Biostatistics and Informatics

Lecture 5:
Estimation & Confidence
10th October 2018

Gergely AGÓCS

Sources: – Herényi L (2016): Statisztika és Informatika (in Hungarian): Chapter 14
– Reiczigel J, Harnos A, Solymosi N (2014): Biostatisztika nem statisztikusoknak (in Hungarian): Chapter 5
– WolframMathWorld: Probability and Statistics: http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html

Goals of this Lecture

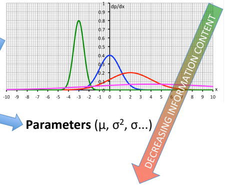
- Understanding the aim, types, and process of **estimation**
  - sample vs. population
  - **point** and **interval** estimation
  - sampling, estimated value and estimator
- What are the properties of a "**Good Estimate**"?
  - unbiasedness, consistence, efficiency, sufficiency
- Understanding **Standard Error** (SE)
- Understanding **Confidence Interval** (CI)
  - proper interpretation of confidence intervals
  - proper indication of confidence intervals
- Learning how to estimate certain parameters including:
  - probability
  - expected value
  - theoretical variance & standard deviation
- Calculation of necessary sample size to reach a desired SE
- Using **Excel** to carry out estimations

The Aim of Estimation

Population:
the (usually infinite) set of the outcomes of all the possible measurements.
Our experiments are always directed toward this set.

E.g.: - human height in cm
- human eye color
- number of live births per woman
- damage after exposition to radiation
- ...

Distribution Functions (CDF, PDF, PMF)



Parameters (μ, σ^2, σ...)

Individual **Probabilities**

But how can we get this information?

The Aim of Estimation

But how can we get this information?

E.g. coin tossing

Create a model
(only theoretical considerations)

P(Heads) = 1/2 & *P*(Tails) = 1/2

- two elementary events that are
- equally probable (= the coin is *assumed* to be regular) and
- universal

This method has
very limited use

Make observations

This means *actually* tossing the coin and using the observed outcome

But we cannot carry out *all* the possible measurements, because:
- impossible (infinity repeats)
- impractical (my cause damage)
- costs too much (money, time...)

The Process of Estimation

Population: - distribution
- parameters

Sample: - distribution
- parameters

RANDOMNESS!

The size of the **population** (often infinite) usually does not allow the observation of all of its elements: **direct examination is not possible**

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

MATH

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

INFERENCE

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

UNCERTAINTY!

Types of Estimations

What % of humans have Rh+ blood type?
We can't examine all the humans.

Take a subset (**sample**) and examine only the elements thereof.

Point estimation:

about
85%

only a value
called **estimate**

How much trust can we put into this *number*?
No idea!

Interval estimation:

between
83% – 87%

with 95% *confidence*

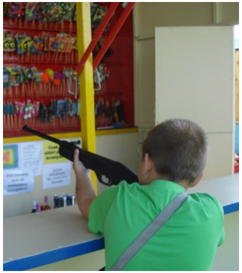
range
(estimator & error)
+ probability

The **level of trust** we can put into this *range* expressed as a **probability**

Point Estimations

Theoretical (population) values:
"**AIM**"

Estimator:
"**SHOT**"



Point Estimations

Theoretical (population) values:
(for discrete random variables)
"**AIM**"

- probability or proportion (P_i)

- expected value ("population mean") of discrete (E(ξ) or μ)

$$E(\xi) = \mu = \sum_{i=1}^{n} p_i \cdot x_i$$

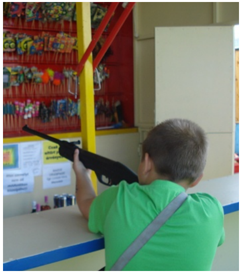
- theoretical variance (Var(ξ) or σ^2)

$$Var(\xi) = \sigma^2 = E\left[\left(\xi - E(\xi)\right)^2\right] = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

- theoretical standard deviation (SD(ξ) or σ)

$$SD(\xi) = \sigma = \sqrt{Var(\xi)} = \sqrt{\sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2}$$

Estimator:
"**SHOT**"



Point Estimations

Theoretical (population) values:
(for discrete random variables)
"**AIM**"

- probability or proportion (P_i)

- expected value ("population mean") of discrete ($E(\xi)$ or μ)

$$E(\xi)=\mu=\sum_{i=1}^{n}p_i\cdot x_i$$

- theoretical variance ($Var(\xi)$ or σ^2)

$$Var(\xi)=\sigma^2=E\left[\left(\xi-E(\xi)\right)^2\right]=\sum_{i=1}^{n}p_i\cdot(x_i-\mu)^2$$

- theoretical standard deviation ($SD(\xi)$ or σ)

$$SD(\xi)=\sigma=\sqrt{Var(\xi)}=\sqrt{\sum_{i=1}^{n}p_i\cdot(x_i-\mu)^2}$$

"Plug-in" Estimator:
"SHOT"

- relative frequency $\hat{p}_i=(k_i/n)$ ✓
  Excel: =COUNTIFS(*data*)/COUNTA(*data*)

- sample mean
  Excel: =AVERAGE(*data*)

$$\bar{x}=\sum_{i=1}^{n}\frac{k_i}{n}\cdot x_i=\frac{1}{n}\sum_{i=1}^{n}k_i\cdot x_i$$ ✓

- "plug-in" variance (s^{**2})

$$s^{**2}=\sum_{i=1}^{n}\frac{k_i}{n}\cdot(x_i-\mu)^2=\frac{1}{n}\sum_{i=1}^{n}k_i\cdot(x_i-\mu)^2$$ ✗

- "plug-in" standard deviation (s^{**}, *SD*)

$$s^{**}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}k_i\cdot(x_i-\mu)^2}$$ ✗

The "Good Estimator" is …

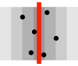
… Unbiased

Example #1: point estimation of a **probability (p_i)** with **relative frequency ($\hat{p}_i$)**

The "plug in" estimator formula we have just learned:

$\hat{p}_i=(k_i/n)$ ✓

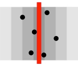
The mean of inifinite repetitions of this estimation will be equal to the probability we are looking for.



Excel: =COUNTIFS(*data*)/COUNTA(*data*)

Example #2: point estimation of an **expected value (μ)** with **sample mean ($\bar{x}$)**

$$\bar{x}=\sum_{i=1}^{n}\frac{k_i}{n}\cdot x_i=\frac{1}{n}\sum_{i=1}^{n}k_i\cdot x_i$$ ✓



Excel: =AVERAGE(*data*)

An estimator is unbiased if the expected value of repeated estimations is equal to the theoretical parameter to be estimated.

The "Good Estimator" is …

… Unbiased

Friedrich **Bessel**
1784–1846

Example #3: point estimation of **theoretical variance (σ^2)**

Look at the "plug in" estimator (s^{**2}) formula we have just learned:

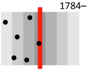
$$s^{**2}=\frac{1}{n}\sum_{i=1}^{n}k_i\cdot(x_i-\mu)^2$$ ✗

μ is a theoretical value, what we don't know. Instead, we have to use the **sample mean**.

Replace the expected value with the sample mean:

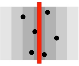
$$s^{*2}=\frac{1}{n}\sum_{i=1}^{n}k_i\cdot(x_i-\bar{x})^2$$ ✗

This formula is now minimal for the *sample mean*, not for the *expected value*. This causes a **bias**: the formula **underestimates** the theoretical variance.



Use a correction factor **n/(n–1)** (called **Bessel's correction**) to **remove this bias:**
(can be proven that it works but we won't prove…)

$$s^2=\frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}k_i\cdot(x_i-\bar{x})^2=\frac{1}{n-1}\sum_{i=1}^{n}k_i\cdot(x_i-\bar{x})^2$$ ✓



Excel: =VAR.S(*data*)

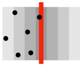
The "Good Estimator" is …

… or *Least* Biased

Example: point estimation of **theoretical standard deviation (σ)**

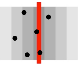
Similar to the variance estimator, the "plug in" standard deviation estimator is biased:

$$s^{*}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}k_i\cdot(x_i-\bar{x})^2}$$ ✗



Here, using Bessel's correction factor **n/(n–1)**
decreases but does not completely eliminate the bias:
(only asymptotically, i.e., for infinitely large samples;
reason: asymmetry of the square root function)

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}k_i\cdot(x_i-\bar{x})^2}$$ ✓



Excel: =STDEV.S(*data*)

We are happy using this less (but not un-) biased estimator.

Point Estimations

Theoretical (population) values: (for discrete random variables) “**AIM**”	**Least Biased Estimators**: **“SHOT”**
- probability or proportion (P_i)	- relative frequency $\hat{p}_i = (k_i/n)$ ✓ Excel: =COUNTIFS(*data*)/COUNTA(*data*)
- expected value (“population mean”) of discrete (μ or E(ξ)) $E(\xi) = \mu = \sum_{i=1}^{n} p_i \cdot x_i$	- sample mean Excel: =AVERAGE(*data*) $\bar{x} = \sum_{i=1}^{n} \frac{k_i}{n} \cdot x_i = \frac{1}{n} \sum_{i=1}^{n} k_i \cdot x_i$ ✓
- theoretical variance (σ^2) $Var(\xi) = \sigma^2 = E\left[\left(\xi - E(\xi)\right)^2\right] = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$	- empirical variance (s^2) Excel: =VAR.S(*data*) $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} k_i \cdot \left(x_i - \bar{x}\right)^2$ ✓
- theoretical standard deviation (σ) $SD(\xi) = \sigma = \sqrt{Var(\xi)} = \sqrt{\sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2}$	- empirical standard deviation (s, SD) Excel: =STDEV.S(*data*) $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} k_i \cdot \left(x_i - \bar{x}\right)^2}$ ✓

Point Estimations

Problem #1: We would like to estimate the proportion of blue-eyed students and the expected value, the theoretical variance, and the theoretical standard deviation of freshmen at Semmelweis.

We took a sample of 15 (see table). Use Excel to give point estimation for the theoretical parameters!

No. of observation	Eye color blue?	Stature in cm
1	FALSE	163
2	FALSE	153
3	FALSE	152
4	FALSE	158
5	FALSE	167
6	TRUE	184
7	TRUE	165
8	FALSE	184
9	TRUE	175
10	FALSE	167
11	FALSE	178
12	FALSE	168
13	FALSE	173
14	TRUE	178
15	FALSE	180

p(blue eye) ≈ $\hat{p}$ =4/15 = 0.2667 = 26.67%

μ(stature) ≈ $\bar{x}$ =AVERAGE(*data*) = 170 cm

σ^2(stature) ≈ s^2 =VAR.S(*data*) = 107.5 cm^2

σ(stature) ≈ s =STDEV.S(*data*) = 10.4 cm

The “Good Estimator” is …

… Efficient

An estimate is efficient if its standard deviation (called *standard error*, *SE*) is minimal.

- Repeated sampling yield a series of **estimates**, which **differ from each other** due to randomness of sampling
- So **the estimate itself is a random variable**, it also has a theoretical distribution, expected value, theoretical standard deviation etc.
- The theoretical standard deviation of an estimator is called **standard error** (SE)

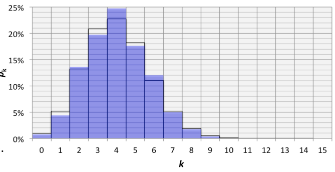
In the following slides we will learn how to calculate the standard error for these two cases.

Theoretical Distribution of Proportions:
Binomial Distribution

Theoretical Distribution of Sample Means:
Student’s t-Distribution (*df* = *n* – 1)

Standard Error of a Proportion

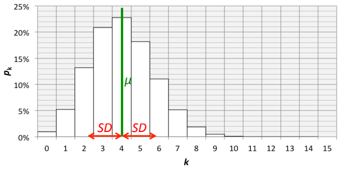
Example #1: We would like to estimate the proportion of blue-eyed students in the population of freshmen. We take a random sample of n = 15 out of which k = 4 turn out to have blue eyes. We have already learned, that the point estimation for population proportion (i.e. probability, p) is the relative frequency ($\hat{p} = k/n$), which in this case is 4/15 = 0.2667 . But what is the error of this estimation?



Suppose that the true value is the same as our estimation: p = 4/15. Using computer simulation, let us take 800 such samples, then prepare a relative frequency distribution of the number of blue-eyed (blue bars). This approximates well the corresponding theoretical distribution: the binomial distribution (black hairlines).

Standard Error of a Proportion

Use the formulae learned in previous lessons to calculate the parameters (μ, σ^2, and σ) of this binomial distribution.



Expected Value

$\mu = np = 4$

Variance

$\sigma^2 = np(1-p) = 44/15 = 2.933$

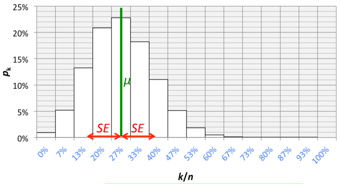
Standard Deviation

$$SD = \sqrt{np(1-p)} = 1.713$$

Standard Error of a Proportion

Since we are estimating proportions, the *k* variable needs to be converted to *k*/*n* proportions, that is, we have to rescale the horizontal axis by dividing the number by the sample size *n*.

We have to do the same to the calculated parameters μ and σ of the binomial distribution.



Expected Value

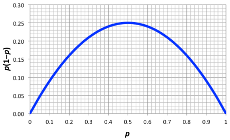
$\mu = p = 4/15$

Standard Deviation = Standard Error of Proportion

$$SE = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}} = 0.1142$$

$$SE_{prop} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$$

Standard Error of a Proportion



Theoretical Distribution of Proportions:

Binomial Distribution

(normalized to sample size)

$$\max\left(SE_{prop}\right) = \frac{1}{\sqrt{4n}}$$

What is the maximal ***SE*** of any proportion in a given sample with *n* size?

The $p(1-p)$ product is maximal, if $p = 1-p = 0.5$ In this case, $p(1-p) = 0.25$ so the *SE* is:

$$\max\left(SE_{prop}\right) = \sqrt{\frac{0.5(1-0.5)}{n}} = \sqrt{\frac{0.25}{n}} = \frac{1}{\sqrt{4n}}$$

If a study contains several estimations of proportions based on the same sample (at least same size of sample), often just the maximum *SE* is given using the above mentioned formula.

E.g. the maximum *SE* for proportion in case of a sample with *n* = 100 elements is 0.05.

Standard Error of a Proportion

Problem #2: What is the standard error for the estimation of the *prevalence* (proportion in the population) of sickle cell disease carriers in Nigeria, if 43 out of 172 examined persons carried the trait of the disease?

Proportion of carriers in the sample: $\hat{p}$ (SC) = *k*(SC)/*n* = 43/172 = 0.25 .

SE can be calculated using the formula: $SE_{prop} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.25\cdot(1-0.25)}{172}} = 0.033 = \underline{3.3\%}$

Problem #3: We would like to carry out a study to determine the prevalence of a set of chronic diseases in Budapest. What would be the recommended minimum sample size if we want to keep the error of estimation below 1%?

Since no actual probabilities or frequencies are know, we have to calculate with 0.5 probability which yields the highest SE for a certain sample size: $SE_{prop} \le \frac{1}{\sqrt{4n}}$

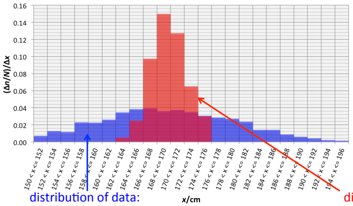
From this, we can find n:

$$n \le \frac{1}{4\left(SE_{prop}\right)^2} = \frac{1}{4\cdot 0.01^2} = \underline{2500}$$

That is, a sample size of 2500 will guarantee, that even in case of an eventual 0.5 prevalence our standard error won't exceed 1%. (The ≤ sign means that for lower or higher prevalence even less sample size would suffice to keep the SE below 1%.)

Standard Error of a Mean

Example #2: We would like to estimate the stature (body height) of freshmen. We take a random sample of n = 15 for which the mean height is $\bar{x}$ = 170 cm and the corrected standard deviation is s = 10 cm. Use computer simulation again to create the distribution of the data (outcomes) and the distribution of the means of samples of n = 15 elements.



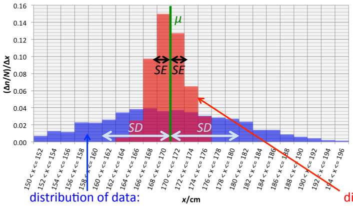
The simulation of 200 samples (3000 elements) clearly shows that the distribution of means is much narrower than the distribution of the elements themselves. But how much?

distribution of data:
$\bar{x}$ = 170.01 cm
SD = 9.947 cm

distribution of means:
$\bar{x}$ = 170.01 cm
SE = 2.507 cm

Standard Error of a Mean

The standard deviation of the data divided by the square root of the sample size n yields the standard deviation of the means. The latter is called the standard error of the mean (*SE*). The distribution of the means is (at least nearly) normal due to the Central Limit Theorem.



$$SE_{mean} = s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

distribution of data:
$\bar{x}$ = 170.01 cm
SD = 9.947 cm

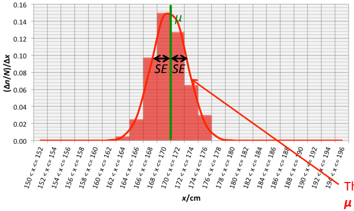
SD/n$^{0.5}$ = 9.947 cm/3.872 = 2.568 ≈

distribution of means:
$\bar{x}$ = 170.01 cm
SE = 2.507 cm

Standard Error of a Mean

However, if *SE* is calculated from sample *SD*, the distribution is **Student's t-distribution** with n–1 degrees of freedom (*df*). This distribution is similar to the normal distribution for large *df*-s but has heavy tails for small *df*-s.

William S. **Gosset**
1876–1937
"Student"



$$SE_{mean} = s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

Theoretical model: t-distribution
$\boldsymbol{\mu} \rightarrow \bar{x}$ = 170.01 cm
$\boldsymbol{\sigma} \rightarrow$ ***SE*** = 2.507 cm
df = n–1 = 14

Standard Error of a Mean

Problem #4: We wanted to estimate the expected value of mass of banana. We measured 5 bananas, the results are: 134 g, 152 g, 158 g, 141 g, 170 g. Give the *SE* of the estimation.
n = 5
$\bar{x}$ = 151 g
SD = 14.14 g
SE_{mean} = **6.32 g**

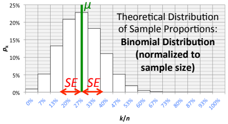
Problem #5: In a scientific article the following was written: "*... the average mass of the rats used in the study was 420 g (SE = 20 g) and their average age was 5 months ...*" The number of rats, however, is not mentioned. How many rats do you guess have been used in the study if we know from elsewhere that the standard deviation of rat body mass at this age is approx. 40 g?
The *SE* of the mean is the *SD* of the random variable divided by square root of n. We can transpose this formula to express n:
$n = (SD/SE)^2 = (40\text{ g}/20\text{ g})^2 = 2^2 =$ **4**

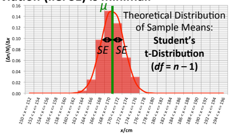
Note: this is a very low number of subjects which can explain why the authors did not share this information. This brings the reliability of the whole article into doubt...

The “Good Estimator” is …

… Efficient

An estimator is efficient if its *standard deviation* (i.e. SE) is minimal.





Standard Error
of Probability Estimation:

$$SE_{prop} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$$ ✓

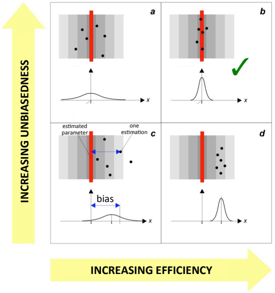
Theoretical Distribution
of Sample Means:

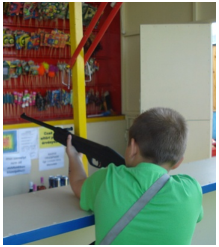
$$SE_{mean} = s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$ ✓

We can say in general, that the square of the *SE* is directly proportional to the the variance of the statistical variable and inversely proportional the sample size.
That is, to double the efficiency, the sample must be four times bigger!

The “Good Estimator” is …

… Unbiased & Efficient



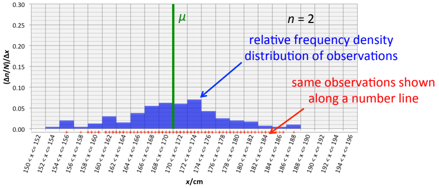


cf. Biophysics Lab Manual
17. MEASUREMENT TECHNIQUES
Fig. 3.: Measurement Accuracy and Precision

The “Good Estimator” is …

… Consistent

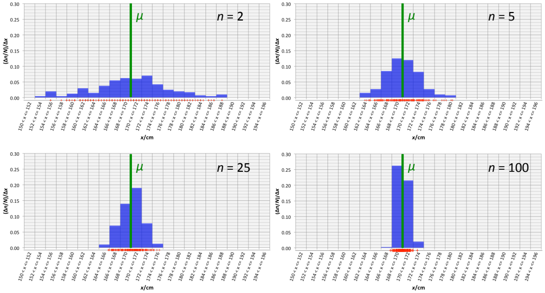
Imagine a sequence of estimations repeated with higher and higher sample sizes (*n*).
An estimator is consistent, if it tends to deviate from the estimated theoretical value less and less with increasing *n*.
In other words: higher sample size will yield less bias and error.



Distribution of 200 estimations of the expected value of the height
with sample size (*n*) = 2, 5, 25, and 100

The “Good Estimator” is …

… Consistent



The “Good Estimator” is …

… Consistent

Consistency should be assessed by strict mathematical derivation for each estimation method, however, we won’t do this. Instead, let’s just show qualitatively that our probability and expected value estimation methods are consistent.

Estimation of Probability

$\hat{p}_i = (k_i/n)$ ✓

Point Estimation is Unbiased

Estimation of Expected Value

$$\bar{x} = \sum_{i=1}^{n} \frac{k_i}{n} \cdot x_i = \frac{1}{n} \sum_{i=1}^{n} k_i \cdot x_i$$ ✓

$$SE_{prop} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$$ ✓

$$SE_{mean} = s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$ ✓

Error goes to 0 with increasing sample size (*n* is in the denominator)

The “Good Estimator” is …

… Sufficient

Sufficiency of an estimator means that it contains **all the information** that can be obtained from a sample relevant for the estimated theoretical (population) parameter.

Example: In case of a statistical variable measured at least on interval scale, the mean is a sufficient estimator of the expected value because it uses all the observed values of the sample. That is, knowing the whole sample does not provide more information than just knowing its mean.

Counter-Example: In the same case, the median would use only the rank of the observed values.

Counter-Counter-Example: However, if the statistical variable is measured on an ordinal scale, the median becomes a sufficient estimator for a central tendency (since mean cannot be used).

The “Good Estimator” is …

… Unbiased

The mean of estimates obtained from many many samples is equal (or very close) to the true value.

… Efficient

The standard error of the estimation (i.e. standard deviation of the estimates) is low.

… Consistent

Bigger sample sizes yield estimations which deviate less from the true value.

… Sufficient

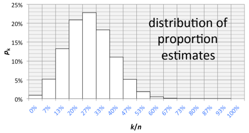
It contains all the information that the whole sample could provide.

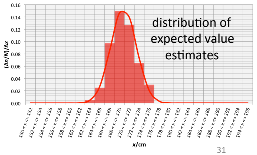
Interval Estimations

Interval estimation processes produce intervals (a pair of lower and upper limits) for the true value. The interval is called **Confidence Interval** (symbol: **CI**) and it is assigned with a probability called **confidence level** (symbol: 1–α) that reflects the **reliability of the process**.

The typical steps of generating a CI are:

- sampling
- calculating the point estimate
- determining the probability distribution of the point estimate
- calculating the standard error of this distribution (optional)
- use these values to determine an interval where the estimate can be “reliably” found





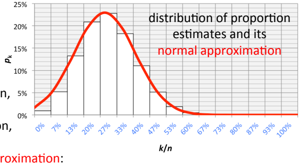
Interval Estimations ...

... of Proportions

What is the probability that a randomly chosen student has blue eyes?

The steps of generating a CI are:

- sampling: n elements with k blue-eyed
- point estimator: $p_k \approx k/n = \ldots$
- probability distribution of the estimator: binomial distribution
- *SE*: $$SE_{prop} \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}} = \ldots$$
- the values give a binomial distribution, which we have to normalize. It is possible to calculate with this function, (Exact method) but complicated
- it is easier to use instead normal approximation: Wald-interval (simple but unreliable method)



distribution of proportion estimates and its normal approximation

Interval Estimations ...

... of Proportions

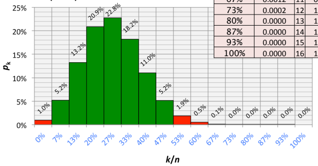
Method #1: The Exact Method

- calculate the probability of each individual outcome (k) given your estimation for p and order them by descending probability
- add these probabilities up beginning with the most probable, then the second most and so on
- when your sum exceeds some preset limit, stop
- the range of outcomes included in the final sum is the exact CI, the final sum is the confidence level (1–α)

k/n	p_k	#	SUMIF
0%	0.0095	9	0.9933
7%	0.0520	6	0.9134
13%	0.1324	4	0.7510
20%	0.2087	2	0.4364
27%	0.2277	1	0.2277
33%	0.1821	3	0.6185
40%	0.1104	5	0.8614
47%	0.0516	7	0.9650
53%	0.0188	8	0.9838
60%	0.0053	10	0.9986
67%	0.0012	11	0.9998
73%	0.0002	12	1.0000
80%	0.0000	13	1.0000
87%	0.0000	14	1.0000
93%	0.0000	15	1.0000
100%	0.0000	16	1.0000

Example #3:

- use the data of example #1
- see the calculated table and the graph representing **the 95% (actually, 96.5%) CI: 7% – 47%**

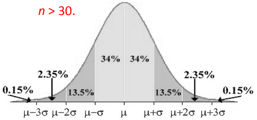


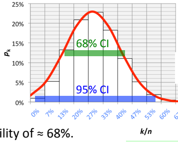
Interval Estimations ...

... of Proportions

This approximation works “well” when $n > 30$.

Method #2: The Normal Approximation (Á. Wald)





Ábrahám **Wald**
1902–1950

To the interval between $\mu \pm \sigma$ belongs a probability of ≈ 68%. The corresponding estimation interval is called **68% Confidence Interval (*CI*)**, the probability **confidence level (1–*α*)**.
For example #3: 68% CI = 15% – 38%. (see figure to the right)

$$68\%\,CI \approx \frac{k}{n} \pm \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$$

To the interval between $\mu \pm 2\sigma$ belongs a probability of ≈ 95%. This corresponds to the **95% Confidence Interval**, used very extensively in health sciences. For small sample sizes the CI can even stretch out from the [0,1] range! –> trimming is needed.
For example #3: 95% CI = 4% – 50%. (see figure to the right and cf. to exact values on previous slide.)

$$95\%\,CI \approx \frac{k}{n} \pm 2 \cdot \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$$

Interval Estimations ...

... of Proportions

Problem #6: We would like to estimate the *prevalence* (proportion in the population) of the Rhesus factor among Budapest citizens. We randomly chose 42 people and determined their blood group: 35 of them proved to be Rh+.
a) Give a point estimation for the prevalence of the Rhesus factor.
p(Rh+) ≈ $\hat{p}$(Rh+) = k(Rh+)/n = 35/42 = **0.833**
b) Give the 95% CI using the Wald-interval method.
First find the *SE*:

$$SE \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}} = \sqrt{\frac{\frac{35}{42}\left(1-\frac{35}{42}\right)}{42}} = 0.00331$$

The 95% CI is approx. $\hat{p}$(Rh+) ± 2*SE* = **0.833 ± 0.006** or **0.826 – 0.839** .

Problem #7: Give the 95% confidence interval for the prevalence of blue eye color among first year students, if a sample of 10 students contained 2 blue-eyed students.
Using the learned formulae: $\hat{p}$(Rh+) = 0.200 and *SE* = 0.126. This would yield the following 95% CI: –0.052 – 0.452 . However, since we are estimating a probability, the CI cannot stretch out from the [0,1] interval, so after trimming the CI is **0 – 0.452**. Here the confidence level is for sure not anywhere close to 95% anymore. This is an example to show how unreliable the Wald interval is – still we can use it for not too small samples as a quick and easy estimation.

Interval Estimations ...

... of Expected Values

Using Student’s t-distribution (W. S. Gosset)

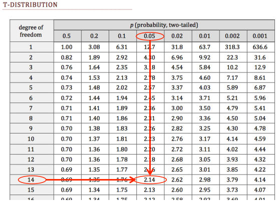
Similar intervals can be defined as in the case of proportions, here, however, we are using the Student *t*-distribution with n–1 degrees of freedom (*df*). Therefore we should not use the previous approximative ranges ($\mu \pm \sigma$ and $\mu \pm 2\sigma$).

Instead, we have to use a look-up table or the Excel command for two-tailed *p*: =T.INV.2T(probability, deg_freedom)

Example #4:

- use data of example #2
- find the *t*-value for 95% CI:
  - in Excel: use the =T.INV.2T(5%, 15-1) function, which yields 2.1445
  - (in the Formula Collection use the look-up table: 2.14)

this is α, i.e., 1 – confidence level



T-DISTRIBUTION

degree of freedom	*p* (probability, two-tailed)							
	0.5	0.2	0.1	0.05	0.02	0.01	0.002	0.001
1	1.00	3.08	6.31	12.7	31.8	63.7	318.3	636.6
2	0.82	1.89	2.92	4.30	6.96	9.92	22.3	31.6
3	0.76	1.64	2.35	3.18	4.54	5.84	10.2	12.9
4	0.74	1.53	2.13	2.78	3.75	4.60	7.17	8.61
5	0.73	1.48	2.02	2.57	3.37	4.03	5.89	6.87
6	0.72	1.44	1.94	2.45	3.14	3.71	5.21	5.96
7	0.71	1.41	1.89	2.36	3.00	3.50	4.79	5.41
8	0.71	1.40	1.86	2.31	2.90	3.36	4.50	5.04
9	0.70	1.38	1.83	2.26	2.82	3.25	4.30	4.78
10	0.70	1.37	1.81	2.23	2.76	3.17	4.14	4.59
11	0.70	1.36	1.80	2.20	2.72	3.11	4.02	4.44
12	0.70	1.36	1.78	2.18	2.68	3.05	3.93	4.32
13	0.69	1.35	1.77	2.16	2.65	3.01	3.85	4.22
14	0.69	1.35	1.76	2.14	2.62	2.98	3.79	4.14
15	0.69	1.34	1.75	2.13	2.60	2.95	3.73	4.07
16	0.69	1.34	1.75	2.12	2.58	2.92	3.69	4.01

Figure: Finding the *t*-value in a look-up table if α = 5% and *df* = 14.

Interval Estimations ...

... of Expected Values

Using Student’s t-distribution (W. S. Gosset)

In the Statistics Lab, we will use Excel to determine the *t*-value. However, *t* has a standardized distribution (μ = 0 and σ = 1), so we have to use the $x = \bar{x} + t \cdot SE$ formula to find the corresponding *x*-values (*x* is our statistical variable, in this case, heigth). Use –t to find the lower limit of the CI (x_L) and +t to find the upper limit of the CI (x_U):

$x_L = \bar{x} - t \cdot SE$ = 170 cm – 2.1445 * 2.5 cm = 164.6 cm

$x_U = \bar{x} + t \cdot SE$ = 170 cm + 2.1445 * 2.5 cm = 175.4 cm

The 95% CI is 164.6 cm – 175.4 cm.

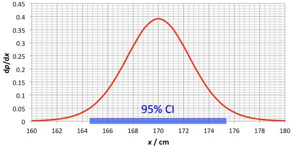


Figure: A graph showing the theoretical distribution of sample means (a non-standard *t*-distribution with *df* = *n*–1, $\mu = \bar{x}$ and $\sigma = SE$) and the 95% CI that we have just calculated.

Interval Estimations ...

... of Expected Values

Problem #8: We would like to estimate the expected value of blood cholesterol level and give a 95% confidence interval for our estimation. We took a sample of 8, the observed values (in mg/dL) are summarized in the table.

No. of Observation	Cholesterol Level in mg/dL
1	174
2	161
3	139
4	168
5	143
6	149
7	120
8	169

We will use Excel to do calculations.

The point estimator of the expected value is the sample mean:

$\mu \approx \bar{x}$ = AVERAGE(data) = **152.9 mg/dL**

Now find the SE using the methods learned before:

n =COUNT(*data*) = 8

SD =STDEV.S(*data*) = 18.47 mg/dL

SE =*SD*/SQRT(*n*) = 6.53 mg/dL

The sample means follow a t-distribution, so we have to find the limits of the 95% CI:

df = *n* – 1 = 7

t =T.INV.2T(5%,*df*) = 2.365

$x_L = \bar{x} - t * SE$ = **137.4 mg/dL**

$x_U = \bar{x} + t * SE$ = **168.3 mg/dL**

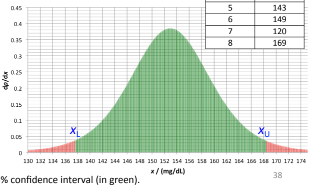


Figure: 95% confidence interval (in green).

Interval Estimations ...

... of Expected Values

Problem #9: We would like to determine the 95% CI for the mean height of freshmen, but we want the CI to be not wider than 1 cm. What should be the sample size? We know from literature that the variance of SD of human height in general is 5 cm and we suppose this value applies for freshmen, too.

Here we need to think “backward”:

– We have a CI that is between a (yet) unknown x_L and x_U.

– The width of the CI is the difference between these limits: *width* = $x_U - x_L$

– Substitute here their formulae: *width* = $(\bar{x}+t*SE) - (\bar{x}-t*SE) = 2*t*SE$

– Now we have a problem: the *t*-value itself depends on the sample size (actually, on degrees of freedom)! We have to make an **assumption** here: suppose that the **sample will be big** enough and just **use *t* = 2**; the formula will look like this: *width* = 4*SE*.

– We know that *SE* = $SD/n^{0.5}$, insert this into the above formula: *width* = 4*SD*/n0.5.

– Transpose the formula to get the necessary sample size: n = $(4SD/width)^2$

– Plug in our data: *n* = $(4*(5\text{ cm})/(1\text{ cm}))^2$ = **400**

Interval Estimations

Confidence level ($1 - \alpha$): the probability that confidence intervals determined by a certain method contain the true value. It does not tell if any *actual* CI contains the true value. **It characterizes the estimation process in general**, not the actual outcome. You cannot tell whether the true value is in your actual CI, only if you somehow *know* the true value – in which case the whole estimation process would be pointless.

Significance level (α): the complementary probability, e.g. if the confidence level is 95%, the significance level is 5%. Which one is used depends on context.

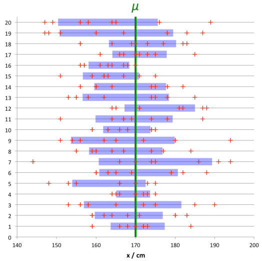
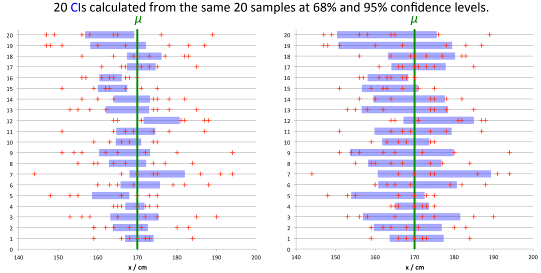


Figure: simulation of 20 estimations of the expected value (μ) of stature (body height) using the same process: taking a random sample of eight units (red +), calculating the mean and the standard error and then the CI is: mean ± 2.36 × SE (blue ribbon).
The true values are: μ = 170 cm, σ = 10 cm

Interval Estimations

20 CIs calculated from the same 20 samples at 68% and 95% confidence levels.



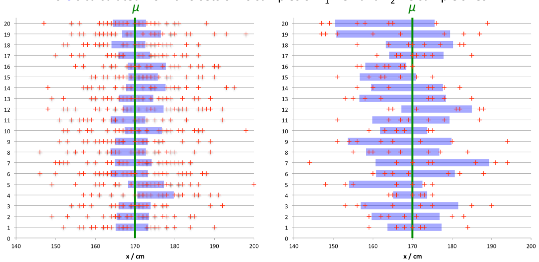
n = 8 (df = 7), μ = 170 cm, σ = 10 cm
$1 - \alpha$ = 68%

n = 8 (df = 7), μ = 170 cm, σ = 10 cm
$1 - \alpha$ = 95%

Higher confidence level =
lower chance of missing the true value but less information content.

Interval Estimations

20 CIs calculated from two sets of 20 samples of n_1 = 34 and n_2 = 8 sample sizes.



n = 32 (df = 31), μ = 170 cm, σ = 10 cm
$1 - \alpha$ = 95%

n = 8 (df = 7), μ = 170 cm, σ = 10 cm
$1 - \alpha$ = 95%

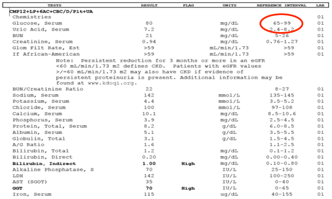
Bigger sample size at same confidence level = narrower CI.

Appendix: Normal Range

Normal range, reference range, or reference interval is an interval of the statistical variable that contains a randomly chosen **element with 95% probability**.

The normal range is also a kind of estimation, but here we estimate the spread of the statistical variable, not a parameter of its distribution. In other words: the normal range is a 95% CI for the data themselves. This is also the range indicated in diagnostic lab reports.

In case of normally distributed data the normal range can be estimated by $\bar{x} \pm 2SD$. (more precisely: $\bar{x} \pm 1.96SD$)



TESTS	RESULT	FLAG	UNITS	REFERENCE INTERVAL	LAB
CMP12+LP+6AC+CBC/D/Plt+UA					
Chemistries					01
Glucose, Serum	80		mg/dL	65-99	01
Uric Acid, Serum	7.2		mg/dL	2.4-8.2	01
BUN	21		mg/dL	5-26	01
Creatinine, Serum	0.94		mg/dL	0.76-1.27	01
Glom Filt Rate, Est	>59		mL/min/1.73	>59	01
If African-American	>59		mL/min/1.73	>59	01
Note: Persistent reduction for 3 months or more in an eGFR <60 mL/min/1.73 m2 defines CKD. Patients with eGFR values >/=60 mL/min/1.73 m2 may also have CKD if evidence of persistent proteinuria is present. Additional information may be found at www.kdoqi.org.					
BUN/Creatinine Ratio	22			8-27	01
Sodium, Serum	142		mmol/L	135-145	01
Potassium, Serum	4.4		mmol/L	3.5-5.2	01
Chloride, Serum	100		mmol/L	97-108	01
Calcium, Serum	10.1		mg/dL	8.5-10.6	01
Phosphorus, Serum	3.9		mg/dL	2.5-4.5	01
Protein, Total, Serum	8.2		g/dL	6.0-8.5	01
Albumin, Serum	5.1		g/dL	3.5-5.5	01
Globulin, Total	3.1		g/dL	1.5-4.5	01
A/G Ratio	1.6			1.1-2.5	01
Bilirubin, Total	1.2		mg/dL	0.1-1.2	01
Bilirubin, Direct	0.20		mg/dL	0.00-0.40	01
Bilirubin, Indirect	**1.00**	**High**	mg/dL	0.10-0.80	01
Alkaline Phosphatase, S	70		IU/L	25-150	01
LDH	142		IU/L	100-250	01
AST (SGOT)	35		IU/L	0-40	01
GGT	**70**	**High**	IU/L	0-65	01
Iron, Serum	115		ug/dL	40-155	01

Problem #10: Calculate the μ, σ, *and* σ^2 for serum glucose level using the lab report normal range.

$x_L = \mu - 2\sigma$ = 65 mg/dL
$x_U = \mu + 2\sigma$ = 99 mg/dL

$x_U + x_L = (\mu+2\sigma) + (\mu-2\sigma) = 2\mu$
$\mu = (x_U + x_L)/2$ = **82 mg/dL**

$x_U - x_L = (\mu+2\sigma) - (\mu-2\sigma) = 4\sigma$
$\sigma = (x_U - x_L)/4$ = **8.5 mg/dL**

σ^2 = **72.25 (mg/dL)^2**

Follow-up Questions

– What is a population?

– What is a sample?

– How can we obtain information about a statistical variable?

– What are the types of estimations?

– What are the steps of an estimation?

– What is a plug-in estimator?

– What is a point estimation?

– What are the disadvantages of point estimations?

– What are the point estimators of probability, expected value, theoretical variance and theoretical standard deviation?

– What are the properties of a “good estimator”?

– What is unbiasedness? Illustrate with examples.

– What is efficiency? How can it be mathematically expressed?

– What distribution does the estimate of proportion follow?

– What distribution does the estimate of the expected value follow?

– What is standard error?

– How can we calculate the stadard error of a proportion?

– How can we calculate the standard error of a mean?

– How does standard error depend on the standard deviation of the variable?

– How does standard error depend on the sample size?

– We want to triple the efficiency of our estimation. How should we change the sample size?

– What is the maximum standard error for proportion, if we take a sample of 25?

Follow-up Questions

– What are the properties of a consistent estimator?

– What does it mean that an estimate is sufficient?

– What is Bessel’s correction? Where do we use it and what is its purpose?

– What is the meaning of CI?

– What is the meaning of confidence level?

– How can we give an exact CI for proportion?

– What is the ground of a Wald-interval and what is its advantage and disadvantage?

– How does the CI change if I increase the confidence level?

– How does the CI change if I increase the sample size?

– How does the CI change if the standard deviation of the variable becomes less?

– How do we obtain a CI for the estimation of the expected value?

– What is a reference range?

– How can we give the reference range for a normally distributed statistical variable?

– We can see the reference ranges for different blood tests in our diagnostic lab report. How can we obtain from these the expected values and the standard deviations used by the hospital?

– What is the relationship between a confidence level and the corresponding significance level?

– Why is it not possible to tell if a CI actually contains the true value?