Principles of Biostatistics and Informatics

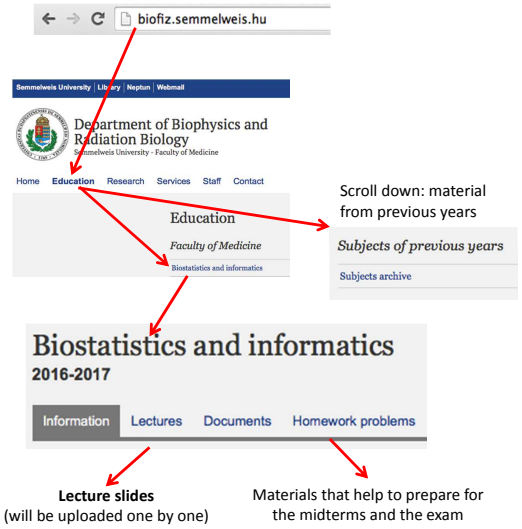
Lecture 1: An Introduction

6th September 2016

Gergely AGÓCS

How to Get Prepared?

- university = **autonomous learning**
- sources:
  - **your notes** made during lectures *(Tuesdays, 14^{00}–14^{45}; 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 in corridor "B")*
  - **consultations** (Wed: 18^{30}–20^{00}; *1st floor in the EOK building, computer labs in corridor "B")*
  - "Medical Biophysics Practices" 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



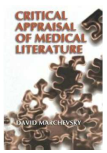
Science and Non-science

Presumption of innocence: „Everyone who has been charged shall be presumed innocent until proved guilty according to law." *CHARTER OF FUNDAMENTAL RIGHTS OF THE EUROPEAN UNION, Article 48 (1)*

„**Presumption of ineffectiveness**": Every treatment and remedy shall be presumed ineffective until proved effective according to scientific requirements.

Treatment of patients

Evidence based medicine (scientific medicine)



1) Decision making is based on objective evidence in all branches and at all levels of medical care. Evidences serving as base of medical care should be accessible.
2) Medical practitioners should be able to correctly judge the quality of scientific publications, as well as to critically read and understand them.

3) The development of health care requires continuous research.

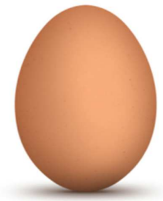
Alternative or complementary medicine (non-scientific medicine, „quackery")

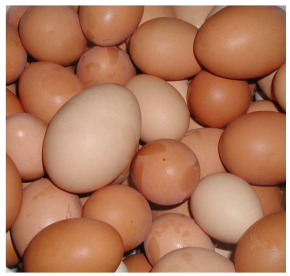
It is based not on evidences but on tradition and belief, e.g.: traditional Chinese medicine, acupuncture, naturopathy, homeopathy, iridology, osteopathy, cupping, bioresonance, chiropractic, etc.

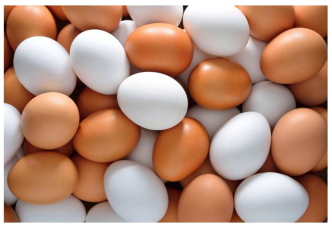
How Does Statistics Help Us?

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

Descriptive statistics → Inferential statistics

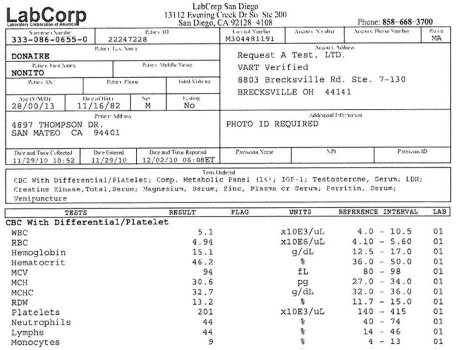






What Type of Data do We Deal with?

Data to be processed show a high degree of variation ...



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

333-086-0655-0 | 22247228 | M304481191 | MA

DONAIRE
NONITO
28/00/13 | 11/16/82 | M | No
4897 THOMPSON DR.
SAN MATEO CA 94401

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

PHOTO ID REQUIRED

11/29/10 10:52 | 11/29/10 | 12/02/10 08:08ET

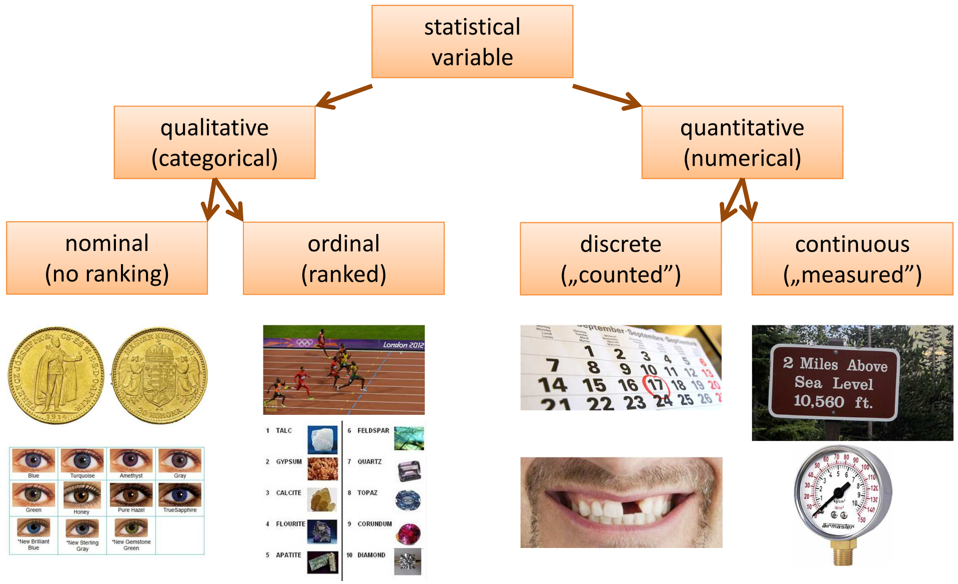
CBC With Differential/Platelet; Comp. Metabolic Panel (14); IGF-1; Testosterone, Serum; LDH; Creatine Kinase,Total,Serum; Magnesium, Serum; Zinc, Plasma or Serum; Ferritin, Serum; Venipuncture

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	–

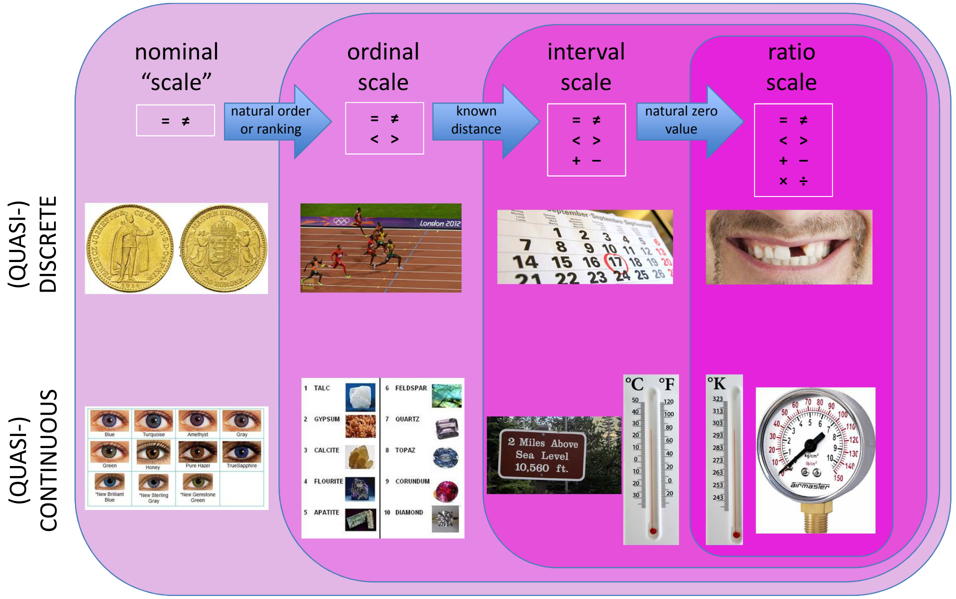
Types of Statistical Variables (I):

First Approach



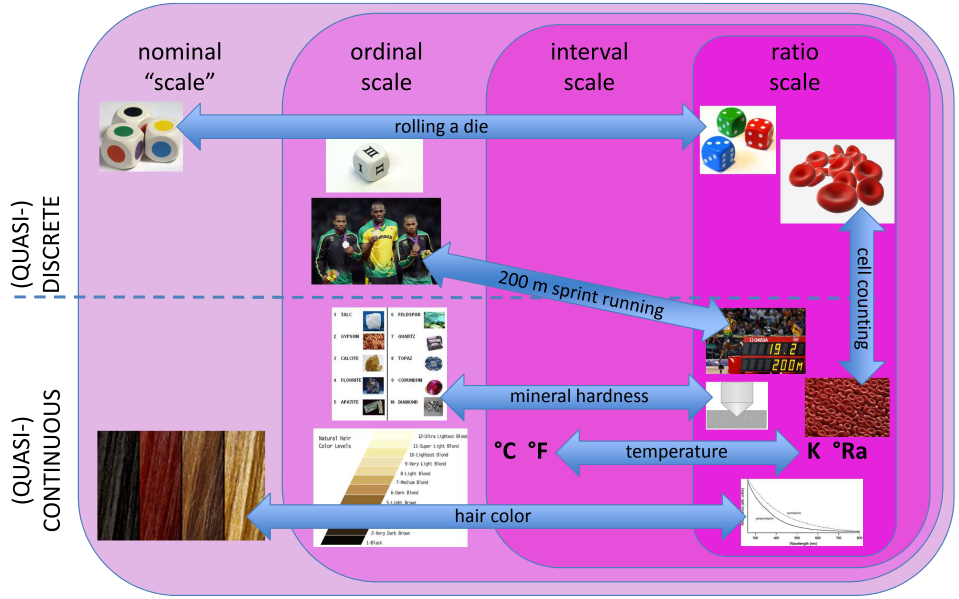
Types of Statistical Variables (II):

Levels of Measurement (by S. S. Stevens)

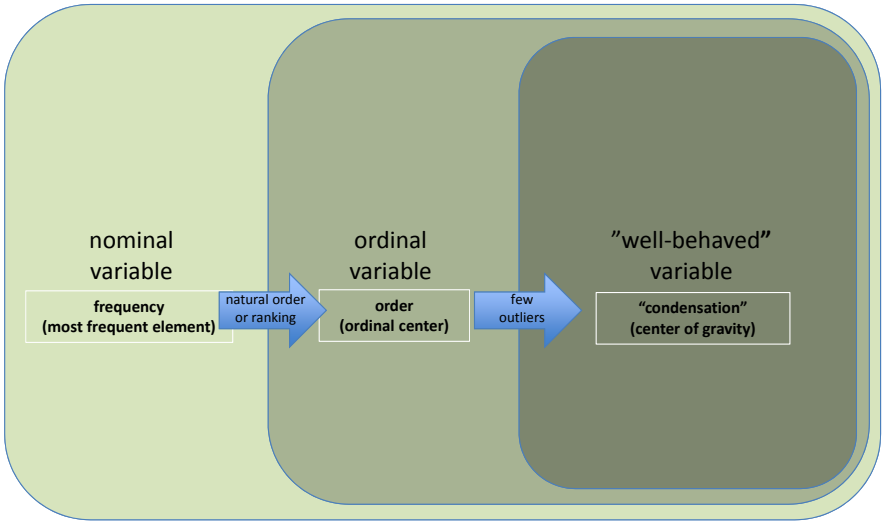


Types of Statistical Variables (II):

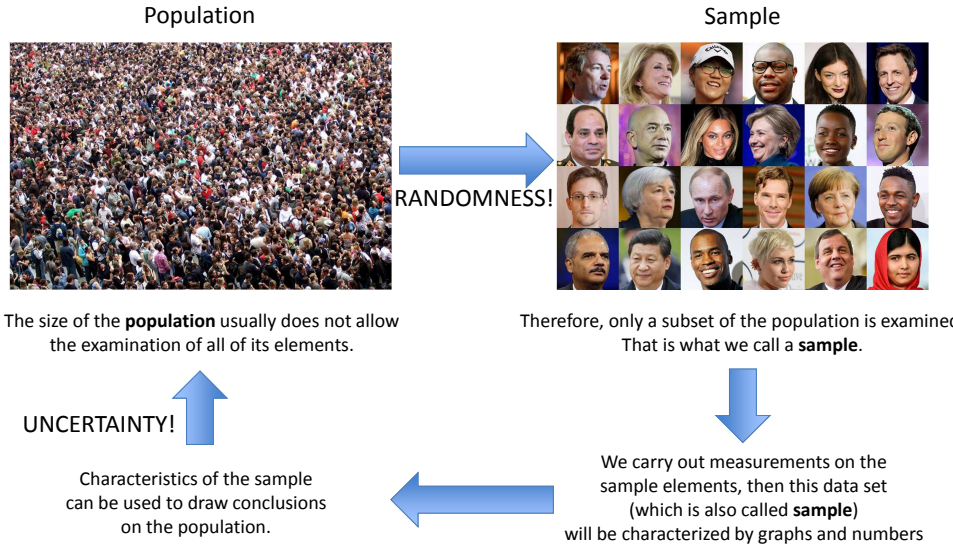
The Importance of the Context



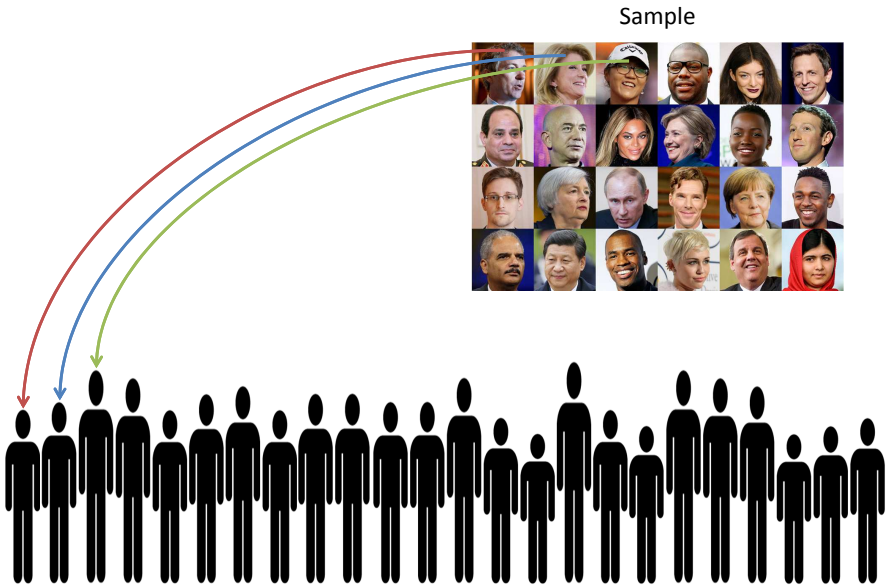
Statistical Variable Types (III):
The Basis for Statistical Comparison



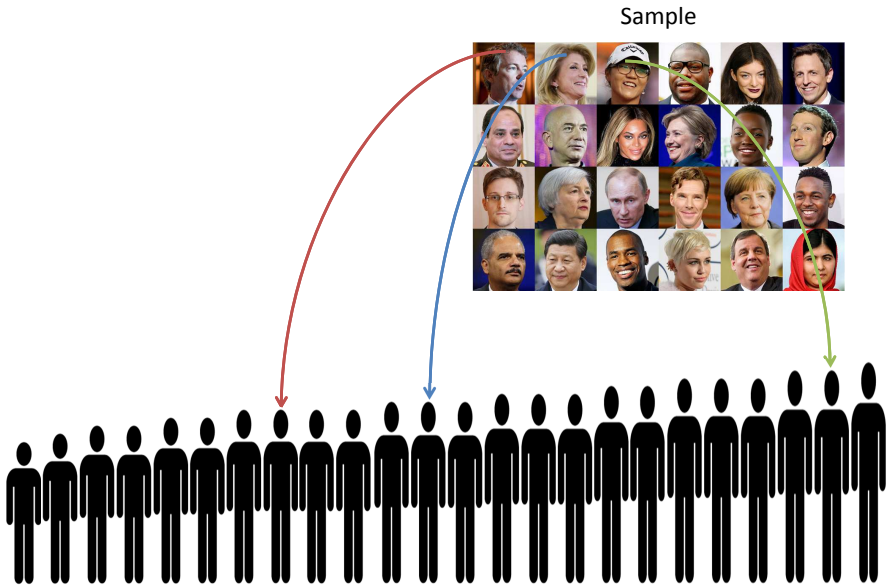
Population and Sample



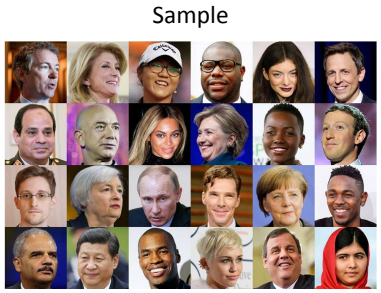
Composition of the Data

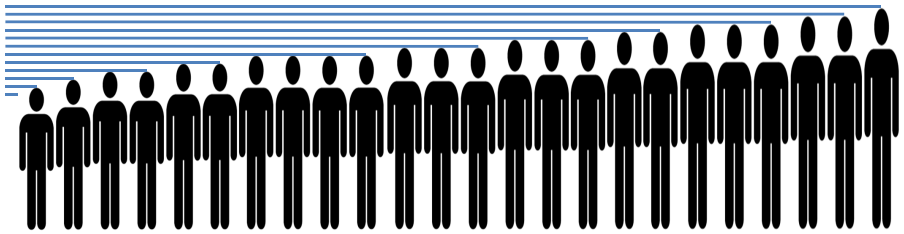


Composition of the Sample



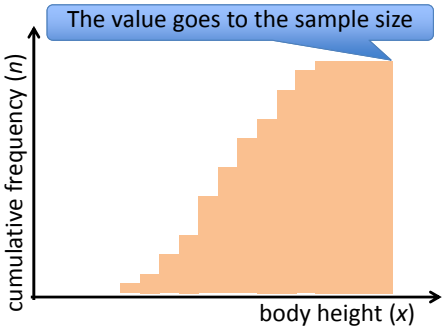
Composition of the Sample

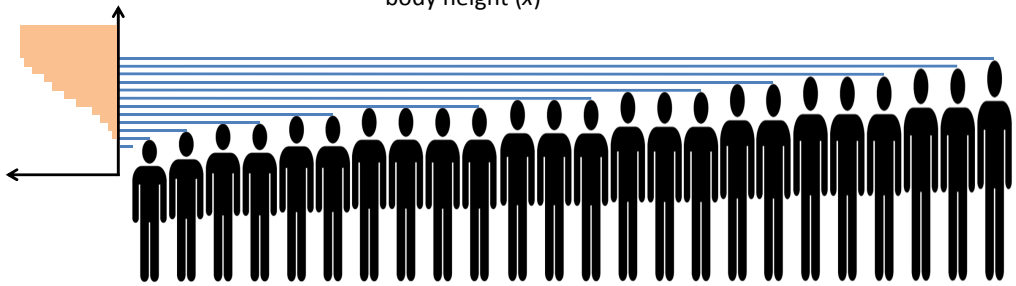




Cumulative Distribution Function

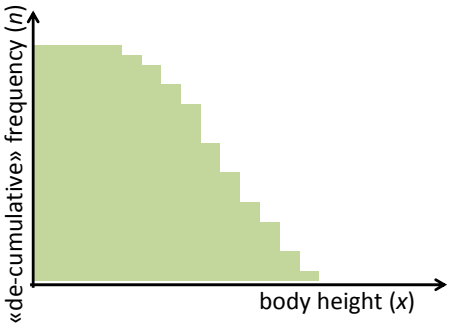
How many elements are **less** than a given *x* value?

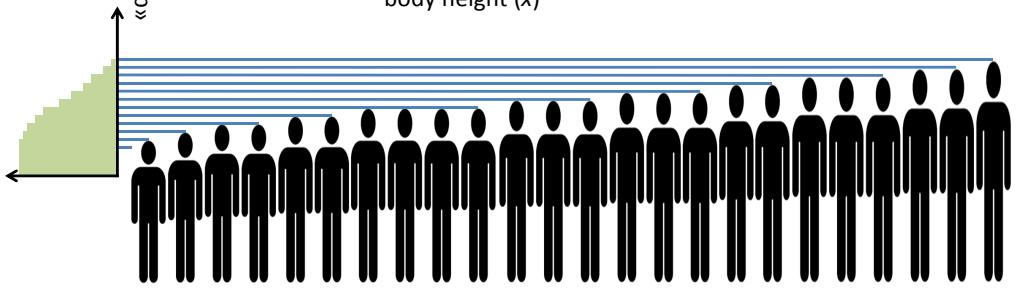




Integral Discrimination Function

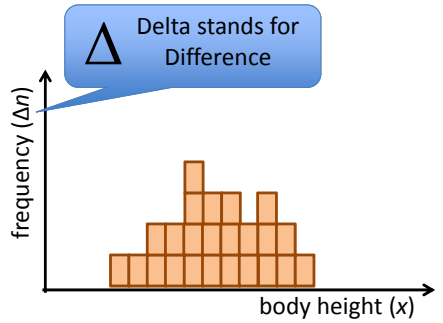
How many elements are **greater** than a given *x* value?

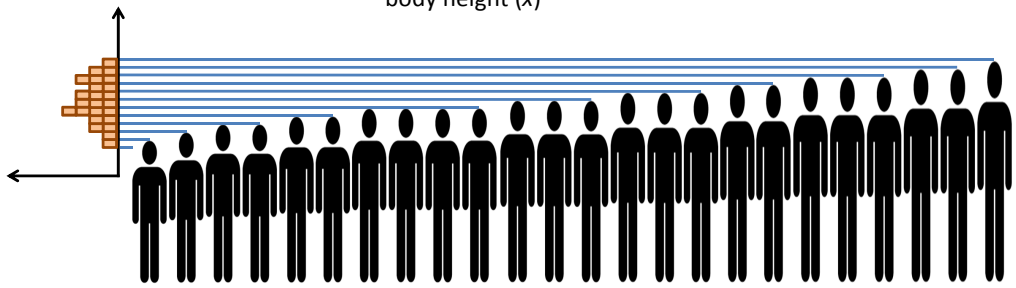




Frequency Distribution Function

How many elements fall **within** a bin of Δ*x* width?





Representation of Data (I)

1) *A simple list of data*
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 ($[\Delta n/n]/\Delta x$)

list: an enumeration of results of all experiments

(absolute) frequency: number of experiments with the given outcome

relative frequency: the proportion of the given outcome within the sample

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	
8	B	
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

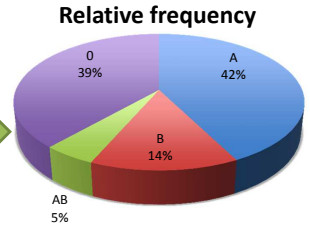
gyakorisági eloszlások (differenciáldiszkriminációs függvények)

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		

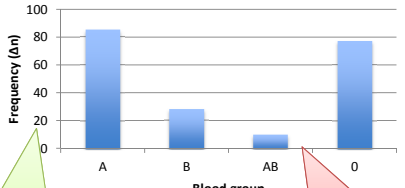
Representation of Data (II)

3) *Representation of frequencies in case of **qualitative** variables*

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



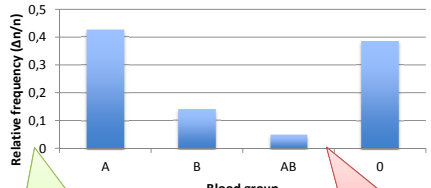
Frequency "distribution"



dependent variable: (absolute) frequency

independent variable: nominal categories (categorical „axis")

Relative frequency „distribution"



dependent variable: relative frequency

independent variable: nominal categories (categorical „axis")

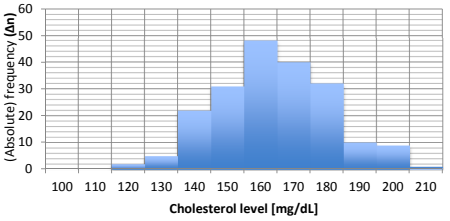
Representation of Data (III)

3) *Representation of frequencies in case of **qualitative** variables*

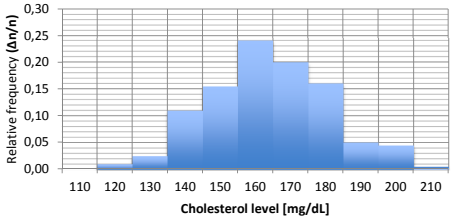
normalized to sample size (n)

normalized to bin width (Δx)

(Absolute) frequency distribution

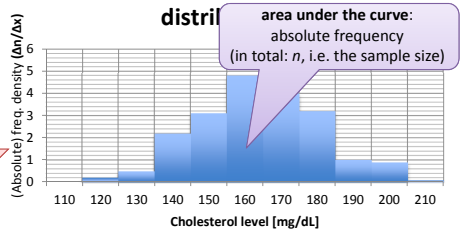


Relative frequency distribution



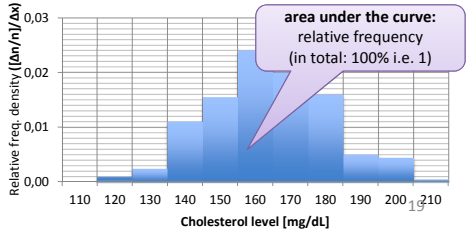
(absolute) freq.density distri

area under the curve: absolute frequency (in total: n, i.e. the sample size)



Relative freq. density distribution

area under the curve: relative frequency (in total: 100% i.e. 1)



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?

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 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?