Biophysics I.
for dentistry students

Lecture 1st:

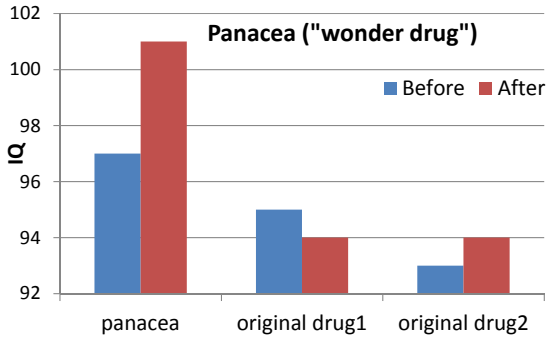
Biostatistics I.

2019. September 9.

Dániel Veres

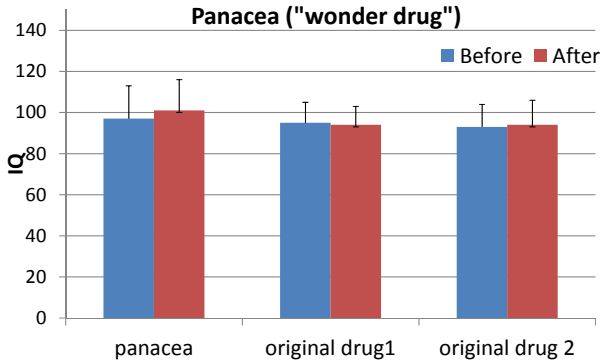
Biostatistics – why to learn?

- „ To decide whether we should believe in something we are reading or to see where the mistake is, that is to say, do not fall so easily into statistical „juggling", artifacts and mistakes.



Biostatistics – why to learn?

- „ To decide whether we should believe in something we are reading or to see where the mistake is, that is to say, do not fall so easily into statistical „juggling", artifacts and mistakes.



+Examples

- *Causality?*

(Ananas – tumor frequency, height – sleeping problems)

☺eg: http://www.fastcodesign.com/3030529/infographic-of-the-day/hilarious-graphs-prove-that-correlation-isnt-causation

- *☺eg: Chocolate Helps Weight Loss*

https://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800

Biostatistics – why to learn?

- „ To decide whether we should believe in something we are reading or to see where the mistake is, that is to say, do not fall so easily into statistical „juggling", artifacts and mistakes. (see excel – panacea,…)
- „To judge better whether we were lucky or not – or none of them"
- „ To judge better what is worth , whether it is worth for risking it" (eg. risk of a treatment)

Biostatistics – why to learn?

- „ To decide whether we should believe in something we are reading or to see where the mistake is, that is to say, do not fall so easily into statistical „juggling", artifacts and mistakes. (see excel – panacea,…)
- „To judge better whether we were lucky or not – or none of them"
- „ To judge better what is worth , whether it is worth for risking it" (eg. risk of a treatment)

- „ So that we can do our best to design and evaluate our own statistics in our work (diploma…)."
- „I got an interested, unexpected result? I just discovered something or just the game of chance I see?"
- „ To make our results more understandable and effective, we can highlight the essence. "
- „ To have a clear understanding of the literature. "

(J. Reiczigel, A. Harnos, N. Solymosi – Biostatisztika nem statisztikusoknak)

Keywords in Statistics

VARIABILITY

Stochastic

Random

Tastitsticsss? What's that?

Tastitsticsss? What's that?

Statistics describes **random mass** phenomena.

Tastitsticsss? What's that?

Statistics describes **random mass** phenomena.

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

Descriptive Statistics

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

Inferential Statistics
(Inductive)

Tastitsticsss? What's that?

Statistics describes **random mass** phanomenons.

- Data Collecting (Sampling)
- Data Organization

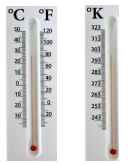
Descriptive Statistics

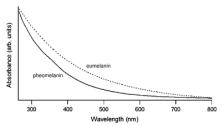
- Data Analysis
- Conclusion

Inferential Statistics
(Inductive)

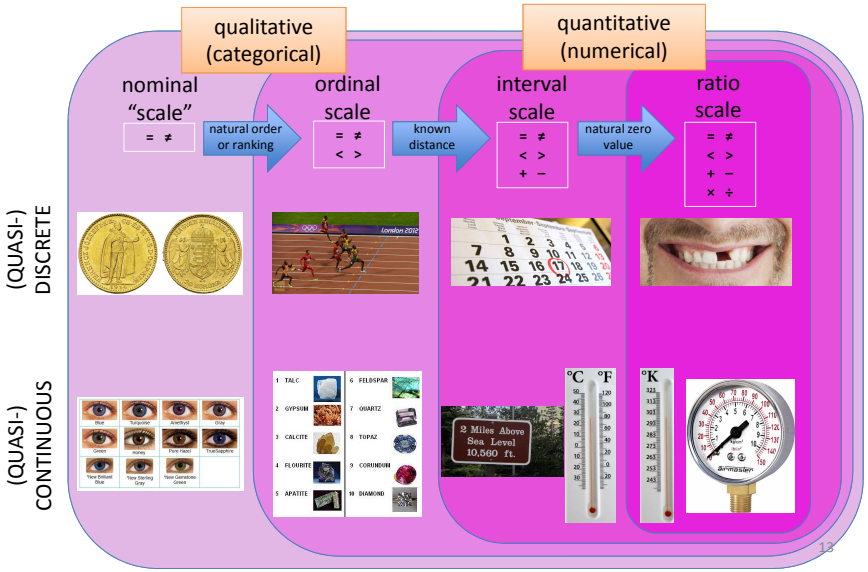
Variables, outcomes

Could be measured or observed





Variable Types:
Levels of Measurement



Description of Nominal Variables I.

Numerical (analytical)

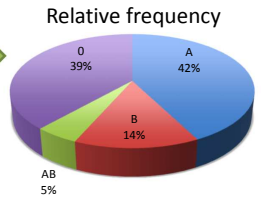
List

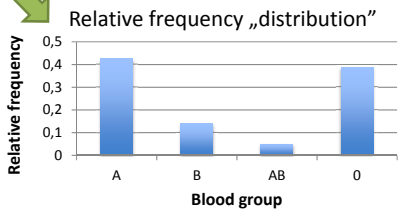
patient №	blood group (ABO)	cholesterol level (mg/dL)
1	B	148
2	AB	147
3	B	169
4	B	159
5	B	150
6	B	167
7	A	144
8	B	158
9	AB	177

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





Description of Nominal Variables II.

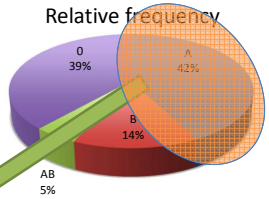
Numerical

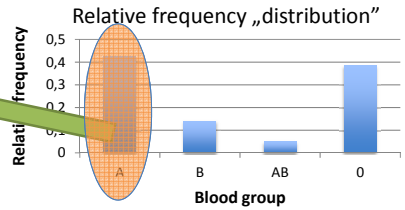
Frequency table

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***

Description of Ordinal Variables I.

Numerical

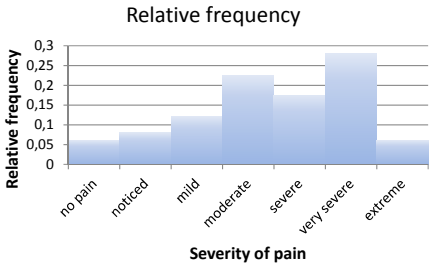
Frequency table

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	

Indicator:
Mode

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

Graphical



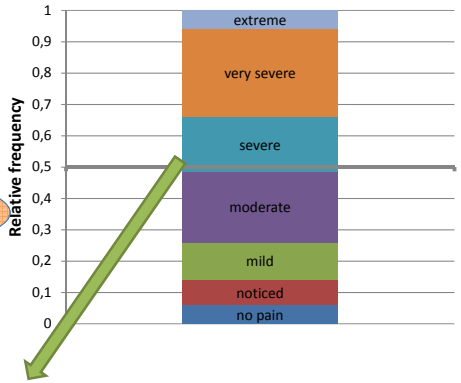
Description of Ordinal Variables II.

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}

Description of Quantitative Variables I.

Numerical (analytical)

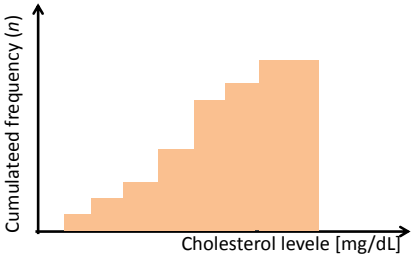
Frequency tables

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

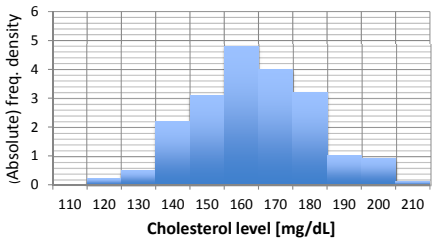
Organizing data – with **loss of information**

Determination of bin width:
- technical and aesthetic concerns
- statistical concerns

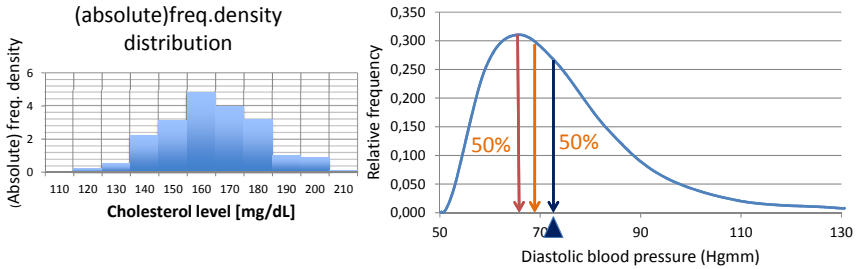
Graphical



(absolute)freq.density distribution



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***

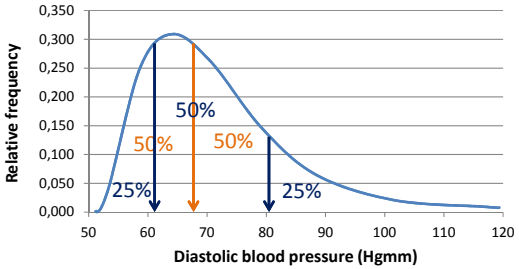
Formulas: in the formula collection...

Remark

Average ≠ Mean

In statistics the average could mean:
- mode,
- median,
- means – arithmetic, geometric, harmonic... mean

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

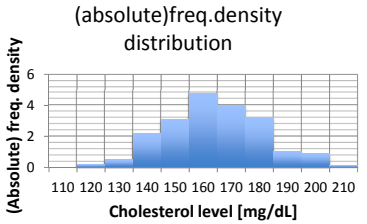
Outliers…

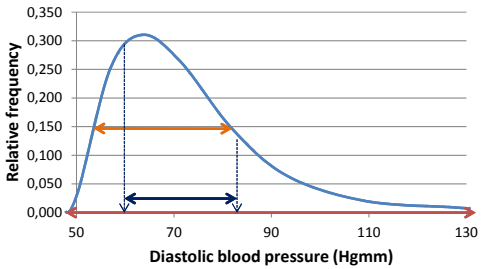
Day	Waiting time (min)		
1	1,27	median	8,48
2	3,3	lower quartile	3,59
3	3,44	mean	7,72
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	12,98		

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?

Description of Quantitative Variables III.



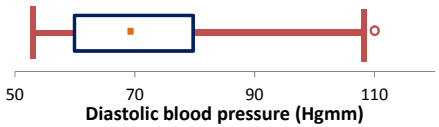


Measures of spread:

- ***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”

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***

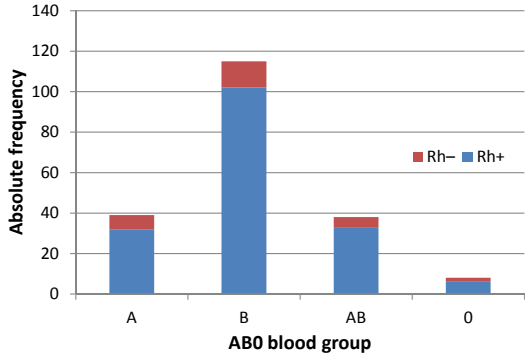
Trimmed mean: mean calculated without outliers

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

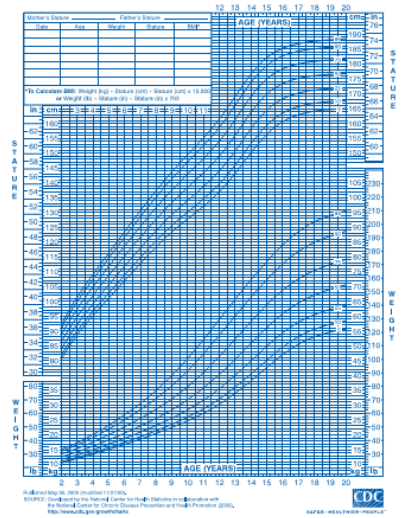
Graphical: ***stacked bar chart***

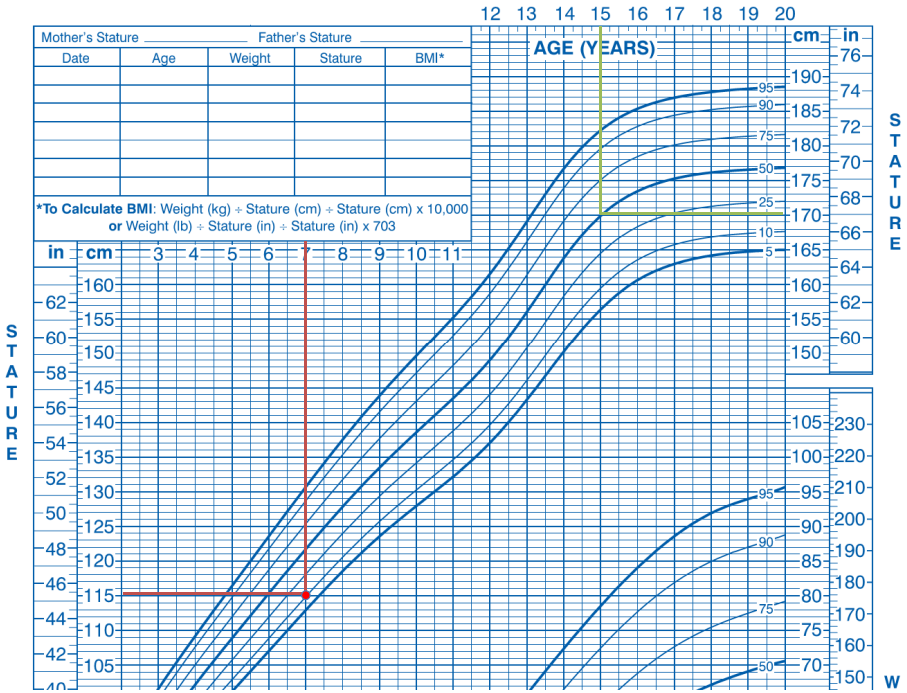


Quantitative Bivariate Description

Graphical: ***percentile curves***

Percentile: quantile expressed as percentage





Yoshino K *et al.*

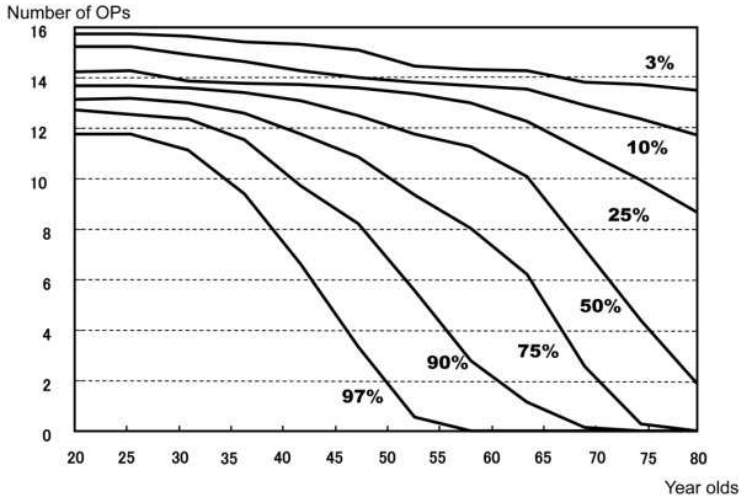
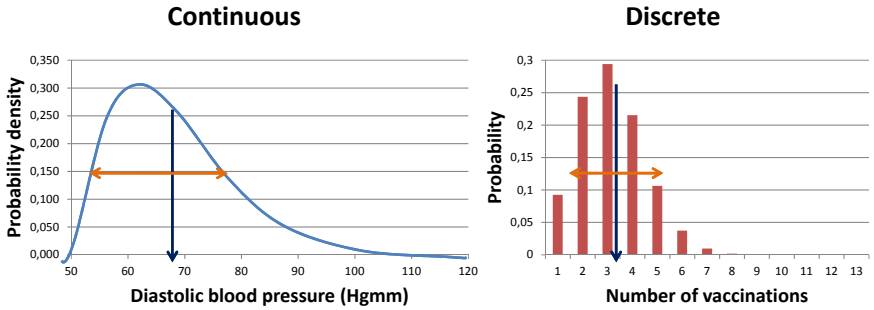


Fig. 1 Percentile curves of occluding pairs in males (n = 1,535)

Parameters of Theoretical Distributions



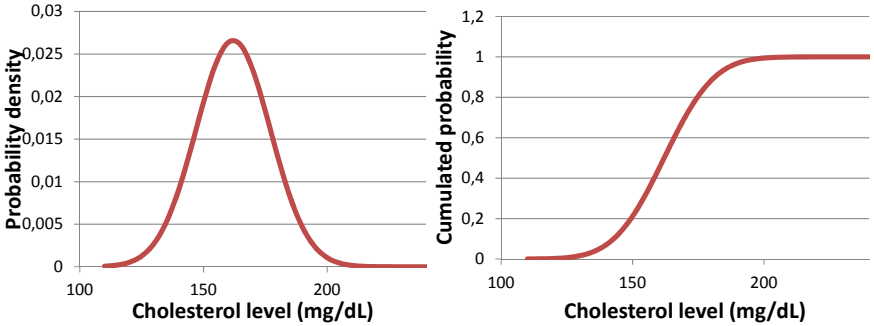
- ***Expected value(E) (location parameter)***

$$E(\xi)=\int_{-\infty}^{\infty} p_i \cdot x_i \qquad E(\xi)=\sum_{i=1}^{m} p_i \cdot x_i$$

- Theoretical variance (Var, D^2) (scale parameter)

$$Var(\xi)=E\left[\left(\xi-E(\xi)\right)^2\right]$$

Normal (Gaussian) Distribution I.



Cholesterol level, glucose level.....
Height, BMI...
Diastolic blood pressure of adults
.....

$$E(\xi)=\mu$$
$$Var(\xi)=\sigma^2$$
$$P=\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Normal (reference) range: 95% of data is in this range: $\sim \mu \pm 2*\sigma$

Gaussian Distribution II.

Central limit theorem (on variables): for given conditions, adding a large number of independent variables yields a normally distributed variable.

Central limit theorem (on sampling): for given conditions, sampling with large sample size (n) the distribution of the sample means is normal with:

$$\mathrm{Var}_{\mathrm{normal}}=\frac{\mathrm{Var}_{(\mathrm{original})}}{\mathrm{n}}$$

Tastitsticsss? What's that?

Statistics describes **random mass** phanomenons.

- Data Collecting (Sampling)
- Data Organization

Descriptive Statistics

- Data Analysis - estimations
- Conclusion

Inferential Statistics (Inductive)

Population and Sample

Population

The size of the **population** usually does not allow the examination of all of its elements.

Population and Sample

Population

RANDOMNESS!

Sample

The size of the **population** usually does not allow the examination of all of its elements.

Therefore, only a subset of the population is examined. That is what we call a **sample**.

Population and Sample

Population

RANDOMNESS!

Sample

The size of the **population** usually does not allow the examination of all of its elements.

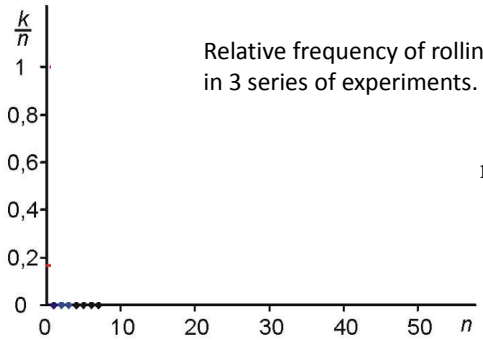
Therefore, only a subset of the population is examined. That is what we call a **sample**.

UNCERTAINTY! (P and E)

Characteristics of the sample can be used to draw conclusions on the population.

We carry out measurements on the sample elements, then this data set (which is also called **sample**) will be characterized by graphs and numbers

Probability I.

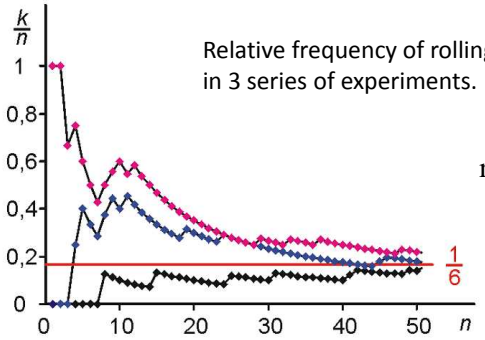


Relative frequency of rolling 6 with a die in 3 series of experiments.

$$\text{rel. freq.} = \frac{N_{favorable}}{N_{all}}$$

We experience that **relative frequencies** – although with fluctuations – **tend to a certain value** independetly from the actual series of experiments if we **increase the number of the experiments**.

Probability I.

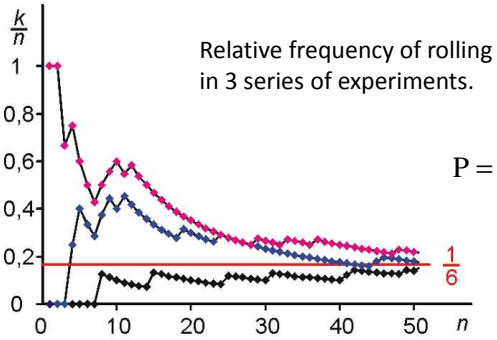


Relative frequency of rolling 6 with a die in 3 series of experiments.

$$\text{rel. freq.} = \frac{N_{favorable}}{N_{all}}$$

We experience that **relative frequencies** – although with fluctuations – **tend to a certain value** independetly from the actual series of experiments if we **increase the number of the experiments**.

Probability as a Quantity



Relative frequency of rolling 6 with a die in 3 series of experiments.

$$\text{P} = \text{rel. freq.} = \frac{N_{favorable}}{N_{all}}$$

$$N_{all} \to \infty$$

Law of large numbers (on relative frequencies): the relative frequency in an infinite sequence tends to a certain value.
We assign that **certain value** to an **event**: ***1/6*** to ***rolling 6*** with a die.
This value is called the ***probability of an event***.
This is an *empirical law – cannot be proven* by logical sequence.

Probability of Events I.

Axioms on probability of events (Kolmogorov):
1. **0 ≤ P(A) ≤ 1**

Probability of Events I.

Axioms on probability of events (Kolmogorov):
1. **0 ≤ P(A) ≤ 1**

2. **P(*sure*) = 1** *(The patient will die sooner or later)*
P(*impossible*) = 0 *(I'm 310 cm tall)*

Probability of Events I.

Axioms on probability of events (Kolmogorov):
1. **0 ≤ P(A) ≤ 1**

2. **P(*sure*) = 1** *(The patient will die sooner or later)*
P(*impossible*) = 0 *(I'm 310 cm tall)*

3. Mutually exclusive events (i.e. P(AandB)=0)
P(AorB)=P(A)+P(B)
(probability of being pregnant or male)

Probability of Events I.

Axioms on probability of events (Kolmogorov):
1. **0 ≤ P(A) ≤ 1**

2. **P(*sure*) = 1** *(The patient will die sooner or later)*
P(*impossible*) = 0 *(I'm 310 cm tall)*

3. Mutually exclusive events (i.e. P(AandB)=0)
P(AorB)=P(A)+P(B)
(probability of being pregnant or male)

And a theorem:
+4. *Independent events*: **P(AandB)=P(A)*P(B)**
(probability that our first patient is male and the second one is female)

Probability of Events II.

Conditional events calculation:
general form: **P(A|B)=P(AandB)/P(B)**
Special cases:
I. Independent events:
Probability that our second patient is male
if *the first one is female*

P(A|B)=P(AandB)/P(B)
P(A|B)=P(A)*P(B)/P(B)
P(A|B)=P(A)

Probability that our second patient is male
if *the first one is female = Probability that our second patient is male*

Probability of Events II.

II. event A is a subset of event B
Probability that a patient has a flu
if suffering from a viral infection

P(A|B)=P(AandB)/P(B)
P(A|B)=P(A)/P(B)

Calculation:
The probability that a patient coming to our office has viral infection is 8% =P(B)
The probability of occurrence of flu infections at our office is 2% = P(A)
The probability that a patient suffering from a viral infection has actually flu is: P(A|B) = 2% / 8 % = 25%.

Probability Calculus

Permutations,
Variations,
Combinations

Probability Calculus Example

During last year's flu epidemic 402 out of the total 2989 patients who turned up at a doctor's office required vaccination. Based on last year's data what is the probability that 4 vaccines will be sufficient (exactly, i.e. no vaccines will be left), if we are expecting a total number of 25 patients?

Probability Calculus Example

During last year's flu epidemic 402 out of the total 2989 patients who turned up at a doctor's office required vaccination. Based on last year's data what is the probability that 4 vaccines will be sufficient (exactly, i.e. no vaccines will be left), if we are expecting a total number of 25 patients?

$$P=\binom{n}{k}\cdot(p)^{k}\cdot(1-p)^{(n-k)}=\binom{25}{4}\cdot\left(\frac{402}{2989}\right)^{4}\cdot\left(1-\frac{402}{2989}\right)^{(25-4)}\approx 0{,}2$$

How to calculate (in excel)? How to read out from a graph, table?
Which equation, table, excel function should we use?

Human thinking and probability...

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

a) Linda is a teacher in a secondary school
b) Linda works in bookstore and participates in yoga courses
c) Linda is a member of the league of women voters
d) Linda is a bank teller.
e) Linda is an insurance agent
f) Linda is a bank teller and is active in the feminist movement.

Error

Population
- Real value (estimand)

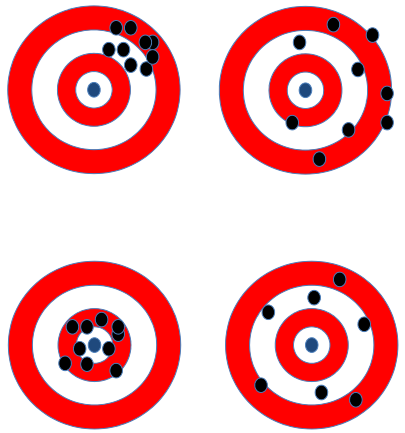
Estimation

Sample
- Estimates

Error

Error – 2 dimension

„Average of differencies" (bias)
Systemic error



Good estimation, if:
Unbiased:
the expected value of the estimates is equal with the estimand
Effective:
the variability is small
(Konsistent:
the variability of estimates decreases with increased sample size)

„Variability" (standard deviation);
Random error

Test Questions #1

- 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 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.
- Define the median(s) of a dataset.
- 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.
- Which central tendency sensitive to outliers?
- What is the advantage of indicators versus distribution functions?
- What are the measures of location?
- Define the p-quantile.
- Define the lower quartile.
- Give the definition of probability based on relative frequences.
- What is the law of large numbers?
- How tends the relative frequencies to the probability? [fluctuations, infinite seuence]
- How we can prove the law of large numbers?
- How to calculate P(A) if P(A|B) and P(B) is given?
- What are the Kolmogorov's axioms?
- What is the central limit theorem?
- What does unbiased esetimation mean?
- What does effective esetimation mean?