Introduction online course on Mathematics and Statistics

Preparatory Course for M.Sc. Integrated Natural Resource Management

2 Introduction to the Basics of Statistics

Department for Agricultural Economics | Resource Economics Group

Syllabus

1 Introduction to the basics of Statistics

- 1.1 Definition and application of statistics
- 1.2 Basic notions
- 1.3 Variables and types of data
- 1.4 Scales of measurement
- 1.5 Statistical data analysis
- 1.6 Data collection and sampling techniques
- 1.7 Observational and experimental studies
- 1.8 Organization of the data
- 1.9 Data description
  - 1.9.1 Measures of center
  - 1.9.2 Measures of variation
- Questions of Repetition
- Answers to the exercises
- References

1.1 Definition and application of statistics

Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data [1].

- Statistics is used to analyse the results of surveys and as a tool in scientific research to make decisions based on controlled experiments. Other uses of statistics include operations research, quality control, estimation, and prediction [1].
- Why to study statistics ? [1]

1. To be able to read and understand the various statistical studies performed in different fields of knowledge → to understand the vocabulary, symbols, concepts, and statistical procedures used in these studies.

2. To conduct research in the field of knowledge, since statistical procedures are basic to research → to be able to design experiments; collect, organize, analyse, and summarize data; make reliable predictions or forecasts for future, and to communicate the results of the study in your own words.

3. To use the knowledge gained from studying statistics to become better consumers and citizens.

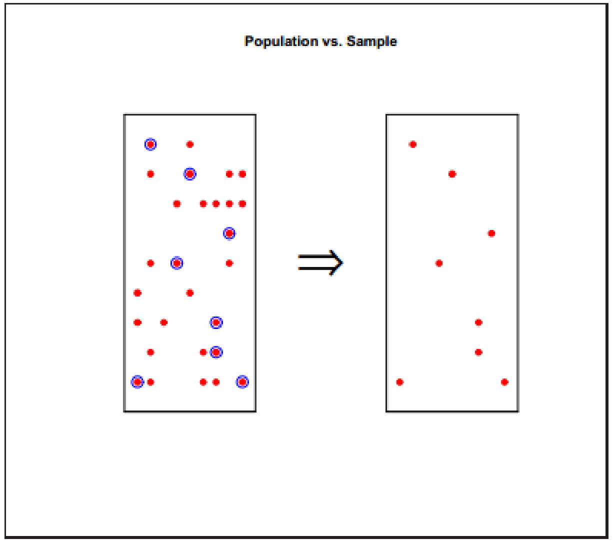
1.2 Basic notions

A ***variable*** is a characteristic or attribute that can attain different values [1].

Data are the values (measurements or observations) that the variables can attain [1].

- Variables whose values are determined by chance are called ***random variables*** [1].
- A collection of data values forms a ***data set***. Each value in the data set is called a ***data value*** or a ***datum*** [1].
- Two main areas of statistics depending on how data are used [1]:

1. ***Descriptive statistics*** consists of the collection, organization, summarization, and presentation of data.
2. ***Inferential statistics*** consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.



Population and sample [8].

A ***population*** consists of all subjects (human or otherwise) that are being studied [1].

A ***sample*** is a group of subjects selected from a population [1].

1.2 Basic notions

An area of inferential statistics called ***hypothesis testing*** is a decision-making process for evaluating claims about a population, based on information obtained from samples [1].

Example: *For example, there is a need to test to which extent organic fertilizers might reduce soil salinity. For this study, two groups of farmers would be selected. One group would use organic fertilizers (such as manure, for example), and the other would use chemical fertilizers. Later, measures of soil salinity on various spots during some period of time would be taken, a statistical test would be done, and a conclusion would be made about the effectiveness of organic fertilization in terms of soil salinity decrease.*

Exercise #1: Read the following and answer the questions.

A study conducted by Spanish scientists showed that school children in Barcelona who had more contacts with green and blue spaces (beaches) have seen their physical activities improved; stress resistance, depression and anxiety decreased; and social contacts increased. These results are based on data on time spent in green spaces and beaches and on Strengths and Difficulties Questionnaires from children parents [2].

Based on this information, attendance of green and blue spaces and physical activity and stress reduction are related. The longer is the school children's playing time in green or blue areas, the better are their physical and mental health and well-being.

1. What are the variables under study?
2. What are the data in the study?
3. Which type of statistics is used?
4. What is the population under study?
5. Was a sample collected? If so, from where?

1.2 Basic notions

A ***parameter*** is an unknown numerical summary of the population.
A ***statistic*** is a known numerical summary of the sample which can be used to make inference about parameters [1].

- A *statistic* is a characteristic or measure obtained by using the data values from a sample [1].
- A *parameter* is a characteristic or measure obtained by using all the data values from a specific population [8].

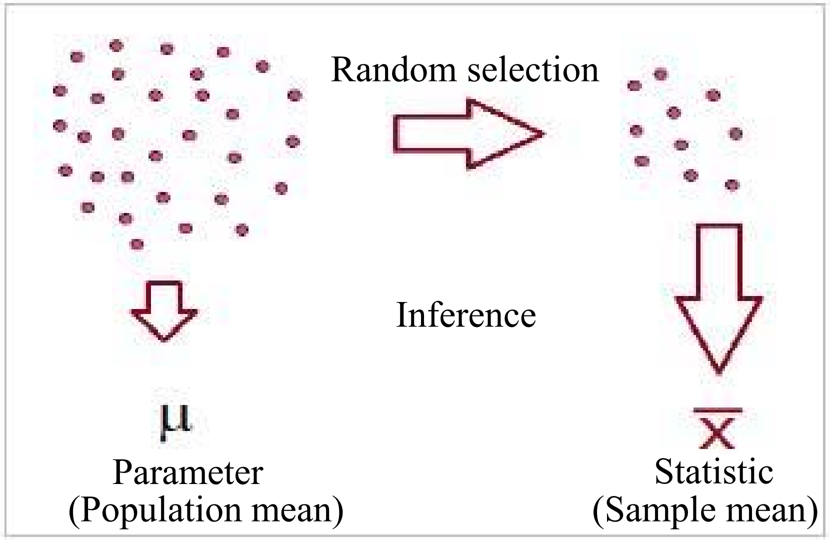


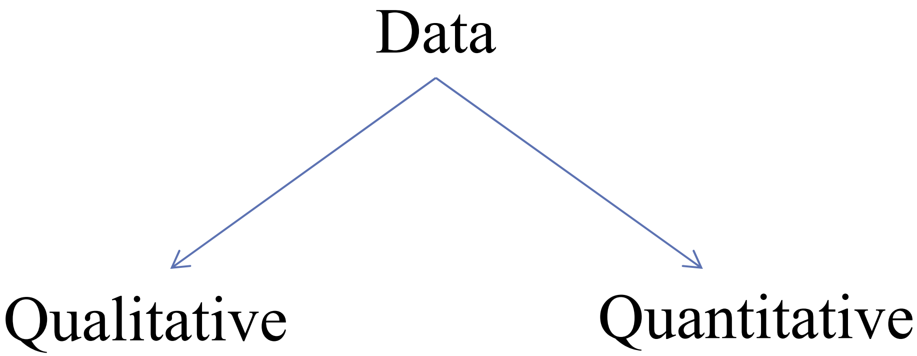
Illustration of the relationship between samples and populations [11].

Example: *Consider the research problem of finding out what percentage of farmers in European Union is practicing energy saving as a measure to reduce greenhouse gas emissions.*
• Parameter: the proportion of farmers in EU practicing energy saving.
• Statistic: the proportion of farmers in EU practicing energy saving calculated from the sample of farmers in EU.

1.3 Variables and types of data

Qualitative variables are variables that can be placed into distinct categories, according to some characteristic or attribute [1].

If subjects are classified according to gender (male or female), then the variable gender is qualitative. Other examples of qualitative variables are religious preference and geographic locations [1].



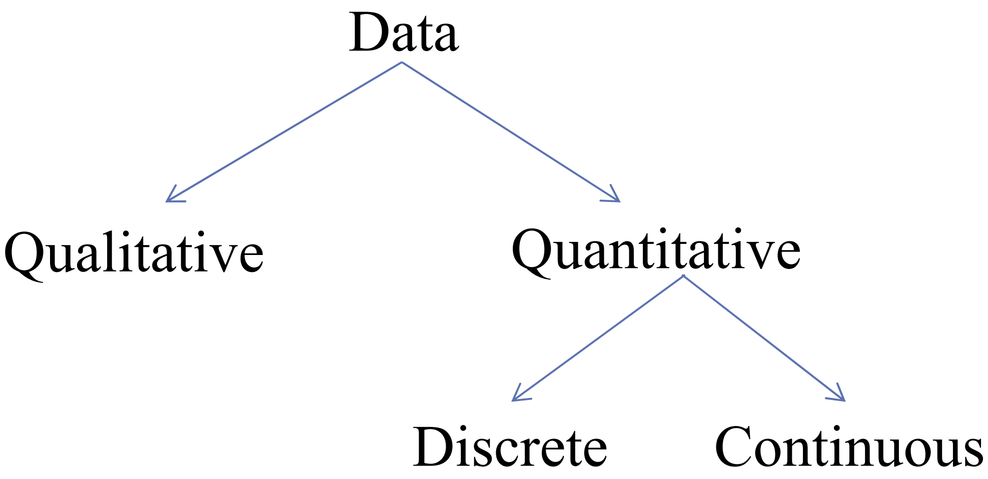
Quantitative variables are numerical and can be ordered or ranked [1].

The variable age is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, and body temperatures [1].

1.3 Variables and types of data

Discrete variables can be assigned values such as 0, 1, 2, 3 and are said to be countable [1].

Examples of discrete variables are the number of children in a family, the number of students in a classroom, and the number of calls received by a switch board operator each day for a month.



Data

Qualitative

Quantitative

Discrete

Continuous

Continuous variables, by comparison, can assume an infinite number of values in an interval between any two specific values [1].

Temperature, for example, is a continuous variable, since the variable can assume an infinite number of values between any two given temperatures [1].

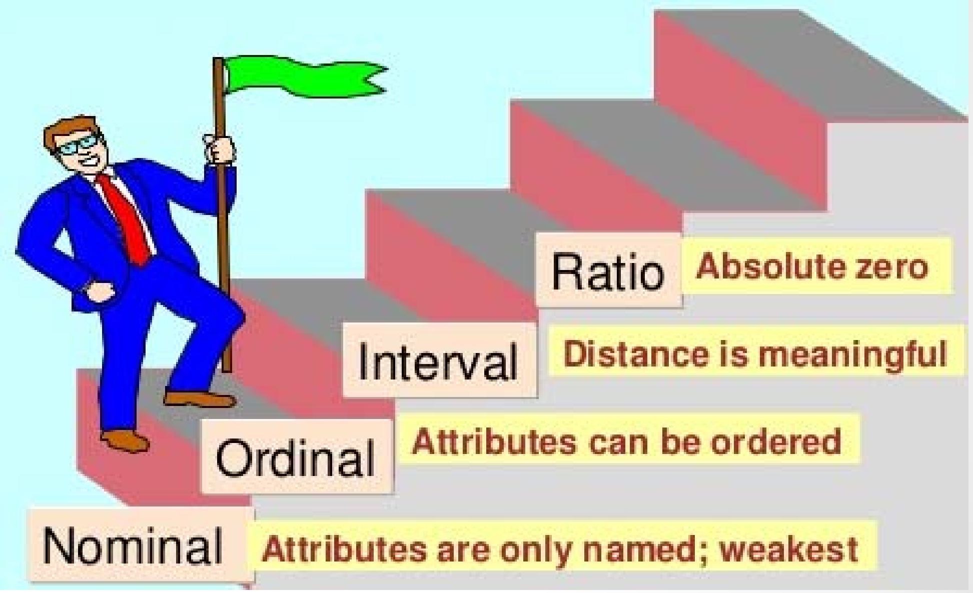
1.4 Scales of measurement

The ***nominal level*** of measurement classifies data into mutually exclusive (non-overlapping), exhausting categories in which no order or ranking can be imposed on the data [1].

Example:

- *Gender (male/female) → dichotomous scale (when there are only two categories);*
- *Nationality*
- *Biological species*
- *Language*

→ *Qualitative data*



- A nominal level is the lowest level of measurement.
- Even if number are assigned to the categories, there is no mathematical meaning in that [1].

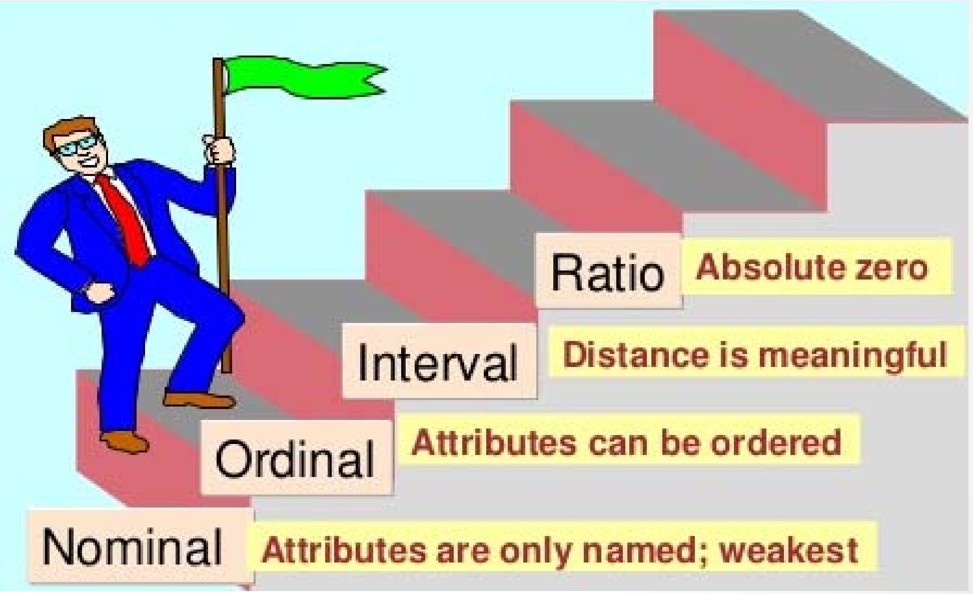
1.4 Scales of measurement

The ***ordinal level*** of measurement classifies data into categories that can be ranked; however, precise differences between the ranks do not exist [1].

Example:

- *Grades*
- *Rating scale*
- *Judging*

→ *measures of non-numeric concepts*



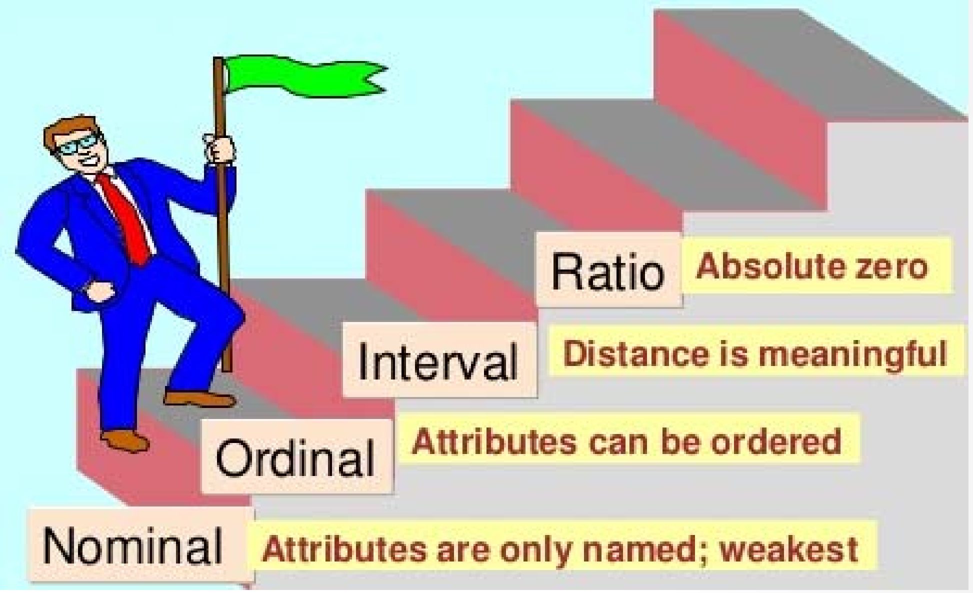
How satisfy are you with new campus services?
1 – Totally satisfied
2 – Somewhat satisfied
3 – Neutral
4 – Completely unsatisfied

1.4 Scales of measurement

The ***interval level*** of measurement ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero [1].

Example:

- *Temperature scale*
- *IQ*
- *Directions measured in degrees*

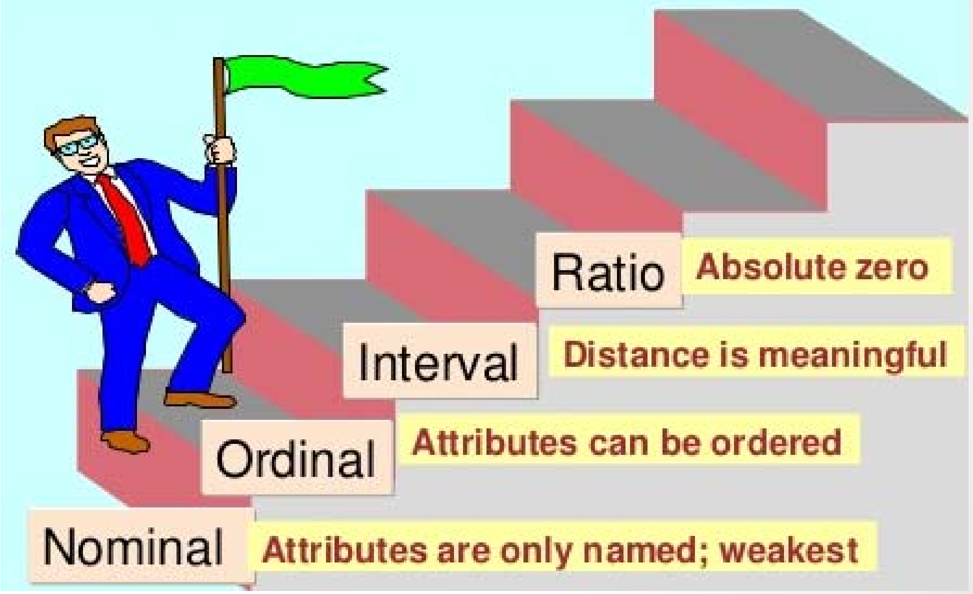


1.4 Scales of measurement

The ***ratio level*** of measurement possesses all the characteristics of interval measurement, and there exists a true zero [1].

- Height
- Weight
- Area

Example: *if one person can lift 100 kilograms and another can lift 50 kilograms, then the ratio between them is 2 to 1. Put another way, the first person can lift twice as much as the second person.*



- The highest level of measurement.
- Ratio scales have differences between units and a true zero.
- In addition, the ratio scale contains a true ratio between values [1].

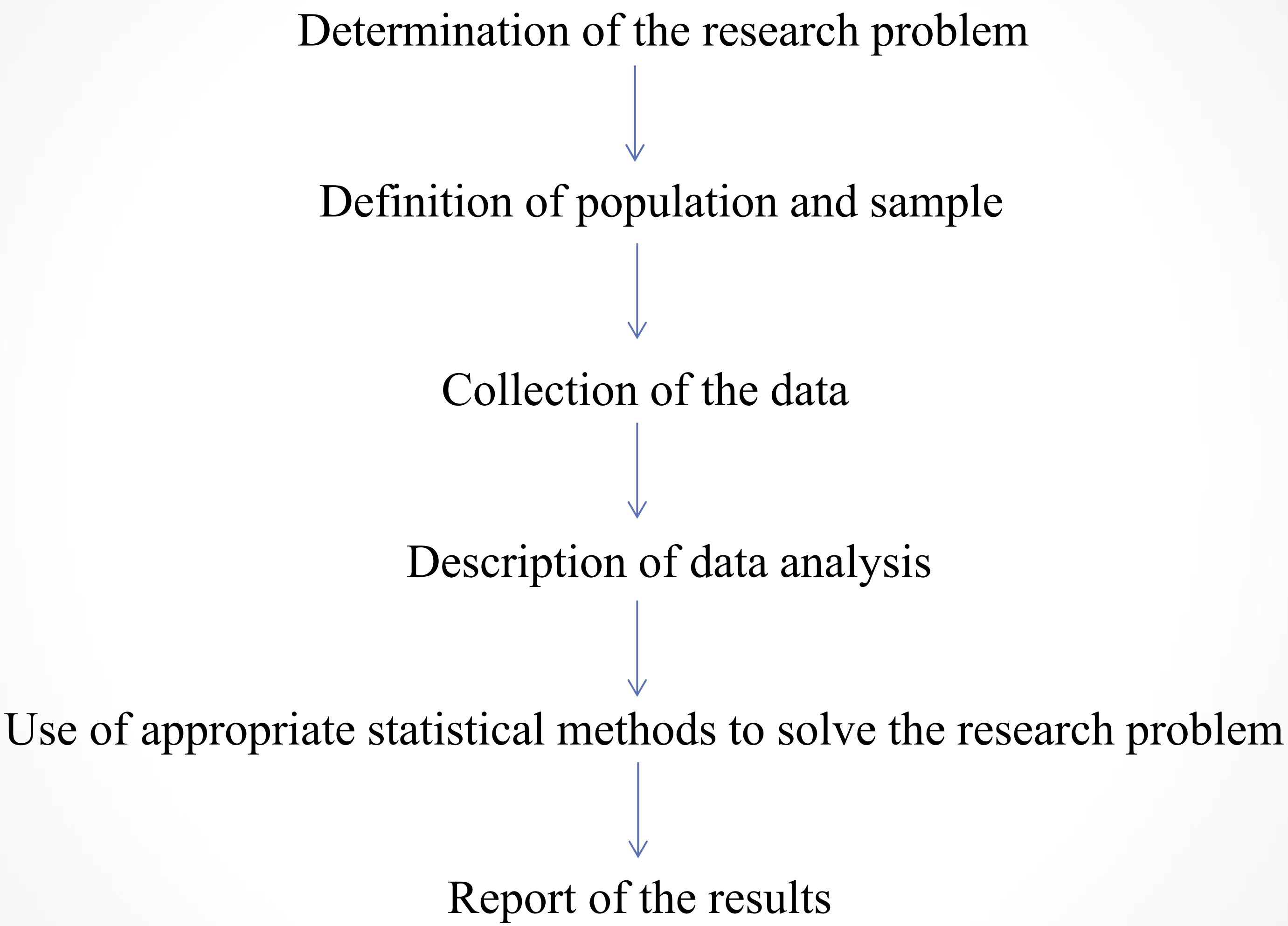
Exercise #2: Read the following information about livestock husbandry in agricultural businesses and answer the questions [7].

Livestock husbandry in agricultural businesses, May 2009

Livestock	Farms (×1000)
Cattle	183.0
of which dairy cows	97.4
Pigs	67.6
of which breeding sows	22.9
Sheep	27.9
Poultry	92.2

1. What are the variables under the study?
2. Categorize each variable as quantitative or qualitative.
3. Categorize each quantitative variable as discrete or continuous.
4. Identify the level of measurement for each variable.
5. The cattle are shown as the most widespread type of livestock in the country. Does that mean that the amount of animals on cattle farms is the greatest? Explain.
6. From the information given, comment on the relationship between the variables.

1.5 Statistical data analysis



Determination of the research problem

Definition of population and sample

Collection of the data

Description of data analysis

Use of appropriate statistical methods to solve the research problem

Report of the results

1.6 Data collection and sampling techniques

Data collection through the use of ***surveys***, which can be done by a variety of methods [1]:

1) Telephone survey:
- less costly
- people can be straightforward and sincere as there is no direct face-to-face contact
- disadvantages: not all have phones and people can simply ignore the calls
- the tone of the interviewer is very important → sets the pace of the survey

2) Mailed questionnaire:
- allows to cover more geographical areas
- costs saving
- anonymity of the respondents
- disadvantages: low rate of responses and difficulties in understanding the questions → inappropriate answers to the addressed questions

3) Personal interview surveys:
- allows getting more profound responses to the questions
- disadvantages: more costly and selection of respondents

Data can also be collected in other ways, such as ***surveying records*** or ***direct observation*** of situations [1].

1.6 Data collection and sampling techniques

1) ***Random samples*** are selected by using chance methods or random numbers [1].
- Generation of random numbers with a computer or calculator

2) ***Systematic samples*** are selected by numbering each subject of the population and then selecting every *k*th subject [1].

Example: *Consider the situation, where there are 100 subjects in the population and a sample of 10 subjects was needed. Since* 100/10= *10, then the coefficient for systematic selection would be 10, and every 10th subject would be selected; however, the first subject (numbered between 1 and 10) would be selected at random. Suppose subject 8 was the first subject selected; then the sample would consist of the subjects whose numbers were 8, 18, 28, etc., until 10 subjects were obtained.*

3) ***Stratified samples*** are selected by dividing the population into groups (called strata) according to some characteristic that is important to the study, then sampling from each group [1].
- Samples within the strata should be randomly selected.

Example: *For example, in order to identify the problems of the use of common pool resources (resources, such as aquifers, fisheries, forests, which are subjected to depletion as exclusion mechanisms are lacking), the data might be collected by respective selection of samples from each group of the resource users (farmers, industry, local citizens, etc.).*

1.6 Data collection and sampling techniques

4) Cluster samples: here the population is divided into groups called clusters by some means such as geographic area or schools in a large school district, etc.

→ Cluster sampling is used when the population is large or when it involves subjects residing in a large geographic area [1].

Example: *If one wants to do a study on food preferences involving young people in the universities in Berlin, it would be very costly and time-consuming to try to obtain a random sample of students taking into account a great number of students in the capital. Instead, a few universities could be selected at random, and the students there would be interviewed in a cluster.*

Exercise #3: Assume you are writing a paper and you have become increasingly concerned about the meat consumption nowadays as there are a lot of debates on its environmental impact. You set up a plan and would like to conduct a survey on how people believe the modern culture (education, media, science, society, etc.) influences meat consumption. You are planning to survey 50 adults and adolescents from around the country. Answer the following questions about your survey.

1. What type of survey would you use (phone, mail, or interview)?
2. What are the advantages and disadvantages of the chosen surveying methods?
3. What type of scores would you use? Why?
4. Would you use a random method for deciding who would be in your sample?
5. Which of other methods (stratified, systematic, cluster,) would you also use? Why?

1.7 Observational and experimental studies

- In an ***observational study***, the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.
- In an ***experimental study***, the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

Independent variable (explanatory variable) in an experimental study is the one that is being manipulated by the researcher [1].

The resultant variable is called the ***dependent variable*** or the ***outcome variable*** [1].

The ***outcome variable*** is the variable that is studied to see if it has changed significantly due to the manipulation of the independent variable [1].

Example: *For example, two different groups of crop plots being treated exactly the same except for the fact that one group of plots is under no-till farming practice. This mainly includes minimization of disturbances of the ground surface during planting and minimization of the removal of crop residues during harvesting. Generally no-tillage results in reduction of erosion and runoff and in increased soil infiltration and soil organic matter content. Therefore, when comparing soil performance of the two plot groups, the one under no-tillage will have more fertile and resilient soils. Hence, the independent variable in this case would be farming practice (conventional or no-tillage). The dependent variable, then, would be the resultant variable, that is, soil quality. If the differences in the dependent or outcome variable are large and other factors are equal, these differences can be attributed to the manipulation of the independent variable. The plot group that is under special, no-tillage, farming practice is called the treatment group while the other – the control group.*

1.7 Observational and experimental studies

	Experimental studies	Observational studies
Advantages	• The researcher can decide how to select subjects and how to assign them to specific groups. • The researcher can control or manipulate the independent variable.	• Natural setting • It can be done in situations where it would be unethical or downright dangerous to conduct an experiment. • Observational studies can be done using variables that cannot be manipulated by the researcher.
Disadvantages	• Unnatural settings, such as laboratories and special classrooms. • Hawthorne effect: the subjects who know they are participating in an experiment actually can change their behaviour in ways that affect the results of the study. • Confounding of variables. A confounding variable is one that influences the dependent or outcome variable but was not separated from the independent variable.	• Since the variables are not controlled by the researcher, a definite cause-and-effect situation cannot be shown since other factors may have had an effect on the results. • Observational studies can be expensive and time-consuming. • Since the researcher may not be using his or her own measurements, the results could be subject to the inaccuracies of those who collected the data.

Advantages and disadvantages of experimental and observational types of statistical studies [1].

1.8 Organization of the data

- Each individual piece of data is called an ***observation*** and the collection of all observations for particular variables is called a ***data set*** or ***data matrix***. Data set are the values of variables recorded for a set of sampling units [1; 8].

Example: *For example, area used for agriculture might be coded by letting 1,2,3, and 4 denote a culture type being arable land, permanent grassland, fruit growing or vines. However, this coded data still continues to be nominal data. Coded numerical data do not share any of the properties of the numbers we deal with ordinary arithmetic. With regards to the codes for type of agricultural land, we cannot affirm that* $3 > 2$ *or* $2 - 1 = 4 - 3$. *This shows how important it is always to check whether the mathematical treatment of statistical data is really valid.*

$$\begin{array}{cc} & \text{Variables} \\ \text{Sampling units} & \left[\begin{matrix} x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3k} \\ \dots & & & & \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{matrix}\right] \end{array}$$

where x_{ij} is a value of the j:th variable collected from i:th observation, i = 1, 2, …, n and j = 1, 2, …, k [8].

1.9 Data description.
1.9.1 Measures of center

The sample mode of a qualitative or a discrete quantitative variable is that value of the variable which occurs with the greatest frequency in a data set [1].

Definition of Mode: Obtain the frequency of each observed value of the variable in a data and note the greatest frequency [1]:

1. If the greatest frequency is 1 (i.e. no value occurs more than once), then the variable has no mode.
2. If the greatest frequency is 2 or greater, then any value that occurs with that greatest frequency is called a sample mode of the variable.

Example: *For example, the data show the number of sheep stock (×million) in Uganda for every decade from 1961 to 2011 [9]. Find the mode.*

0.9 0.8 1.3 0.8 1.1 1.9

***Solution**: Since the value 0.8 occurs twice, the mode is 0.8.*

The mode for grouped data is the modal class. The **modal class** is the class with the largest frequency [8].

1.9.1 Measures of center

The Mode [1]**:**

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode can be used when the data are nominal, such as religious preference, gender, or political affiliation.
4. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

<ins>Example:</ins> *Consider an example: A study showed the age distribution of farm managers in the organic sector in the EU-27 in 2010 [5]. Find the mode.*

<35 years	*10.4%*
35-44 years	*21.7%*
45-54 years	*29.7%*
55-64 years	*19.9%*
>=65 years	*12.2%*

Solution: *Since the age category with the highest frequency is 45-64 years, the most common age of organic farm managers is from 45 up to 64.*

1.9.1 Measures of center

The sample median of a quantitative variable is that value of the variable in a data set that divides the set of observed values in half, so that the observed values in one half are less than or equal to the median value and the observed values in the other half are greater or equal to the median value [1].

Definition of Median. Arrange the observed values of variable in a data in increasing order.

1. If the number of observation is odd, then the sample median is the observed value exactly in the middle of the ordered list.

2. If the number of observation is even, then the sample median is the number halfway between the two middle observed values in the ordered list.

→In both cases, if we let *n* denote the number of observations in a data set, then the sample median is at position *n+1/2* in the ordered list [1; 8].

Example: *the cattle stock (in thousand heads) in Central Asian countries (Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan) in 2011 was 6 185, 1 339, 1 912, 2 200, and 9 094 [6]. Find the median.*

Solution:

Step 1 Arrange the data in order.

1 339, 1 912, 2 200, 6 185, 9 094.

Step 2 Select the middle value.

1 339, 1 912, 2 200, 6 185, 9 094.

↑

Median

→ *Hence, the median is 2 200 thousand heads.*

1.9.1 Measures of center

The symbol for the median is MD.

Example: *Wheat production (in thousand tonnes) in Central and Eastern Europe (Bulgaria, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia) in 2011 was 4 458, 4 913, 360, 4 107, 937, 1 869, 9 339, 7 132, 1 639, and 154 [6]. Find the median.*

Solution:

Step 1 Arrange the data in order.
154, 360, 937, 1639, 1 869, 4 107, 4 458, 4 913, 7 132, 9 339.

Step 2 Select the middle value.
154, 360, 937, 1639, 1 869, 4 107, 4 458, 4 913, 7 132, 9 339.
↑
Median

MD = (1 869+4 107)/2 = 2 988

Hence, the median is 2 988 thousand tonnes of wheat.

The Median [1]

1. The median is used to find the center or middle value of a data set.
2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
3. The median is used for an open-ended distribution.
4. The median is affected less than the mean by extremely high or extremely low values [1].

1.9.1 Measures of center

The sample mean of the variable is the sum of observed values in a data divided by the number of observations [1].

The mean is the sum of the values, divided by the total number of values. The symbol $\overline{X}$ represents the sample mean [1].

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n} = \frac{\Sigma X}{n}$$

where *n* represents the total number of values in the sample.

For a population, the Greek letter *μ* is used for the mean and *N* is used to represent the total number of values in population [1].

Example: *The data shown represent the amount of water resources per capita* (m^3*/yr/cap) for five countries in Central Asia (Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan) in 2010 [6]. Find the mean.*

6 839, 9 177, 2 323, 4 903, 1 837.

Solution:

$$\overline{X} = \frac{\sum X}{n} = \frac{6\ 839 + 9\ 177 + 2\ 323 + 4\ 903 + 1\ 837}{5} = 5\ 015{,}8$$

The mean for the water distribution per capita in Central Asian countries in 2010 is 6868.3.

1.9.1 Measures of center

The Mean [1]

1. The mean is found by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for the data set is unique and not necessarily one of the data values.
5. The mean cannot be computed for the data in a frequency distribution that has an open-ended class.
6. The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.

The ***midrange*** is a rough estimate of the middle [1].

The midrange is defined as the sum of the lowest and the highest values in the data set, divided by 2. The symbol MR is used for the midrange [1]:

$$\mathrm{MR} = \frac{\text{lowest value} + \text{highest value}}{2}$$

Example: *The data presented below are on eggs production (in thousand tonnes) in some European countries (Cyprus, Luxembourg, Malta, Iceland, Ireland, Norway) in 2011 [6]. Find the midrange.*

8, 2, 4, 3, 45, 60

Solution:

$$MR = \frac{2+60}{2} = 31$$ → *Hence, the midrange is 31.*

1.9.1 Measures of center

Exercise #4: The following data represent production of renewable energy from agriculture (in thousand tonnes) in some countries of the EU in 2011 [4], [1].

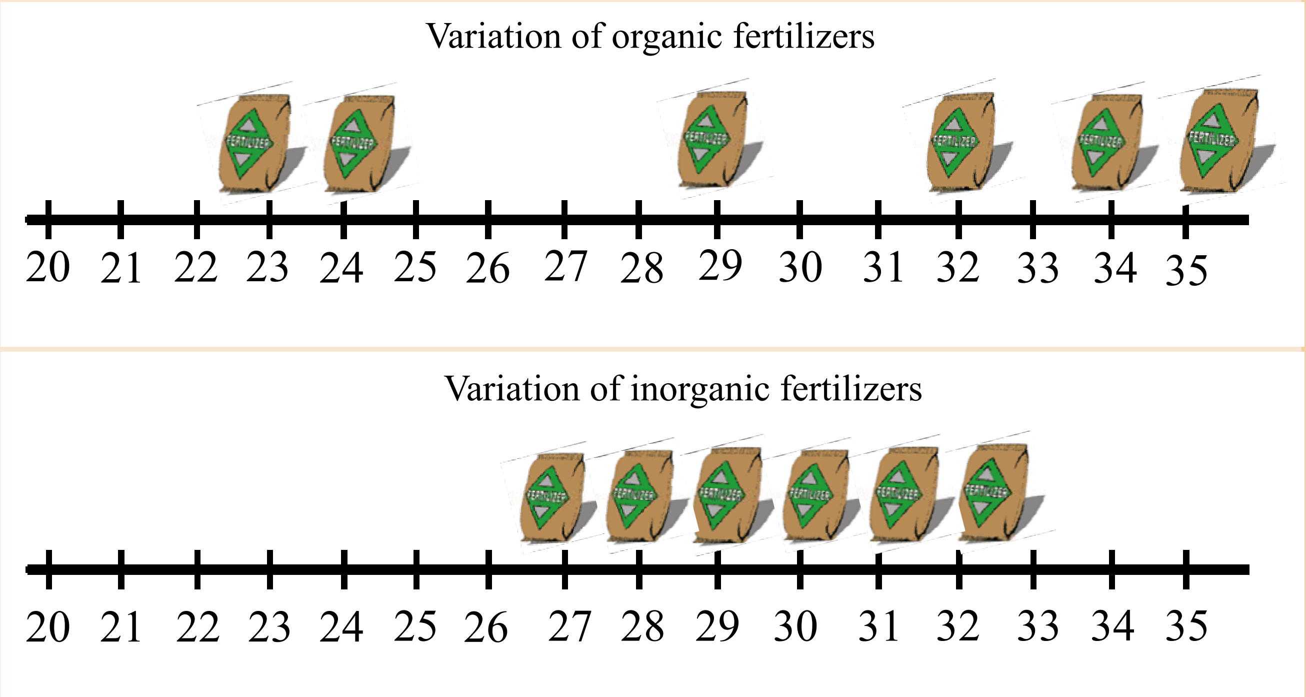
696.2 370.4 7 724.8 33.6 848.5 1 158.6 63.8
11.3 673.2 426.2 124.9 185.5 353.2 436.3

1. Compute the mean, median, and mode.
2. In 2011 renewable energy production is seen to be declined, thereby changing the upward trend that had been observed before that. Assume the need to promote more green energy production in agricultural sector. Use the best measure of central tendency to support this position.
3. How can outliers (extreme values: maximum and minimum) be used to support such position.
4. Which measure of central tendency can be misleading when a data set contains outliers?

1.9.2 Measures of variation

Example: *Imagine the study conducted on the application of different types of fertilizers (organic and inorganic) with the goal to bring about comparable yields. Here are possible results representing wheat yields (in dt/ha/year) under the treatment of organic and inorganic fertilizers over 6 years. Find the mean of each group of fertilizers* [1].

Organic	Inorganic
32	32
29	29
34	27
24	28
35	31
23	30



Examining data sets graphically: Variation of organic and inorganic fertilizers.

Solution:

The mean for organic fertilizers is $\mu = \frac{\sum X}{N} = \frac{177}{6} = 29{,}5$ *dt/ha/year*

The mean for inorganic fertilizers is $\mu = \frac{\sum X}{N} = \frac{177}{6} = 29{,}5$ *dt/ha/year*

1.9.2 Measures of variation

The ***range*** is the highest value minus the lowest value. The symbol *R is used for the range.*
R = highest value - lowest value [1]

Example: *Referring to the previous example, find the ranges for the fertilizers.*

Solution:
For organic fertilizers, the range is R = 35 - 24 = 11 dt/ha/year.
For inorganic fertilizers, the range is R = 32 - 27 = 5 dt/ha/year.

The range for organic fertilization shows that 11 dt/ha/year separate the largest data value from the smallest data value. For inorganic fertilization, 5 dt/ha/year separate the largest data value from the smallest data value, which is almost a half of organic fertilization's range [1].

1.9.2 Measures of variation

Example: *Find the variance and standard deviation for the data set for organic fertilizers: 32, 29, 34, 24, 35, 23.*

Solution:

Step 1 *Find the mean for the data.* $\mu = \frac{\sum X}{N} = \frac{32+29+34+24+35+23}{6} = 29{,}5$

Step 2 *Subtract the mean from each data value.*

32 – 29,5 = +2,5 34 – 29,5 = +4,5 35 – 29,5 = +5,5

29 – 29,5 = -0,5 24 – 29,5 = -5,5 23- 29,5 = -6,5

Step 3 *Square each result.*

$(+2{,}5)^2 = 6{,}25 \quad (+4{,}5)^2 = 20{,}25 \quad (+5{,}5)^2 = 30{,}25$

$(-0{,}5)^2 = 0{,}25 \quad (-5{,}5)^2 = 30{,}25 \quad (-6{,}5)^2 = 42{,}25$

Step 4 *Find the sum of the squares.*

6,25 + 0,25 + 20,25 + 30,25 + 30,25 + 42,25 = 129,5

Step 5 *Divide the sum by N to get the variance.*

Variance = 129,5/6 ~ 21,58

Values X	X - μ	(X - μ)²
32	+2,5	6,25
29	-0,5	0,25
34	+4,5	20,25
24	-5,5	30,25
35	+5,5	30,25
23	-6,5	42,25

Step 6 *Take the square root of the variance to get the standard deviation. Hence, the standard deviation equals √21,58, or ~ 4,65.*

It is helpful to make a table, where the first column contains the raw data X, the second column contains the differences X - μ obtained in step 2; finally, the last column contains the squares of the differences obtained in step 3 [1].

1.9.2 Measures of variation

The ***variance*** is the average squared difference of the scores from the mean [1].

The symbol for the population variance is σ^2 (σ is the Greek letter sigma). The formula for the population variance is [1]:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

where X is an individual value, μ is a population mean and N represents the population size.

The ***standard deviation*** is the square root of the variance [1].

The symbol for the population standard deviation is σ. The corresponding formula for the population standard deviation is [1], [10]:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

1.9.2 Measures of variation

The variance for a sample: $\boxed{\frac{\sum(X-\bar{X})^2}{n}}$, where $\bar{X}$ is the sample mean and *n is the sample size* [1].

This formula is not usually used, however, since in most cases the purpose of calculating the statistic is to estimate the corresponding parameter. For example, the sample mean is used to estimate the population mean μ [1].

The expression $\frac{\sum(X-\bar{X})^2}{n}$ does not give the best estimate of the population variance because when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by *n, variance of the sample is computed by dividing by n - 1, giving a slightly larger value and an unbiased estimate of the population variance* [1].

The formula for the sample variance, denoted by s^2, is

$$\boxed{s^2 = \frac{\sum(X-\bar{X})^2}{n-1}}$$

where $\bar{X}$ is a sample mean and *n* is a sample size [1].

1.9.2 Measures of variation

The shortcut formulas for computing the variance and standard deviation for data obtained from samples [1]:

- Variance

$$s^2 = \frac{n(\sum X^2) - (\sum X)^2}{n(n-1)}$$

- Standard deviation

$$s = \sqrt{\frac{n(\sum X^2) - (\sum X)^2}{n(n-1)}}$$

Example: *Find the sample variance and standard deviation for the amount of common wheat yield (100 kg/ha) in Germany for a sample of 4 years shown: 78.2, 72.2, 70.2, 73.4 [5].*

Step 1 *Find the sum of the values:*

$\sum X = 78.2 + 72.2 + 70.2 + 73.4 = 294$

Step 2 *Square each value and find the sum:*

$\sum X^2 = 78.2^2 + 72.2^2 + 70.2^2 + 73.4^2 = 21643.68$

Step 3 *Substitute in the formulas:*

$s^2 = 4(21543.68) – 294^2 / 4(4\text{-}1) = 86754.72\text{-}86436 / 4(3) = 138.72 / 12 = 11.56$

The variance is then 11.56.

$s = \sqrt{}11.56 = 3.4$

Hence, the sample standard deviation is 3.4 [1].

1.9.2 Measures of variation

Exercise #5:
The dairy yields (kg/cow) in 5 selected European countries in 2009 are as follows [5]:
5722, 5067, 4816, 4770, 6050.
The barley yield (in 100kg/ha) for these same 5 countries in 2009 is listed here [5]:
34.3, 35.2, 34.4, 30.9, 34.2
Which set is more variable?

Exercise #6: Find the variance and standard deviation for the data set for inorganic fertilizers in the example that was considered. The wheat yields (in dt/ha/year) under the treatment of inorganic fertilizers were 32, 29, 27, 28, 31, 30.

Questions of repetition

- What is statistics and why is it important to get to know some basics of it?
- What is the difference between the population and the sample?
- Why is it important to distinguish between descriptive and inferential statistics?
- What is variable? Which types of variables do exist and how can they be differentiated?
- How many scales of measurement do exist? What are their distinctive features?
- What is the difference between a parameter and a statistic?
- Which means can be used for data collection?
- Which sample techniques do you know?
- What are advantages and disadvantages of observational and experimental studies?
- How do dependent and independent variables correlate?
- Which measures of centre do exist? What are the differences? When these measures can be used?
- What is the relationship between variance and standard deviation?
- Why might the range not be the best estimate of variability?

Answers to the exercises

Exercise #1:

1. The variables are contact with/attendance of green and blue spaces and physical activities, stress resistance, depression, social cohesion/physical and mental health and well-being.
2. The data consist of different indices of behavioral indicators (stress resistance, anxiety and social contact, etc.) and time spent in green and blue spaces.
3. These are descriptive statistics.
4. The population under study is school children in Barcelona.
5. While not specified, the data is probably from a sample of school children in Barcelona.

Exercise #2:

1. The variables are livestock and number of farms.
2. The type of livestock is a qualitative variable, while the number of farms is quantitative.
3. The number of job-related injuries is discrete.
4. Type of livestock is nominal, and the number of farms is ratio.
5. The cattle farms are the most common farms, however, it does not directly mean that a cattle stock is the highest.
6. Answers will vary. One possible answer is that the number of poultry farms are twice less than cattle farms.

Answers to the exercises

Exercise #3:

Answers may vary, so these are one possible answers.

1. I would use a telephone survey.
2. The advantage would be that this is a relatively inexpensive survey method (although more expensive than using the mail). Interviewing by phone would allow me to use follow-up questions and to clarify any questions of the respondents at the time of the interview. However, interviewing is very labor- and cost-intensive. The disadvantage to my survey method is that I cannot include people without a telephone. Moreover, quite often people don’t have enough time, are busy or cannot speak at the very moment.
3. I could use ordinal data on a scale of 1 to 5: 1 - strongly disagree, 2 - disagree, 3 - neutral, 4 - agree, 5 - strongly agree.
4. The random method that I could use is a random dialing method.
5. To include people from each administrative unit, I would use a stratified method; and the random one while selecting respondents.

Exercise #4:

1. The sample mean is 936,18 thousand tonnes, the sample median is 398,3 thousand tonnes, and there is no mode.
2. The sample mean would be a good measure of center to report.
3. With the outliers removed, the sample mean is 447.5 thousand tonnes, the sample median and the sample mode will be the same. → The mean is greatly affected by the outliers.
4. The mean can be misleading in the presence of outliers, since it is greatly affected by these extreme values.

Answers to the exercises

Exercise #5:

Dairy yields: range – 1280
variance – 327226
standard deviation – 572

Barley yields: range – 4,3
variance – 2,785
standard deviation – 1,67

Hence, the dairy yields are more variable.

Exercise #6: variance – 2,9
standard deviation – 1,7

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Resource Economics

References:

[1] Allan G. Bluman, 2009. *Elementary Statistics: A Brief Version*, 7th Edition, New York: McGraw-Hill.

[2] Amoly E., 2014: Green and Blue Spaces and Behavioral Development in Barcelona Schoolchildren: The BREATHE Project. Environmental Health Perspectives, volume 122, number 12.

[3] European Commission: Agriculture and Rural Development. Statistics and indicators. Rural development statistics, 2011. Retrieved from http://ec.europa.eu/agriculture/statistics/rural-development/2012/full-text_en.pdf

[4] European Commission: Agriculture and Rural Development. Statistics and indicators. EU agriculture - Statistical and economic information – 2013. Retrieved from http://ec.europa.eu/agriculture/statistics/agricultural/2013/pdf/d01-1-41_en.pdf

[5] Facts and figures on organic agriculture in the European Union. European Commission. Retrieved from http://ec.europa.eu/agriculture/statistics/agricultural/2013/pdf/d01-1-41_en.pdf

[6] FAO Statistical Yearbook, 2014: Europe and Central Asia: food and agriculture. Food and Agriculture Organization of the United Nations Regional Office for Europe and Central Asia, Budapest.

[7] German Agriculture. Facts and Figures, 2010: Federal Ministry of Food, Agriculture and Consumer Protection/Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz, BMELV.

[8] Jarkko Isolato, Basics of Statistics. Retrieved from http://www.mv.helsinki.fi/home/jmisotal/BoS.pdf

[9] Leliveld A. et al., 2013: Agricultural dynamics and food security trends in Uganda. Research Report 2013-ASC-2. Developmental Regimes in Africa (DRA) Project ASC-AFCA Collaborative Research Group:Agro-Food Clusters in Africa (AFCA), London/Leiden.

[10] Online Statistics Education: A Multimedia Course of Study. Project Leader: David M. Lane, Rice University. Retrieved from http://onlinestatbook.com/

[11] http://www.cliffsnotes.com/math/statistics/sampling/populations-samples-parameters-and-statistics