Principles of Biostatistics and Informatics

1st Lecture: An Introduction

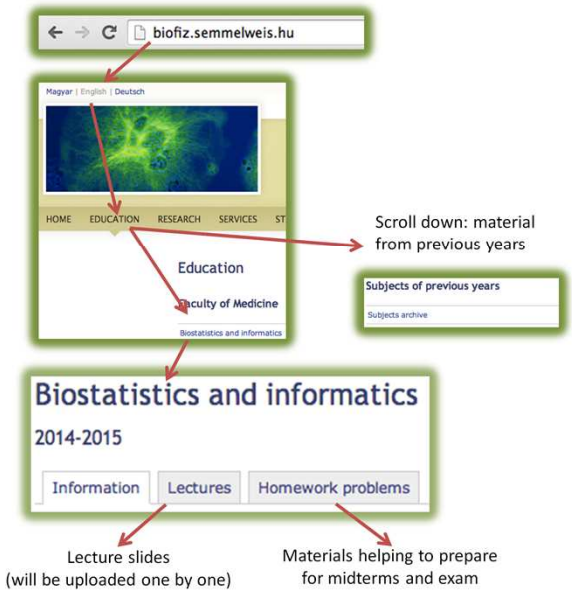
11th September 2014

Gergely AGÓCS

1

How to Get Prepared?

- university = **autonomous learning**
- sources:
  - **your** notes made during lectures *(Thursday 17^{25}–18^{10}; EOK "Szent-Györgyi Albert" lecture hall)*
  - **your** notes made during computer lab classes *(once a week, 90 minutes, 1st floor in the EOK building, computer labs from corridor "B")*
  - consultations (Tuesdays and Thursdays, 18^{30}–20^{00}; *1st floor in the EOK building, computer labs from corridor "B")*
  - "Lab Manual of Medical Biophysics" lab practice book:
    - Biostatistics chapter *(40 page summary of theory)*
    - Problems chapter *(problems 71–77)*
  - homepage: biofiz.semmelweis.hu
    - subject requirements
    - lecture schedule and slides
    - lab schedule
    - homework problems
    - matrial from previous years

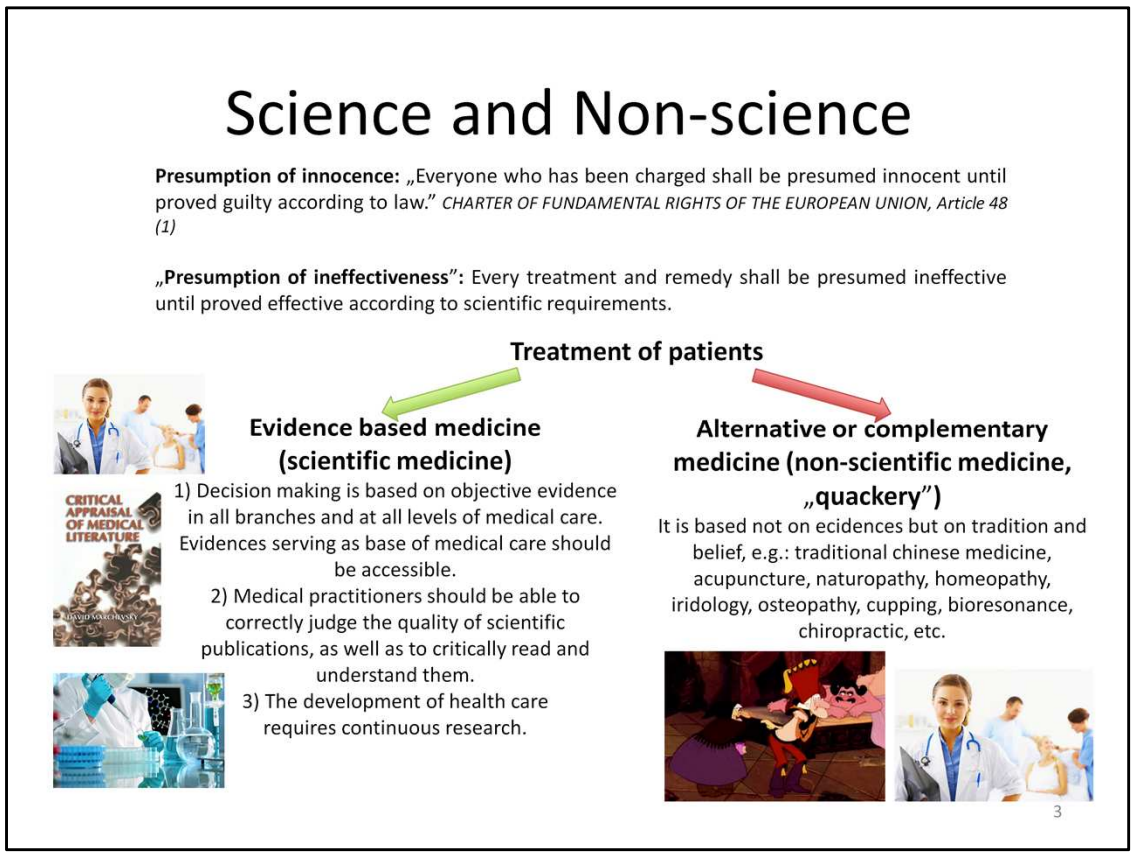


2

Under each slide I make a summary of important things to know. This can be supplemented by your own notes as well as information from your statistics lab teacher.

The most important sources of information are your own notes made at lectures and at computer lab practices. Additionally, lecture slides can also be used, but these are rather an aid for demonstration NOT a substitute of your own notes. I, however, attach these scripts as a help for the first lecture. Your other written source is the "Biostatistics" chapter of the Biophysics Lab Manuals, which is a 40-pages summary of the theory of statistics. You can also find some calculation problems in the "Problems" chapter. These calculation problems are, however, too few, extra problems for practice will be listed under the "Homework" at the homepage of the Department.

Lecture schedule, lecture slides (optionally with supplementary materials), practice schedule, homework, and other info can be found at the homepage of the Department as explained above. It can also be useful to look up lecture slides from last year in the subject archives, e.g. when the lab practice covers other topics than the actual lecture.



The basis of medical training at our university is teaching scientific medicine. This requires the understanding of the word "science". A statement may be called scientific if the observation of similar phenomena leads independently to the same conclusion. That is, scientific results can be controlled, reproduced, they do not depend on the individual observing the phenomenon or making the conclusion – therefore it is fundamentally different from a mere opinion, belief, or tradition. The aim of science is to discover the reality, the reality that exists independently from us.

If we want to condense the main point of scientific statements in one sentence, we can draw a parallel with law: a statement cannot be considered valid until its trueness is proven.

Historically, medicine was not based on science, partly because of lack of knowledge on the side of medical practitioners, partly because of defenselessness and lack of information on the side of patients. Although scientific, i.e. evidence based medicine became the standard in the world by now, practices without scientific foundations still receive a significant role. Some of these practices were the subject of extensive studies, which could not produce any evidence of efficiency. To highlight how disputed the situation is, two examples are shown:

1) One of them is the "German Acupuncture Studies" (GERAC-Studien) – one of the most extensive studies on the efficiency of acupuncture involving several thousands of patients. The efficiency of acupuncture was investigated in 4 different conditions but it could not be shown in any of them. Nevertheless, in two conditions acupuncture

became one of the treatments covered by the German public health insurance.

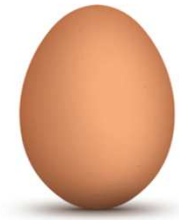
2) The other one is a publication that appeared in the scientific journal Nature about an experiment on “water memory” carried out by Jacques Benveniste (the “water memory” theory is one of the key concepts in the [pseudo]scientific explanation of the way of working of homeopathy). The experiment could not be reproduced by anyone else, not even in the laboratory of Benveniste. Nevertheless, abusing the superficially informed public, these experiments are actually cited to support (!) homeopathy by its proponents.

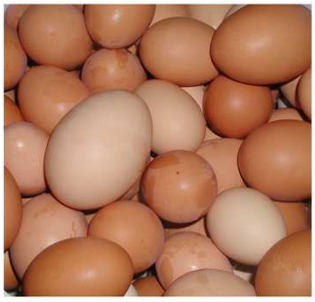
Therefore, we have to understand the importance of scientific approach at the first time we encounter medical sciences. In contrast with medical practices based on belief or traditions, scientific medicine uses methods supported by scientific evidence. To gather the necessary scientific knowledge it is essential to be able to critically read scientific literature. Moreover, the expansion of the range of medical treatments requires constant research for scientifically supported methods. Statistics is the tool of creating new scientific knowledge and to understand already existing knowledge. Statistics is fundamentally a mathematical science mostly based on probability calculus and logic.

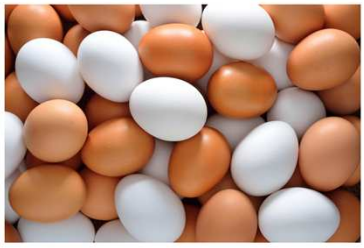
How Does Statistics Help Us?

Statistics deals with the collection, organization, analysis of data, and drawing conclusions

Descriptive statistics → Inferential statistics







4

Statistics is, again, the tool of creating and understanding science. However, scientific questions may be very complex, so it is necessary to simplify them, to break down them into elementary questions. These elementary questions usually point to the DESCRIPTION and COMPARISON of things.

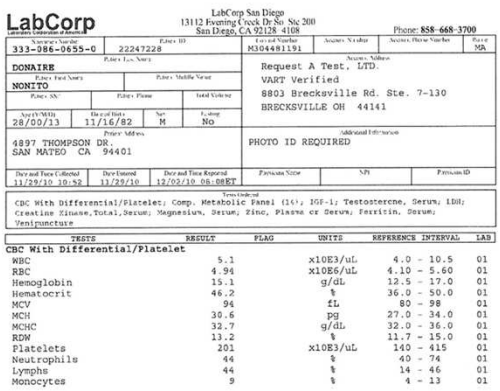
Let us consider for example an egg. We can determine the mass, length, chemical composition of an egg. But knowing a single egg does not help us to make GENERAL statements on eggs, we need to examine more of them for that. On the image in the middle, there are many eggs, which are quite similar but differences e.g. in size are also clearly visible. In the image on the right we can again see a lot of eggs – the egg on the left could also come from here –, which are also similar but have different colors.

So we can see that it generally does not make much sense to study a single object (a single value), if we decide, however, to work with many data then we need to accept that it will show some degree of variation or fluctuation. It will be the common traits and the degree of tolerated deviations that will help us to create the general (abstract, theoretical) concept of “egg”.

Statistics deals exactly with the quantification of collective characteristics of many things and the comparison of these quantified data. Consequently statistical statements always refer to not a single thing but to a group of things.

What Type of Data do We Deal with?

Data to be processed show a high degree of variation …



LabCorp

LabCorp San Diego
13112 Evening Creek Dr So. Ste 200
San Diego, CA 92128-4108
Phone: 858-668-3700

Acct #	Patient ID	Control Number	Account Number	Account Phone Number	Page
333-986-0655-0	22247228	M304481191			NA

Patient Last Name: DONAIRE
Patient First Name: NONITO
Patient Middle Name:
Patient SS#:
Patient Phone:
Total Volume:
Age(Y/M/D): 28/09/13
Date of Birth: 11/16/82
Sex: M
Fasting: No
Patient Address: 4897 THOMPSON DR. SAN MATEO CA 94401

Account Address: Request A Test, LTD. VART Verified 8803 Brecksville Rd. Ste. 7-130 BRECKSVILLE OH 44141

Additional Information: PHOTO ID REQUIRED

Date and Time Collected	Date Entered	Date and Time Reported	Physician Name	NPI	Physician ID
11/29/10 10:52	11/29/10	12/02/10 06:08ET			

Tests Ordered: CBC With Differential/Platelet; Comp. Metabolic Panel (14); IGF-1; Testosterone, Serum; LDH; Creatine Kinase,Total,Serum; Magnesium, Serum; Zinc, Plasma or Serum; Ferritin, Serum; Venipuncture

TESTS	RESULT	FLAG	UNITS	REFERENCE INTERVAL	LAB
CBC With Differential/Platelet					
WBC	5.1		x10E3/uL	4.0 - 10.5	01
RBC	4.94		x10E6/uL	4.10 - 5.60	01
Hemoglobin	15.1		g/dL	12.5 - 17.0	01
Hematocrit	46.2		%	36.0 - 50.0	01
MCV	94		fL	80 - 98	01
MCH	30.6		pg	27.0 - 34.0	01
MCHC	32.7		g/dL	32.0 - 36.0	01
RDW	13.2		%	11.7 - 15.0	01
Platelets	201		x10E3/uL	140 - 415	01
Neutrophils	44		%	40 - 74	01
Lymphs	44		%	14 - 46	01
Monocytes	9		%	4 - 13	01

The physicist measures …	The physician measures …	The medical student measures …
length	height	diameter of red blood cells (2)
frequency	heart rate	pulse frequency (22)
concentration	blood sugar level	protein conc. in blood plasma (4)
voltage	ECG-signal	ECG-signal (27)
sound intensity	hearing threshold	hearing threshold (25)
electric impedance	impedance-plethysmograph (volume)	skin impedance (24)
pressure	blood pressure	–
speed	speed of blood flow	–

5

A physician does not deal with eggs but data collected from patients but the process is pretty much the same. Let us have an overview of data that can be found e.g. on a lab report: it contains patient personal data, health insurance number, name of physician, concentration of different substances dissolved in the blood, count of different formed elements of the blood, pH, blood group, etc. It is clear at first glance that data are very diverse, e.g. some are expressed with numbers, others with text.

The table on the right allows us to compare some of a physician's everyday tasks with what a physicist measures and what a medical student measures during biophysics lab practices. E.g. the physicist measures length in general, while a physician or student measures a more specific length to answer an exact question in their field of interest.

Statistics: Basic Concepts

Example: rolling a die

- **phenomenon**: *the die is rolled*
- **experiment**: *we count the number of dots on top when the die comes to a halt*
- example for an **outcome** (or **elementary event**): *number one is seen on top*
- **sample space**: *it consists of the following outcomes: {1; 2; 3; 4; 5; 6}*
- example for a (non elementary) **event**: *the number on top is even*
- example for a **sample**: *the die is rolled five times, the outcomes are: {4; 2; 2; 5; 6}*
- **population**: *infinite in this case for infinite number of rolls is possible*
- **statistical variable**: *the numbers on top in each experiment*
- example for a **model**: *the different values of the statistical variable occur with the same frequency*

6

Statistics like every science has its own vocabulary – the most important elements of which have to be learned for these indicate the most important concepts.

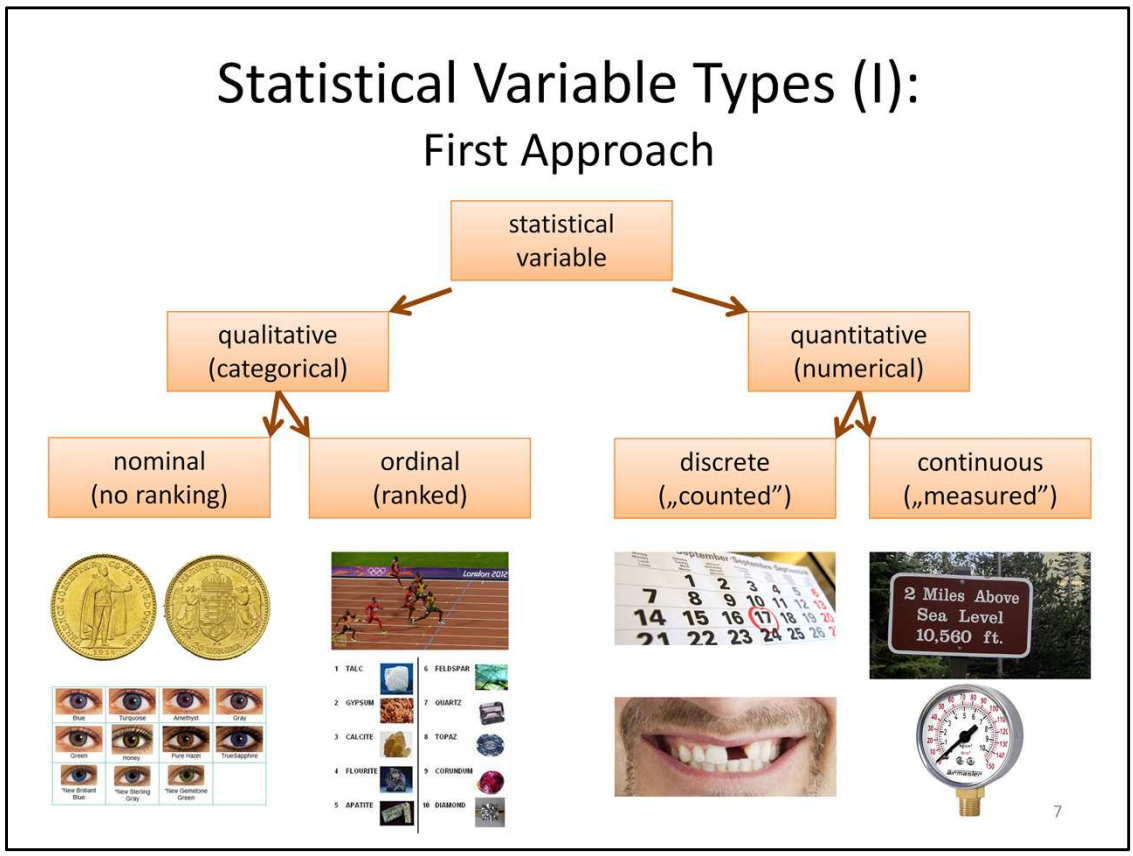
What do we investigate with statistics?
data (singular: datum): a fact helping to know someone or something, may be qualitative or quantitative
signals: carriers of data

In the slide I am using the die rolling as an example to explain the practical meaning of the different concepts. Below one can find the definitions (as precise as necessary for us):

phenomenon: everything that may reoccur among practically similar circumstances, that can be observed, or that can be subject to „experiments" (e.g. a rolling die is a phenomenon, but a historical event is not for it is not reproducible. Remember: statistics deals with mass phenomena, reproducible examinations)
experiment (observation, measurement): an activity to gain data. In case of a phenomenon more than one kind of observation is possible leading to different kind of data (e.g. when rolling a die: the number on top when the die halts or the duration of rolling).
outcome (simple or elementary event): an event with a single result of an experiment (a subset of the sample space with a single element, it may also be the subset of multiple events). In the stricter sense, we may call an outcomes elementary events only if the sample space is finite.

event: a statement which either occurs during the measurement or not. It is a subset of the sample space containing multiple outcomes. An event can be looked upon as an arbitrary set of outcomes.
sample space: it is the set of all possible outcomes of a given experiment, a universal set. It is all the possible outcomes collectively.
statistical variable: a variable the value of which is the event (or outcome) observed during a series of experiments involving a certain phenomenon. In case of rolling a die, the statistical variable may have the values 1, 2, 3, 4, 5, and 6. Chance has a leading role in creating the actual value of the variable.
population: a set containing the outcomes of all theoretically possible experiments. It can be finite or infinite.
sample: a subset of the population, the set of outcomes of experiments that were actually carried out.

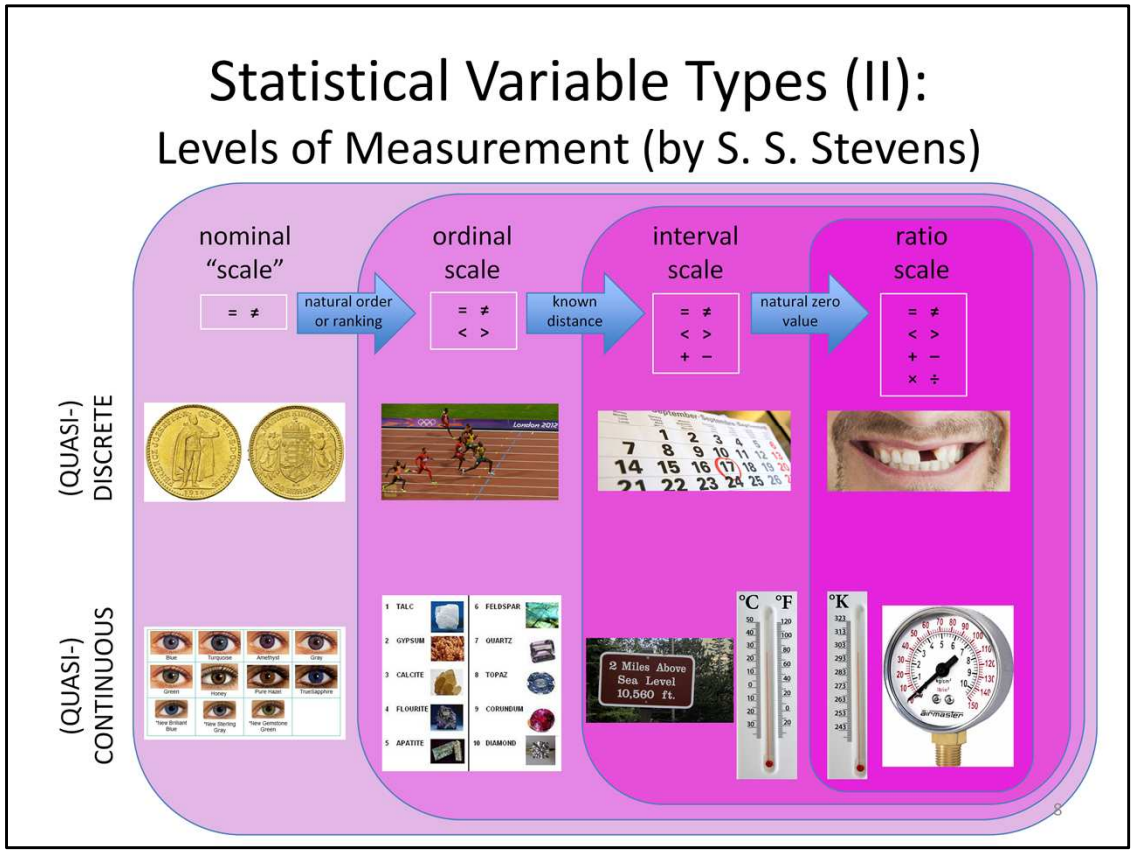
model: a system used to describe reality, which is simplified (i.e. it considers only a part of reality) and abstarct (i.e. it does not describe specific things but their general or common properties); e.g. ideal gas, absolute black body, mass point, species
mathematical model: a mathematical description of the rationale of a model



There are many ways to group statistical variables depending on our aim. As a first approach, we can group variables into the qualitative (categorical) and quantitative (numerical) types. It is characteristics of qualitative variables that we set up categories (these can be either "natural" like the outcomes of a coin toss, or rather arbitrary like hair or eye color, where shades cannot be exactly distinguished), and the values taken up by the variable are assignd to these categories. The categories may either have equal rank, in this case the variable is considered nominal (e.g. blood group, coin toss outcomes), or nonequal, in this case they are called ordinal, and the values of the variable can be set in order (indicated by an ordinal number).

Quantitative or numerical variables may take either discrete values (usually in case of counting, e.g. number of teeth) or continuous values (usually when measured, e.g. blood preassure or body heigth)

The characterisation of different variables, the possible mathematical operations and ways of graphical representation also depend on the type of the variable.



The previous intuitive classification could be fine tuned depending on what we want to do with the variables. Stanley Smith Stevens published his article titled "On the Theory of Scales of Measurement" in Science in 1946; here he described his system on the different types of measurement scales and the hierarchy among them. Although – as it is normal in the scientific scene – his theory was contested by some, it gives a clear guide on the measurement scales used in practice and the mathematical operations that can be carried out on them.

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 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,

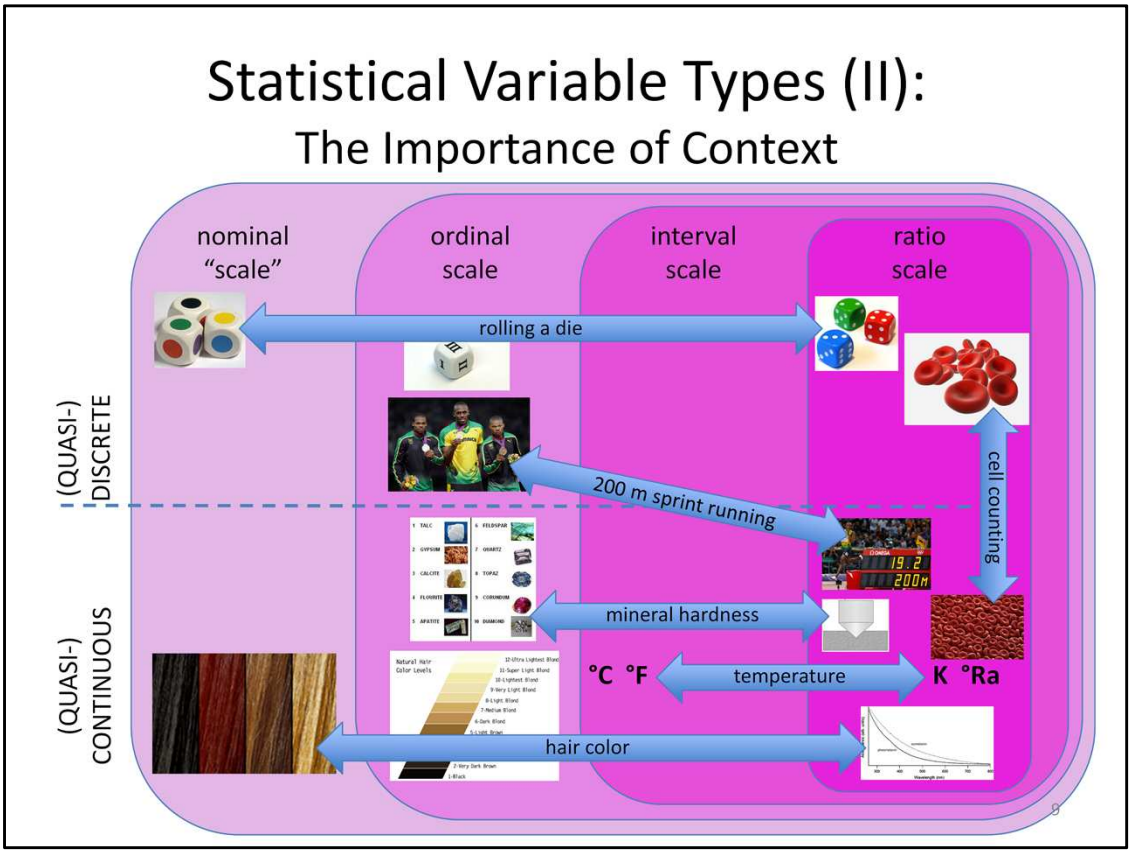
examples are school notes, severity grade of diseases or injuries, or the Mohs scale of mineral hardness. Consequently, on a nominal scale not only identity can be defined, but “less than”/”greater than” relationships as well. Scale elements are usually denoted by ordinal numbers, which has to be kept 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, temporal duration etc.

It is possible to distinguish between more or less discrete and continuous variables at all scale levels. In case of nominal categories e.g. coin tossing is discrete but the limit between different eye colors is rather blurred and it only depends on us how many categories are defined.

We will later encounter the name of S. S. Stevens in biophysics, when discussing psychophysical laws in the second semester.



There are many variables which cannot be unambiguously assigned to one of the categories. Others could be, but for historical or practical reasons are not assigned to the category they would belong to according to their nature.

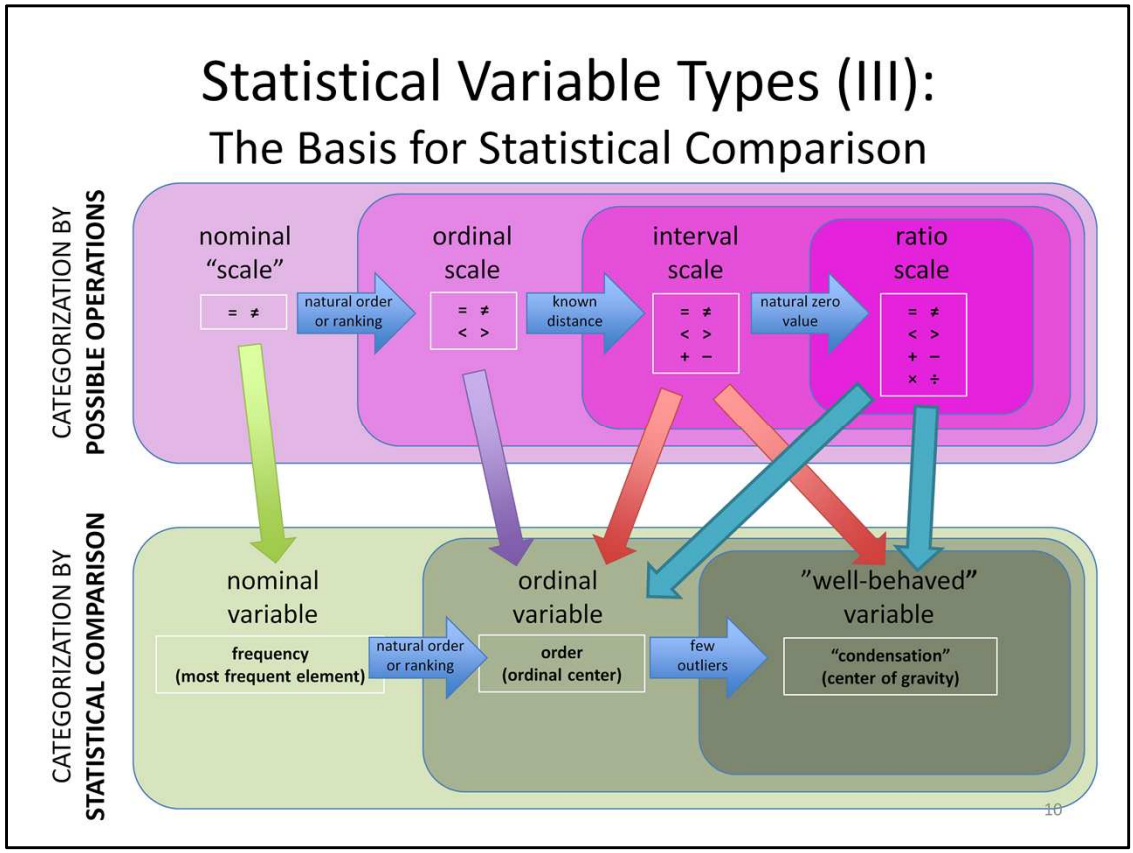
For example, the first temperature scales were constructed before the existence of an absolute zero pint was proven. Therefor, although temperature could be measured on a ratio scale by nature, we use an interval scale for everyday purposes. The height over sea level, on the other hand, is a different case: here a natural zero point does not exist even theoretically. (The Earth is not a perfect sphere, so it does not have for example a well defined "center point"). This also true for all "potential"-type quantities.

Judgment of the outcome of rolling a die as a statistical variable is even more problematic. It is a categorical variable by nature because there is no natural order at all between the faces of a cube as a geometrical body. Using colors to label the faces makes them distinguishable, but this in itself does not create an order. We can assign ordinal numbers to the faces, which would make an ordinal variable. If we draw dots on the faces – which are countable – we will get a variable expressible on an interval scale, moreover, although zero cannot be rolled, in case of counting the dots, the zero point of the scale is also well defined (no dots). So the fact that no zero outcome can be observed during the measurement does not exclude that the scale itself has a – even natural – zero point. Together with these, the range of possible mathematical operations also widens according to the explanation above. But it has to be kept in mind, that the "extra meaning" of the variable raised to a higher level of measurement scale does not reflect the properties of the cube, but that of the rules assigned by us to the cube.

Now consider hair (or eye) color, as a statistical variable. In everyday conversations we usually just say that the hair color of someone is brown, black, blonde or red. If we want to be more precise, we can make the categories finer like: dark brown, chestnut brown, fire red, light brown etc, but this does not influence the level of the measurement scale, it only elucidates the continuous nature of the variable and the arbitrary nature of the categories. However, it is possible to put these categories in order along a dark-light scale, or, using precise measurements, even a color scale may be assigned to hair color. That is, the everyday categorical hair color may be turned into a continuous numerical variable, which corresponds to reality much more precisely.

It is also a practical question whether the red blood cell count is considered a discrete or a continuous variable. If there are only a few cells to be counted, e.g. in an electron micrograph, than it is rather handled as a discrete variable, but in a larger blood sample there can be several millions of blood cells, which makes the variable practically continuous.

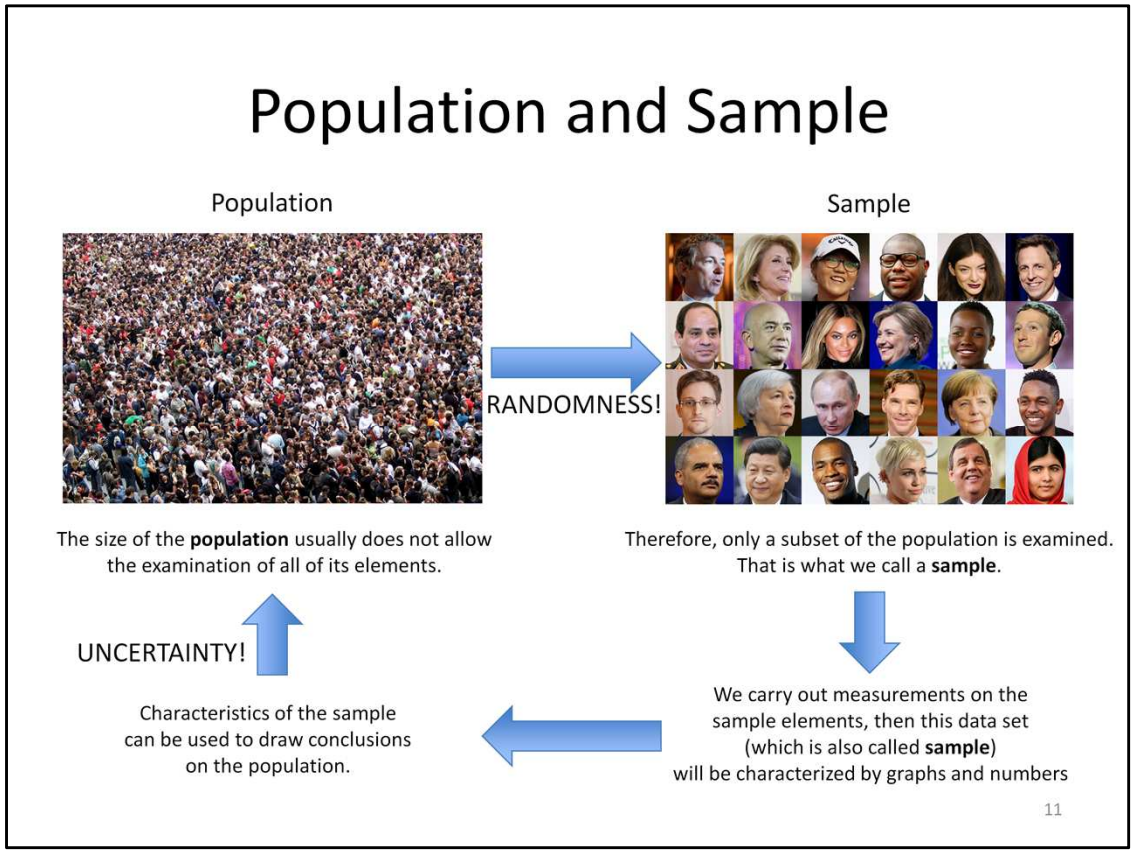
It is widely know that Usain Bolt won the 200 meter run in the 2012 London Olympics, but it is quite sure that only a few remember he actually won it with 19.32 seconds. Although the time elapsed as a continuous quantity measured on a ratio scale carries much more information than the place expressed on an ordinal scale, remembering the latter requires much less effort.



Now let us go back to one of the key task of statistics: the determination of identity and difference. This requires a third type of grouping. All variables may be compared based on the frequency of occurrence of their possible values. For example, the population of two countries may be compared with regard to the percentual occurrence of ABO blood group types. This comparison can be made even simpler if we just heck whether the most frequent blood group is the same or not.

If the elements can be set in order, then a midpoint ("halving point") may be determined such that as many elements are less than it as greater than it. Sets containing elements can be compared based upon their midpoint.

The comparison may be further sophisticated if instead of the midpoint we consider the "center of gravity", that is, from which the number of elements *weighted with distance* is similar. This point (which is named arithmetical mean or simply average, and will be more precisely introduced later) therefore does not only depend on the number of elements with less and greater value but also on *how far* they are, this makes this point much more sensitive than the simpler midpoint. Because of this sensitivity, its use is limited to cases when the variable is, so to say, "well-behaved", that is, there are not too many outliers that would which would distort the information represented by the center of gravity. Therefore those data sets which come from measurements on interval or ratio scales but contain many outliers (that is, the are non "well-behaved") will be handled like ordinal variables.



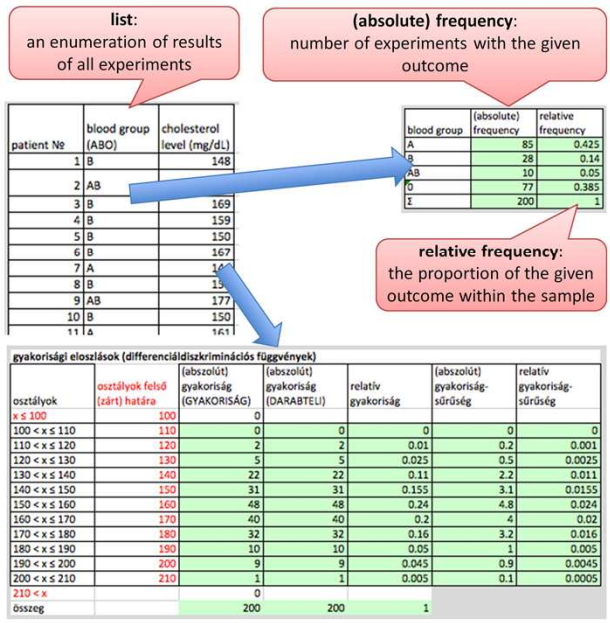
After overviewing the multitude and possible classifications of statistical variables, let us clarify the meaning of population and sample.

As we mentioned before, statistics examines repeatable phenomena. This means that during examination of a phenomenon many, if not infinitely many measurements would be possible. The set containing the outcomes of all these theoretically possible measurements is called **population**. Theoretically, the complete understanding of a statistical variable would require the execution of all the possible measurement, but of course it is not possible. Consequently, we only observe a subset of the population, which is called **sample**. The most evident way of generating this subset is **random selection**.

We carry out measurements on the sample, the set of measurement results is also called **sample**. (That is: in less precise way the sample may be a group of students [individuals, objects] of the university as a population. In more precise way, the population is the height of all people at the university, the sample is the set of height values for a group that was actually measured.) The sample might be characterized graphically or numerically, then the properties learned that way may be extrapolated to the population. E.g. if 25% of people in a group have blood type “A”, we may expect the same from the whole population. Since the sample is chosen randomly, it will not necessary represent the population, the frequency of occurrence of different values within the population perfectly. As a result, every conclusion drawn from a sample carries a burden of **uncertainty**.

Representation of a Sample (I)

1) *A simple list of sample elements*
2) *Summary of frequencies in tables*
   - absolute frequency (Δn) and relative frequency ($\Delta n/n$)
   - categories are evident for qualitative variables [Excel: =COUNTIF() function]
   - categories (bins) are created arbitrarily for quantitative variables [Excel: =FREQUENCY() function can also be used]
   - frequency density ($\Delta n/\Delta x$) and relative frequency density ($[\Delta n/n]/\Delta x$)



patient №	blood group (ABO)	cholesterol level (mg/dL)
1	B	148
2	AB	
3	B	169
4	B	159
5	B	150
6	B	167
7	A	14
8	B	15
9	AB	177
10	B	150
11	A	161

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

gyakorisági eloszlások (differenciáldiszkriminációs függvények)

osztályok	osztályok felső (zárt) határa	(abszolút) gyakoriság (GYAKORISÁG)	(abszolút) gyakoriság (DARABTELI)	relatív gyakoriság	(abszolút) gyakoriság-sűrűség	relatív gyakoriság-sűrűség
x ≤ 100	100	0				
100 < x ≤ 110	110	0	0	0	0	0
110 < x ≤ 120	120	2	2	0.01	0.2	0.001
120 < x ≤ 130	130	5	5	0.025	0.5	0.0025
130 < x ≤ 140	140	22	22	0.11	2.2	0.011
140 < x ≤ 150	150	31	31	0.155	3.1	0.0155
150 < x ≤ 160	160	48	48	0.24	4.8	0.024
160 < x ≤ 170	170	40	40	0.2	4	0.02
170 < x ≤ 180	180	32	32	0.16	3.2	0.016
180 < x ≤ 190	190	10	10	0.05	1	0.005
190 < x ≤ 200	200	9	9	0.045	0.9	0.0045
200 < x ≤ 210	210	1	1	0.005	0.1	0.0005
210 < x		0				
összeg		200	200	1		

12

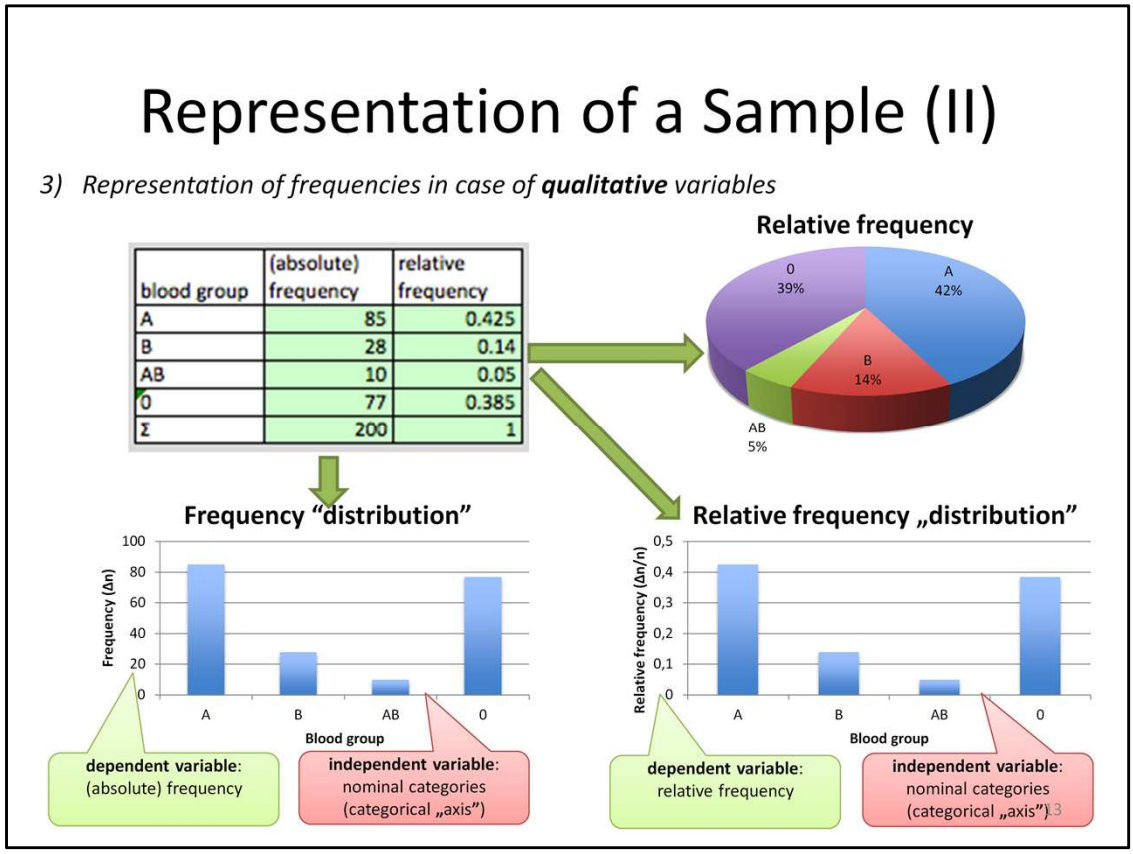
The easiest way of representing a sample is to list its elements (in the order of measurement). This sort of representation does not show much more than the existence of of sample elements. We use this “method” to store the elements of even big samples, but it is not fit for representation.

It is more demonstrative to sort the elements into categories, then list the frequency of occurrences for each category. Incase of qualitative variables, the categories are self-evident (the possible outcomes) although merging of categories may be necessary (if, for examples, there would be too few elements in a category). In case of numerical data we need to make arbitrary categories ourselves, these are called intervals or bins. There are practical aspects of setting up bins, which are related to sample size and range of values (see later). In case of greater samples more bins may be set up with narrower width. But keep in mind: setting up interval delimiters will influence the frequencies.

The number of elements in a category or bin is called absolute frequency or simply frequency. To make comparison of differently sized samples is easier, if we express the relative frequencies (i.e. the percentual proportion within the sample). That way it is easy to compare the distribution of blood groups within the populace of a town and that of a country.

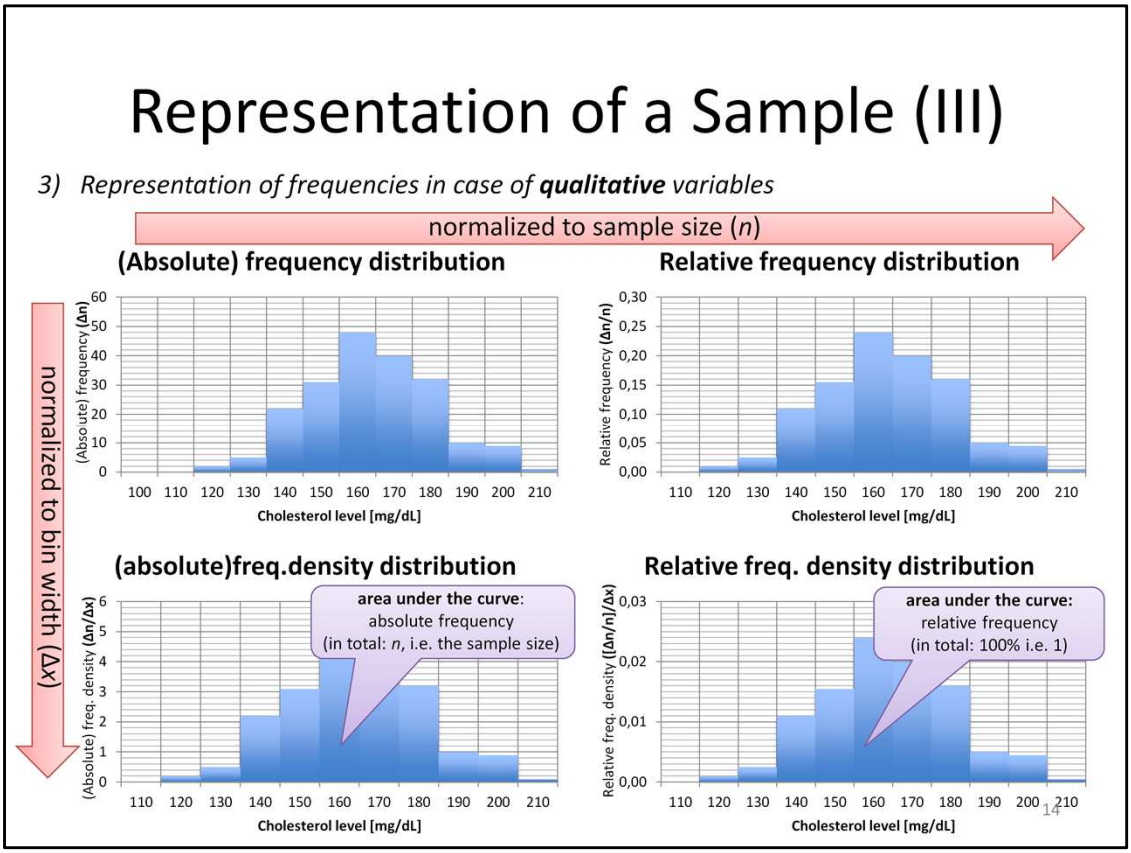
In case of numerical data not only the sample size my differ significantly but the delimiters separating intervals as well. Obviously, using wider bins will change frequency values even for the same sample. To make comparison simpler, we can relate frequencies or relative frequencies to bin width (i.e. divide by bin width). The quantities

we get that way are [absolute] frequency density and relative frequency density.



Qualitative data may be represented by bar charts and pie charts. Bar charts appear similar to the coordinate system used to represent functions between numerical data, but it is important that the horizontal "axis" shows a categorical (in the above example, nominal) variable. A nominal "scale" – according to the reasons explained before – there is no natural order of variables, and the distance is also unknown between them. The order is set up only based on practical considerations (e.g. alphabetical order, order by frequency etc.) and aesthetical considerations. The dependent variable is either the absolute or the relative frequency (the latter is often indicated in percent).

The whole pie in a pie chart represents the complete sample, the size of the sectors (slices) represents the different frequencies within the sample. The order of sectors is again set up according to practical reasons without any particular mathematical meaning.



In case of quantitative variables, the x axis represents the interval scale divided into arbitrary categories (bins).

The y axis may represent absolute frequency in which case the height of each column represents the number of elements in the corresponding bin. If relative frequencies are measured on the y axis, then the column height will give the proportion of elements of the bin within the sample.

If absolute frequency is related to the bin width, then we get the absolute frequency density. If this is measured on the y axis, the area of the columns (rectangles) will correspond to the number of elements in the given bin. The total area will, therefore, be equal to the sample size. In case of relative frequency density (relative frequency divided by bin width), column areas correspond to proportion within the sample, while the total area is equal to 1 (= 100%).

Note: The word “density” means that we relate to some unit of length, area, or volume. A three dimensional example is mass density, that is, mass per unit volume. In two dimension we can mention population density: number of people per surface area; or surface density: mass of a square centimeter paper. In case of frequency density, we essentially relate frequency to the Δx section length (= bin width), so it can be considered a one dimensional density.

Test questions #1

- What are the most important characteristics of science?
- What are the properties of scientific thinking?
- How does scientific medicine differ from quackery?
- How can we make sure that a given medical method is based on scientific evidence?
- Give examples for non-scientific medical methods!
- Who has to carry the burden of proof: Who states that a new method is efficient or who sais it is not?
- What is statistics?
- What is the aim of statistics?
- Which branches of mathematics does statistics rely on?
- What is the aim of model making?
- What is the relationship between a model and reality?
- What should be proven about a medicine or treatment: whether it is efficient or it is inefficient?
- What is the problem with the following statement: “No one has proven that this method is inefficient, therefore it would be unjust to limit its use.”
- How can the variables acquired during the examination of a patient be grouped?
- Name some nominal variables.
- Name some ordinal variables.
- Name some discrete numerical variables.
- Name some continuous numerical variables.
- Name some “well-behaved” statistical variables.
- Name some non-”well-behaved” statistical variables.
- Give examples for nominal “scale”.
- Why is the word “scale” in the term “nominal scale” a misnomer?

15

Tho followin 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

- Give examples for ordinal scale.
- 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?
- Show the importance of the context of the statistical variable with examples.
- How can the statistical variables be grouped from the point of view of comparison?
- What is the relationship between Stevens' levels of measurements and the hierarchy of variable comparison?
- What does it mean, that a variable is "well-behaved"?
- How can the different levels defined in the hierarchy of variable comparison be characterized?
- Is a quantity measured on (Stevens') interval scale always "well-behaved" from the aspect of statistical comparison?
- What is a population?
- What is a sample?
- How do we take a sample from the population?
- Why does the sample not perfectly represent the population?
- What is the reason of the uncertainty burden on conclusions an the populations drawn from the sample?
- What are the ways of characterization of the sample?
- What is the meaning of frequency and absolute frequency?
- If we just say "frequency" what does it refer to: absolute or relative frequency?
- How can the frequency values of a set of data be summarized in a table?

Test questions #3

- What do we need to pay attention to during the the tabular summary of frequencies of numerical variables?
- How can we make frequencies directly comparable, if samples differ in size?
- How can we make frequencies directly comparable, if bins differ in width?
- Why is relative frequency “relative”?
- What does the word “density” refer to in the term frequency density?
- What are the practical ways of graphical representation for qualitative variables?
- What is represented on the horizontal axis of a column chart constructed from qualitative data?
- What is the meaning of the “categorical axis” of a column chart constructed from qualitative data?
- What and how can be read out from the graph of a frequency distribution?
- What and how can be read out from the graph of a relative frequency distribution?
- What and how can be read out from the graph of a frequency density distribution?
- What and how can be read out from the graph of a relative frequency density distribution?
- What is the total area under the curve of a frequency and a relative frequency density distribution?