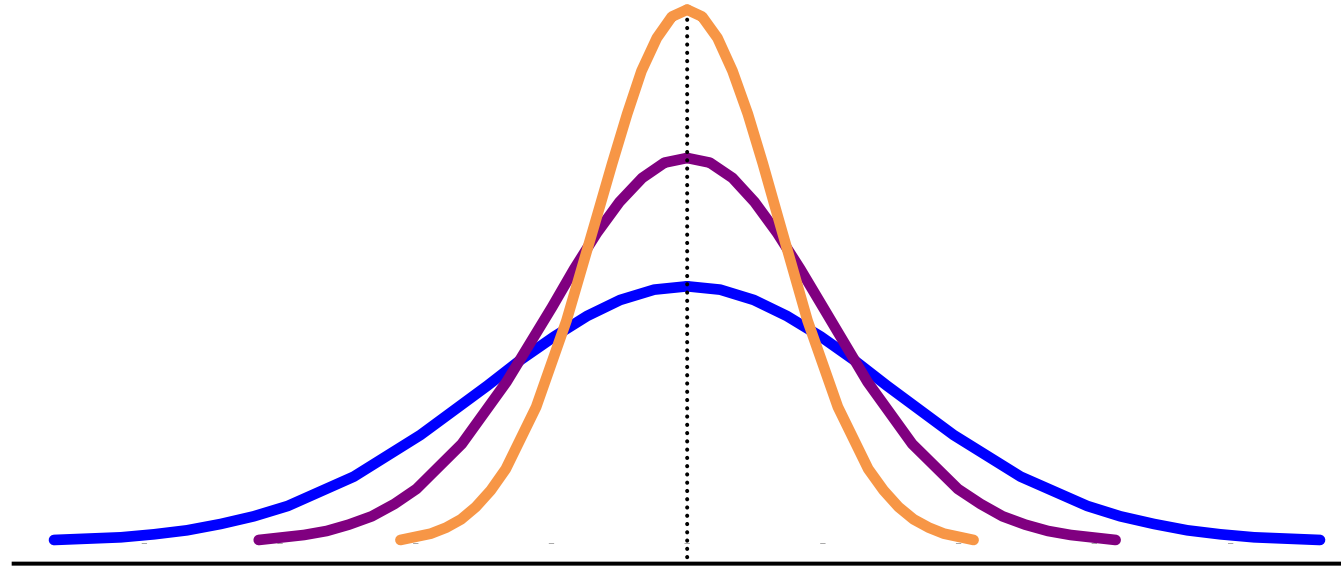




Author: PhD. MsC. Bondor Cosmina-Ioana

Confidence intervals

- A** ALWAYS
- S** SEEK
- K** KNOWLEDGE

Objectives

Sampling distribution

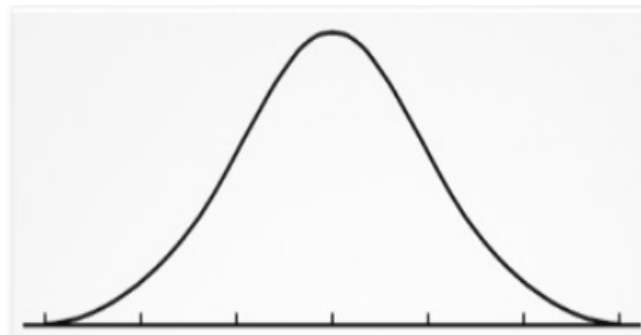
Estimation

- Point estimation

- Confidence intervals estimation

Remember

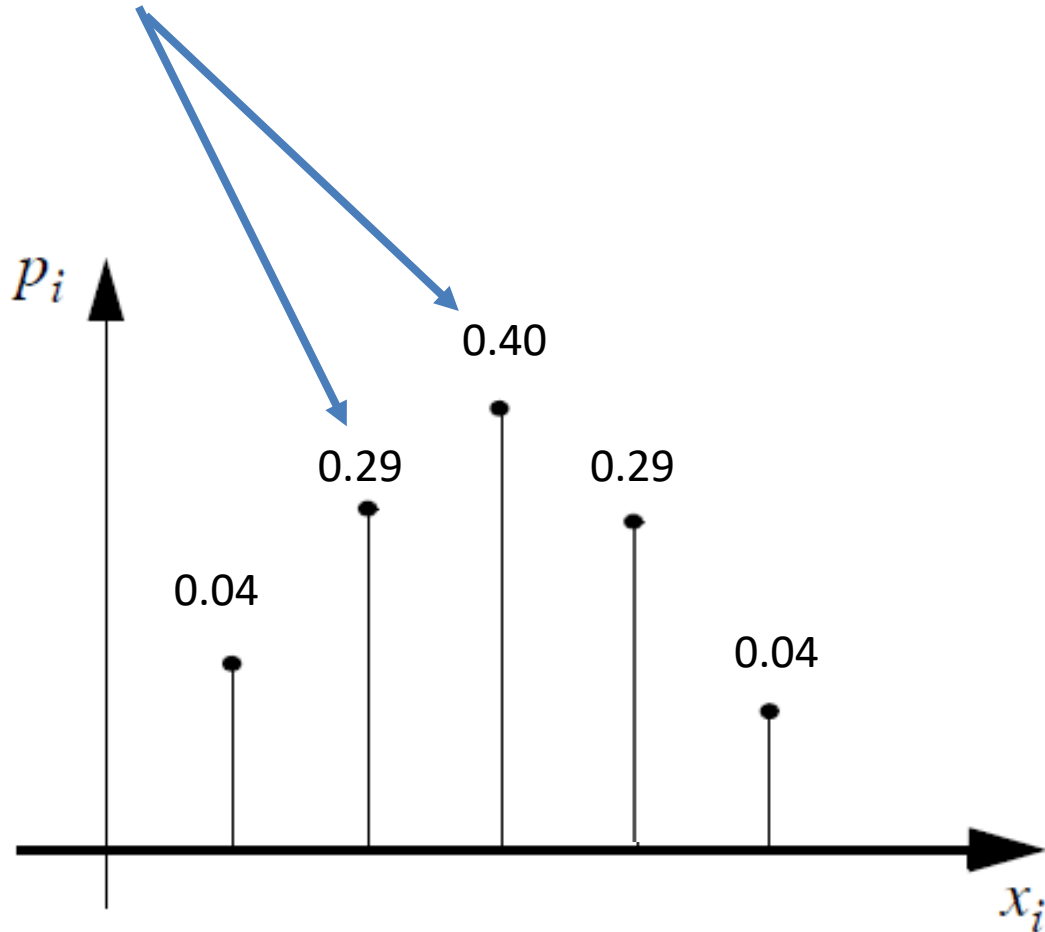
- arithmetic mean of the sample $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$
- standard deviation of the sample $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}}$
- standard error of the sample $SE = \frac{s}{\sqrt{n}}$
- normal distribution



Remember

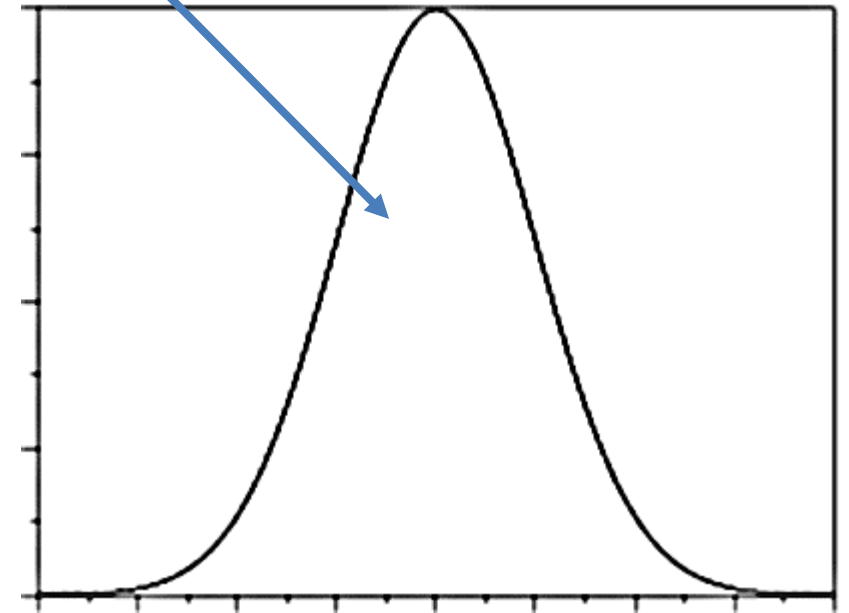
Finite

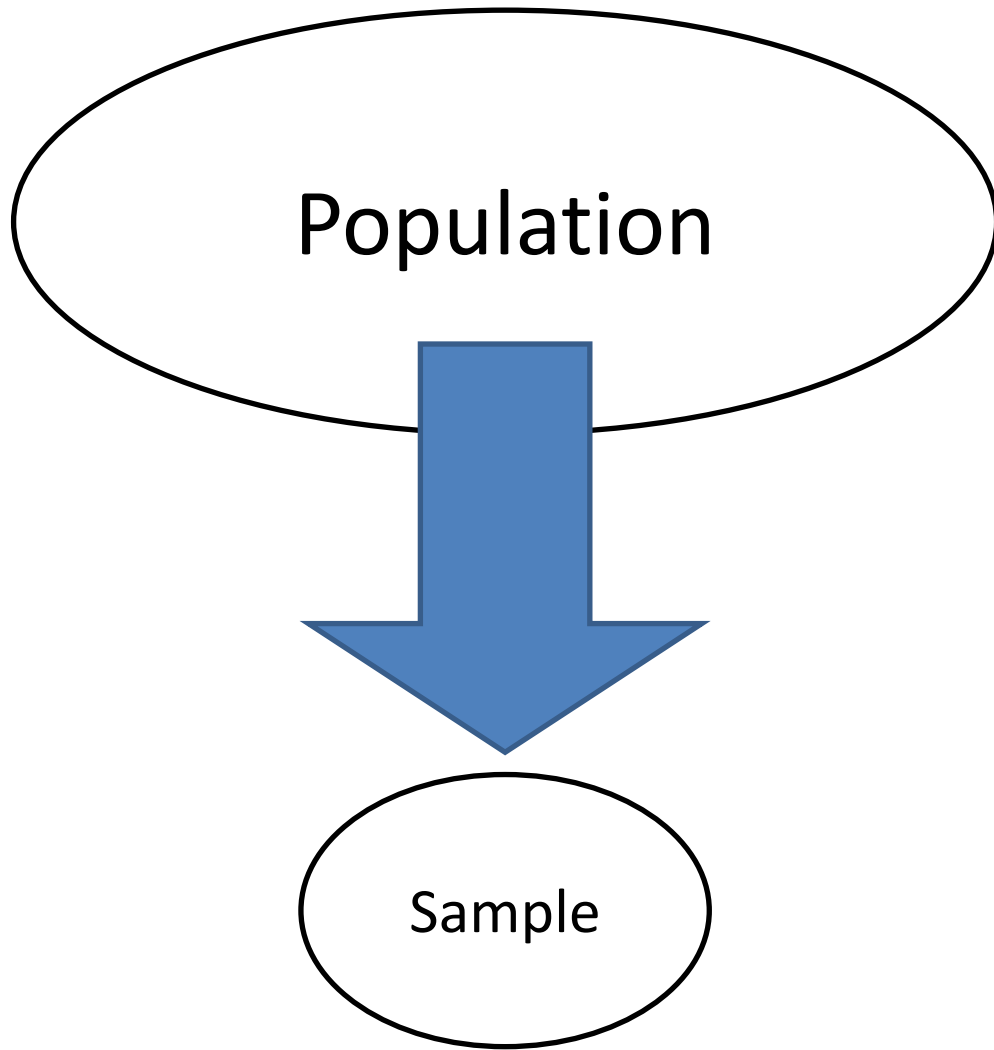
Sum of probabilities = 1



Infinite

Area under the curve = 1





Unknown μ (inaccessible)



statistical inference =
we want to estimate

We are in this
point

Known $\bar{X}$ (accessible)

population P

– we are interested in a characteristic:

1. We select a random sample from the population

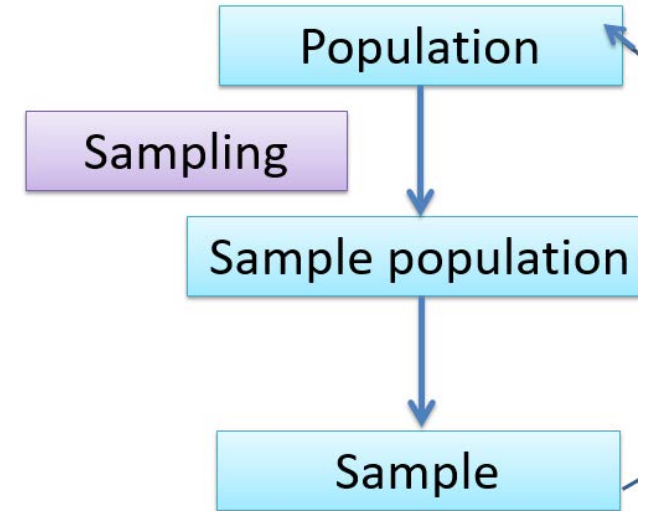
2. By means of descriptive statistics calculate:

- Qualitative: observed frequency

or

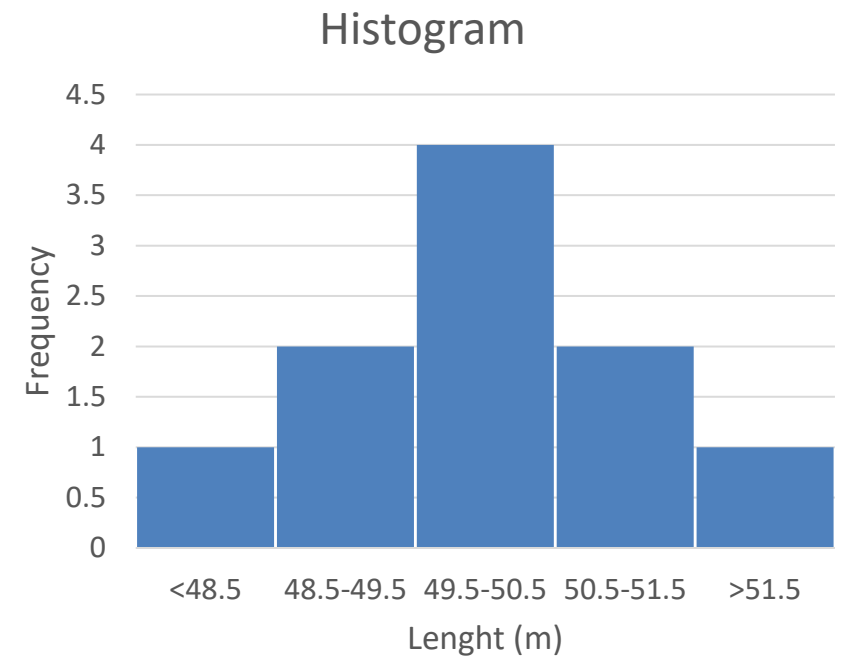
- Quantitative: mean and standard deviation

- What will happen if we repeat the study?



The wall

- A very long wall
- A tape of 5m
- We measure the wall → 50.1 m
- We measure the wall again → 49.9 m
- again → 50.3 m
- again → 50 m
- again → 50.1 m
- again → 50 m
- ...



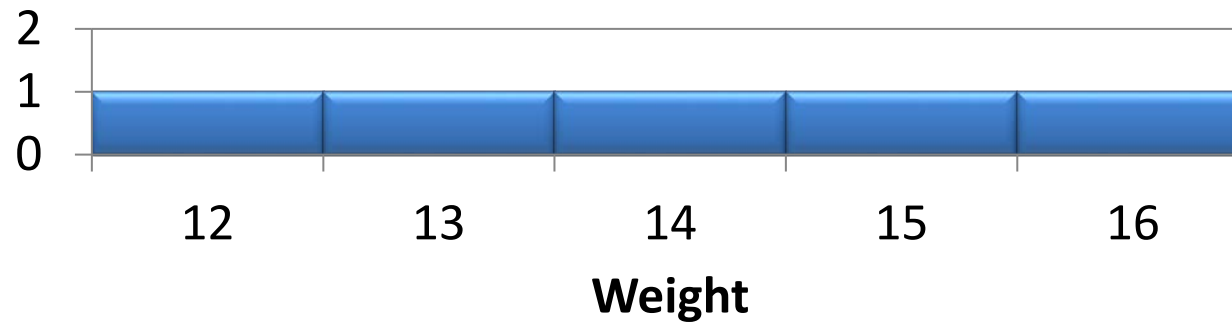
Normal distribution of the repeated measures

Sampling distribution

Even in population of 5 we get normal distribution of repeated measurements

Population - 5 boys two years old

- their weights: 12, 13, 14, 15, 16 kg
- This is the **population**

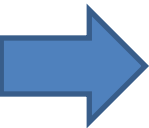


all possible samples of 2 subjects

- Population 5 boys
- How many samples of 2 boys can be taken?
- 11 21 31 41 51
- 12 22 32 42 52
- 13 23 33 43 53
- 14 24 34 44 45
- 15 25 35 45 55

All possibilities of samples of two boys from 5 = 25

Sample no.	First boy	Second boy	Weight for the first boy	Weight for the second boy	Arithmetic Mean for the weights (kg)
1	1	1	12	12	12
2	1	2	12	13	12.5
3	1	3	12	14	13
4	1	4	12	15	13.5
5	1	5	12	16	14
6	2	1	13	12	12.5
7	2	2	13	13	13
8	2	3	13	14	13.5
9	2	4	13	15	14
10	2	5	13	16	14.5
11	3	1	14	12	13
12	3	2	14	13	13.5
13	3	3	14	14	14
14	3	4	14	15	14.5
15	3	5	14	16	15
16	4	1	15	12	13.5
17	4	2	15	13	14
18	4	3	15	14	14.5
19	4	4	15	15	15
20	4	5	15	16	15.5
21	5	1	16	12	14
22	5	2	16	13	14.5
23	5	3	16	14	15
24	5	4	16	15	15.5
25	5	5	16	16	16

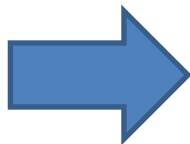
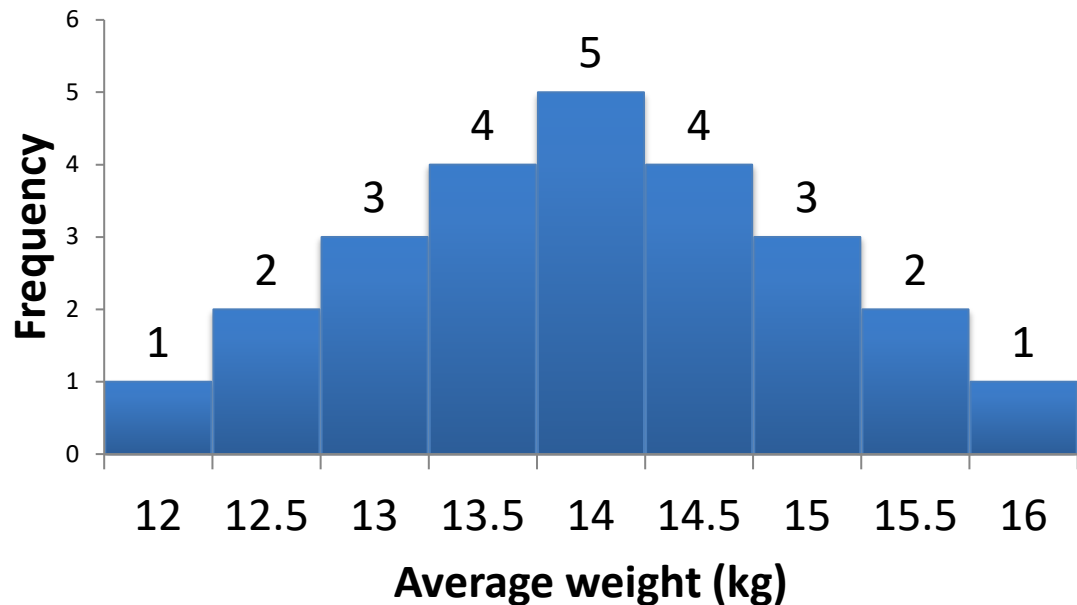


Arithmetic Means

12
12.5
12.5
13
13
13
13.5
13.5
13.5
13.5
13.5
14
14
14
14
14
14.5
14.5
14.5
14.5
15
15
15
15.5
15.5
16



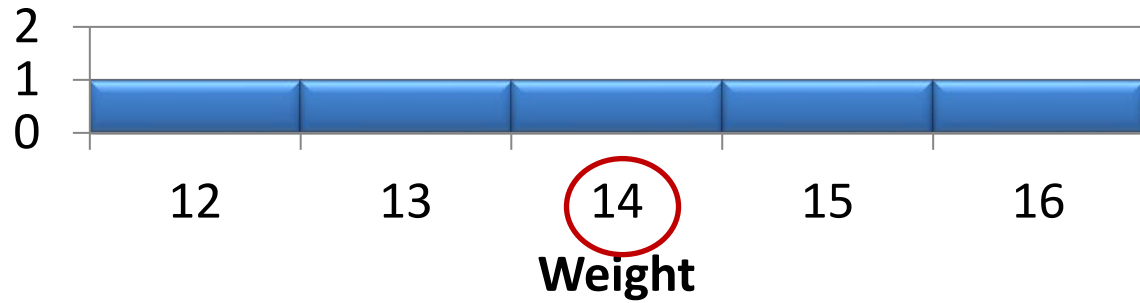
	Frequency
12	1
12.5	2
13	3
13.5	4
14	5
14.5	4
15	3
15.5	2
16	1



this is the **sampling distribution of the means**
(for all possible samples of 2 boys)

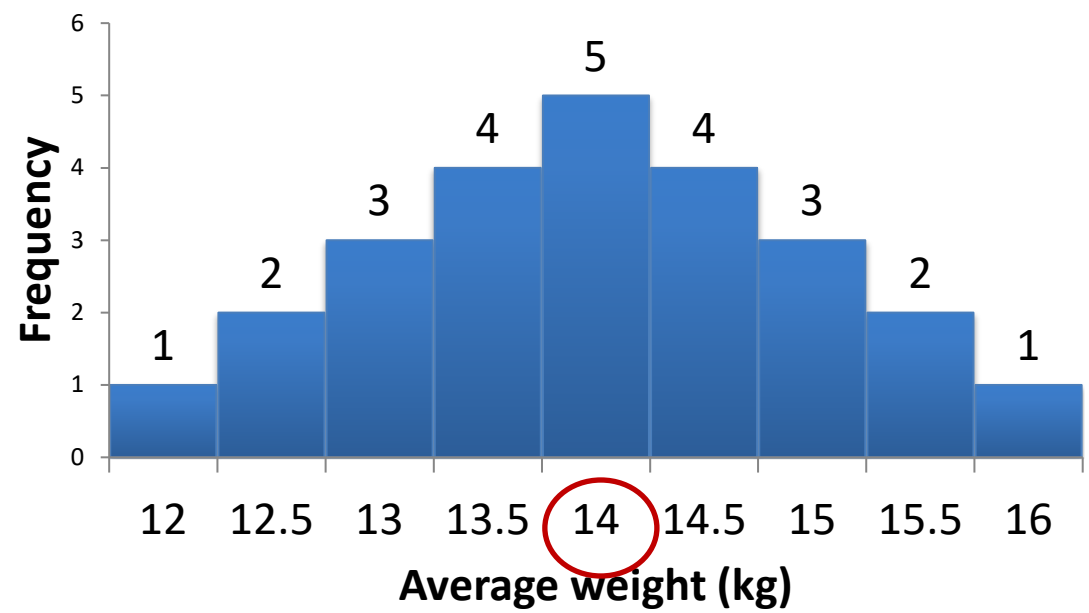
Arithmetic
Mean

- 12
- 12.5
- 12.5
- 13
- 13
- 13
- 13.5
- 13.5
- 13.5
- 13.5
- 14
- 14
- 14
- 14
- 14
- 14
- 14.5
- 14.5
- 14.5
- 14.5
- 15
- 15
- 15
- 15.5
- 15.5
- 16



Population (all 5 boys) distribution
Mean of the population = 14

Sampling distribution had the same arithmetic mean as the population

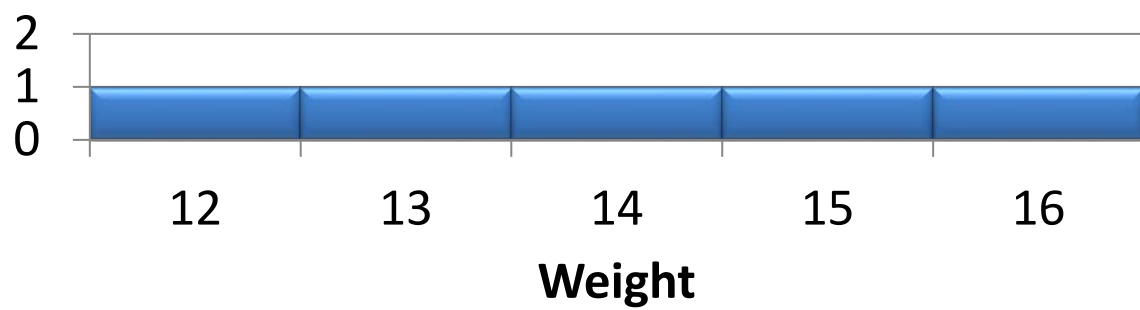


Sampling distribution of the means (all possible samples of 2 boys)
Mean (of the means) of the sampling distribution = 14

These are the means from the last two slides

Arithmetic
Mean

12
12.5
12.5
13
13
13
13.5
13.5
13.5
13.5
14
14
14
14
14
14
14.5
14.5
14.5
14.5
15
15
15
15
15.5
15.5
16



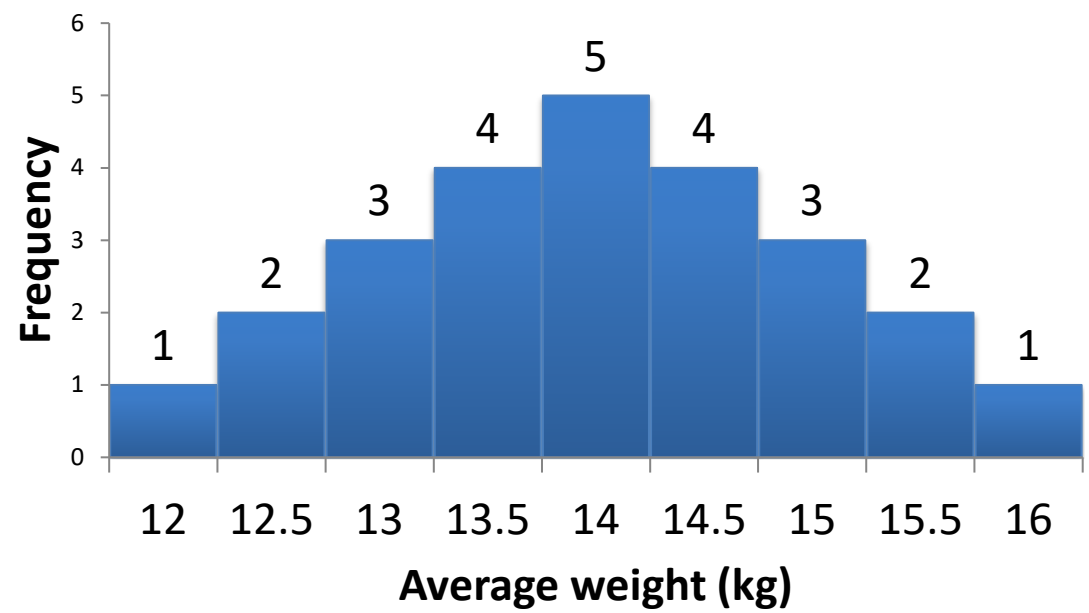
Population (all 5 boys) distribution

Mean of the population = 14

Standard deviation of the population = 1.58

Standard error of the population = 1.02

Standard deviation of the sampling distribution is the standard error of the population



Sampling distribution of the means (all possible samples of 2 boys)

Mean (of the means) of the sampling distribution = 14

Standard deviation of the sampling distribution = 1.02

These are the means from the last two slides

Commonly used symbols in inference and statistics



	Population Parameters Symbol	Sample Parameters Symbol
Mean	μ	$\bar{X}$
Standard deviation	σ	$SD (s)$

Central limit theorem

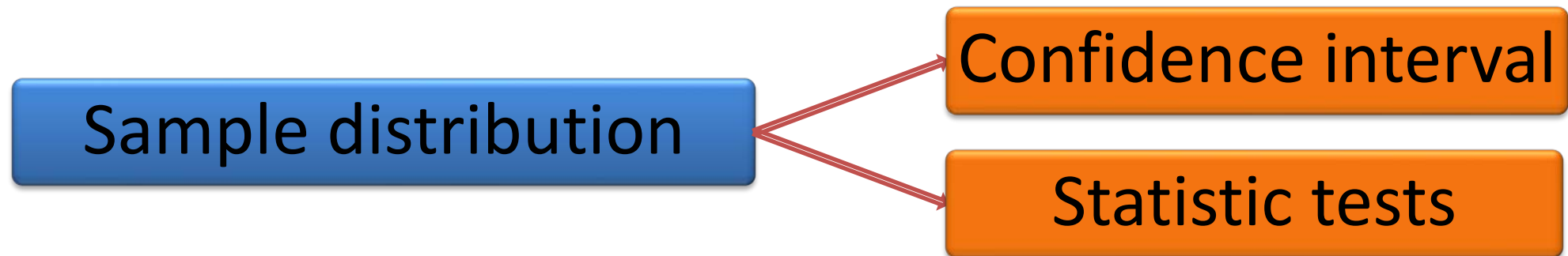
Given a population mean μ and standard deviation σ , a sampling distribution based on repeated sampling of size n has the properties:

- Mean of the sampling distribution is μ = average population
- Standard deviation of the sampling distribution = $\sigma / \sqrt{n}$, that is the standard error of the population
- If the population distribution is normal, then the sampling distribution is normal.
- If the sample is large enough, the sampling distribution approaches normal distribution.



Sampling distribution

- The most important concept in Inferential Statistic
- Sampling distribution is the distribution of the means if we repeat the study

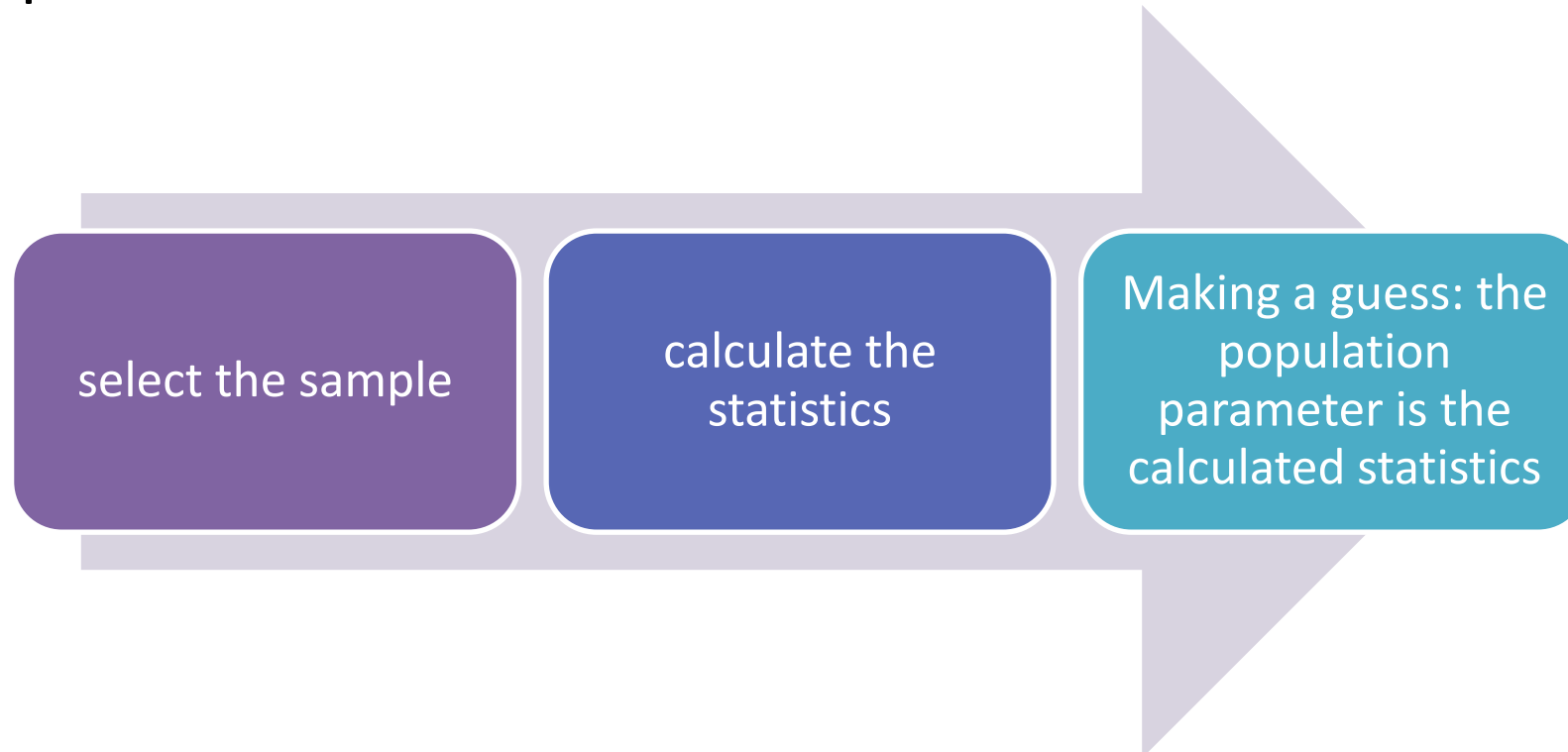


- **point estimation** - we can estimate the arithmetic mean μ of the population based on the known sample $\bar{X}$
- **confidence interval estimation** - we can construct an interval that in 95% of all possible studied samples, the population arithmetic mean μ will lie within this interval.



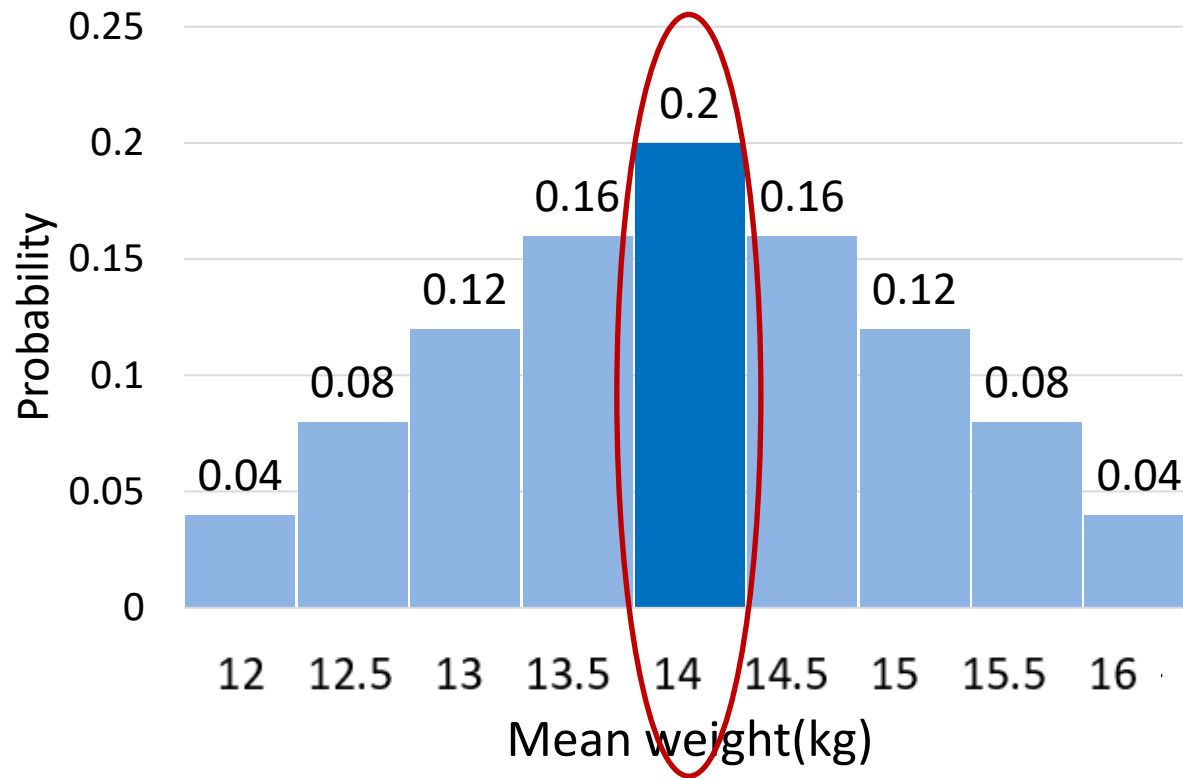
Point (punctual) estimation

- we will estimate (approximate)
 - μ (mean in the population) by the sample average $\bar{X}$,
 - σ (standard deviation in the population) by the sample standard deviation s
 - ϕ (frequency in the population) by the sample frequency of the disease f



In our example we select a random sample – one from the 25th possible.

we estimate that the mean population is the same as the sample mean



!!! In 20% of the possible selection we estimate correct $\mu = 14$

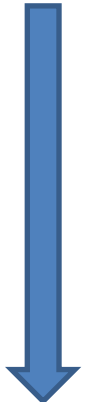


we get 0.2 probability of correct estimation

In 16% we estimate $\mu = 14.5$, small error
In 16% we estimate $\mu = 13.5$, small error

...

in 4 % we estimate $\mu = 16$, larger error
in 4% we estimate $\mu = 12$, larger error



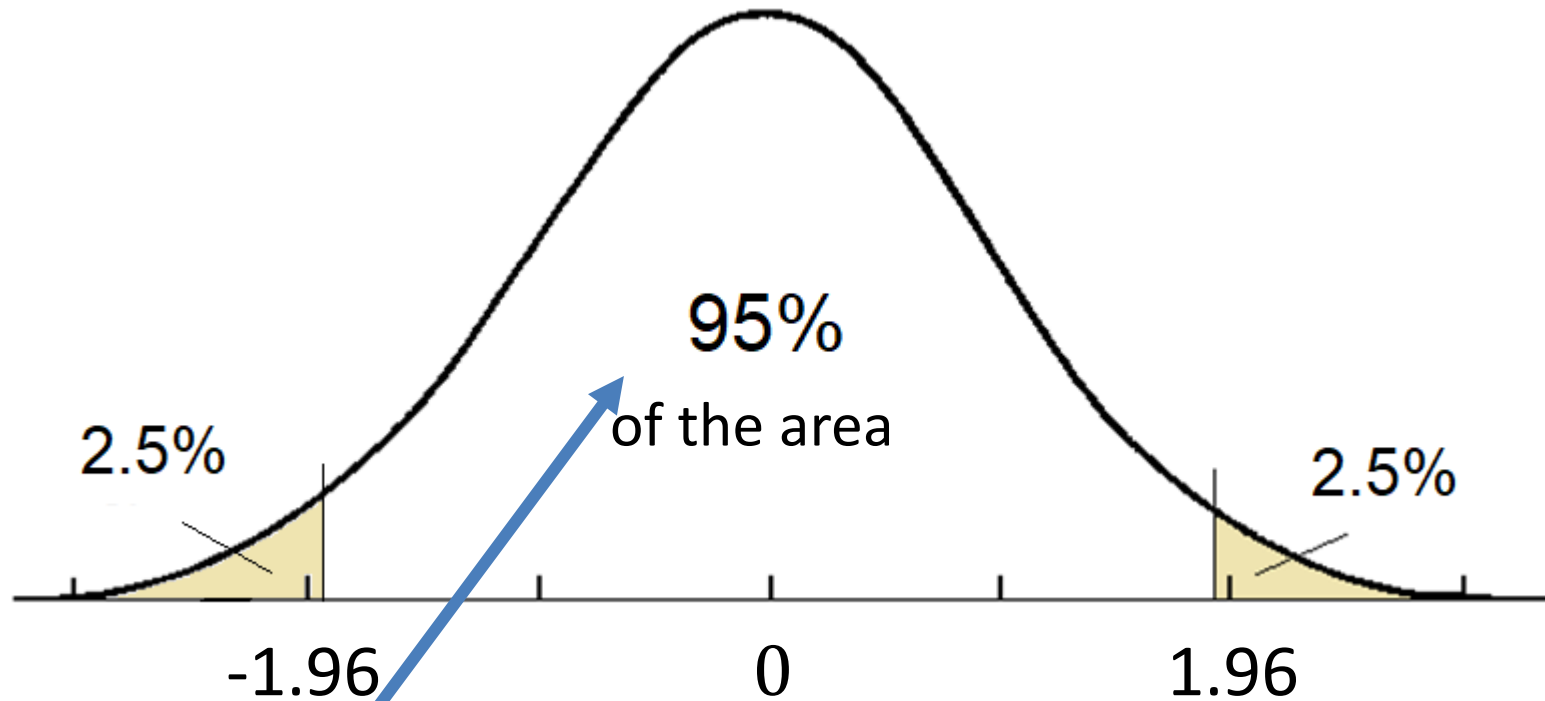
95 % Confidence interval estimation

construct an interval that in 95% of all possible studied samples, the population arithmetic mean μ will lie within this interval.



Standard Normal Distribution

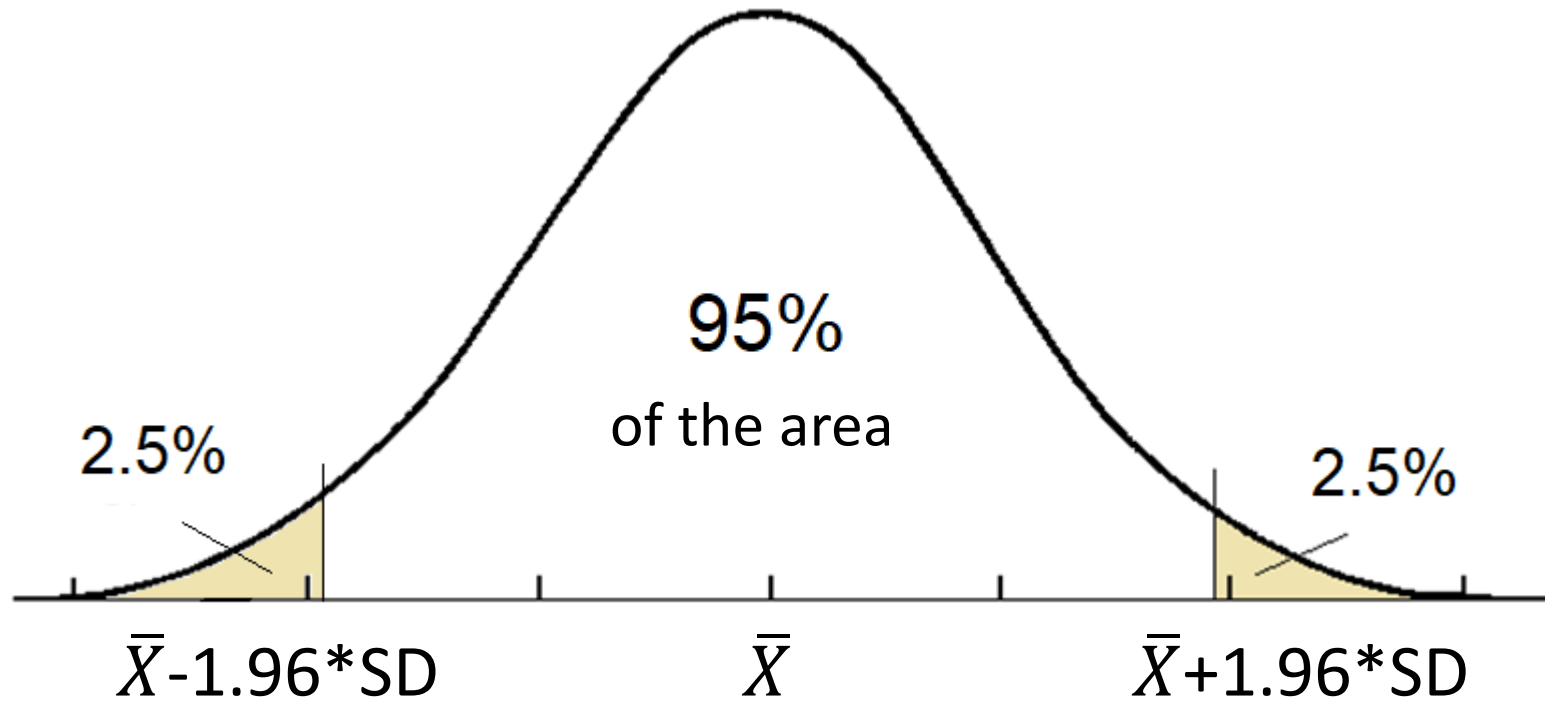
mean=0, standard deviation=1



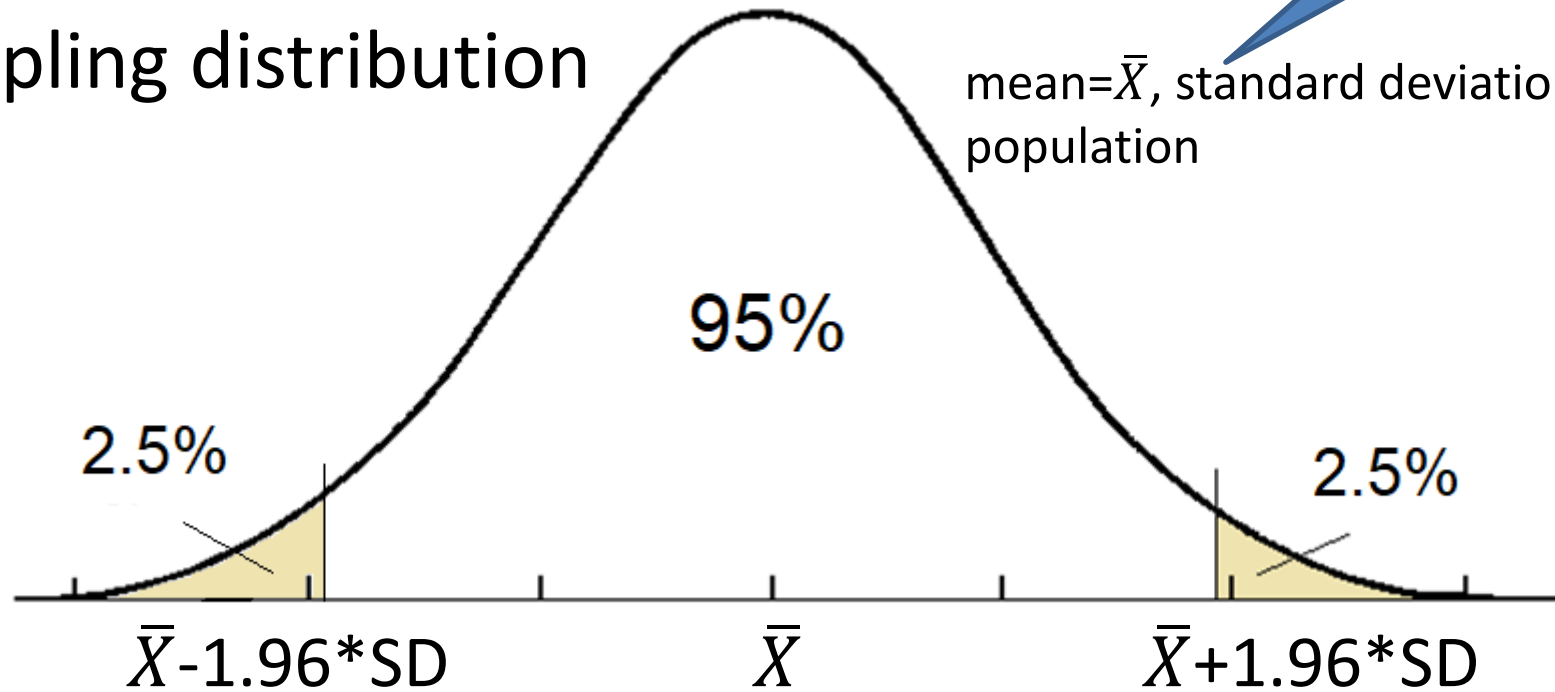
exactly 95% correspond to the interval -1.96; +1.96



Any normal distribution



sampling distribution



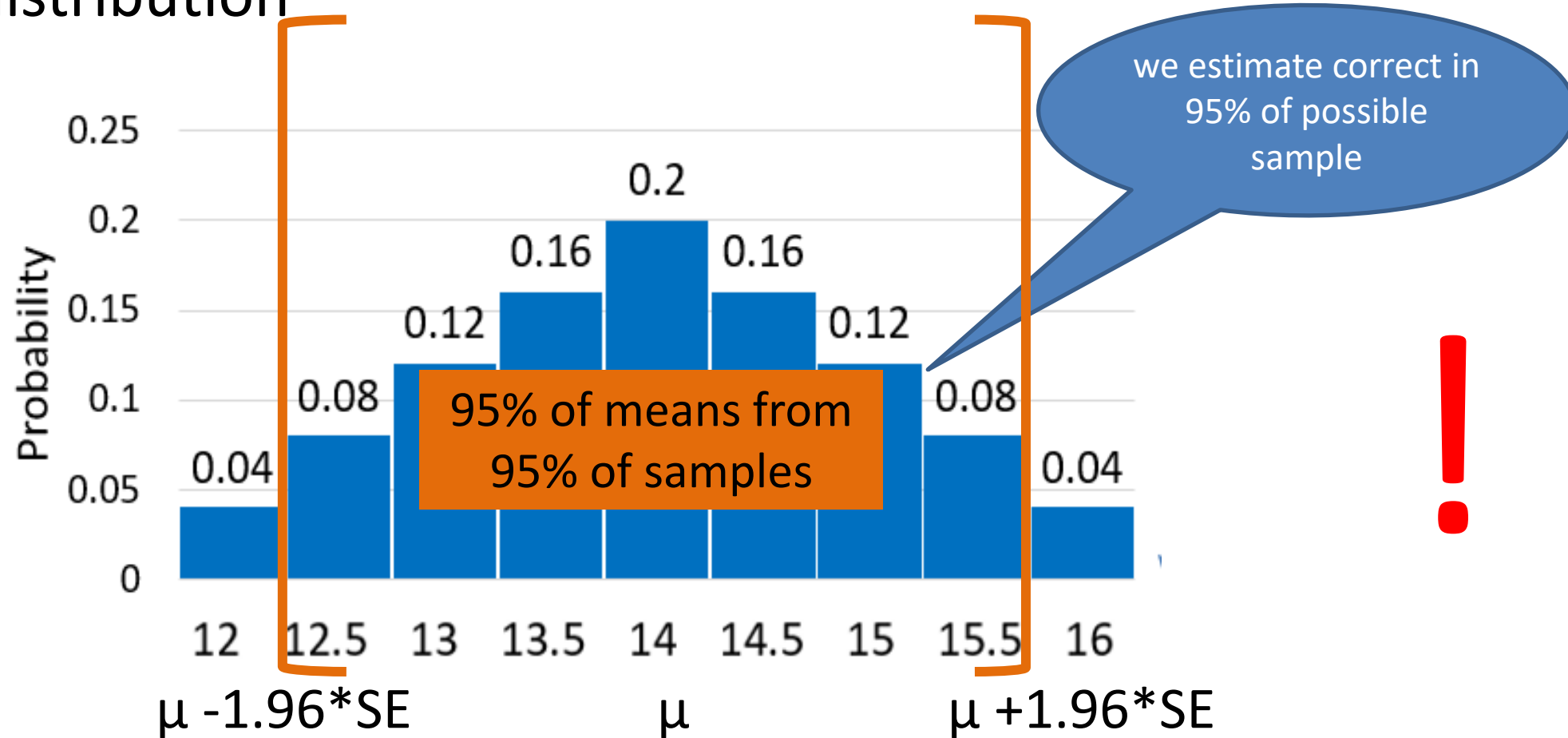
this is the distribution of the means if we repeat the study



SD for sampling distribution
= Es-standard error of the
population

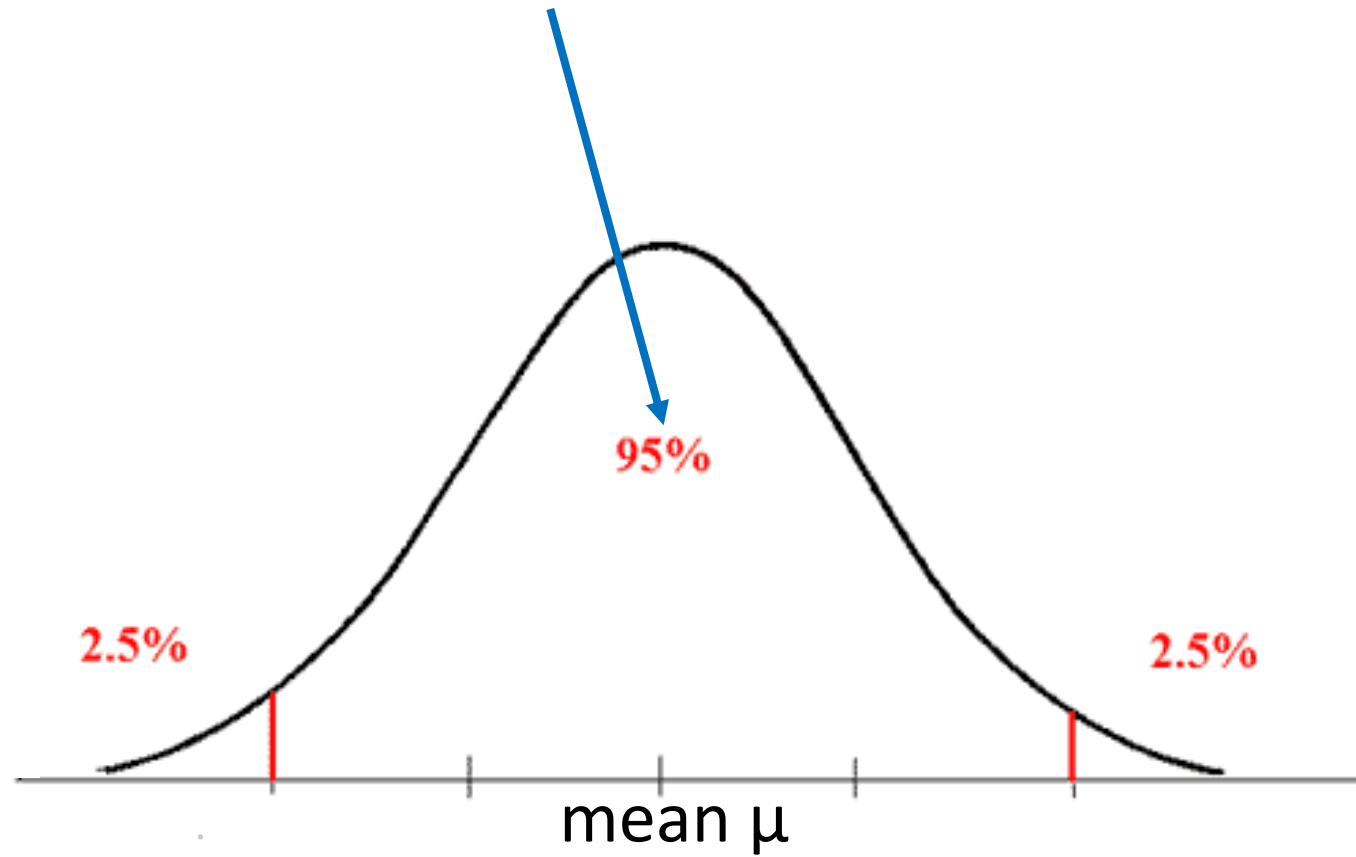
Exactly 95% of the weight means of the replicated studies on all possible samples of 5 boys will be between $\mu - 1.96SE$ to $\mu + 1.96SE$, where μ is the arithmetic mean of the weight in the population and SE is the standard error in the population

Sampling distribution

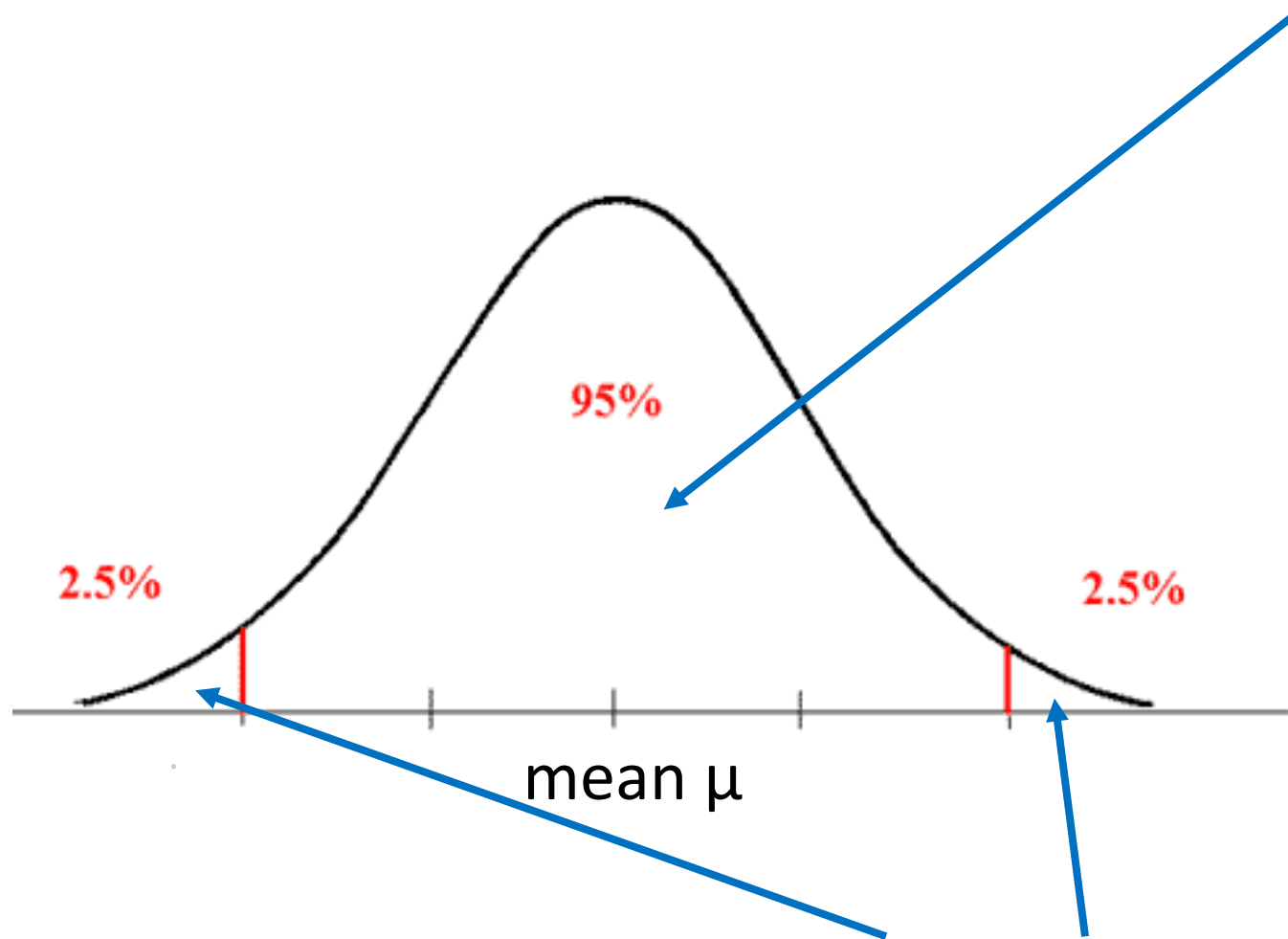


precision of 95% (probability 0.95)

= to be 95% sure that the population parameter is in the estimated interval



Error $\alpha=0.05 \leftrightarrow$ confidence 0.95



Error $\alpha = 0.05$ i.e. 5% = 2.5%+2.5%

!

1- α Confidence interval for the arithmetic mean μ when σ is known

$$\left[\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

where

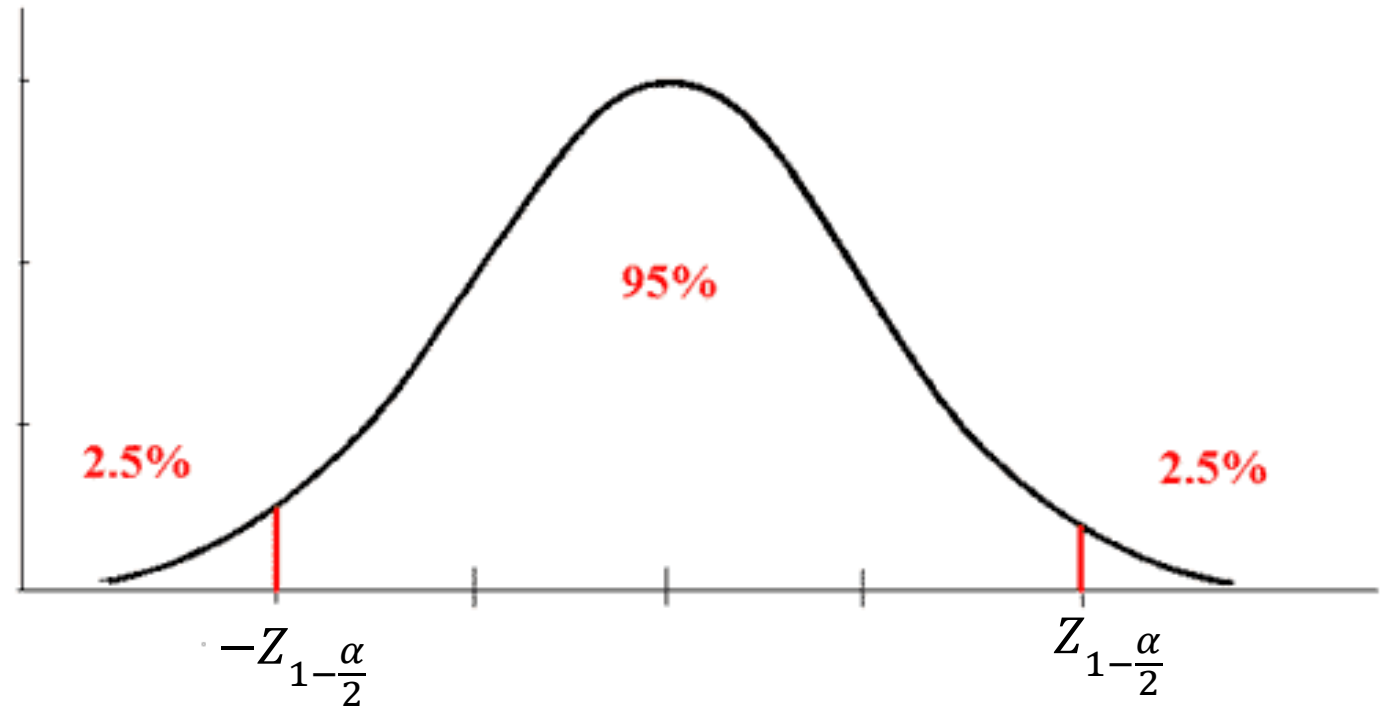
$\bar{X}$ – the sample arithmetic mean

σ – the population standard deviation

n – sample size

α – the error

$Z_{1-\frac{\alpha}{2}} = Z$ critic



1- α Confidence interval for the arithmetic mean μ when σ is unknown

If σ unknown we estimate it with S

$$\left[\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}, \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \right]$$

where

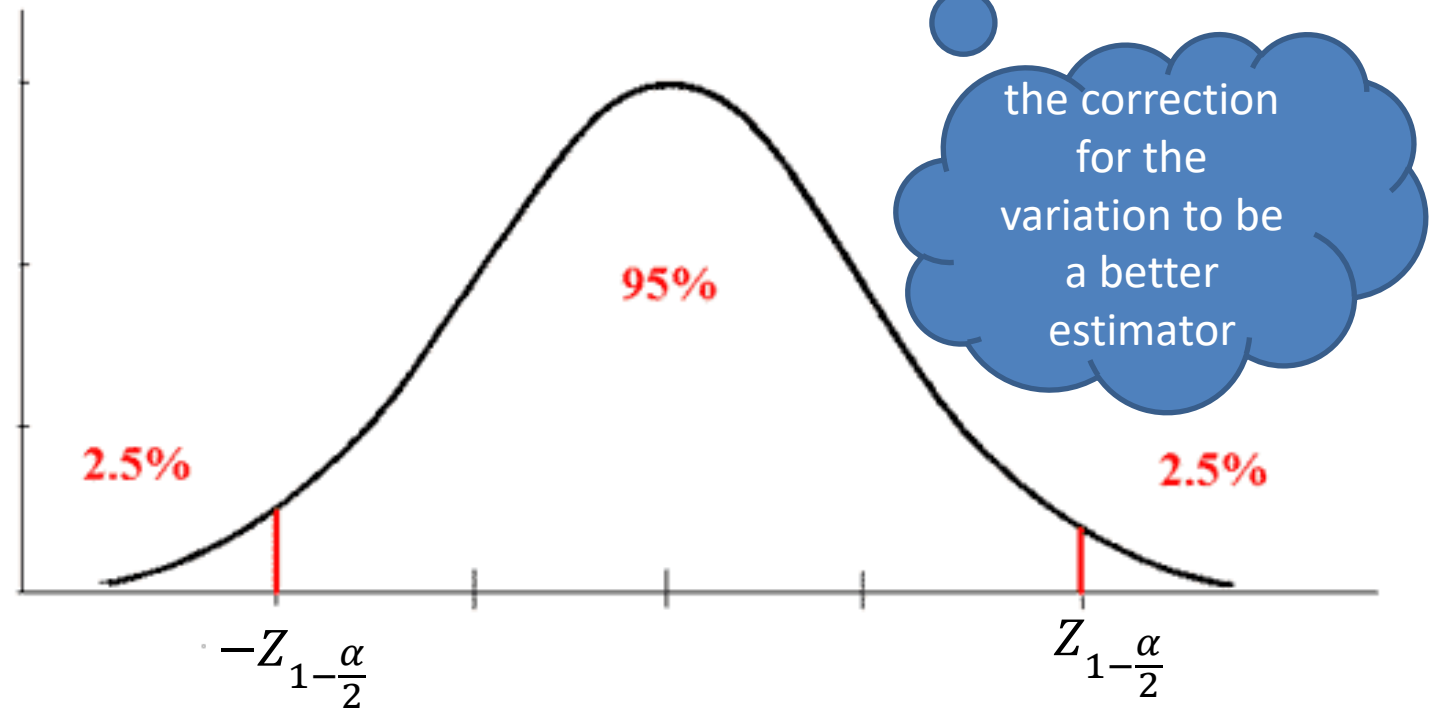
$\bar{X}$ – the sample arithmetic mean

S – the sample standard deviation

n – sample size

α – the error

$Z_{1-\frac{\alpha}{2}} = Z$ critic



When $\alpha=0.05$, estimate the 95% confidence interval for the arithmetic mean μ when σ is unknown and sample size $n \geq 30$

If σ unknown we estimate it with S

$$\left[\bar{X} - 1.96 \frac{S}{\sqrt{n-1}}, \bar{X} + 1.96 \frac{S}{\sqrt{n-1}} \right]$$

where

$\bar{X}$ – the sample arithmetic mean

S – the sample standard deviation

n – sample size



When $\alpha=0.05$, estimate the 95% confidence interval for the arithmetic mean μ when σ is unknown and sample size $n \geq 30$

If σ unknown we estimate it with s

$$[\bar{X} - 1.96 \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \frac{s}{\sqrt{n}}]$$

where

$\bar{X}$ – the sample arithmetic mean

s – the corrected sample standard deviation

n – sample size



The **95% confidence interval** of the arithmetic mean in the case of $n \geq 30$ and σ unknown

$$(\bar{X} - 1.96 \cdot SE; \bar{X} + 1.96 \cdot SE)$$

where

$\bar{X}$ - arithmetic mean of the sample,

$SE = \frac{S}{\sqrt{n-1}}$ standard error of the sample,

S- standard deviation,

n – sample size (number of people)

σ – standard deviation of the population

$$\left(\bar{X} - 1.96 \frac{S}{\sqrt{n-1}}, \bar{X} + 1.96 \frac{S}{\sqrt{n-1}} \right)$$



Example

The average incubation period in days of a random sample of 43 cases of hepatitis A from an epidemic source in city X in 1992 is $\bar{X} = 28.93$ and standard deviation $s=3.68$

95% confidence interval of the mean incubation period:

$$(\bar{X} - 1.96 \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \frac{s}{\sqrt{n}})$$

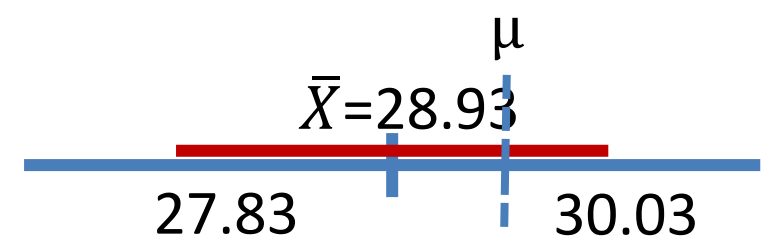
$$\left(28.93 - 1.96 \frac{3.68}{\sqrt{43}}; 28.93 + 1.96 \frac{3.68}{\sqrt{43}} \right)$$

$$(28.93 - 1.10; 28.93 + 1.10)$$

$$(27.83; 30.03)$$

Interpretation: the population mean incubation period is between 27.83 and 30.03 with a 5% error

Example



The average incubation period in days of a random sample of 43 cases of hepatitis A from an epidemic source in city X in 1992 is $\bar{X} = 28.93$ and standard deviation $s=3.68$

95% confidence interval of the mean incubation period:

$$(\bar{X} - 1.96 \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \frac{s}{\sqrt{n}})$$

$$\left(28.93 - 1.96 \frac{3.68}{\sqrt{43}}; 28.93 + 1.96 \frac{3.68}{\sqrt{43}} \right)$$

$$(28.93 - 1.10; 28.93 + 1.10)$$

$$(27.83; 30.03)$$

Interpretation: the population mean incubation period is between 27.83 and 30.03 with a 5% error

Example

Objectiv: to estimate the arithmetic mean μ of cholesterol in the population

- a sample **n=101**
- arithmetic mean in the sample
- standard deviation in the sample

$$\bar{X} = 120 \text{ mg/dl}$$

$$S=16$$

$$[\bar{X} - 1.96 \frac{S}{\sqrt{n-1}}, \bar{X} + 1.96 \frac{S}{\sqrt{n-1}}]$$

$$[120 - 1.96 \frac{16}{\sqrt{101-1}}; 120 + 1.96 \frac{16}{\sqrt{101-1}}]$$
$$[120 - 3.14; 120 + 3.14]$$

[116.86; 123.14] – the 95% confidence interval

Result: with a 95% confidence we estimate that the population (true) arithmetic mean μ of cholesterol is between 116.86 and 123.14 mg/dl

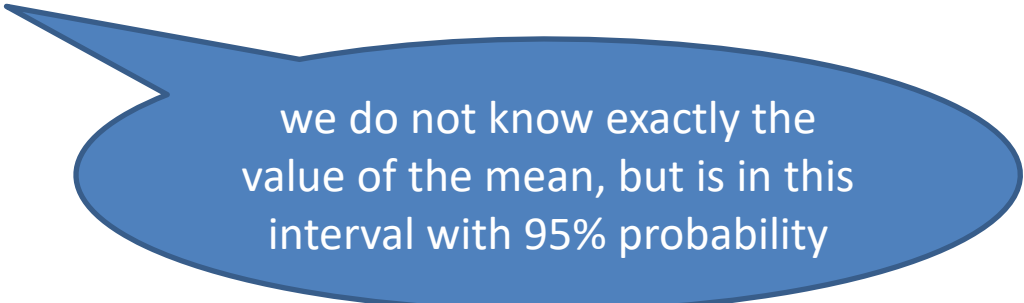
interpretation of the 95% confidence interval !

= 95% of the calculated means of all possible sample of n subjects taken from the population are included (if we repeat the study)

- or

= we are 95% confident that the population (true) mean is between the lower and upper margins of the interval

= interval where the population (true) mean is included with a 5% level of error



we do not know exactly the value of the mean, but is in this interval with 95% probability

1- α confidence intervals for a frequency

- The formula:

$$\left(f - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}; f + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}} \right)$$

Where f – frequency, $f < 1$, n – total number of subjects

When $\alpha=0.05$, estimate the 95% confidence intervals for a frequency



- The formula:

$$\left(f - 1.96 \sqrt{\frac{f(1-f)}{n}}; f + 1.96 \sqrt{\frac{f(1-f)}{n}} \right)$$

Where f – frequency, $f < 1$, n – total number of subjects

Example

- We want to estimate the frequency of mouth cancer in population with age higher than 90 years old.
- In a study with 1000 participants, 300 had memory problems

$$f = \frac{300}{1000} = 0.3$$

95% confidence intervals for the frequency of mouth cancer

$$\left(f - 1.96 \sqrt{\frac{f(1-f)}{n}}; f + 1.96 \sqrt{\frac{f(1-f)}{n}} \right)$$
$$\left(0.3 - 1.96 \sqrt{\frac{0.3(1-0.3)}{1000}}; 0.3 + 1.96 \sqrt{\frac{0.3(1-0.3)}{1000}} \right)$$

$$(0.3 - 0.03; 0.3 + 0.03)$$

$$(0.27; 0.33)$$

Interpretation: with a 5% error we estimate that the population frequency (the true frequency) of memory problem in the population over 90 years old is between 27% and 33%

If we select a smaller sample size?

- We want to estimate the frequency of mouth cancer in population with age higher than 90 years old.
- In a study with 100 participants, 30 had memory problems

$$f = \frac{30}{100} = 0.3$$
$$\left[0.3 - 1.96 \sqrt{\frac{0.3(1-0.3)}{100}}; 0.3 + 1.96 \sqrt{\frac{0.3(1-0.3)}{100}} \right]$$
$$[0.3 - 0.09; 0.3 + 0.09]$$

[0.21; 0.39] – interval for n=100

frequency of memory problems between 21% și 39% with a 95% probability

[0.27; 0.33] – interval for n=1000

frequency of memory problems between 27% și 33% with a 95% probability

Answer: **smaller the sample → wider the interval**
an increasing n at the denominator, an opposite effect

If the sample increase → it increases the measurement precision by decreasing the interval required for estimation

If the probability decrease 95% → 80%

- We want to estimate the frequency of mouth cancer in population with age higher than 90 years old.
- In a study with 100 participants, 30 had memory problems

For 80% confidence, Z critic = 1.29

$$f = \frac{30}{100} = 0.3$$

$$\left[0.3 - 1.29 \sqrt{\frac{0.3(1-0.3)}{100}}; 0.3 + 1.29 \sqrt{\frac{0.3(1-0.3)}{100}} \right]$$

$$[0.3 - 0.06; 0.3 + 0.06]$$

[0.24; 0.36] – interval for 80%

frequency of memory problems between 24% și 36% with a 95% probability

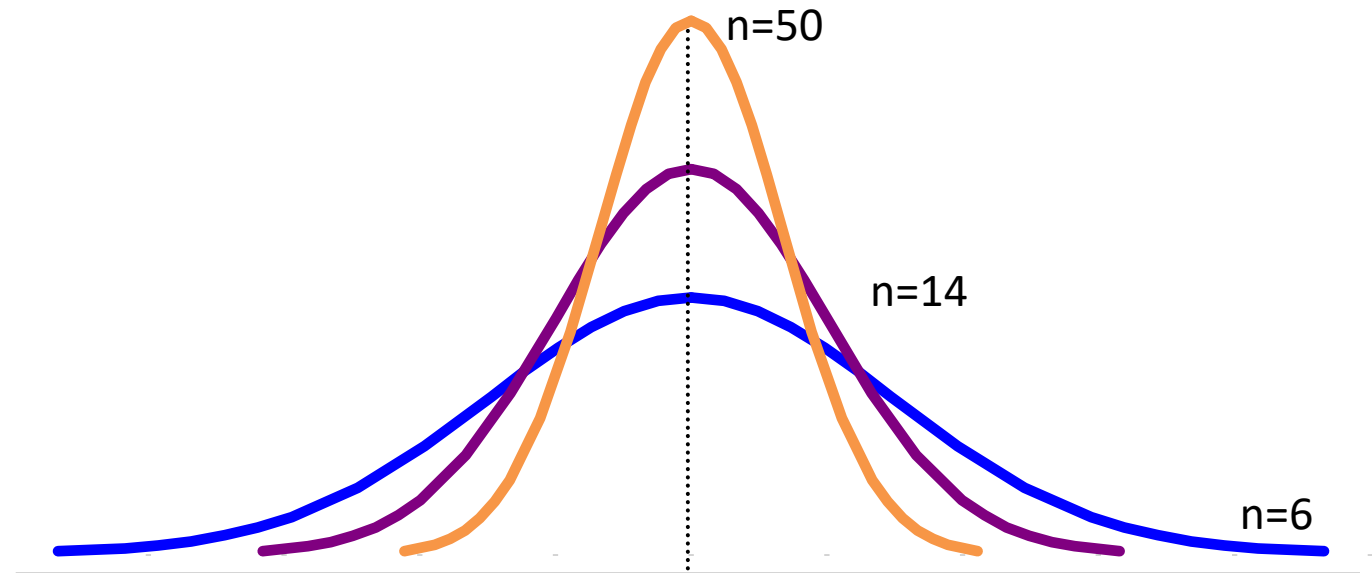
[0.21; 0.39] – interval for 95%

frequency of memory problems between 21% și 39% with a 95% probability

Answer: the smaller the probability – the narrower the interval
Z at numerator

Increasing precision → interval increase

- If $n < 30$
 - in this case the sampling distribution no longer follows the normal law
 - the sampling distribution follows Student's law – very similar
 - Student's law varies with sample size
 - depends on the degrees of freedom
 - Student's law is close to the normal law if $n \geq 30$



Degree of freedom = df

The number of components that are "free" to vary in a data set

- $df = n - 1$

Ex: 5 measurements, the average = 6

- we know the first four values: 8, 9, 10, 11
- the fifth can be calculated

--> the data set has only 4 degrees of freedom

1- α confidence interval for the average μ in the case of small samples $n < 30$ with σ unknown

- using Student t distribution

$$\left(\bar{X} - t_{n-1, 1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1, 1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right)$$

where

$\bar{X}$ - sample arithmetic mean,

s – sample standard deviation,

n – sample size,

$t_{n-1, 1-\frac{\alpha}{2}}$ critical t for $n-1$ degree of freedom

$1 - \alpha$ level of confidence



- $n=6 \rightarrow df = n-1 = 5$
 - confidence interval 95% $\rightarrow t = \pm 2.571$
- $n=10 \rightarrow df = 9$
 - confidence interval 95% $\rightarrow t = \pm 2.262$
- $n=30 \rightarrow df = 29$
 - confidence interval 95% $\rightarrow t = \pm 2.042$
- The increase of n causes the value of t to approach 1.96, $\rightarrow$ the curve tends towards the normal distribution.
- The differences in calculations can be considered negligible for samples of size $n \geq 30$

VALUES OF t FOR 90%, 95%, AND 99% CONFIDENCE INTERVALS

df	90%	95%	99%
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861
20	1.725	2.086	2.845
30	1.697	2.042	2.750
40	1.684	2.021	2.704
60	1.671	2.000	2.660
120	1.658	1.980	2.617
∞	1.645	1.960	2.576

Example

Study of mobility by extension of the lumbar spine in individuals aged between 30 and 39 years

$n=17$, mean= 40° and $t_{critic}=2.12$,
 $s=2.36^\circ$

$$\left[40 - 2.12 * \frac{2.36}{\sqrt{17}}; 40 - 2.12 * \frac{2.36}{\sqrt{17}}\right]$$

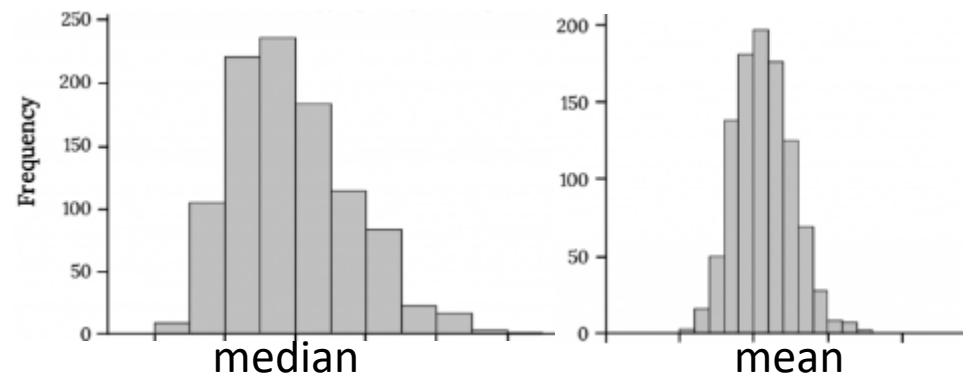
$$[40 - 1.21; 40 + 1.21]$$

$$[38.79^\circ; 41.21^\circ]$$

Answer: mobility of the lumbar spine in young people is between 38.79° and 41.21° with a 5% error

Quality of the estimator

- to be without errors
 - sometimes corrections are needed for a better estimate
 - e.g. if we use the standard deviations as an estimator of σ we need to divide by $n-1$ instead of n otherwise the estimate produces a value too small
- have high stability (the variance of the sampling distribution should not be large)
 - the median is not a stable estimator compared to the mean
 - if possible, we will use the mean as the estimator, not the median



The precision of the estimation of a population parameter

depend on

- the estimator
 - ex. mean is more precise than median
- the sample size
 - law of big numbers – bigger sample size – bigger precision
- variation of the data
 - less variation – bigger precision
- needed precision
 - bigger precision – wider interval



- Thank you !!!