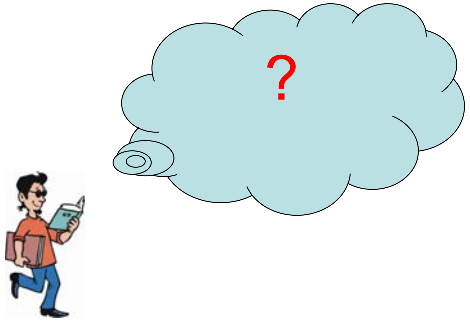
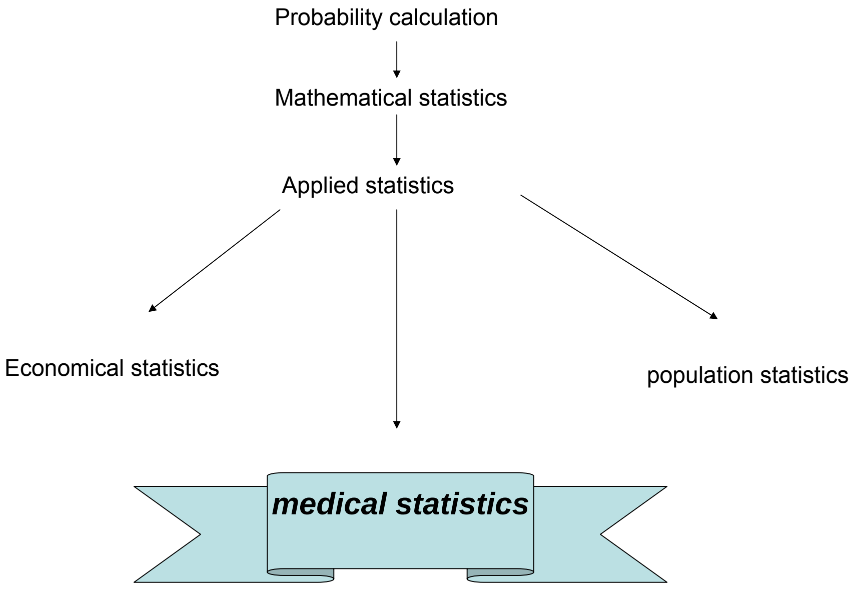
Probability calculation
and
statistics





Probability calculation

Mathematical statistics

Applied statistics

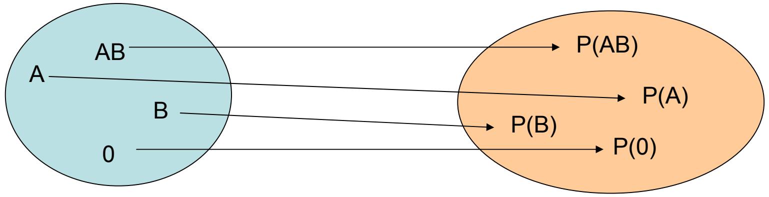
Economical statistics

population statistics

medical statistics

etc.

Example: blood type



AB

A

B

0

P(AB)

P(A)

P(B)

P(0)

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:

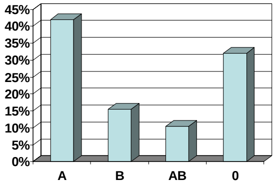
probability = P(A)+P(B)

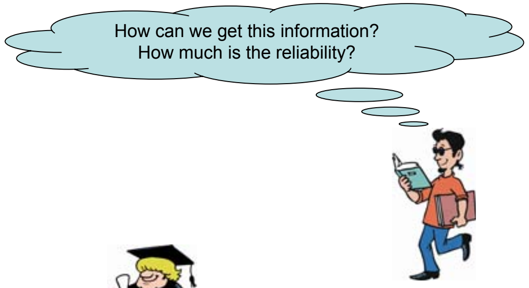
Is there anything to do?

?

Distribution

Distribution of blood types in Hungary





How can we get this information?
How much is the reliability?

Theoretical way (rare)
(e.g. throw of dice:
probability of a elementary event: 1/6.)

Experimental way
(experiment or trial.
trial: measurement,
observation, asking,
etc.)

Trial

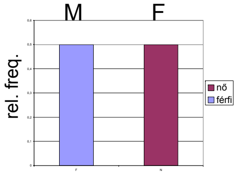
Example: the ratio of the male and female.
The sample space has 2 elements: male, female
probabilities: P(M) és P(F).
It's true: P(F)+P(M) = 1

What is the gender?

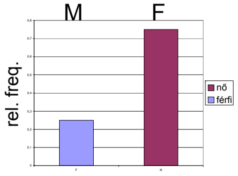
outcome:
M(ale) or F(emale)

no. of elements: 1
(1 trial)

P(F) = 0.5
P(M) = 0.5



P(F) = 0.75
P(M) = 0.25



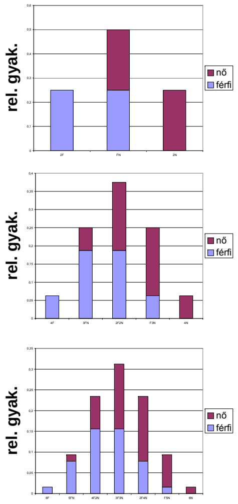
P(F) = 0.5
P(M) = 0.5

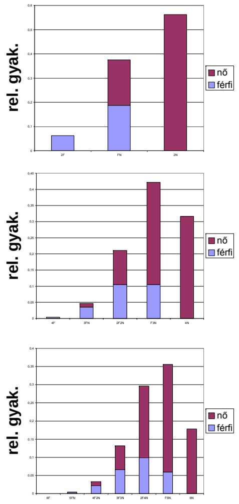
P(F) = 0.75
P(M) = 0.25

no. of elements: 2

no. of elements: 4

no. of elements: 6





Principle of sampling

Conclusion
larger no. of elements – smaller differences, 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 must be random.

Population and sample

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.

ssible cases.

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 survay error e.g.: response error, processing error etc.

Gynecology

Next please!

An extreme example:
Non-random sampling!
(previous example)

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 reliabilty).

warrant of caption
…
hight: 170-175 cm
…

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 squered error of the estimation from the paramater has minimum.

Consistent: Increasing the sample size increases the probability of the estimator being 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 int he case of the normal distribution).

Categorical quantity

trial: select a people and do a test!

outcome:
A or B or AB or 0.

Select enough large no. of people!

n: no. of elements.

Sample: *n* people from the population.

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·P(A) .

Estimation of n·P(A) on the base of the sample: k_A

Point estimation of P(A) : $\mathbf{k_A / n}$.

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

The error of the relative frequency

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

(Oop! Probability calculation?)

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

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!

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

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 cofidence 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,
- Infinit acurracy 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.

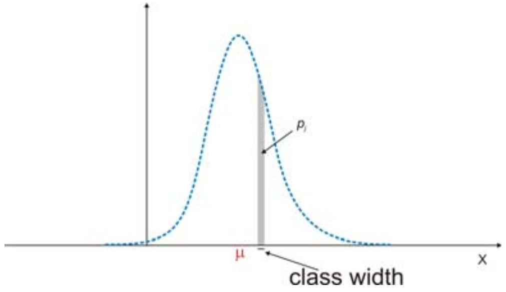
Sampling in the case of continuous quantity

Exact statement:

Height (x):

$171.5 \le x < 172.5\,\text{cm}$

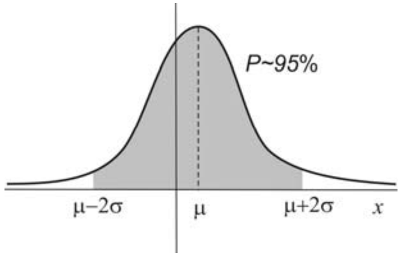
Instead of a concrete value we use an interval so-called **class**.
(We can use them as the discrete values)



p_j – probabilty, that *x* is in the given class.

μ and σ

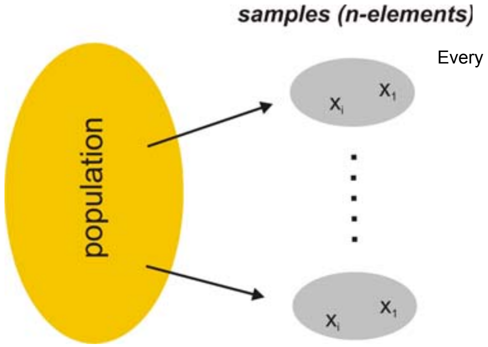
σ characterizes the deviation of the data around the μ. About 68% of the data are around the μ in the 2σ wide interval.



$$(\mu \pm \sigma) \approx 68\%$$
$$(\mu \pm 2\sigma) \approx 95\%$$

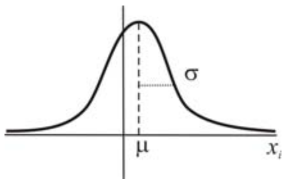
$$(\mu \pm \infty) = ?$$

Sampling distribution



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.



$M(x_i) = \mu$ and $D^2(x_i) = \sigma^2$

The expected value of the average and it's variance

This is a simple sum.

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

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

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

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

Estimation in the case of the continuous quantity

parameters of the distribution: expected value and theoretical sd. Definitions:

Expected value: $M(x) = \int xf(x)dx \longrightarrow \sum_j p_j x_j$

Theoretical *sd*: $D^2(x) = \int [x - M(x)]^2 f(x)dx \longrightarrow \sum_j p_j (x_j - \mu)^2$

Estimation of the expected value

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

Point estimation: average.

Unbiased:

approximate p_j by k_j/n!

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

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

Estimation of σ

$\sum_j p_j (x_j-\mu)^2$

approximate p_j by k_j/n!

$$\sum_j p_j(x_j-\mu)^2 = \sum_j \frac{k_j}{n}(x_j-\mu)^2 = \frac{1}{n}\sum_j k_j(x_j-\mu)^2 = \frac{1}{n}\sum_i (x_i-\mu)^2$$

μ= ?

normally unknown, only the average is known.

$$\frac{1}{n}\sum_i (x_i-\bar{x})^2 \quad ?$$

Good estimation?

Previously was proved:

$$\frac{1}{n}\sum_i (x_i-\mu)^2 > \frac{1}{n}\sum_i (x_i-\bar{x})^2$$

$\downarrow$ increase n! $\downarrow$

limit: $\sigma^2 > \frac{1}{n} M\left[\sum_i (x_i-\bar{x})^2\right]$

Calculate the average of several samples containing n elements! (expected value)

This is a biased estimation!

Corrected empirical *s_d*

The difference derives from the difference of μ and average.

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

$$\downarrow$$

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

$$\downarrow$$

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

increase n!

This is the variance of the samples.

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

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

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

$$\downarrow$$

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

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

Standard error

variance of the average:

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

sd of the average:

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

s is a good estimation of σ.

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

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

But σ is normally unknown.

Confidence interval of the expected value

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

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

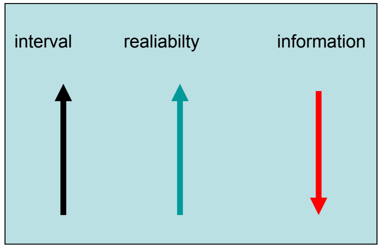
This interval includes μ about with 68% confidence .

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*))

Relation among parameters

n		∞
sample		population
average	→	μ
sd		σ
standard error		0

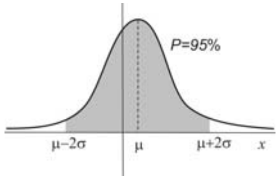
But, frequently I don't known?...

That is the reason to use statistics!

Reference or normal range

What dose it mean?

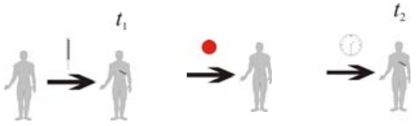
	male	female
K	3,5-5 mmol/l	3,5-5 mmol/l
WBC	$4\text{-}10 \cdot 10^9$/l	$4\text{-}10 \cdot 10^9$/l
HCT	42-54 %	38-50 %



In the case of quantity having normal distribution see the figure! (Instead of μ and σ normally we use their estimation from a large sample).
Anyway we use the interval containing 95% of the data.

Problem

experiment



$x = \Delta t = t_2 - t_1$

A possible conclusion:

Δt positive or zero: ineffective
Δt negative: effective.

Effective the medicine or not?

Is it true?
Whait is the situation if we don’t use it?
What is the role of the random fluctuations?