Medical Statistics, Informatics and Telemedicine

Lecture 4

Estimation and Confidence

1st October 2021

Gergely Agócs

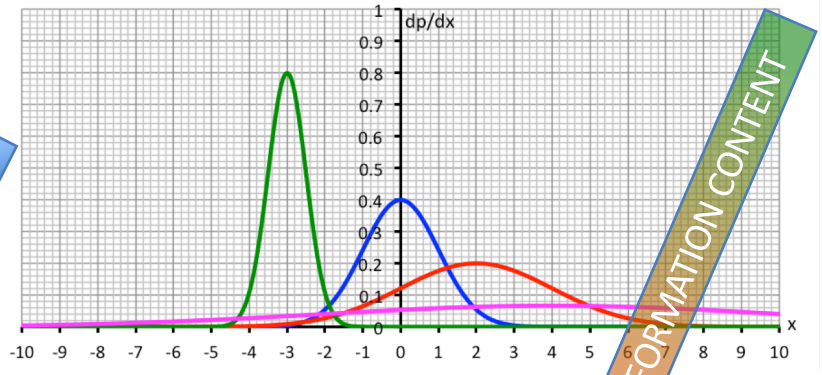
Topics of the Lecture

- The aim, types and process of **estimation**
- The **error of estimation** and its assessment
- The **confidence of estimation** and its assessment

The Aim of Estimation

Population

Distributions



Parameters

DECREASING INFORMATION CONTENT

Individual **Probabilities**

How to determine these?

The Aim of Estimation

How to determine these?

Example: coin tossing

Creating a model
(only theoretical considerations)

Experiment, observation

The Process of Estimation

Population

Sample

Sampling

RANDOMNESS!

MATH

characterization of the sample

INFERENCE

drawing **conclusions**

UNCERTAINTY!

Sampling and Sampling Errors

What is a good sample like?

Typical sampling strategies

Some common problems

Types of Estimation

What percentage of people have an Rh+ blood type?

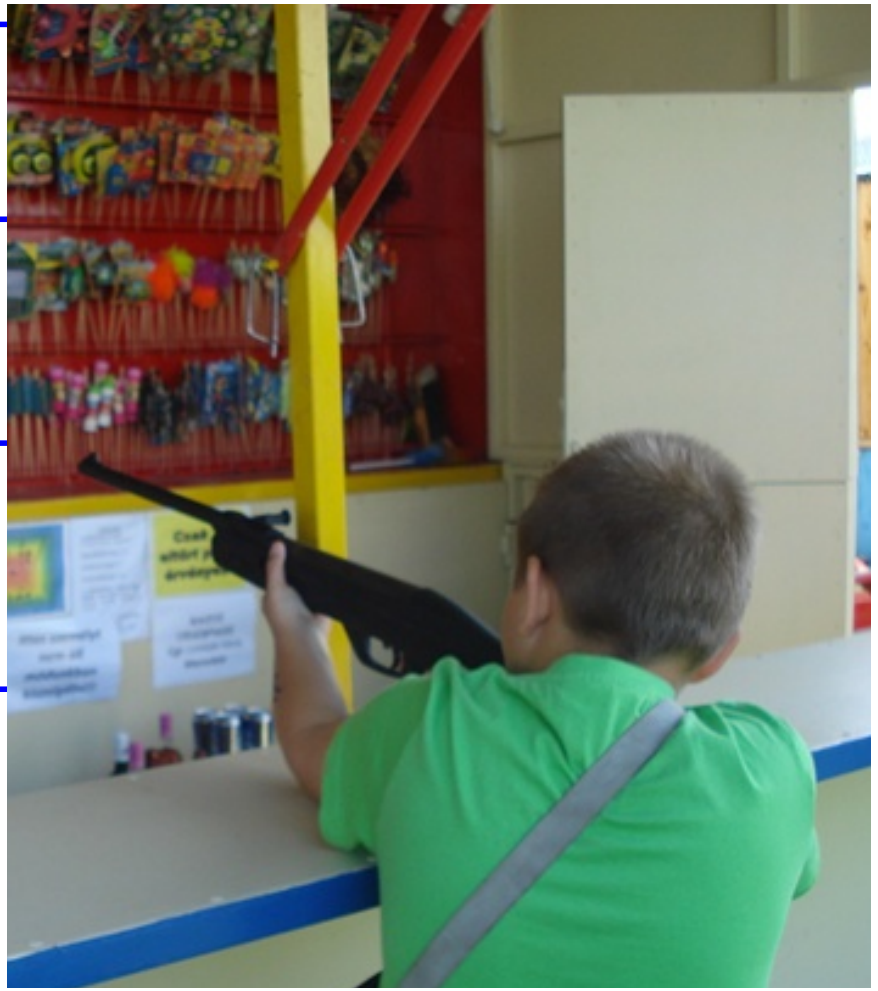
Point Estimation

Interval Estimation

Point Estimation

Theoretical values (of the population):
„AIM"

Estimators (from the sample):
„SHOT"



Point Estimation

Theoretical values (of the population): „**AIM**”	**Estimators (from the sample)**: „**SHOT**”
- probability or proportion (p = *probability*)	- relative frequency ($\hat{p}$ [p hat])
- expected value, „population mean” ($E(\xi)$ = *expected value* or μ [mu])	- sample mean ($\bar{x}$ [x bar])
- theoretical variance ($Var(\xi)$ = *variance* vagy σ^2 [szigma négyzet])	- corrected sample variance (s^2)
- theoretical standard deviation ($SD(\xi)$ = *standard deviation* vagy σ [szigma])	- corrected sample standard deviation (s, SD = *standard deviation*)

The Error of the Estimation

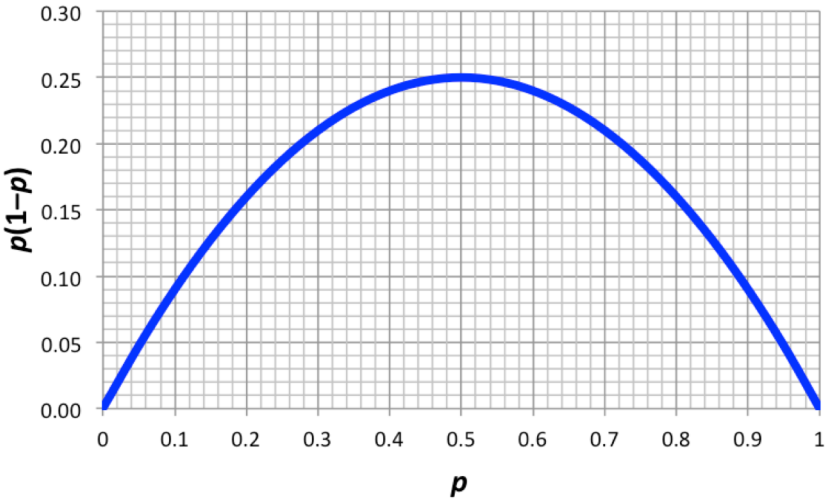
- **the estimator is itself a random variable**

- the theoretical standard deviation of the estimator is called **standard error**

(Addendum: Standard Error of Proportion)

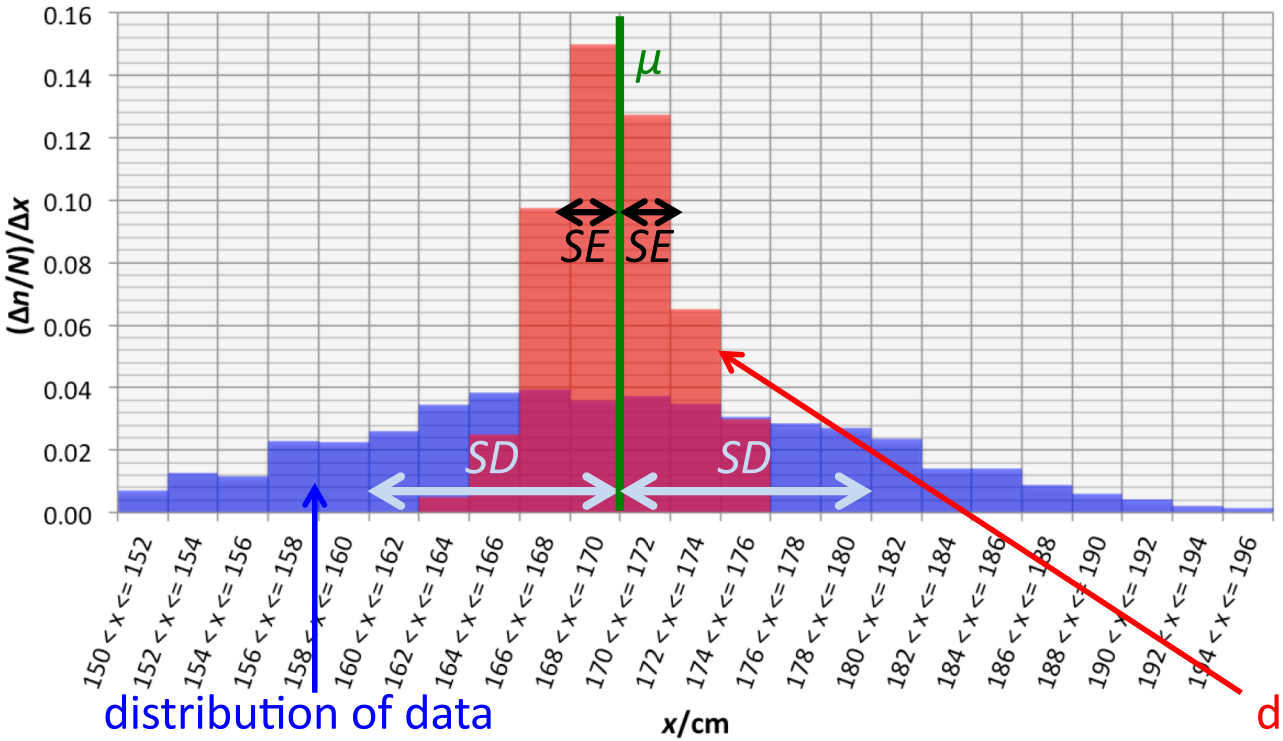
$$SE_{\text{prop}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$$

(Addendum: Standard Error of Proportion)



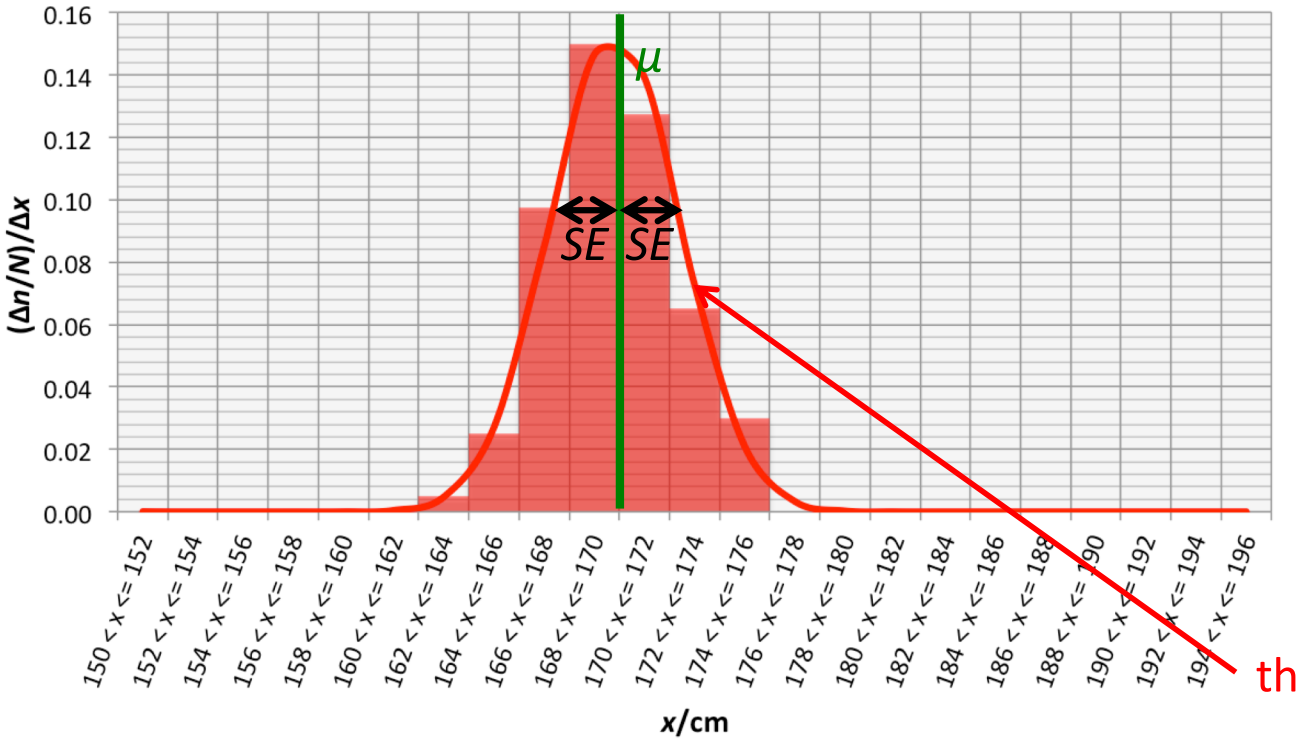
$$\max\left(SE_{\text{prop}}\right) = \sqrt{\frac{0.5(1-0.5)}{n}} = \sqrt{\frac{0.25}{n}} = \frac{1}{\sqrt{4n}}$$

Standard Error of the Mean



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

Standard Error of the Mean

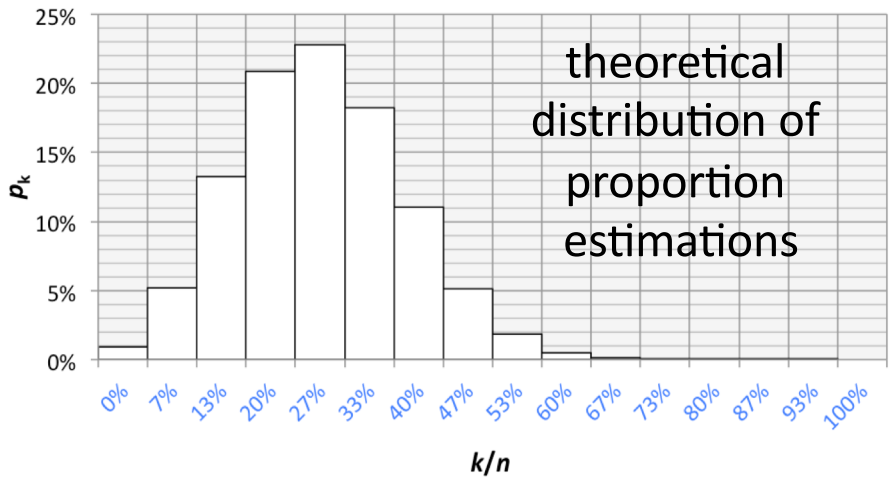


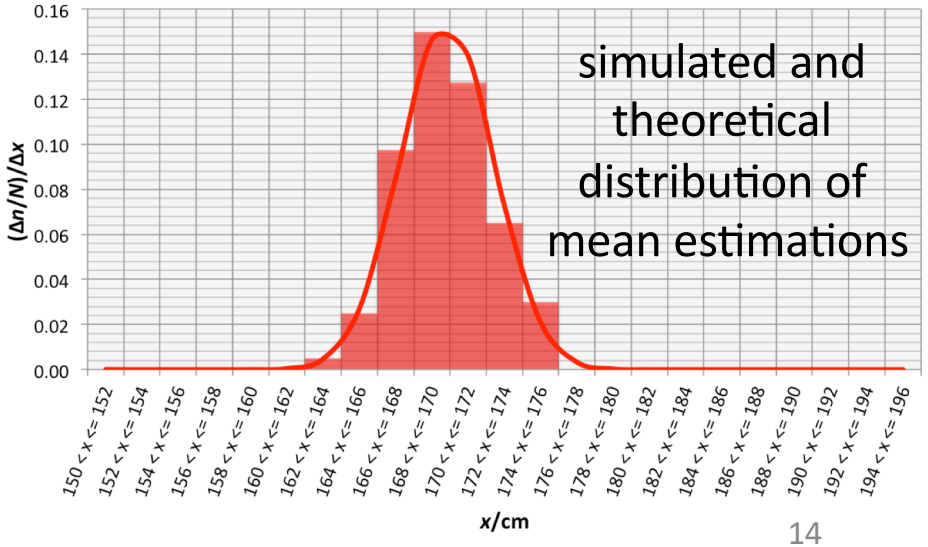
William S. **Gosset**
1876–1937
"Student"

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

theoretical model: *t*-distribution

Interval Estimation

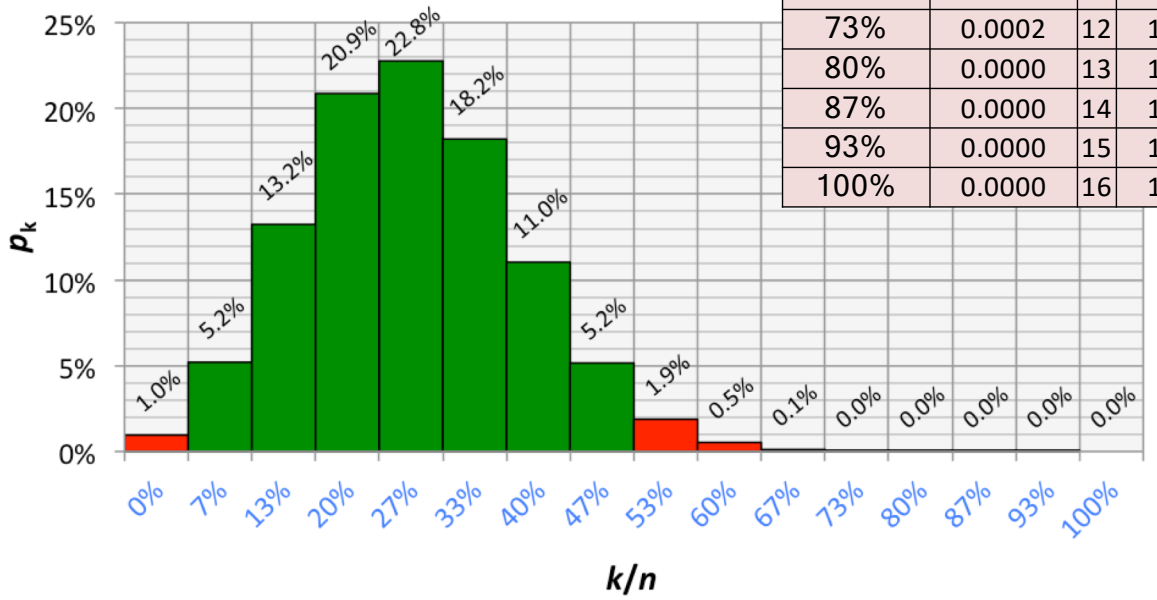




Interval Estimation of Proportion

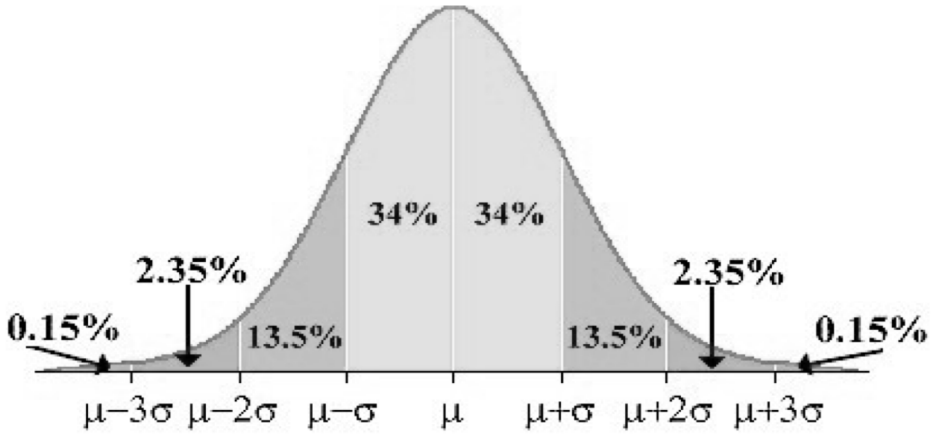
Method 1: exact binomial

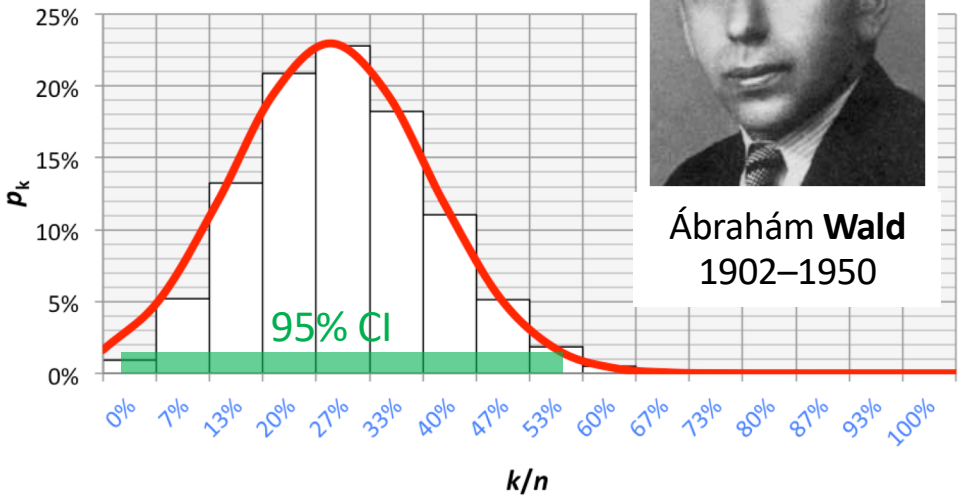
k/*n*	p_k	#	SUMIF
0%	0.0095	9	0.9933
7%	0.0520	6	0.9134
13%	0.1324	4	0.7510
20%	0.2087	2	0.4364
27%	0.2277	1	0.2277
33%	0.1821	3	0.6185
40%	0.1104	5	0.8614
47%	**0.0516**	**7**	**0.9650**
53%	0.0188	8	0.9838
60%	0.0053	10	0.9986
67%	0.0012	11	0.9998
73%	0.0002	12	1.0000
80%	0.0000	13	1.0000
87%	0.0000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



Interval Estimation of Proportion

Method 2: normal approximation





Ábrahám **Wald**
1902–1950

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

Interval Estimation of the Mean

observation	cholesterol level (mg/dL)
1	174
2	161
3	139
4	168
5	143
6	149
7	120
8	169

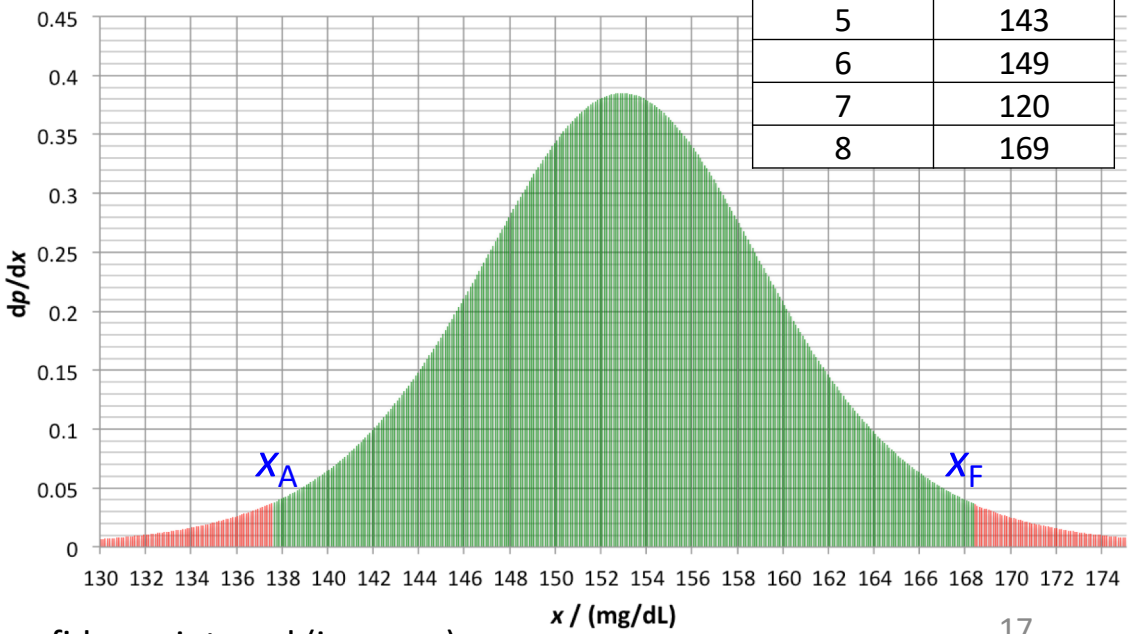


Fig.: 95% confidence interval (in green)

Interval Estimation

Confidence level (1 – α): the probability that a CI calculated in a certain way will contain the estimated value.

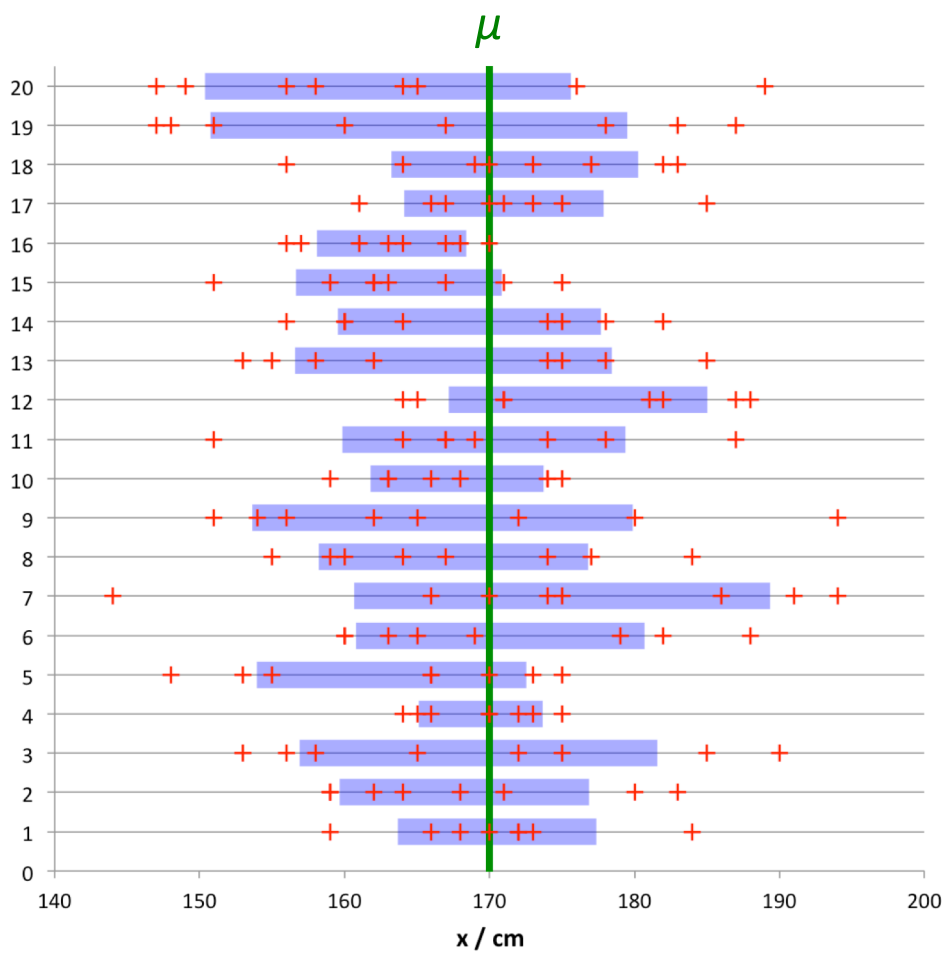
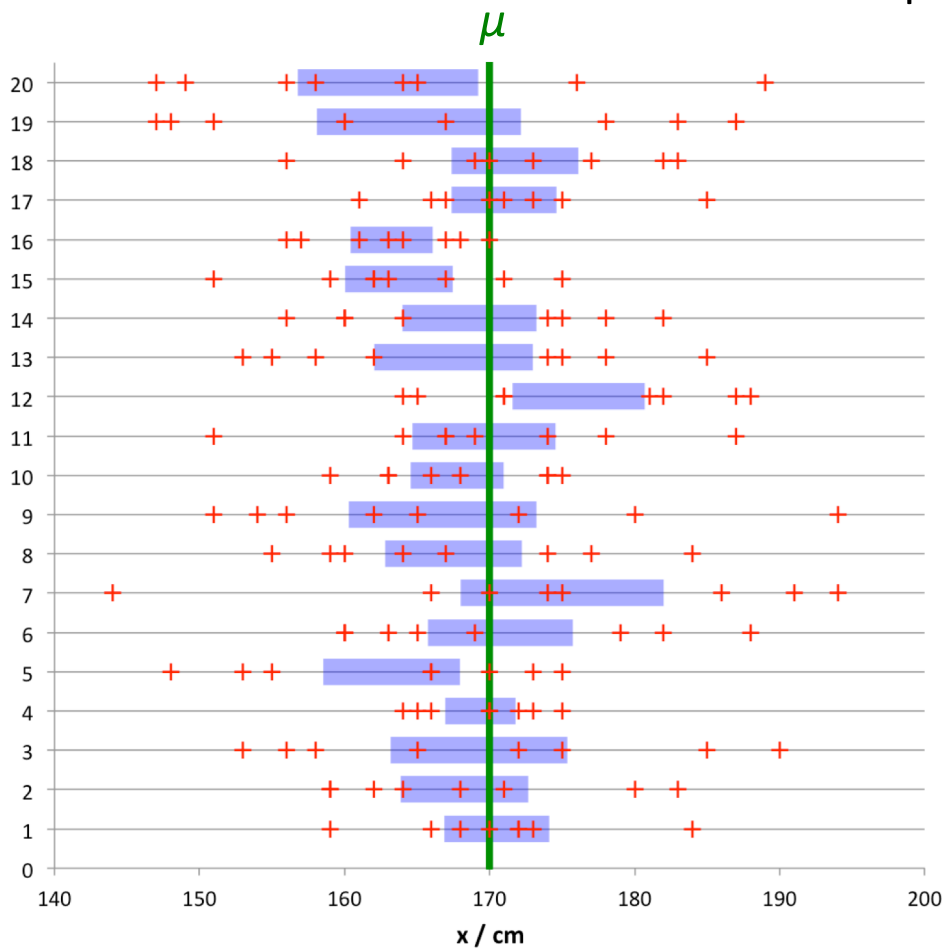


Fig: Simulation of 20 interval estimations for the expected value (μ) of body height using the same method: sample of 8 elements (red +), calculation of the mean and SD, then of 95% CI: *mean* ± 2,36 × *SE* (blue band). The estimated theoretical values are: μ = 170 cm, σ = 10 cm.

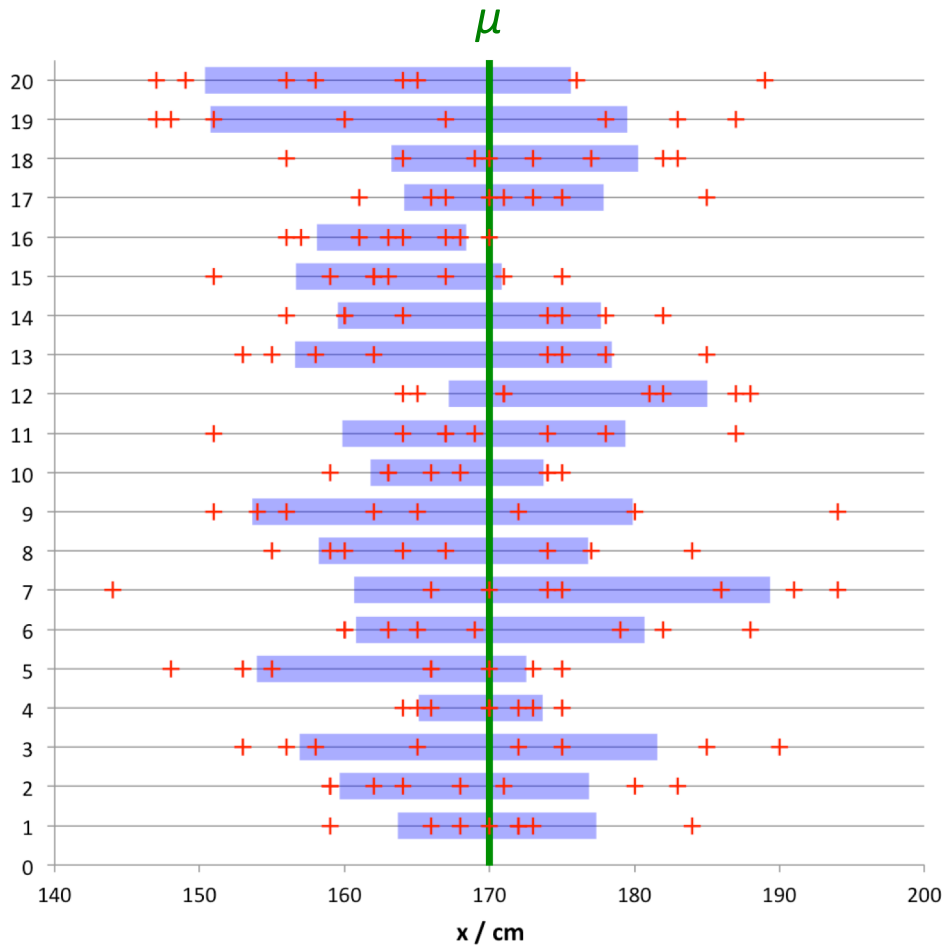
Interval Estimation

20 CI calculated for the same 20 samples at both 68% and 95% confidence levels.



n = 8, μ = 170 cm, σ = 10 cm

$1 - \alpha$ = 68%



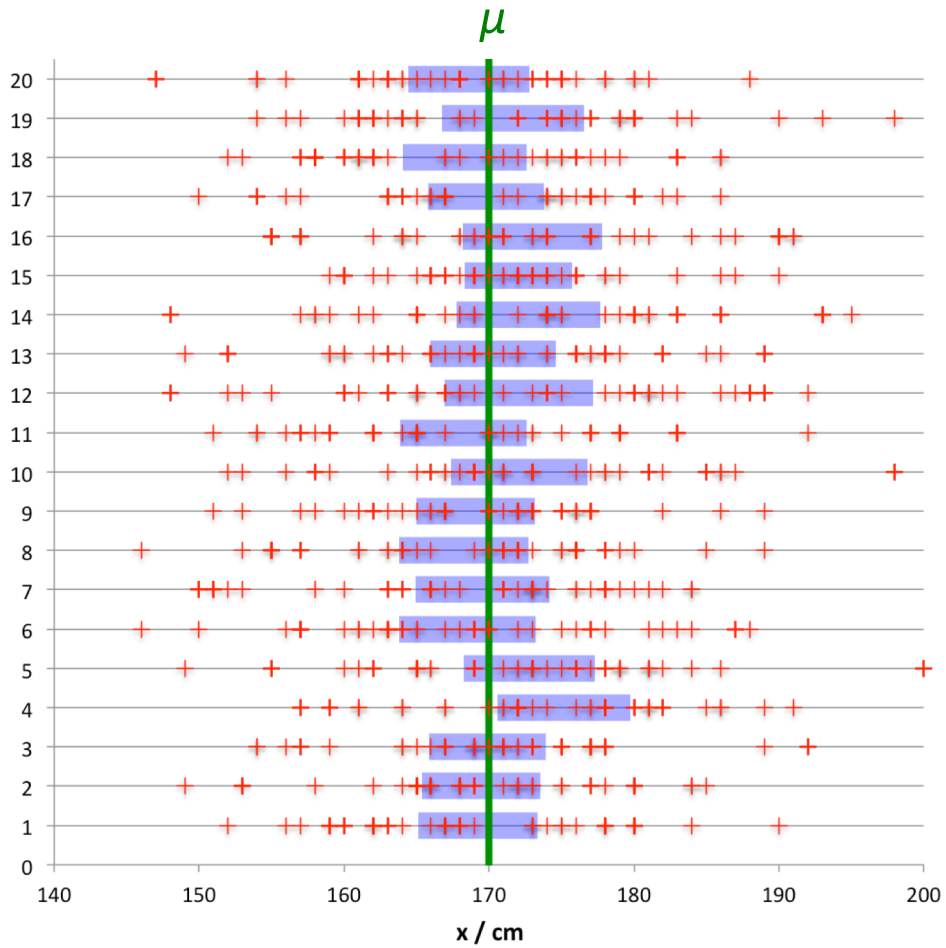
n = 8, μ = 170 cm, σ = 10 cm

$1 - \alpha$ = 95%

Higher level =
lower chance for the CI to miss the estimated value, but less information content.

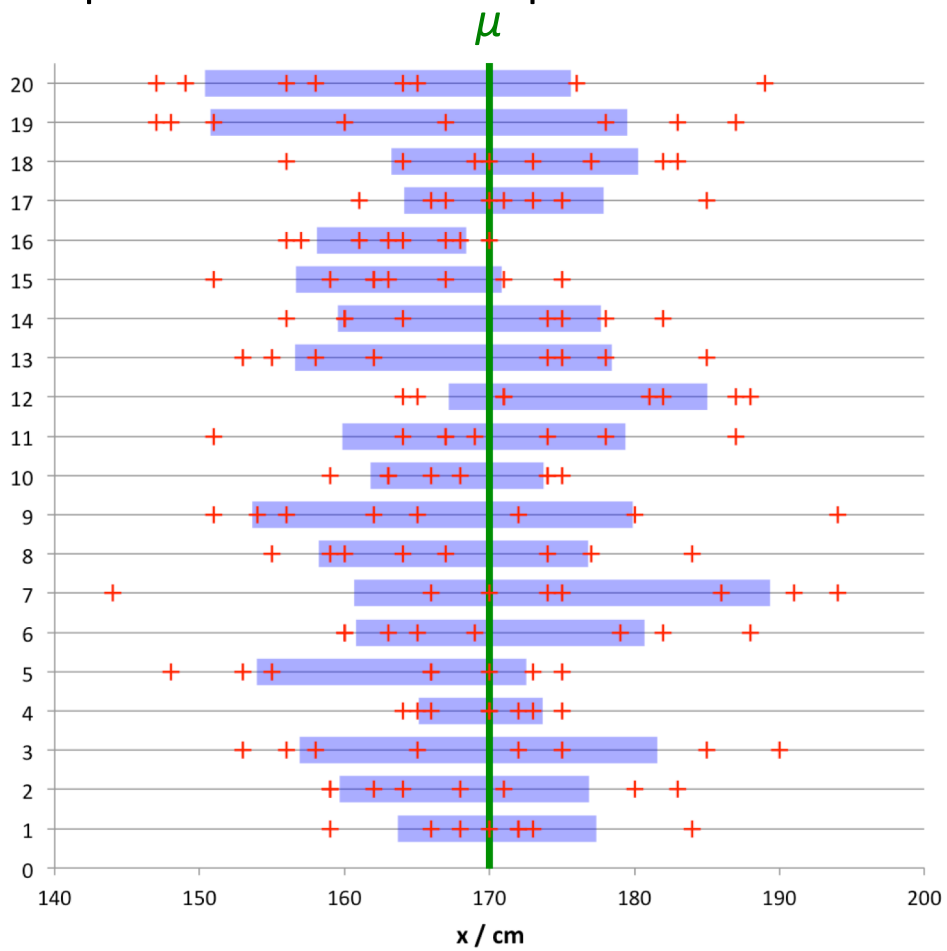
Interval Estimation

calculation of 20–20 CIs with 95% level for 20 samples of 32 and 20 samples 8 elements.



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, same level => narrower CI.
(4 times the sample size => ½ width of CI)

Appendix: Normal Range

Normal range, reference range, or reference interval: a range for the statistical variable, which contains a random element with **95% probability**.

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