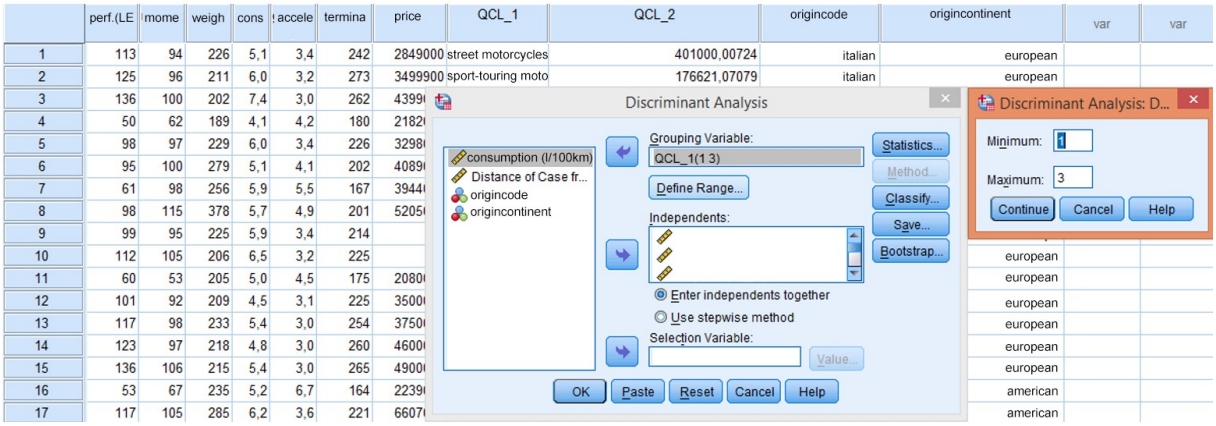
Figure 3/4. Graphic illustration of continents and clusters

Based on the database's data, continents and clusters fit well. The figure shows that as according to lifestyle, motorcycle production has been segmentated on the basis of continents – let us just think about the American habits. One may know the American cars well, and now it has turned out that American people have the same style and habits regarding motorcycles as well.

3.4. Discriminant analysis

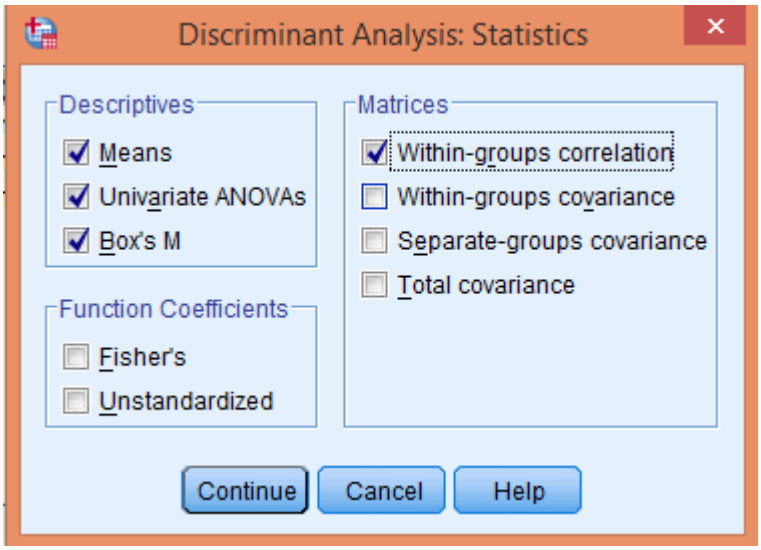
Discriminant analysis is a multivariate data analysis method that is mostly applied in order to separate groups and forecast category. It attempts to explain the values of dependent variables by the ones of independent variables so it is seeking for the answer if the group membership can be estimated in advance and if yes, to what extent (%) it can be estimated with the independent variables. The aim here is not only to explore the connection between the variables but also to forecast the unknown values of the dependent variables based on the values of the independent ones. The method is similar to the variance analysis and the multivariate regression. The similarity to the latter one comes from the line fitting problem. We can test if the discriminant analysis is good by comparing the adjusted group to the real group. Logistic regression is also similar to discriminant analysis the application of which does not require such strict terms to be filled. In the case of discriminant analysis, the dependent variable will be measured on a nominal scale, the independent one will be measured on interval or ratio scale, while in case of logistic regression the independent variable may include nominal and ordinal scale variables.

Let us continue our example by examining if it can be estimated to which cluster (street, sport-touring, cruiser) the motorcycle belongs to, knowing the motorcycles' parameters (engine displacement, performance (kW), performance (LE), moment, weight, consumption, acceleration, terminal velocity, price). Access path to the analysis: ANALYZE / CLASSIFY / DISCRIMINANT. First we add the new clusters as the grouping (dependent) variable, then we define them (Define Range) according to the number of clusters generated. Add one as minimum and three as maximum value. Independent variables have to be moved to the Independents box with the help of the arrow. (Source: motor.sav)



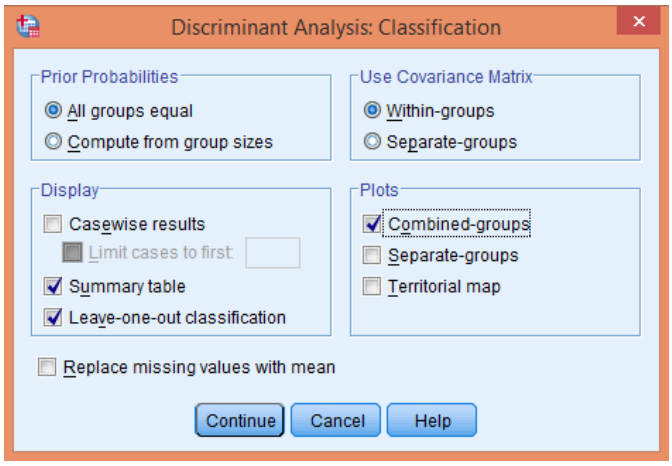
Screen view 3/22. Discriminant analysis settings

Now, let us select all options in STATISTICS/DECREPTIVES since we can test the prerequisites of the analysis.



Screen view 3/23. Prerequisite settings

From the MATRICES options let us select correlation within groups (Within-groups correlation) jelöljük. Finally, the following options have to be selected in the menu CLASSIFY:



Screen view 3/24. Analysis classification settings

Besides the default setting, let us request a Summary table from the Display options which includes information on the cases ordered into their valid groups just like Leave-one-out classification. For the graphic illustration, the option Combined-groups may be requested which displays the situation of groups depending in the gained discriminance functions. Running the analysis results in numerous tables from which we explain the most important ones in details.

The first table (Analysis Case Processing Summary) contains the simple basic statistics like the number of valid (50) and missing (3) cases. The next table (Group Statistics) shows the mean, standard deviation and weight of all variables included in the analysis, both according to their clusters and their total value.

Table 3/19. Basic Statistics

Group Statistics

Cluster Number of Case		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
street motorcycles	displacement (cm3)	930,958	282,2981	24	24,000
	performance (kW)	69,625	21,5786	24	24,000
	performance (LE)	94,750	29,3054	24	24,000
	moment (Nm)	85,833	21,9934	24	24,000
	weight (Kg)	235,958	35,1388	24	24,000
	acceleration (0-100	3,917	1,1309	24	24,000
	terminal velocity (k	216,750	38,1567	24	24,000
	price (Ft)	2448000,000	351868,6761	24	24,000
sport-touring motorcycles	displacement (cm3)	1070,789	160,4215	19	19,000
	performance (kW)	94,474	23,1789	19	19,000
	performance (LE)	128,421	31,4984	19	19,000
	moment (Nm)	107,474	14,3231	19	19,000
	weight (Kg)	234,053	31,9435	19	19,000
	acceleration (0-100	3,274	,7377	19	19,000
	terminal velocity (k	252,158	36,2495	19	19,000
	price (Ft)	3676521,053	499190,3473	19	19,000
road heavy cruisers	displacement (cm3)	1418,429	230,6483	7	7,000
	performance (kW)	62,286	18,8212	7	7,000
	performance (LE)	84,857	25,3734	7	7,000
	moment (Nm)	117,000	22,3159	7	7,000
	weight (Kg)	345,286	43,6452	7	7,000
	acceleration (0-100	5,257	1,1088	7	7,000
	terminal velocity (k	181,286	25,6886	7	7,000
	price (Ft)	6203571,429	705079,8671	7	7,000
Total	displacement (cm3)	1052,340	282,6103	50	50,000
	performance (kW)	78,040	25,1826	50	50,000
	performance (LE)	106,160	34,1637	50	50,000
	moment (Nm)	98,420	22,8492	50	50,000
	weight (Kg)	250,540	51,7649	50	50,000
	acceleration (0-100	3,860	1,1681	50	50,000
	terminal velocity (k	225,240	42,8855	50	50,000
	price (Ft)	3440618,000	1343598,867	50	50,000

In the next table, we can examine to what extent the independent variables contribute to the establishing function. Besides the F-value, the statistics Wilks'- Lambda is also available to test the significance of the variables.

Table 3/20. Results of the F-test

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
displacement (cm3)	,668	11,665	2	47	,000
performance (kW)	,724	8,947	2	47	,001
performance (LE)	,725	8,905	2	47	,001
moment (Nm)	,696	10,263	2	47	,000
weight (Kg)	,443	29,521	2	47	,000
acceleration (0-100 km/h (	,697	10,226	2	47	,000
terminal velocity (km/h)	,678	11,162	2	47	,000
price (Ft)	,117	178,009	2	47	,000

It can be seen that all variables have significant effects. The value of Wilks' Lambda is always between 0 and 1 from which always the variables with a value closer to zero have more significant influence on the discriminant function.

Table 3/21. Multicollinearity test

Pooled Within-Groups Matrices

		displacement (cm3)	perf. (kW)	perf. (LE)	moment (Nm)	weight (Kg)	acceleration (0-100 km/h (s))	terminal velocity (km/h)	price (Ft)
Correlation	displacement (cm3)	1,000	-,058	-,058	,841	,792	,289	-,280	,239
	performance (kW)	-,058	1,000	1,000	,426	-,213	-,822	,933	,049
	performance (LE)	-,058	1,000	1,000	,426	-,214	-,821	,933	,048
	moment (Nm)	,841	,426	,426	1,000	,637	-,145	,173	,252
	weight (Kg)	,792	-,213	-,214	,637	1,000	,432	-,408	,175
	acceleration (0-100	,289	-,822	-,821	-,145	,432	1,000	-,856	,018
	terminal velocity (k	-,280	,933	,933	,173	-,408	-,856	1,000	-,015
	price (Ft)	,239	,049	,048	,252	,175	,018	-,015	1,000

Two assumptions will be tested in the next two tables. The Pooled Within-Groups Matrices table is testing multicollinearity. The next table is testing the homogeneity (homoskedasticity) of the variance-covariance matrices by the Box's M test.

The next important table (Eigenvalues) gives the first information on the establishing function.

Table 3/22. Summary of function values

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	8,603[a]	89,5	89,5	,946
2	1,005[a]	10,5	100,0	,708

a. First 2 canonical discriminant functions were used in the analysis.

As shown by the table, two functions have been generated. The number of functions can be calculated by substracting one from the lower amount from the number of clusters and the number of independent variables. The eigenvalues help the researcher decide on the

importance of the two functions. Based on the eigenvalues of the table and the values of explained variance, the first function is more important for us. Canonical correlation (0.946) means that the function explains a significant part from the total variance. The square of the gained value shows how many percent of the dependent variable is explained by the group of independent variables (89.49%).

Table 3/23. Test of Function(s)

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,052	130,133	14	,000
2	,499	30,604	6	,000

The appearing Wilks' Lambda table includes the significance test of the functions. As shown by the table, both functions are significant but the effect of the first one is more relevant.

With the help of the standardized coefficients in the next table (Standardized Canonical Discriminant Function Coefficients) it can be determined which variables separate the groups the most.

The matrix of the correlation coefficient (Structure Matrix) has to be interpreted similarly as the Component Matrix of the factor analysis since it includes the Pearson linear correlation of the independent variables and the discriminant functions, Pooled within-groups.

Table 3/24 Stucture Matrix

Structure Matrix

	Function	
	1	2
price (Ft)	,932*	,307
displacement (cm3)	,240*	,038
terminal velocity (km/h)	-,106	,613*
performance (LE)	-,032	,610*
performance (kW)	-,031	,609*
acceleration (0-100 km/h (	,150	-,491*
weight (Kg)	,355	-,415*
moment (Nm)	,190	,355*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

As shown by the matrix, the first function contains the price and the engine displacement, while the second one includes all the remaining ones, except the performance in

horsepower (LE). Based on this, the researcher can give a name to the dimensions (similarly to the dactor analysis).

The next table (Functions at Group Cetroids) contains the centroids of the groups.

Table 3/25. Functions at Group Centroids

Functions at Group Centroids

Cluster Number of Case	Function	
	1	2
street motorcycles	-2,030	-,736
sport-touring motorcycles	,132	1,241
road heavy cruisers	6,602	-,843

Unstandardized canonical discriminant functions evaluated at group means

We can state that the first and the third group have high values in the first dimension while the sport-touring motorcycles show high values in the second dimension. The program uses these coordinates in further graphic illustration.

The next section is about classification statistics which is the most important part of our analysis. The first table (Prior Probabilities for Groups) lists the original values.

Table 3/26. Classification results

Prior Probabilities for Groups

Cluster Number of Case	Prior	Cases Used in Analysis	
		Unweighted	Weighted
street motorcycles	,333	24	24,000
sport-touring motorcycles	,333	19	19,000
road heavy cruisers	,333	7	7,000
Total	1,000	50	50,000

As shown by the table, chances of becoming member of a group were 33.3 percent. Next, the graphic illustration follows where the axes are the functions (dimensions) themselves.

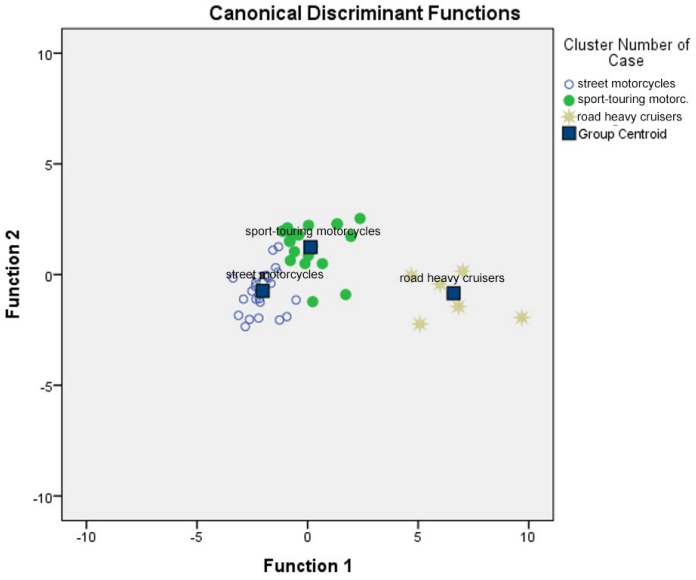


Figure 3/5. Graphic illustration of the discriminant analysis

The figure plots the values of the cases included in the analysis as well as the centroids. The proportion of correctly categorized group memberships are shown by the table called Classification Results.

Table 3/27. Classification Results

Classification Results[c]

		Cluster Number of Ca	Predicted Group Membership: street motorcycles	sport-touring motorcycles	road heavy cruisers	Total
Original	Count	street motorcycles	22	2	0	24
		sport-touring motorcycles	1	18	0	19
		road heavy cruisers	0	0	7	7
	%	street motorcycles	91,7	8,3	,0	100,0
		sport-touring motorcycles	5,3	94,7	,0	100,0
		road heavy cruisers	,0	,0	100,0	100,0
Cross-validated[a]	Count	street motorcycles	21	3	0	24
		sport-touring motorcycles	1	18	0	19
		road heavy cruisers	0	0	7	7
	%	street motorcycles	87,5	12,5	,0	100,0
		sport-touring motorcycles	5,3	94,7	,0	100,0
		road heavy cruisers	,0	,0	100,0	100,0

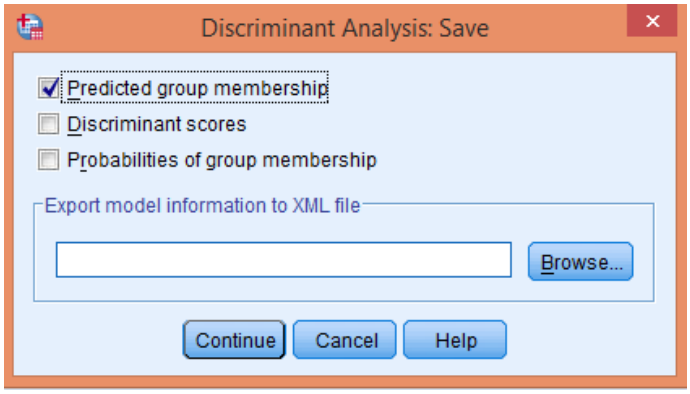
a. Cross validation is done only for those cases in the analysis. In cross validation, each case is the functions derived from all cases other than that case.

b. 94,0% of original grouped cases correctly classified.

c. 92,0% of cross-validated grouped cases correctly classified.

As shown at the bottom of the table, the model was able to categorize by the given independent variable to an extent of 94%. The original group membership and the categorization of discrimination function (Cross-validated) has been compared hete. This means (looking at the diagonal values) that 21 from the 24 street motorcycles have been categorized correctly, and 3 incorrectly which makes 87.5%. 18 cases from the 19 sport-touring motorcycles have been categorized correctly which makes 94.7% while all cruisers

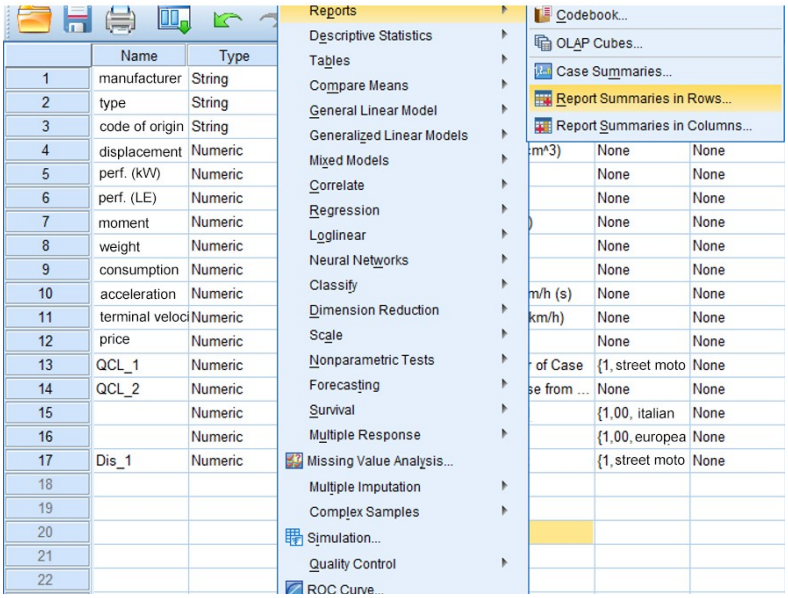
were ordered into their valid group (100%). The three groups altogether have reached a 94% match. The 92% in the third statement under the table refers to the fact that the Leave-One-Out option has been selected in the CLASSIFY menu, which does test the same crossvalidity, too. This percentage used to be smaller than the one above it since its measure is stricter. Its method is to repeat the analysis, always leaving out one case. Let us save (SAVE) the number of clusters estimated by the function.



Screen view 3/25. Saving the number of predicted groups

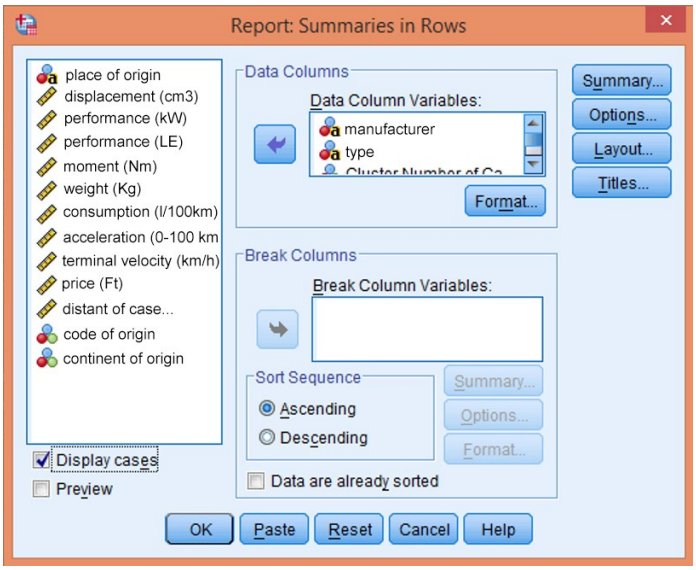
As a result, we get a new variable (Dis_1) in the Data Editor window that we label as "the number of estimated groups".

Now, let us list the original and estimated group memberships. This can be done several ways in ANALYZE / REPORTS. Let us request descriptive statistics in rows (Report Summaries in Rows).



Screen view 3/26. Displaying original and predicted groups

In the next settings we use the arrow to add variables to be displayed in the coulums, i.e. we request a list on the established variables of producer, type, number of clusters, and the number of estimated groups.



Screen view 3/27. Settings of listed variables

Without altering any other options, press OK to get the following results in the Output window:

Table 3/28. Table of summaries

manufacturer	type	number of clusters generated	estimated number of groups
Aprilia	RST 1000 Futura	1	1
Aprilia	RSV mille R	2	2
Benelli	Tornado 900 i.e.	2	2
BMW	F 650 GS	1	1
BMW	R 1100 S	2	2
BMW	R 1150 RT	2	2
BMW	R 1200 C Independent	2	2
BMW	K 1200 LT	3	3
Cagiva	Navigator 1000	.	.
Cagiva	Raptor 1000	.	.
Ducati	Monster 620 Dark	1	1
Ducati	Monster S4	2	2
Ducati	ST 4 S	2	2
Ducati	998	2	2
Ducati	999	2	2
Harley-Davidson	XL 883R Sportster	1	1
Harley-Davidson	VRSCA V-Rod	3	3
Harley-Davidson	Night Train	3	3
Harley-Davidson	Fat Boy	3	3
Harley-Davidson	Road King FLHR	3	3
Harley-Davidson	Electra-Gide Ultra Classic	3	3
Honda	Hornet 600	1	1
Honda	CBR 600 F	1	1
Honda	XRV 750 Africa Twin	1	1
Honda	VFR/CBS-ABS	1	1
Honda	Fireblade	2	2
Honda	VTR 1000 SP-2	2	2
Honda	X-11 CBS	1	1
Honda	CBR 1100 XX	1	2
Honda	GL 1800 Gold Wing	3	3

The part of the results shows that the program gives a simple list based on the variables. A much better list can be created in REPORTS /CASE SUMMARIES since statistics within categories defined by one or more grouping variables can be requested.

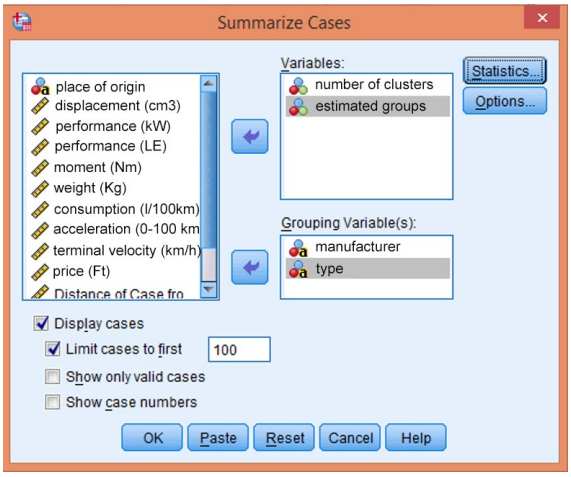


Figure 3/28. Summary table settings

The Variables box should contain the number of established clusters and the number of estimated groups while the Grouping Variable(s) box should include the variables producer and type. The next table shows a part of the table generated like this.

Table 3/29. Case Summaries

Case Summaries[a]

					number of generated clusters	Estimated number of groups
gyártó	Aprilia	type	RST 1000 Futura	1	street motorcycles	street motorcycles
			RSV mille R	1	sport-touring motorcycles	sport-touring motorcycles
	Benelli	type	Tornado 900 i.e.	1	sport-touring motorcycles	sport-touring motorcycles
	BMW	type	F 650 GS	1	street motorcycles	street motorcycles
			R 1100 S	1	sport-touring motorcycles	sport-touring motorcycles
			R 1150 RT	1	sport-touring motorcycles	sport-touring motorcycles
			R 1200 C Independent	1	sport-touring motorcycles	sport-touring motorcycles
			K 1200 LT	1	road heavy cruisers	road heavy cruisers
	Cagiva	type	Navigator 1000	(missing) 9		
			Raptor 1000	(missing) 10		
	Ducati	type	Monster 620 Dark	1	street motorcycles	street motorcycles
			Monster S4	1	sport-touring motorcycles	sport-touring motorcycles
			ST 4 S	1	sport-touring motorcycles	sport-touring motorcycles
			998	1	sport-touring motorcycles	sport-touring motorcycles
			999	1	sport-touring motorcycles	sport-touring motorcycles

From the table it can be easily seen which types of motorcycles have been ordered by the discriminant analysis into a group different than the original one.

4. PUBLICATION AND PRESENTATION OF RESULTS AND RESEARCH REPORTS

Research results are usually published in research reports. The aim of these reports is to interpret results, conclusions and suggestions towards the client or a defined target audience. Readers may vary, thus there are several types of research report, depending on who, where and when will read it. Depending on these factors,the aim of the report can be different:

- account towards the supporters of the research
- information for sport experts (e.g. in professional journal, in the form of an article)
- thesis, final exam or dissertation.

The research report must explain in detail:

1. the research topic and the relevant hypotheses
2. the research method (sample, tools, process of the research)
3. data analysis
4. significance of the research and new findings (expected results, possible application)

Inappropriate length is the most common mistake of research reports. Setting the appropriate length is difficult indeed, as in case the report is too short, the reader may think that the task was not carried out properly. However, if the report is too long, then the reader may simply not read it through at all. "This report, by its very length, defends itself against the risk of being read."[43]

Writers of the report should aim to compile accurate and grammatically correct and understandable sentences, also to ensure authenticity. They should keep in mind that there might not be a single theory that will explain all possible end results or behaviour. Unfortunately, when writing their research report, many researchers put their trust into the correctness of a single theory.

The research report is usually sent to the committee which approved the research; and to those persons who supported it or expressed interest in the execution or the results; or to those to whom it is obligatorily sent (as in case of the university thesis).

[43] Winston Churchill

The first page should contain the title of the work, the researchers' name, address and contact details (the group should agree on the order of authors before executing the research).

Sometimes authors write a foreword which is usally placed around page 3-5. This is where those people are listed who helped in the research but are not mentioned as authors. Researchers express their gratitude here to participants, the technical team and the leaders of the programs, faculties and institutions where they belong. This is also the place where authors explain their ideas according to the given topic and tell why these issues are important. Motivations and reasons for topic choice and writing are described here in a more personal style.

On the next page we can find the description of the actual research, divided into the classic sections: introduction, discussion, conclusion (suggestions). Interpretation of the statements and inventions of the research shall always end with by integrating the explanation and proof or rejection of the hypotheses. Thus this section is closed by an integrated, holistic approach that includes all of the statements and also designates future directions. Right after the conclusion and the suggestions we can find the resource list which inludes all literature that was used during the report. The appendix comes next, including the original questionnaire, photos of the documents and equipments, and all other material that may be important for any reader or researcher interested in further details.

Research reports are usally shortened and edited to be published in journals or books. The most important guideline here is to use more compact wording but keep the main message. If the original report included a lot of analysis, tables, figures, interviews, quotes or observations, then more than one publications can be prepared based on the same material, marked by numbers (I, II, III), to indicate that they are parts of a research report series. The shortened version has the same sections as a longer report, but must only contain the main theoretical explanations, the hypotheses, the methods, the results, the conclusions and the references.

The short summary of the research is called the **abstract**. This contains usally 250-500 words and should be prepared after the report. Its sections are the following: aim of the study, short methodology (research plan), main results, conclusion, and implication. It is useful to mention a few reliable studies that we used as resources. It is highly important to

state whether our results proved our persumptions, and if not, we shall provide some suggestions for future steps.

Research methods are often accompanied by an **oral presentation** that consists of the following steps:

1. Definition of the aim
2. Exploring the audience
3. Definition of the structure
4. Preparing supporting materials
5. Preparing visual tools
6. Testig the presentation
7. Preparing the location of the presentation
8. Preparing the voice
9. Presenting
10. Answering questions
11. Wrapping up

Source: Sajtos L.-Mitev A (2007), p. 381.

Previously a finished research was followed by a simple presentation. First it meant reading from some notes, then with the passage of time technical innovations like boards and overhead projectors became available, with videos a bit later. Nowadays, in our computerised times, the most popular tools are the laptop and the projector. Computerised presentations are prepared by user-friendly softwares such as PowerPoint, which have a few advantages compared to the traditional oral presenting:

- children learn how to use it early on (in primary and secondary schools)
- the completed presentation can be checked and modified easily
- appearance of the texts can be delayed
- its quality is a lot better than the overhead projector's
- presenting diagrams, figure, pictures, videos and animations is easy
- etc.

Oral presentations ususally have time limits, which is about 10-45 minutes on average, depending on the type of the presentation (conference, thesis defence, project closing

presentation, etc), and on the position of the presenter, both of which shall be taken into consideration.

General guidelines on the format of the presentation:

- the same background, font, format and animation should be used on all slides.
- font sized should be customised according to the location, but it should be kept larger than 36 pt so that everybody in the room can read it.
- multimedia elements can be included (animation, graphs, diagrams, tables, movies, sound).
- include lists with one-word items.
- average number of words for a slide is 15-25.
- rule 6x6 (maximum 6 words in maximum 6 lines).

General guidelines for computerised presentations:

- The first slide shall include the title, authors and their contact details.
- Summarise the main theoretical points and hypotheses on slides 1-3. Each slide should contain only a minimum number of words as too much information will destroy the presentation.
- Explain the research plan (graphically, if possible) on slides 4-5. E.g. sampling of the participants, applied tools, execution of the research, etc.
- Explain results on slides 6-8. One slide should contain only one figure or table (preferably figures instead of tables). Figures are easy to understand and are more informative.
- Conlusion, results and why they are new, and suggestions should be explained on slides 9-10.
- Slide 11 should list the resources, contacts and other details (optional).
- Slide 12 shall thank the audience for the attention.

Reading from paper should be avoided during presenting as it impedes the audience from paying attention and questions the competence of the presenter. Graphic visauls should be preferred during oral presentations, but tables and figures with small fonts should be avoided. Apart from the general requirements of format, tables and graphs should include the size of the sample (N). We must pay extra attention to font size, as the presentation can become nonrelevant if fonts are too small.

We shall pick the terms and sentences that define our presentation. We should switch between tones to avoid a monotonous presentation. We must be sure that our research and our results are correct – we must show our interest. ***Rehearse the presentation!***

The oral presentation is one of the most important parts of the research process, thus it needs thorough planning and preparation. In a good presentation the preparedness of the presenter is mirrored by his convincing and clear explanations. A well-prepared presentation should be ready for various persons and interests within the audience. The starting should be emphasised and clear, followed by the presentation of the facts. The presenter should keep up the audience's attention throughout the whole speech, as the listener should take away the summary and the new findings.

In summary we would like to provide a few general guidelines to help prepare and give presentations:

Preparing the slides:

1. Indicate page-numbers on the slides (except the first one), next to the total number of slides – this way the audience will know, how much more is left of the presentation and if there are questions at the end, the listener may refer to the problematic slide more easily.
2. Avoid abbreviations.
3. The slides contain the outline of the presentation, so avoid long and complex sentences. Do not read up exactly what is on the slide.
4. Avoid too much animation as it may be disturbing for the audience and you may also pass the time limit.
5. The presentation shall unambiguously state what the task was. It can be explained on a slide entitled "Tasks" or "What was the problem?"
6. The presentation should unambiguously state the new results. They may be compared to results from other researchers.
7. At least one slide should explain the conclusion or the summary.
8. As usually the last slide is projected for the longest time, we can consider this a key slide. It is a good idea to use this to project an informative summary or suggestions for future research, and not to simply say "Thank you for your attention".

9. The aim of the audience is to understand the problem, the solution and the results, so there is no need to explain every small detail. Also, we shall use the time wisely.
10. The completed slides should be discussed with an expert before presentation. If this is a preparation for the defence of the thesis, then the supervisor must see it before presentation.
11. There might be a few expected questions. Answers for these can be put on slides after the last slide of the core presentation – if someone asks these questions at the end (for example during a thesis defence), then the prepared slide can be shown immediately.

Presentation:

1. Greet the audience at the beginning. Say 'Thank you for your attention' at the end (orally as well) – this will also indicate that the presentation has come to its end.
2. Rehearse the presentation to see whether it fits into the time limit. Further rehearsing can help cutting off another 1-3 minutes. Keeping the time limit is important because in some cases the overrun can be penalised, or the presentation is not allowed to be finished.
3. Don't turn your back to anybody. Instead, turn towards the audience or members of the board, if there is one. Don't keep your hands in your pockets.
4. Wear clothes that fit the occasion.
5. Arrive a lot earlier to have enough time to upload your presentation.
6. Even the best slides will remain dull if the presenter is tired. Get a good night's sleep the night before.

5. APPENDIX (TABLES)

5.1. Standard normal distribution

Density function values

z	0	1	2	3	4	5	6	7	8	9
0,0	0,500	0,496	0,492	0,488	0,484	0,480	0,476	0,472	0,468	0,464
0,1	0,460	0,456	0,452	0,448	0,444	0,440	0,436	0,433	0,429	0,425
0,2	0,421	0,417	0,413	0,409	0,405	0,401	0,397	0,394	0,390	0,386
0,3	0,382	0,378	0,374	0,371	0,367	0,363	0,359	0,356	0,352	0,348
0,4	0,345	0,341	0,337	0,334	0,330	0,326	0,323	0,319	0,316	0,312
0,5	0,309	0,305	0,302	0,298	0,295	0,291	0,288	0,284	0,281	0,278
0,6	0,274	0,271	0,268	0,264	0,261	0,258	0,255	0,251	0,248	0,245
0,7	0,242	0,239	0,236	0,233	0,230	0,227	0,224	0,221	0,218	0,215
0,8	0,212	0,209	0,206	0,203	0,200	0,198	0,195	0,192	0,189	0,187
0,9	0,184	0,181	0,179	0,176	0,174	0,171	0,169	0,166	0,164	0,161
1,0	0,159	0,156	0,154	0,152	0,149	0,147	0,145	0,142	0,140	0,138
1,1	0,136	0,133	0,131	0,129	0,127	0,125	0,123	0,121	0,119	0,117
1,2	0,115	0,113	0,111	0,109	0,107	0,106	0,104	0,102	0,100	0,099
1,3	0,097	0,095	0,093	0,092	0,090	0,089	0,087	0,085	0,084	0,082
1,4	0,081	0,079	0,078	0,076	0,075	0,074	0,072	0,071	0,069	0,068
1,5	0,067	0,066	0,064	0,063	0,062	0,061	0,059	0,058	0,057	0,056
1,6	0,055	0,054	0,053	0,052	0,051	0,049	0,048	0,047	0,046	0,046
1,7	0,045	0,044	0,043	0,042	0,041	0,040	0,039	0,038	0,038	0,037
1,8	0,036	0,035	0,034	0,034	0,033	0,032	0,031	0,031	0,030	0,029
1,9	0,029	0,028	0,027	0,027	0,026	0,026	0,025	0,024	0,024	0,023
2,0	0,023	0,022	0,022	0,021	0,021	0,020	0,020	0,019	0,019	0,018
2,1	0,018	0,017	0,017	0,017	0,016	0,016	0,015	0,015	0,015	0,014
2,2	0,014	0,014	0,013	0,013	0,013	0,012	0,012	0,012	0,011	0,011
2,3	0,011	0,010	0,010	0,010	0,010	0,009	0,009	0,009	0,009	0,008
2,4	0,008	0,008	0,008	0,008	0,007	0,007	0,007	0,007	0,007	0,006
2,5	0,006	0,006	0,006	0,006	0,006	0,005	0,005	0,005	0,005	0,005
2,6	0,005	0,005	0,004	0,004	0,004	0,004	0,004	0,004	0,004	0,004
2,7	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003
2,8	0,003	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002
2,9	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,001	0,001	0,001
3,0	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001

Critical values for different significance levels

	Significance level (α)					
One-tailed	0.1000	0.0500	0.0250	0.0225	0.0100	0.0050
Two-tailed	0.2000	0.1000	0.0500	0.0450	0.0200	0.0100
z	1.280	1.645	1.960	2.000	2.330	2.587

5.2. Student's t-distribution

Critical values of Student's t-distribution for different significance levels

Degrees of freedom	**Significance level**				
	0.1	0,05	0.1	0,01	0.1
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

5.3. χ^2-distribution

Critical values of ${}_n\chi^2$-distribution for different significance levels

Degress 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

5.4. F-distribution

Critical values of F-distribution for one-tailed 5% significance level (two-tailed 10%)

Denominator'	Numerator's degree of freedom															
	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	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,63	8,62	8,58	8,55
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,77	5,75	5,70	5,66
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,52	4,50	4,44	4,41
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,83	3,81	3,75	3,71
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,40	3,38	3,32	3,27
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,11	3,08	3,02	2,97
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,89	2,86	2,80	2,76
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,73	2,70	2,64	2,59
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72	2,65	2,60	2,57	2,51	2,46
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,50	2,47	2,40	2,35
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,41	2,38	2,31	2,26
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,34	2,31	2,24	2,19
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,28	2,25	2,18	2,12
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,23	2,19	2,12	2,07
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,18	2,15	2,08	2,02
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,14	2,11	2,04	1,98
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,11	2,07	2,00	1,94
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,07	2,04	1,97	1,91
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,18	2,10	2,05	2,01	1,94	1,88
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,15	2,07	2,02	1,98	1,91	1,85
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,13	2,05	2,00	1,96	1,88	1,82
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,11	2,03	1,97	1,94	1,86	1,80
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,09	2,01	1,96	1,92	1,84	1,78
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,07	1,99	1,94	1,90	1,82	1,76
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,06	1,97	1,92	1,88	1,81	1,74
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,04	1,96	1,91	1,87	1,79	1,73
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,03	1,94	1,89	1,85	1,77	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,88	1,84	1,76	1,70
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11	1,96	1,88	1,82	1,79	1,70	1,63
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,78	1,74	1,66	1,59
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,73	1,69	1,60	1,52
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,69	1,65	1,56	1,48
75	3,97	3,12	2,73	2,49	2,34	2,22	2,13	2,06	2,01	1,96	1,80	1,71	1,65	1,61	1,52	1,44
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,62	1,57	1,48	1,39
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,72	1,62	1,56	1,52	1,41	1,32

5.5 Critical values of F-distribution for one-tailed 2.5% significance level (two-tailed 5%)

Denominator'	Numerator's degree of freedom															
	1	2	3	4	5	6	7	8	9	10	15	20	25	30	50	100
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,3	14,2	14,1	14,1	14,0	14,0
4	12,2	10,7	10,0	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,66	8,56	8,50	8,46	8,38	8,32
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,27	6,23	6,14	6,08
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,27	5,17	5,11	5,07	4,98	4,92
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,57	4,47	4,40	4,36	4,28	4,21
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,10	4,00	3,94	3,89	3,81	3,74
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,77	3,67	3,60	3,56	3,47	3,40
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42	3,35	3,31	3,22	3,15
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,33	3,23	3,16	3,12	3,03	2,96
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,18	3,07	3,01	2,96	2,87	2,80
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,05	2,95	2,88	2,84	2,74	2,67
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	2,95	2,84	2,78	2,73	2,64	2,56
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,86	2,76	2,69	2,64	2,55	2,47
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,79	2,68	2,61	2,57	2,47	2,40
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,72	2,62	2,55	2,50	2,41	2,33
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,67	2,56	2,49	2,44	2,35	2,27
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,62	2,51	2,44	2,39	2,30	2,22
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,57	2,46	2,40	2,35	2,25	2,17
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,53	2,42	2,36	2,31	2,21	2,13
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,50	2,39	2,32	2,27	2,17	2,09
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,47	2,36	2,29	2,24	2,14	2,06
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,44	2,33	2,26	2,21	2,11	2,02
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,41	2,30	2,23	2,18	2,08	2,00
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,39	2,28	2,21	2,16	2,05	1,97
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,36	2,25	2,18	2,13	2,03	1,94
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,34	2,23	2,16	2,11	2,01	1,92
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,32	2,21	2,14	2,09	1,99	1,90
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,31	2,20	2,12	2,07	1,97	1,88
35	5,48	4,11	3,52	3,18	2,96	2,80	2,68	2,58	2,50	2,44	2,23	2,12	2,05	2,00	1,89	1,80
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,18	2,07	1,99	1,94	1,83	1,74
50	5,34	3,97	3,39	3,05	2,83	2,67	2,55	2,46	2,38	2,32	2,11	1,99	1,92	1,87	1,75	1,66
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,06	1,94	1,87	1,82	1,70	1,60
75	5,23	3,88	3,30	2,96	2,74	2,58	2,46	2,37	2,29	2,22	2,01	1,90	1,82	1,76	1,65	1,54
10	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	1,97	1,85	1,77	1,71	1,59	1,48
20	5,10	3,76	3,18	2,85	2,63	2,47	2,35	2,26	2,18	2,11	1,90	1,78	1,70	1,64	1,51	1,39

6. SOURCES

1. 169/2000. [IX. 29.] és 154/2004. [X. 14.] Kormány Rendelet
2. Ács P. (2007): A területi egyenlőtlenségek feltérképezése során leggyakrabban alkalmazott mérőszámok bemutatása, a sporttehetségek területi elhelyezkedésének példáján, Egy életpálya három dimenziója- Tanulmánykötet Pintér József emlékére, Pécsi Tudományegyetem Közgazdaságtudományi Kar, Pécs, 10- 22. o.
3. Ács P. (2015): Gyakorlati adatelemzés. Pécsi Tudományegyetem Egészségtudományi Kar. Pécs
4. Bíróné. N. E. (2004): Sportpedagógia. Dialóg Campus Kiadó. Budapest- Pécs
5. Dr. Hepp F.- Dr. Nádori L. (1971): Bevezetés a tudományos kutatásba. Kézirat. Tankönyvkiadó. Budapest.
6. Dr. Jánosa A. (2005): Adatelemzés számítógéppel, Perfekt Kiadó. Budapest, 271. o.
7. Falus I. (szerk.) (2000): Bevezetés a Pedagógiai kutatás módszereibe. Pedagógus Könyvek. Budapest. Műszaki Könyvkiadó. 540. o.
8. Gyetvai Gy.- Kecskemétiné Petri A. (1997): Testkultúra elméleti- és kutatásmódszertani alapismeretek. Főiskolai jegyzet. Juhász Gyula Tanárképző Főiskola. Szombathely. 208. o.
9. Hajdu O. (1997): A szegénység mérőszámai. KSH. Könyvtár és Dokumentációs Szolgálat. Budapest
10. Hajdu O. (1987): Sokváltozós statisztikai módszerek gyakorlati alkalmazása. Prodinform Műszaki Tanácsadó Vállalat. Budapest
11. Harsányi L (1998): Jó úton a sporttudomány akadémiai elismerése. Sporttudomány. 1998.2. sz.
12. Harsányi L. (2007): Az irodalomjegyzék készítés, idézés, hivatkozás további szabályai. Kézirat. Pécs. 2007. január 25.
13. Horváth L.- Prisztóka Gy. (2005): A sportpedagógia és sportpszichológia alapkérdései (főiskolai tankönyv) Bessenyei György Könyvkiadó, Nyíregyháza 2005.
14. Hunyadi L. (2001): Statisztikai következtetéselmélet közgazdászoknak. KSH, Budapest
15. Hunyadi L. (2002): Grafikus ábrázolás a statisztikában, Statisztikai Szemle 2002/1. 22-53. old.
16. Istvánfi Cs. (2000): Gondolatok a sporttudományokról. Kalokagathia. 2000. 1-2 sz. 7-18. o.

17. Kaj M.- Csányi T.- Karsai I.- Marton O. (2014): Kézikönyv a Nemzeti Egységes Tanulói Fittségi Teszt (NETFIT) alkalmazásához. Testnevelés Módszertani Könyvek (Csányi T. főszerk.). Magyar Diáksport Szövetség. Budapest
18. Kecskeméty L- Izsó L. (2005): Bevezetés az SPSS programrendszerbe, ELTE-Eötvös Kiadó, Budapest, 460.o
19. Kehl D.- Rappai G. (2006): Mintaelem-szám tervezése Likert-skálát alkalmazó lekérdezésekben. Statisztikai Szemle 84. évfolyam 9. szám. 848- 876. o.
20. Kerlinger F. (1980): Analysis Of Covariance Structure Tests Of A Criterial Referents Theory Of Attitudes, Multivariate Behavioral Research, Volume 15, Issue 4 January 1980 , 403 – 422. o.
21. Mundruczkó Gy. (1981): Alkalmazott regressziószámítás, Akadémiai Kiadó, Budapest
22. Müller A. (2004): Mozgásvizsgálatok a mozgásegyenletesség és a teljesítménykonstancia példáján. Disszertáció Semmelweis Egyetem Doktori Iskola Nevelés- és Sporttudományok Doktori Iskolája (Sport és Társadalomtudomány).
23. Pintér J. – Rappai G. (2001): A mintavételi tervek készítésének néhány gyakorlati megfontolása. Marketing & Menedzsment 2001/4. 4-11. o.
24. Ramanathan R. (2003): Bevezetés az Ökonometriába alkalmazásokkal, Panem Kft. Budapest
25. Rappai G. (2001): Üzleti statisztika Excellel, Központi Statisztikai Hivatal, Budapest
26. Sajtos L.- Mitev A. (2007): SPSS kutatási és adatelemzési kézikönyv, Alinea Kiadó. Budapest, 402. o.
27. Szabó K. (2002): Kommunikáció felsőfokon. Kossuth Kiadó. Budapest. 2.Kiadás. 404 o.
28. Székelyi M.- Barna I. (2005): Túlélőkészlet az SPSS-hez, Typotex Kiadó, Budapest, 455.o.
29. Vass M. (2005): Nevelés a sportban: kompetenciák c. habilitációs nyilvános előadás, Veszprémi Egyetem Interdiszciplináris Bölcsészet- és társadalomtudományok (nyelvtudomány; neveléstudomány) Doktor Iskola, Veszprém, 2005. október 18.
30. www. tanulokozosseg.mindentudo.hu/s_doc_server.php?id=1271
31. Zsolnai J. (1996): A pedagógia új rendszere címszavakban Nemzeti Tankönyvkiadó. Budapest, 227. p.

THE MANUAL HAS BEEN PRODUCED IN THE FRAMEWORK OF A PROJECT REGISTERED AS
TÁMOP-4.1.2. E-15/1/KONV-2015-0003

SZÉCHENYI 2020
MAGYARORSZÁG KORMÁNYA
Európai Unió
Európai Szociális Alap
BEFEKTETÉS A JÖVŐBE