Principles of Biostatistics and Informatics

2nd Lecture: Descriptive Statistics

19th September 2018

Dániel VERES

1

In this course and in this lecture we are going to discuss the basic concepts of statistics. In the first part of the lecture I am going to give simplified definition on statistics, variables and their outcomes. Thereafter we will classify the functions in statistics as descriptive and inferential statistics. After that we will discuss on types of variables. In the core of the lecture we will talk about descriptive statistics elements along the lines of Stevens's measurement scale. At first, we describe univariate and then multivariate samples. On the last few slides we will learn about percentile curves. At last I am going to share some thoughts on data collection and data recording.

Tastitsticsss? What's that?

Statistics describes **random mass** phenomena.

- ***Data Collecting (Sampling)***
- ***Data Organization***

Descriptive Statistics

- ***Data Analysis***
- ***Conclusion***

Inferential Statistics
(Inductive)

One definition of statistics: statistics describes *random mass* phenomena.
To describe the „random mass" – so (several) variables with several outcomes - we perform the next actions: *collecting data* (sampling with other words), *organizing data, analyzing data* and *making conclusions*. The first two activity falls within the scope of *descriptive statistics* and the second two belongs to the *inferential statistics* (called inductive statistics also). Although there is no sharp boundary between the two plots. I have to highlight that descriptive statistics – both data collection and organization – is always needed to create complete statistics and perfect conclusions.

Tastitsticsss? What's that?

Statistics describes **random mass** phanomenons.

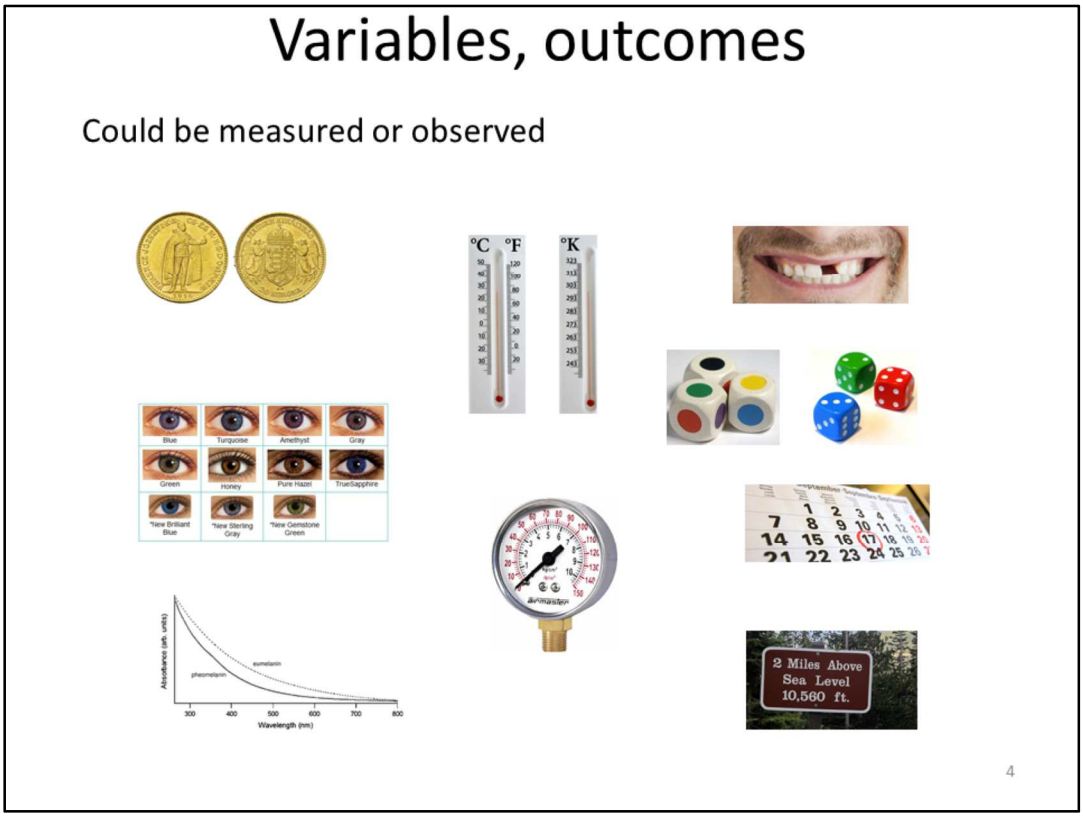
- Data Collecting (Sampling)
- Data Organization

} **Descriptive Statistics**

- Data Analysis
- Conclusion

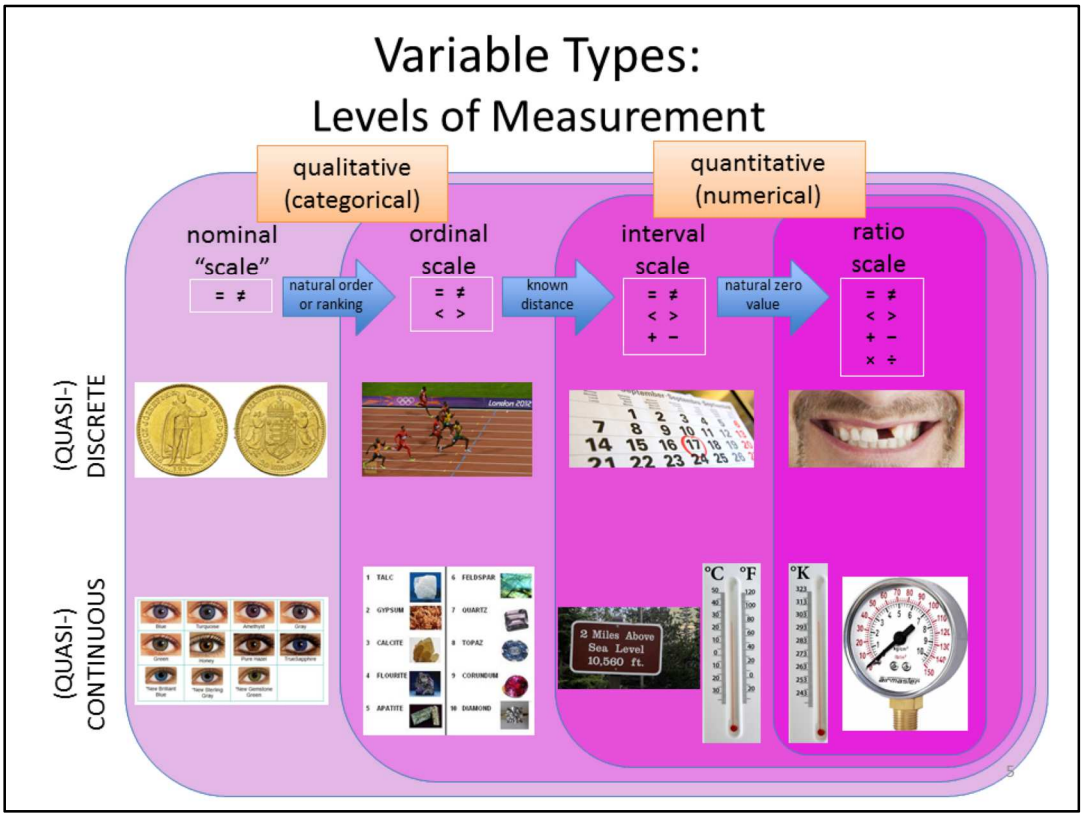
} Inferential Statistics (Inductive)

Let's begin with the descriptive statistics. We will return later to the sampling but now we have to concentrate on data organization.
Data organization helps to describe, show or summarize our data in the sample *in a meaningful way* such as, for example, patterns might emerge from the data.



In statistics data are belong to *variable*s.
I give you here a simplified definition for variables: variable could be anything that we could measure or observe. E.g. tossing a coin, hair or eye color, temperature, blood pressure, rolling a die, etc.
When we are measuring or observing variables in a given circumstances it has possible outcomes.
E.g.: the outcomes of tossing a coin could be tail or head, the eye color could be brown, blue ,green, etc. but we can measure and characterize eye color by its spectra.
It is very import to handle the variable based on its outcome in the given situation.

In this presentation you will find figures from previous lecture to remind you what you have to already know...
For data organization *the level of the measuring scale* (the type) and the number of the examined variables *is crucial*.



There are many ways to group statistical variables depending on our aim.
In a practical view I would like to show a categorization on variables based on their possible *outcomes* so the *levels of measurement*.

As a first approach, we can group variables into the *qualitative (*also called *categorical)* and quantitative (*numerical*) types.

The most primitive scale s the nominal scale, which is at the bottom of the hierarchy of measurement scales. Examples are personal name, blood group, hair or eye color, citizenship etc. The scale is created by defining categories, these categories can be identified by simple naming (hence, "nominal"). During observations it is possible to determine whether two elements are identical or not. There is no natural order among categories, but there may be practical orders set (e.g. alphabetical order, assigned ordinal number), which are used according to tradition or customs, which help comparison. However, these orders do not have any meaning. Therefore, even the name "scale" is sort of misleading (misnomer), it would be more correct to speak about nominal system, which would not let us expect natural (meaningful) order. The delimitation of nominal categories may either be easier (self evident, like in case of coin tossing) or more difficult (arbitrary, e.g. eye color).

The ordinal scale also uses categories, but there's a natural order among them, examples are school notes, severity grade of diseases or injuries, or the Mohs scale of mineral hardness. Consequently, on a nominal scale not only identity can be defined, but "less than"/"greater than" relationships as well. Scale elements are usually denoted by

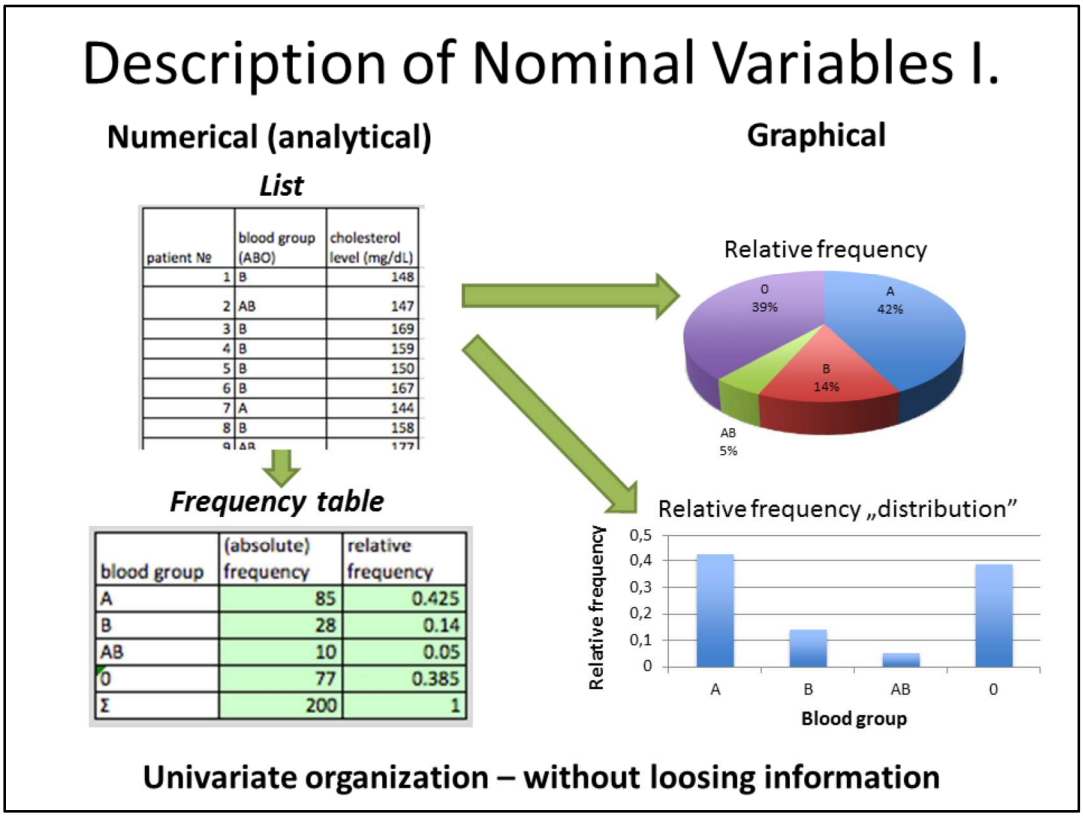
ordinal numbers, which has to be kept in mind since the usual mathematical operations cannot be carried out on them. The difference or distance between the categories of an ordinal scale are either unequal or cannot be determined.

The interval scale is more developed than the ordinal scale because the distance between the possible values is known, so not only the order but the difference and addition can be interpreted. Examples from everyday life are calendar years, temperature in degrees Celsius or Fahrenheit, or the height above see level. It is evident from the examples that the zero value of interval scales is set arbitrarily.

Instead of such arbitrary zero values, ratio scales have natural zero values, actually ratio (and proportion) can be interpreted due to the existence of this natural zero value. So mathematical operations related to proportionality (i.e. multiplication and division) can also be carried out on such scales. Examples are: temperature measured on Kelvin scale, length, blood pressure...

It is possible to distinguish between more or less discrete and continuous variables at all scale levels based on the possible different outcomes. In practice we say continuous if we have at least 20 different outcomes.

In statistics for data organization and for the further evaluation *the level of the measuring scale* (the type) and the number of the examined variables *is crucial* as you will see in the next slides and the further lectures.



Let’s begin with how to organize a variable in a *nominal scale* using blood group type as an example.
In all statistical descriptions there are basically two options: *numerical* (called analytical also) where numbers are used to interpret the data and *graphical* where data are presented on charts.
After data collection we have a *list* that could be compacted to a *frequency table* or could be visualized on *frequency charts*. [As you learned in the first lecture and you will learn in the practices.]
In the case of nominal variables the shown organizations are *without loosing information* that means we can recreate the original dataset on AB0 blood group if we are interested only on blood type, not taking into account which blood type belongs to which patients – so we made univariate description.

For further analyses or comparisons, there are too much „information” – so we have to find a typical value – an indicator that could characterize our dataset.

Description of Nominal Variables II.

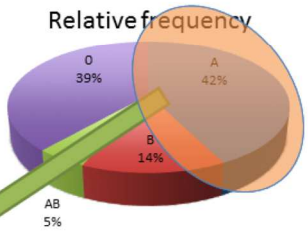
Numerical

Frequency table

blood group	(absolute) frequency	relative frequency
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

Graphical

The brain and the common sense

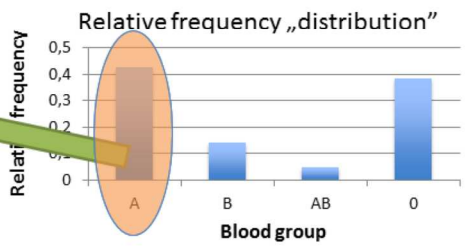


Organization, but **loss of information**

„Typical value" (***indicator***): ~~**Mean?!**~~

***Mode*:** most frequent element(s)

Notation: *Mod*, x_{mod}



Other parameters:

data count (*n*)***, count of categories***

In the case of nominal scale this indicator could be only the *most frequent element, the mode*. (The mean has no meaning here...)

However, this solution has a disadvantage: only knowing the mode we can not restore the original data set – *we lose information*.

Never forget the power of a graphical interpretation: our human brain together with the common sense could easily find meaningful patterns behind the numbers! In the plots we could easily find the mode – that is *A* blood group type in the example. Another benefit of the graphical description that we could also easily see is the „goodness" of the mode in a certain case: in our example we realize that the frequency of *A* and *0* are very close.

There are other important parameters that are essential to the description: the *data count* (count of the data) and the *count of the categories*.

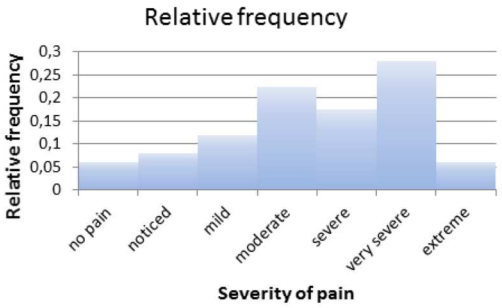
Description of Ordinal Variables I.

Numerical

Frequency table

Severity of pain	Relative frequency	Cumulative relative frequency
no pain	0,06	0,06
noticed	0,08	0,14
mild	0,12	0,26
moderate	0,225	0,485
severe	0,175	0,66
very severe	0,28	0,94
extreme	0,06	1
Σ	1	

Graphical



Indicator:
Mode

Other parameters:
data count (*n*)***, count of categories***

On ordinal scale – for example a severity of pain scale - we can use the same descriptive solutions as in the case of nominal variables: we can use frequency tables, frequency distribution plots and the mode.
But could we give a new indicator that use the advantage of sorting opportunity?
Reminder: in this case there is a meaning to calculate cumulative frequencies. In the pain scale the cumulation shows the (relative) frequency of a given maximal pain.

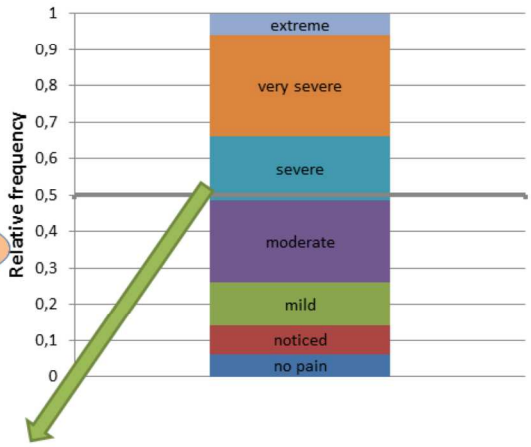
Description of Ordinal Variables II.

Numerical

Frequency table

Severity of pain	Cumulative relative frequency
no pain	0,06
noticed	0,14
mild	0,26
moderate	0,485
severe	0,66
very severe	0,94
extreme	1
Σ	

Graphical



New indicator:
Median: „middle" element(s)
Notation: Me, Med, x_{med}

Based on the ordering ability there is a new indicator that is the *median: the „middle" element(s),* or „middle" point(s) in a sorted dataset. It means that in the sorted dataset 50% of the elements is below of this value and the 50% is over it. In this example the median value is *severe*.
Why I used plural between quotation marks? Could we use the quarter point like middle point? - Later we returns to it...

Description of Quantitative Variables I.

Numerical (analytical)

Frequency tables

frequency distributions (differential discrimination functions)

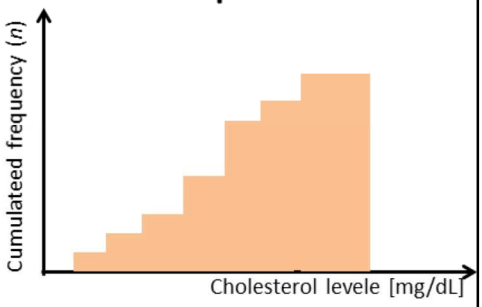
bins (classes, intervals)	(absolute) frequency (FREQUENCY)	relative frequency	(absolute) frequency density	relative frequency density
x ≤ 100	0			
100 < x ≤ 110	0	0	0	0
110 < x ≤ 120	2	0,01	0,2	0,001
120 < x ≤ 130	5	0,025	0,5	0,0025
130 < x ≤ 140	22	0,11	2,2	0,011
140 < x ≤ 150	31	0,155	3,1	0,0155
150 < x ≤ 160	48	0,24	4,8	0,024

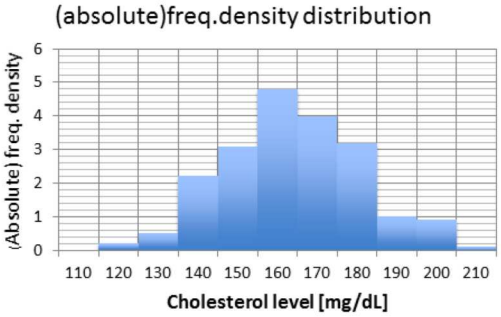
Organizing data – with **loss of information**

Determination of bin width:

- technical and aesthetic concerns
- statistical concerns

Graphical

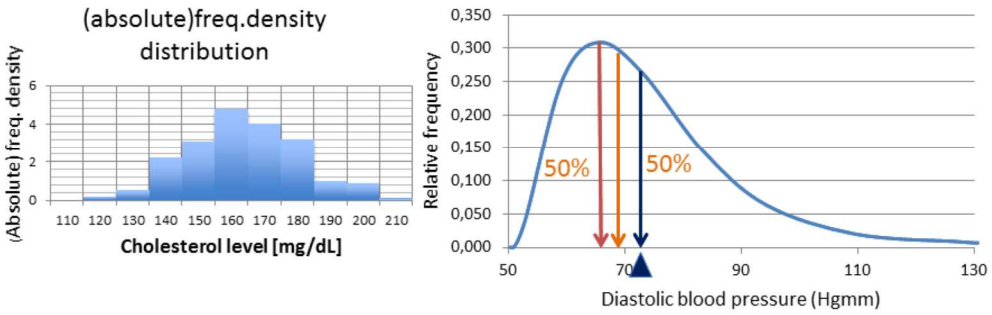




In the next slides we discuss on quantitative (numerical) variables.
In this situation we can preserve all information in case of graphical representation only when we create the cumulated frequency distribution.
Otherwise in the case of quantitative variables to create frequency tables or frequency distribution of the sample we have to *define arbitrary categories* called intervals, bins or classes.
Organizing data in this way *resulted loss of information*. Remark: we don't lose information using cumulated frequency distribution.
How to determine bin width? It is based on *technical and aesthetic concerns together with statistical concerns*.
The statistical determination of the bin width uses often the next formula: bin width = (maximum-minimum)/(square root of data count).
The meaning of the technical and aesthetic concerns are more complicated. For example it has no meaning to use smaller bin width than the smallest measurable difference or use non integer bin width if our measurand is integer. We like to see series like integers or 0,5,10,15 or 10,20,30...
Summarizing the best way to determine the bin width is first use statistical concerns then round it up based on technical and aesthetic aspects.

Reminder 2. In the case of quantiatative variables the difference between the outcomes can be interpreted also and density functions can be created.

Description of Quantitative Variables II.



„Typical values" – ***central tendencies*** (special ***measures of location***):

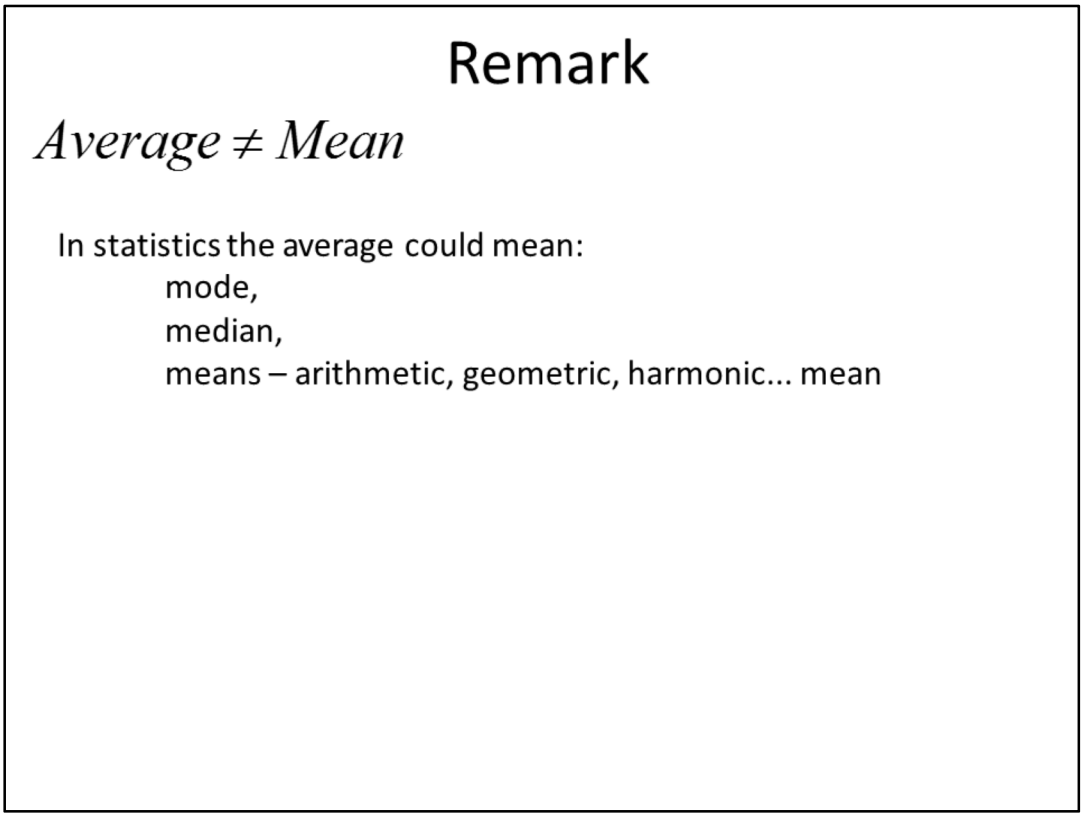
- Mode: most frequent element(s) ?
- Median: „middle" element(s?)
- ***Mean*** (arithmetic mean): „gravity center" , sensitive to „outliers"?

  Notation: x_{mean}, $\bar{x}$

Advantage: compact, ***could be determined from few data***

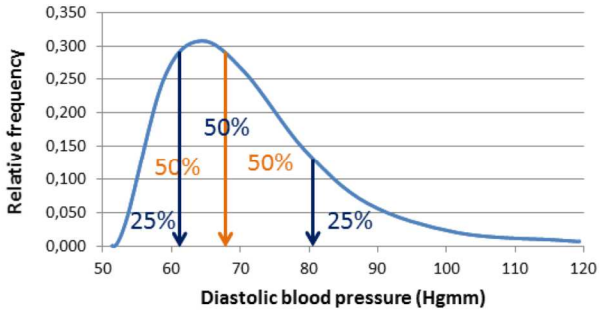
Formulas: in the formula collection...

To describe a quantitative variable we have the same and several new opportunities as before. To „feel the meaning" of the indicators imagine that I could create a frequency distribution with infinitely small bin width. (The measured variable is the diastolic blood pressure of 4 years old boys.)
One kind of „typical values" (indicators) that are very useful called central tendencies that try to describe the center of the distribution. These parameters are special measures of location.
*The mode as the most frequent elemen*t in the dataset belongs to the highest frequency – the peak – in the graph.
The median divide the area under the curve to two same area – 50% of the boys has smaller and 50% of them has higher blood pressure.
The mean is the center of gravity – so if I crop this curve from a paper I could balance this (like a teeter) in the point of the mean value.
The relative position of central tendencies are visible in the graph: in a non symmetric distribution the median and a mean shifts to the tail respectively.
The advantage of the central tendencies against frequency distributions that these parameters could be determined from few data too.
I will return to the question (?) marks later.



Remark 1. In statistics average has different meaning than mean. The average could mean mode, median, means – arithmetic, geometric, harmonic... mean. Use the term *mean* in statistics not the *average*!

Quantiles I.



Other measures of location:

- Median: 50-50% (Q_2)
- ***Quartile***: lower quartile(Q_1): 25-75%; upper quartile (Q_3): 75-25%

General

p-quantile(s): is the number to which the count of data are smaller is maximum *n*p* and to which the count of data are larger is maximum *n*(1 – p)*,

where p is between 0 and 1, and *n* is the count of data

Now we definine other measures of location. Like the median as a midpoint we can determine „quadrant points" that divide the area in 25-75% proportion . This point (value) called *quartile* (from latin quartus that means ¼), more precisely *lower or upper quartile*.

We could generalize it and give a general dividing point (value) called quantile. *p-quantile(s)*: is the number to which the count of data are smaller is maximum n*p and to which the count of data are larger is maximum n*(1 – p), where p is between 0 and 1, and n is the count of data.

Using this general definition we could say that the median is the 0.5-quantile. The lower quartile is the 0.25-quantile, because ¼ = 0.25. It is called first quartile (Q1) also because 1 is divided by four. The upper quartile is the 0.75-quantile, or third quartile, because ¾ of all data is smaller than its value. With the same terminology we could call the median to second quartile (2/4).

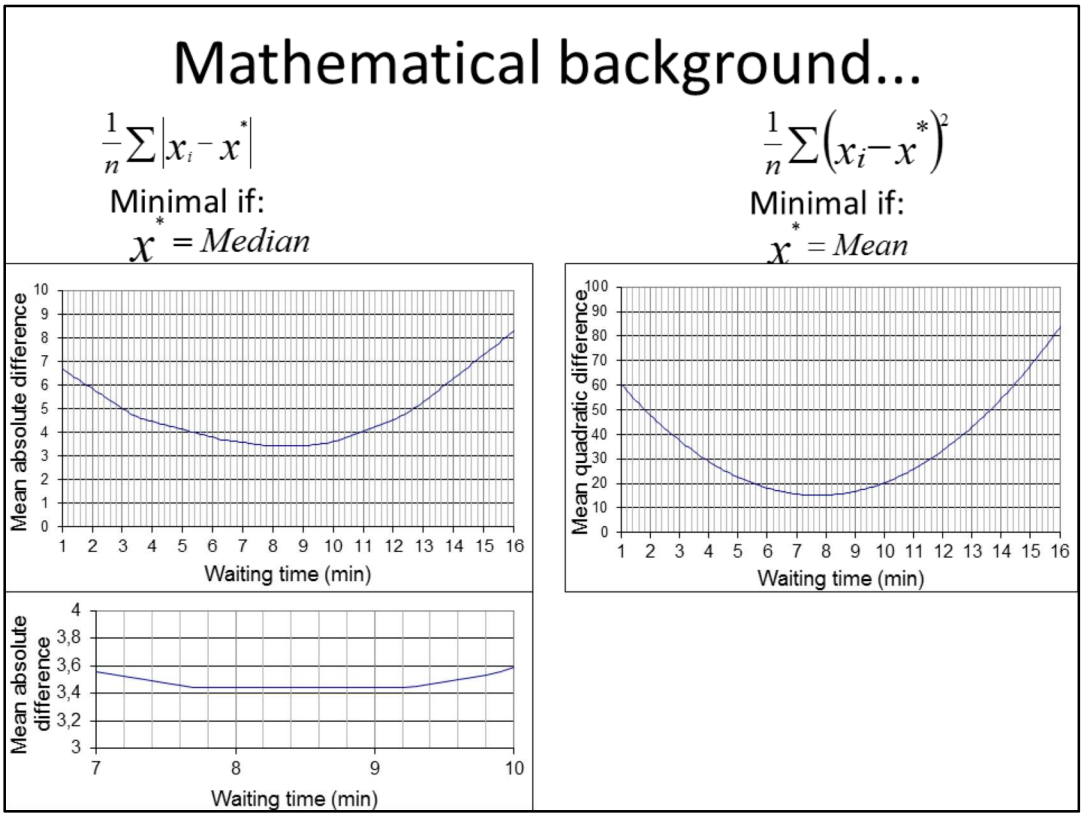
Outliers…

Day	Waiting time (min)		
1	1,27	median	8,48
2	3,3	lower quartile	3,59
3	3,44	mean	7,72
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	12,98		

Day	Waiting time (min)		
1	1,27	median	8,48
2	3,3	lower quartile	3,59
3	3,44	mean	8,31
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	20		

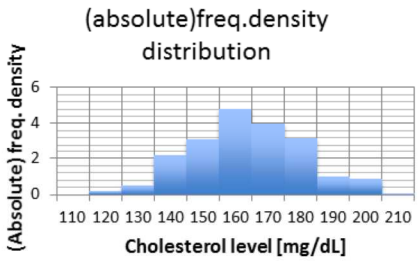
Median, quantiles could differ in theory and practice.
Mean is sensitive to the outliers, but quantiles not (...).
Mode?

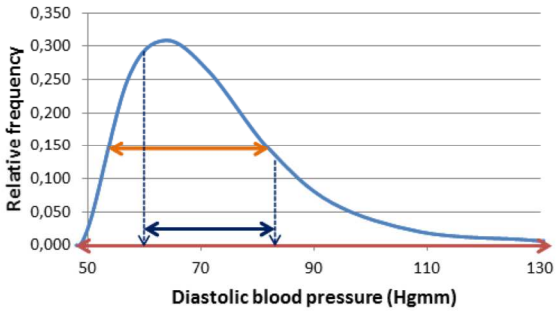
Remark 2. In this slide I try to explain some of the points that mentioned before with ?,(s),”” marks.
Our example dataset is the waiting time in the public transport. In the slide I show you the sorted dataset.
At first about why I used plural for medians, quartiles and quantiles. For the median: 50% (p=0.5) of 12 is 6. It means 6 data is below the median and 6 is over the median. Based on the definition (theory) the median is the value between 7.72 and 9.23 – so all the numbers between! We have the same situation for the quartiles and any kind of quantiles. In practice excel calculate only one number using inverse proportions between the two numbers „in the border”. For example in our dataset the lower quartile (25-75% smaller and larger) is theoretically between 3.44 and 3.65. The difference of this two numbers (the range between them) is 0.2. In practice the quartile value will be 3.44+(0.75*0.2) = 3.59.
Secondly, observe what will happen with the median and the mean if we have an „outlier” – a value that is far form the others (we define it later). In our example I changed the highest value, the 12.98, to 20. We could realize that there is no change in the median while the mean changed greatly. In statistics we say that the mean is sensitive and the median is not sensitive to outliers.
Finally: what about the mode? Is there any, or all of them in our dataset? In the case of a continuous variable it is hard and meaningless to define a mode in the sample. (We may give a range if it is necessary.)



Remark 3. A little math.
The mean of absolute differences from a certain value is minimal for the median. (medians)
The mean of quadratic differences from a certain value is minimal for the mean.

Description of Quantitative Variables III.





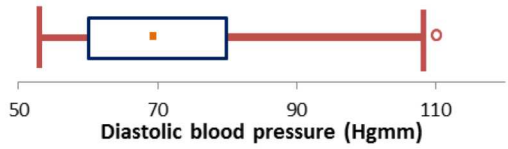
Measures of spread:

- ***Range***: the difference between the maximum and the minimum
- ***Variance (s^2):*** the average of the squared distance from the mean (corrected - sample, uncorrected - population)
- ***Standard deviation (s, sd, SD)***: the square root of the variance the width of the curve
- ***Interquartile range (IQR)*** : the difference between the upper and the lower quartile – not sensitive to the „outliers"

An other kind of indicators are that try to describe the width of the distribution – so the variations within the dataset. These parameters called measures of spread.
One of them is the *range*: the difference between the maximal and minimal values.
The *variance* is the average of the squared difference from the mean. We use the (Bessel) corrected variance if we describe the sample and without correction if we describe our population. The formulas for them are available in the formula collection.
The *standard deviation* is the square root of the variance.
The *interquartile range* is the difference between the upper and lower quartiles.
The variance and the standard deviation is sensitive for outliers while the interquartile range is not.

Description of Quantitative Variables IV.

Graphical: Box plot



Middle point: mean, or *median*
Box: 2*standard deviation, or *interquartile range,* p-quantile range
Whisker: 3*SD, minimum and maximum, 0.05 and 0.95 quantiles, p-quantiles, *1.5*IQR*...
out of whiskers: ***outliers***

Trimmed mean: mean calculated without outliers

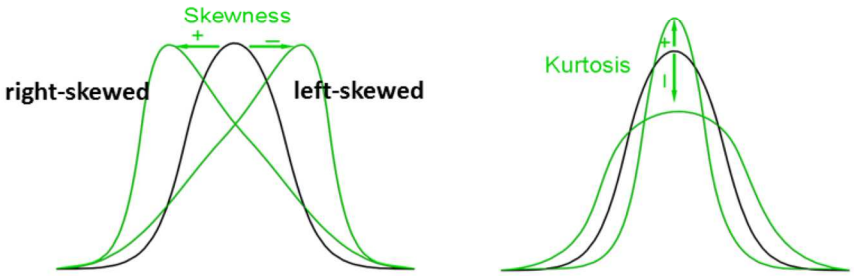
There is a very effective graphical representation of a quantitative variables using the mentioned indicators. That is the *Box plot*, also called Whisker plot.
It consists of a *middle point* that is typically the median, but sometimes the mean. (Now I represented the median.) We may use mean if we have a symmetrical distribution without outliers.
It has a *box* that represent typically the interquartile range, but it could show the standard deviation, standard error too. We use standard deviation or standard error (if we have few data) if we use the mean as a middle point. We use the interquartile range if we have the median as middle (as in the example).
And it has *whiskers*. If the dataset doesn't contains values that are „very different" we could use the minimum and maximum for whiskers. Otherwise we use the multiple times of the SD (usually 2 times) or IQR (1.5 times typically, as in our example) for the mean or median respectively. The 1.5*IQR is commonly called non outlier range.
The outliers are the values that are out of the outlier range.
As you see there is a lot of possibility how we can construct our box plot, I only gave a recommendation that you have to know.
It is important to show what we use in the certain case.
There is an other variant of mean called trimmed mean that is the mean calculated without outliers.

Description of Quantitative Variables V.

Other parameters:

- ***moment***:
  the k-th moment: $\Sigma(x_i)^k/n$
- ***central moment***:
  the k-th central moment: $\Sigma(x_i - \mu)^k/n$

- ***skewness***,
- ***kurtosis*** } ***measures of shape***



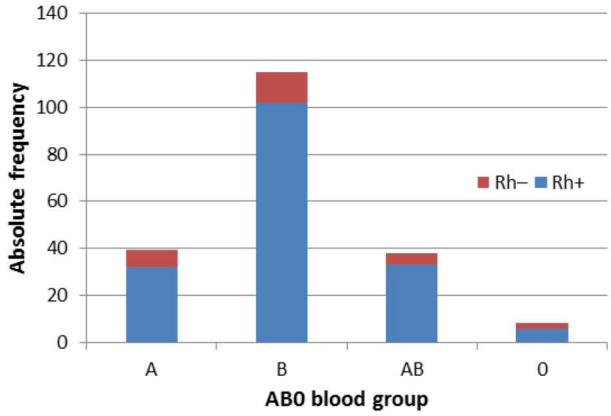
There are other parameters that you have to know how to calculate: *moments* and *central moments*.
The third main indicator group of describing a quantitative variables is the measures of shape. These indicators describe the „position of the bulk of the values". The *skewness* shows the horizontal shift from a symmetrical distribution, while *kurtosis* gives the peakedness of the distribution in the case of a simple monomodal distribution. The direction of the skeweness named based on the tail (part with smaller frequency): if the distribution has a tail on the left side we call it left-skewed. The distribution with a negative kurtosis called platykurtic or platykurtotic (flat peak), while the distribution with positive kurtosis called leptokurtic or leptokurtotic.

Qualitative Bivariate Description

Numerical: ***contingency*** table

	A	B	AB	0	Σ
Rh+	32	102	33	6	173
Rh–	7	13	5	2	27
Σ	39	115	38	8	200

Graphical: ***stacked bar chart***

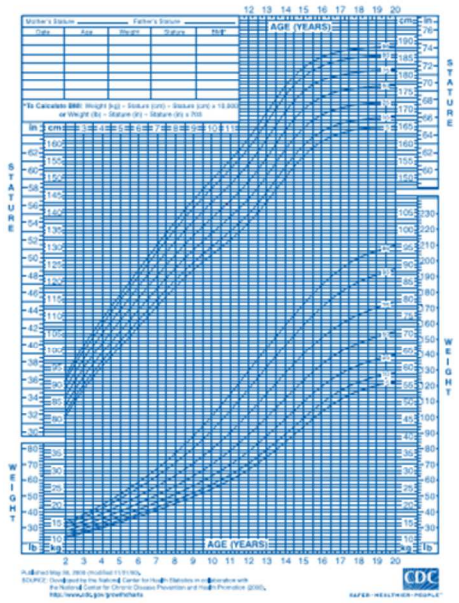


Describing more than one variable together is more difficult.
Now I give only some example for them. To organize two qualitative variables we usually use *contingency tables* – that is a 2 way frequency table. For graphical representation we can use *stacked bar charts*. Here the power of graphs for our mind is very obvious.

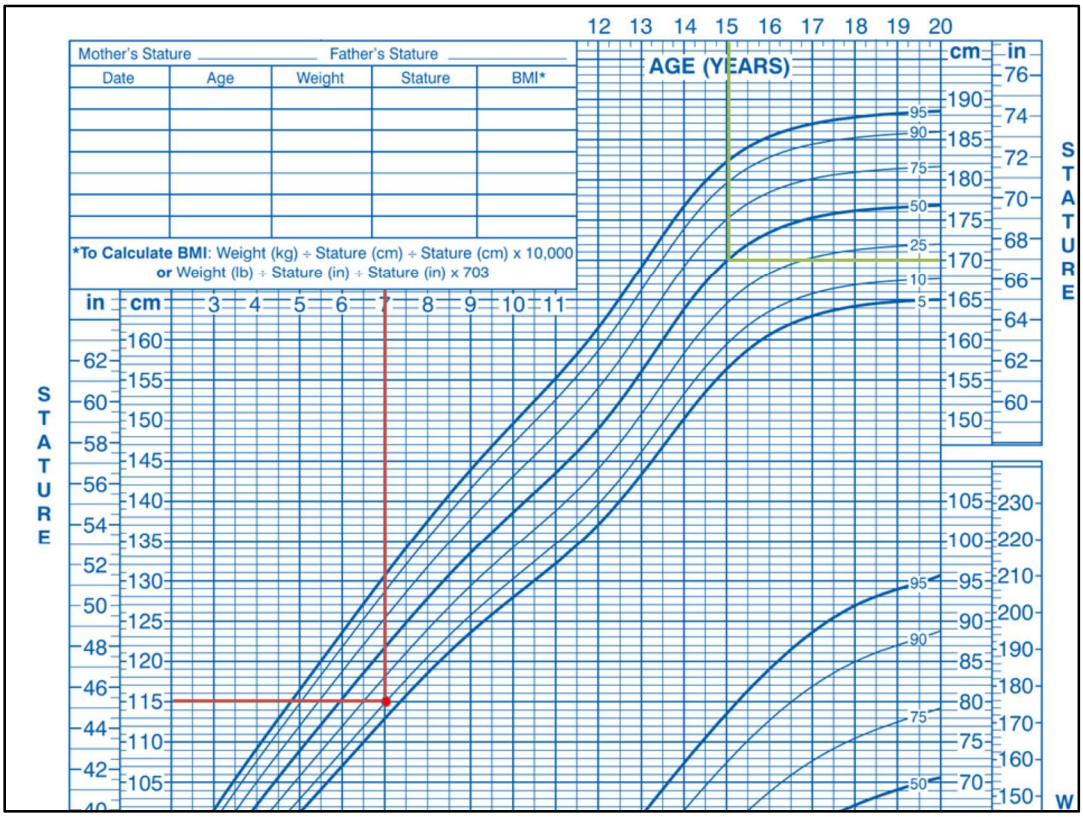
Quantitative Bivariate Description

Graphical: ***percentile curves***

Percentile: quantile expressed as percentage

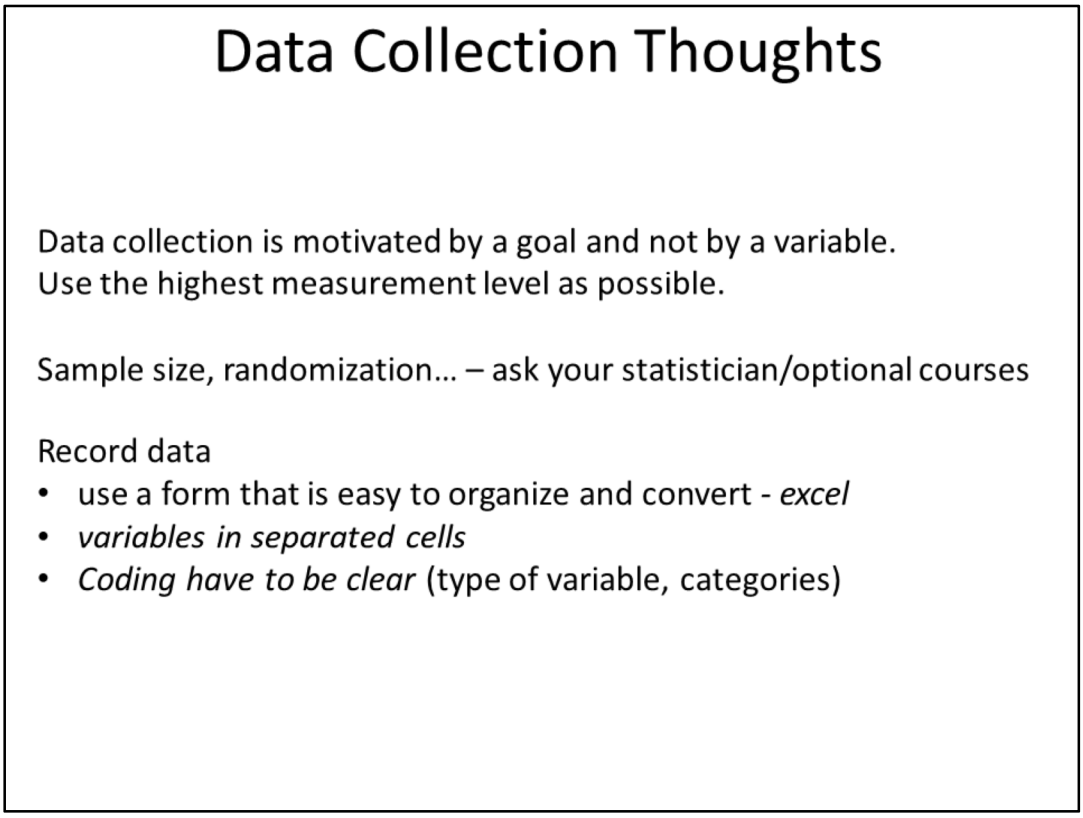


For two quantitative variable we usually use *scatterplots*.
But in medical practice there is an other representation for two quantitative variables called *percentile curves*. These are very common in pediatrics. The percentile is a quantile expressed as percentage.



How to use this curves? What is the meaning of them? What is the meaning of the red point?

In our example the red point – that is in the 10th percentile curve – means that 10% of the 7 years old boys are less than 115 cm.



I would like to highlight some ideas in data collection and data recording that are very important, but sometimes are forgotten.
Data collection (sampling) is motivated by a goal and not by a variable.
Use the highest Stevens' measurement level as possible.

Record data

- use a form that is easy to organize and convert - excel
- separate the variables in different cells
- coding have to be clear (e.g. a nominal variable won't be numerical if you coded with 0 and 1)

There are many other factors to consider in sampling, eg. the sample size, how to randomize, ethics… From these you can learn in advanced level (optional) courses or you have to ask your statistician.

Test Questions #1

- Give the four actions of statistics.
- Give the two part of descriptive statistics.
- Give the two part of inferential statistics.
- Name some ordinal variables, scales.
- Name some discrete numerical variables, scales.
- Name some continuous numerical variables, scales.
- What is the substantial difference between a nominal and an ordinal scale?
- Give example for interval scale.
- What is the substantial difference between an ordinal and an interval scale?
- Give examples for ratio scale.
- What is the substantial difference between an interval and a ratio scale?
- Why is it important to define a statistical variable properly?
- What are the two way as we could describe a variable?
- What are the indicators that we can use to describe a nominal variable?
- What are the indicators that we can use to describe an ordinal variable?
- What are the indicators that we can use to describe a numerical variable?
- Define the mode(s) of a dataset.
- What is the notation of mode?
- Define the median(s) of a dataset.
- What is the notation of median?
- In which type of measurement scale do we loose information usually?
- How we can determine the bin width?
- What is the equation we have to use to determine the bin width?
- What are the central tendencies in case of a numerical variable?
- What is the „meaning" of the mode in a diagram?
- What is the „meaning" of the median in a diagram?
- What is the „meaning" of the mean in a diagram?
- Define the mean of a dataset.
- What is the notation of mean?
- Which central tendency sensitive to outliers?
- What is the advantage of indicators versus distribution functions?
- What is the difference between average and mean?
- What are the measures of location?
- Define the p-quantile.
- Define the lower quartile.
- What is the difference between the second quartile and the median?
- Show how we could calculate the lower quartile of a dataset in theory and in practice.
- What is the value that for the sum of the absolute differences are minimal?
- What is the value that for the sum of the squared differences are minimal?

The following questions may be answered using lecture material, consultation with practice teacher, or your own investigation (on the library or the internet). These test questions are examples for multiple choice items that may occur in the midterm and exam tests.

Test Questions #2

- What are the measures of spread?
- What are the measures of shape?
- Define the variance.
- Define the standard deviation.
- Define the skewness.
- Define the kurtosis.
- Define the interquartile range.
- What is the notation of interquartile range?
- What is a box plot?
- What are the parts of a box plot?
- What we could use as a middle point of a box plot?
- What we could use as a box of a box plot?
- What we could use as a whisker of a box plot?
- What is recommended middle point in a box plot if we have a non symmetrical distribution with outliers?
- What is recommended box boundary in a box plot if we have a non symmetrical distribution with outliers?
- What is recommended box boundary in a box plot if we used a median as a middle point?
- What is the trimmed mean?
- How we define the outlier range commonly?
- What are the moments?
- What are the central moments?
- What is the first central moment?
- What is the first moment?
- What is the second central moment?
- What are the percentiles?
- What we could read out from a percentile curve?