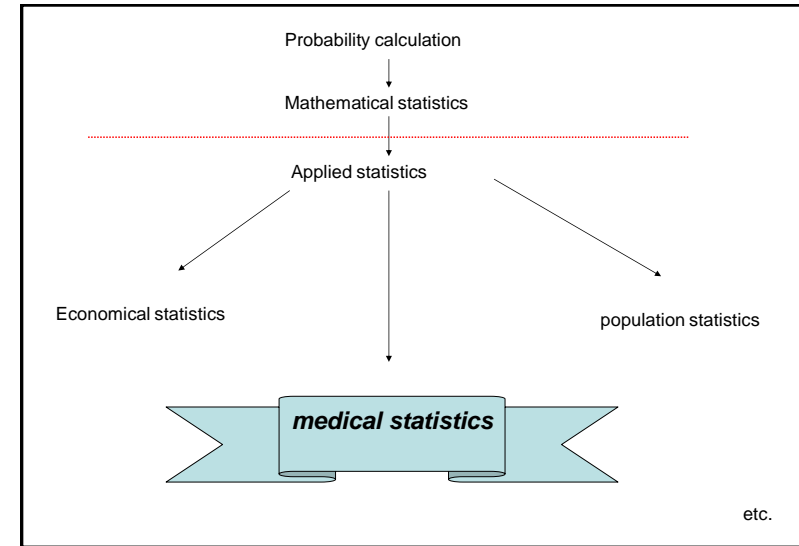
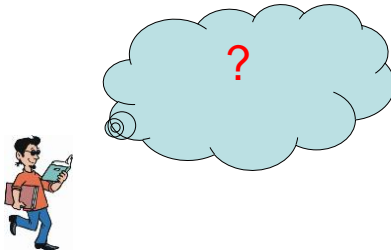
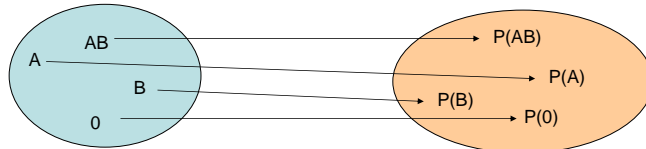


## Probability calculation and the statistics!



## Example: blood type



Elementary events: A, B, AB, 0  
Sample space:  
Probabilities:  
 $P(A)$ ,  $P(B)$ ,  $P(AB)$ ,  $P(0)$   
Exclusive events:

$$P(A) + P(B) + P(AB) + P(0) = 1$$

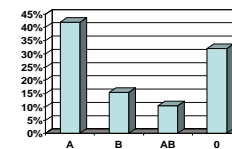
An event for example:  
A antigen is present:  
probability =  $P(A) + P(AB)$   
one and only one antigen is present:  
probability =  $P(A) + P(B)$

Is there anything to do?

?

## Distribution

Distribution of blood types  
in Hungary



How can we make such information?  
How much is the credibility?

**Theoretical way** (rare)  
(e.g. dice rolling:  
probability of an elementary event:  $1/6$ .)

**Experimental way**  
(Experiment or trial.  
Trial: measurement,  
observation,  
questioning, etc.)

## Population and the sample

Ideal: if we observe every possible cases.

**Population (statistical universe):** A large set or collection of items that have a common observable property or properties. This may consist of finite or infinite no. of items. Theoretical universe is also possible with potentially observable elements.

**Sample:** A small portion of the population selected according to a certain rule or rules.

## Principle of sampling

Conclusion:  
larger no. of elements – less difference, more reliable result.

•No. of elements: as large as possible.  
(Within the bounds of reason.)

•**Random sampling.**

•**In medicine:**

If there is no exclusive occasion, then it must be random.

## Error sources

### Sampling error

Origin: we deal with the sample only (a smaller part of the statistical universe).

We are not able to avoid but we are able to analyze and take into the consideration using statistical methods!



### Non-sampling error

Sample survey error e.g.: response error, processing error etc.

An extreme example:  
Non-random sampling!

Gynecology



Next please!

## Estimation

How high is the tree?

About 7 m.



**Estimation:** such kind of procedure, that orders a value to a variable or to a case on the base of incomplete, empirical data.

## Type of the estimation

### Point estimation

Estimation by one value.



warrant of caption  
...  
height: about 175 cm  
...

### Interval estimation

Estimation by interval (it is inside the range with high reliability).



warrant of caption  
...  
height: 170-175 cm  
...

## Main properties of a good estimation

- Unbiased:** The expected value of the estimation is the required parameter in the case of every possible no. of elements.
- Efficient:** The squared error of the estimation from the parameter has minimum.
- Consistent:** Increasing the sample size increases the probability of the estimator to be close to the population parameter.
- Sufficient:** Contains every information that possible to get from the sample (E.g. a mean and standard deviation are sufficient in the case of the normal distribution).

## Categorical quantity

trial: select a people and do the test!



Select enough large no. of people!

$n$ : no. of elements.

**Sample:**  $n$  people from the population.

outcome:  
A or B or AB or 0.

Blood type	frequency
A	$k_A$
B	$k_B$
AB	$k_{AB}$
0	$k_0$

## Estimation of a probability

$P(A)$  probability of the A blood type.  
The expected value of the frequency of A:  $n \cdot P(A)$ .

Estimation of  $n \cdot P(A)$  on the base of the sample:  $k_A$

Point estimation of  $P(A)$  :  $k_A / n$ .



O.K., but another sample results different value.  
How much is the reality of this value?

## The error of the proportion



Binomial distribution.  
expected value:  $np$   
variance:  $np(1-p)$

(Oop! Probability calculation?)

$n$  elements:  
 $k$  elements have  $A$  blood type,  
 $(n-k)$  not.

$s_{k/n}$  is the sd of  $k/n$ , or  
**standard error of the proportion.**



Estimation of the sd of the  $k_A$  value:

$$s_k = \sqrt{nP(A)(1-P(A))}$$

Estimation of the sd of the  $k_A/n$  value:

$$s_{k/n} = \frac{\sqrt{nP(A)(1-P(A))}}{n} = \sqrt{\frac{P(A)(1-P(A))}{n}}$$

Instead of  $P(A)$  use  $k_A/n$ !

## Confidence interval

Using this value we are able to determine an interval.  
(interval estimation)

$$\left( \frac{k}{n} \pm s_{k/n} \right)$$

**68% confidence interval,**  
68% **confidence level** belongs to  
this.



Meaning:

If we repeat the observation many  
times, about the 68% of the  
confidence intervals contains the  
 $P(A)$ .

The reliability of the interval  
estimation is about 68%.

## Continuous quantity

Example: height

height: 172 cm.



Is it correct?



No!

- Exact measurement is impossible,
- Infinite accuracy were required.

Sample space infinitely large!

Finite no. of elements in the  
sample.  
Theoretically there is no two equal  
elements.  
(frequencies: 1 or 0)

False conclusion,  
Can't be used.

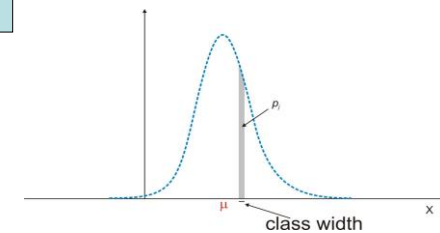
## Sampling of a continuous quantity

Exact statement:

Height ( $x$ ):  
 $171.5 \leq x < 172.5$  cm



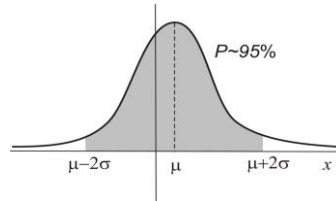
Instead of a concrete value we use an  
interval so-called **class**.  
(We can use them as the discrete  
values.)



$p_i$  – probability, that  $x$  is in the given class.

## $\mu$ and $\sigma$

$\sigma$  characterizes the deviation of the data around the  $\mu$ .  
About 68% of the data are around the  $\mu$  in the  $2\sigma$  wide interval.



$$(\mu \pm \sigma) \approx 68\%$$

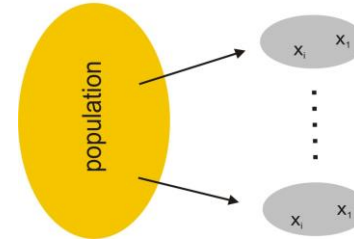
$$(\mu \pm 2\sigma) \approx 95\%$$

$$(\mu \pm \infty) = ?$$



## Sampling distribution

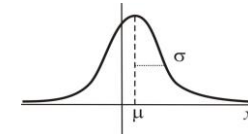
samples ( $n$ -elements)



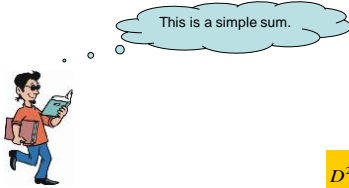
Every  $x_i$  element in the samples differ from each other.  
Distribution is used to describe.

The distribution of  $x_i$  values is same as the distribution of the population.

$$M(x_i) = \mu \quad \text{and} \quad D^2(x_i) = \sigma^2$$



## The expected value of the average and it's variance



$$\bar{x} = \frac{1}{n} \sum_i x_i$$

$$M(\bar{x}) = \frac{1}{n} \sum_i M(x_i) = \frac{1}{n} (n\mu) = \mu$$

$$D^2(\bar{x}) = \frac{1}{n^2} \sum_i D^2(x_i) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}$$



The expected value of the averages is equal to the  $\mu$  of the population, variance is  $n$  times smaller.

## Estimation of the expected value

$$M(x) = \sum_j p_j x_j$$

Point estimation: average.

approximate  $p_j$  by  $k_j/n$ !

Unbiased:

$$\sum_j \frac{k_j}{n} x_j = \frac{1}{n} \sum_j k_j x_j = \frac{1}{n} \sum_i x_i = \bar{x}$$

$$M(\bar{x}) = \mu$$



## Corrected empirical $s_d$

The difference derives from the difference of  $\mu$  and average.

$$(\bar{x} - \mu)^2$$

$$M[(\bar{x} - \mu)^2]$$

$$\frac{\sigma^2}{n}$$

This is the variance of the samples.

increase n!

$$\sigma^2 = M(s^2) + \frac{\sigma^2}{n}$$

$$\sigma^2 = \frac{n-1}{n} M(s^2)$$

$$s^{*2} = \frac{n}{n-1} s^2$$

$$s^{*2} = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

We use  $s$  to symbolize the corrected empirical sd.

## Standard error

variance of the average:

$$\frac{\sigma^2}{n}$$

sd of the average:

$$\frac{\sigma}{\sqrt{n}}$$

But  $\sigma$  is normally unknown.

$s$  is a good estimation of  $\sigma$ .

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

This is the sd of the average or it's standard error.

## Confidence interval of the expected value

If we know the standard error we are able to determine the confidence interval of the expected value.

$$[\bar{x} \pm s_{\bar{x}}]$$

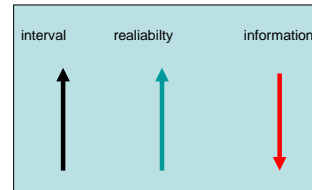
This interval includes  $\mu$  about with 68% confidence.

## Properties of the interval estimation

68%? Isn't too small?

We can increase, e.g.: in this case about 95% is the confidence level, but the information content is less.

$$[\bar{x} \pm 2 \cdot s_{\bar{x}}]$$

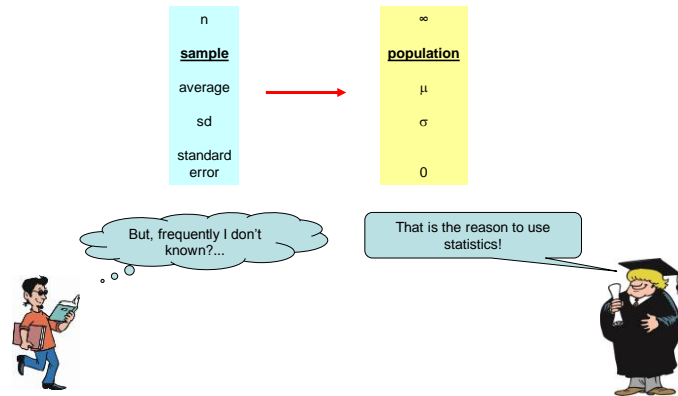


exact formula:

$$[\bar{x} \pm t_p \cdot s_{\bar{x}}]$$

where  $t_p$ : value of the  $t$ -distribution with  $(n-1)$  degree.  
(confidence level  $(1-p)$ )

## Relations between parameters



## Reference or normal range

