Point estimators & Confidence intervals

Objectives

- Inferential statistics: definition
- Point estimators (definition)
- Confidence intervals:
  - definition
  - for one mean
  - for the difference between two means
  - for one proportion
  - for the difference between two proportions

Glossary

Inferential statistics = the process of making guesses about the truth on the population by examining a sample extracted from the population

Sample statistics = summary measures calculated from data belonging to a sample (e.g., mean, proportion, ratio, correlation coefficient, etc.)

Population parameter = true value in the population of interest

Point estimation involves the use of sample data to calculate a single value (known as a statistic) which is to serve as a "best guess" or "best estimate" of an unknown (fixed or random) population parameter.

Glossary

- A population distribution: the variation in the larger group that we want to know about.
- A distribution of sample observations: the variation in the sample that we can observe.
- A sampling distribution:
  - a normal distribution whose mean and standard deviation are unbiased estimates of the parameters and allows one to infer the parameters from the statistics.
- Researchers do not typically conduct repeated samples of the same population. Instead, they use the knowledge of theoretical sampling distributions to construct confidence intervals around estimates

Point estimator

Point estimation provide one value as an estimate of the population parameter (e.g., the sample mean is a point estimator for population mean)

We are interested in the mean of height of 10-years-old boys and girls in Romania. It would be impossible to measure the height of all 10-years-old boys and girls height so we will investigate a random sample of 30 boys and a random sample of 30 girls of 10-years-old. The sample mean for boys is 140 cm and for girls is 132 cm.

- The sample mean of 140 cm is a point estimator of boys population mean
- The sample mean of 132 cm is a point estimator of girls population mean

Point vs. Interval estimation

Interval estimation: provide a range of values (an interval) that contain with a high probability the unknown parameter

Confidence interval: the interval that contains an unknown parameter (such as the population mean) with a certain degree of confidence

It is recommended to estimate a theoretical parameter by using a range of value, not a single value

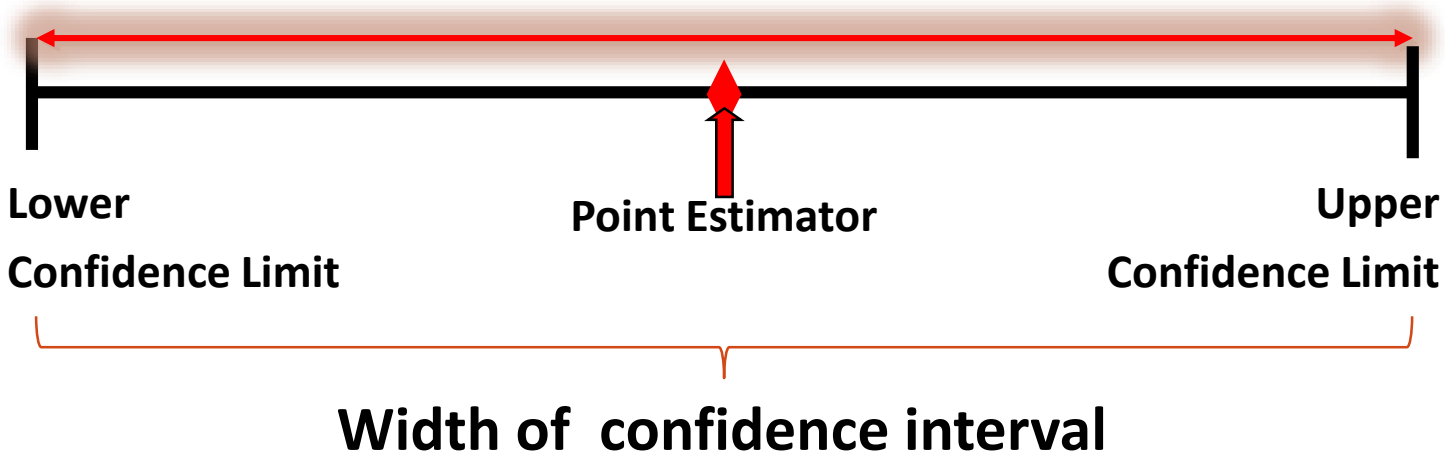
- It is called confidence interval.
- The estimated parameter belongs to the confidence intervals with a high probability.

Point vs. interval estimation

Point estimator = one value obtained on a sample

- How much uncertainty is associated with a point estimator of parameter?

An interval provides more information about a population characteristic that does a point estimator ⟶ confidence interval



Interval estimation

An interval gives a range of values:

- Takes into consideration variation in sample statistics from sample to sample
- Based on observations from one sample
- Provides information about closeness to unknown population parameters
- Stated regarding the level of confidence. (Can never be 100% confident)

The general formula for all confidence intervals is equal to:

$$\text{Point Estimator} \pm (\text{Critical Value}) \times (\text{Standard Error})$$

Table value

Margin of error

Interval estimation

Point Estimator $\pm$ (Critical Value)×(Standard Error)

Margin of error

The margin of error, and hence the width of the interval, get smaller the as the sample size increases.

The margin of error, and hence the width of the interval, increases and decreases with the confidence.

Interval estimation

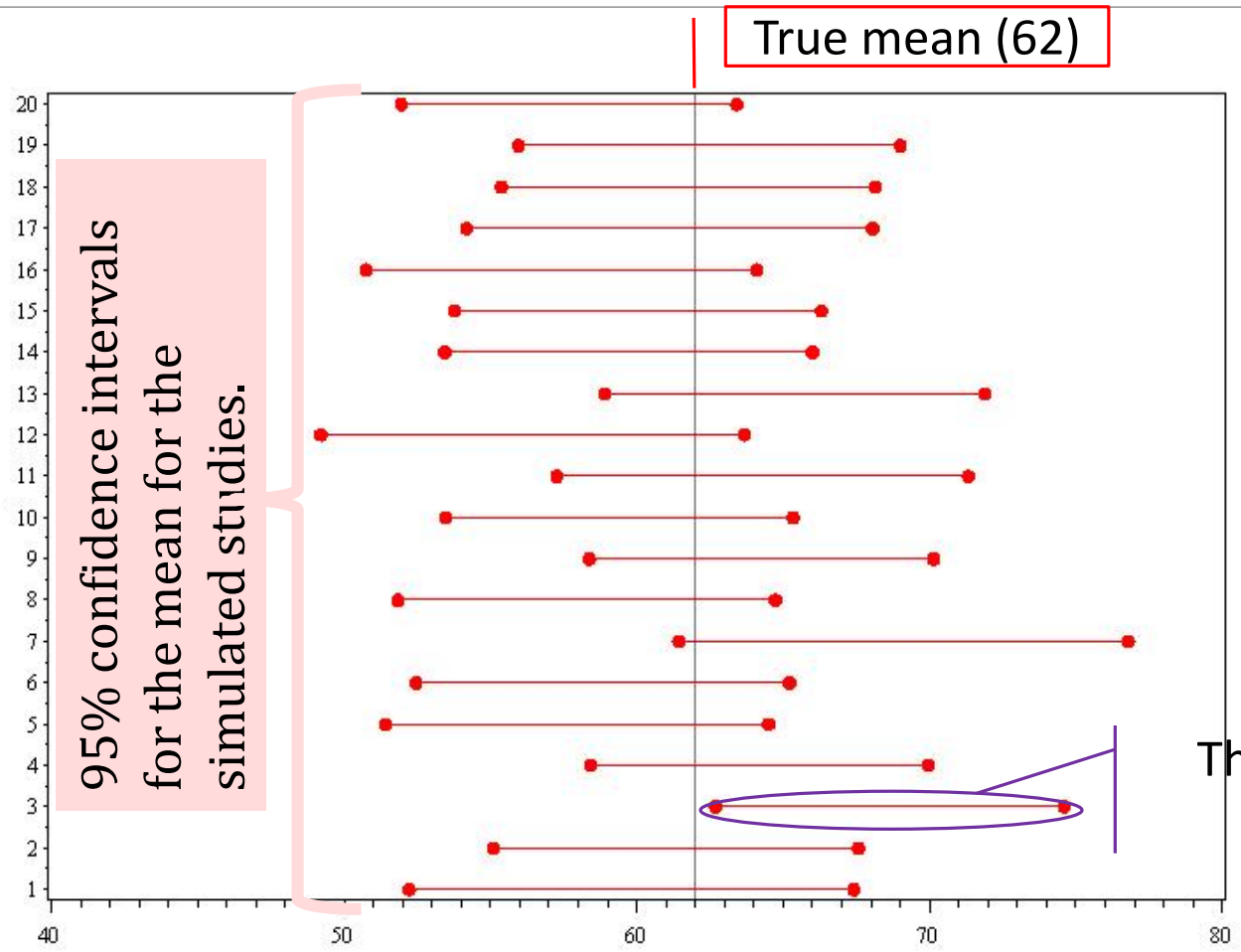
Significance level α = 5% → 95% confidence interval (CI)

CI = (1 - α) = 0.95

Interpretation:

- If all possible samples of size *n* are extracted from the population, and their means and intervals are estimated, 95% of all the intervals will include the **true (real) value of the unknown parameter**
- A specific interval either will contain or will not contain the true parameter (due to the 5% risk)

Interval estimation



This CI did not include the true value

Confidence intervals

Provides:

- A plausible range of values for a population parameter.
- The precision of a point estimator.
  - When sampling variability is high, the confidence interval will be wide to reflect the uncertainty of the observation.
- Statistical significance.
  - If the 95% CI does not cross the null value, it is significant at 0.05.

Confidence intervals

Are calculated taking into consideration:

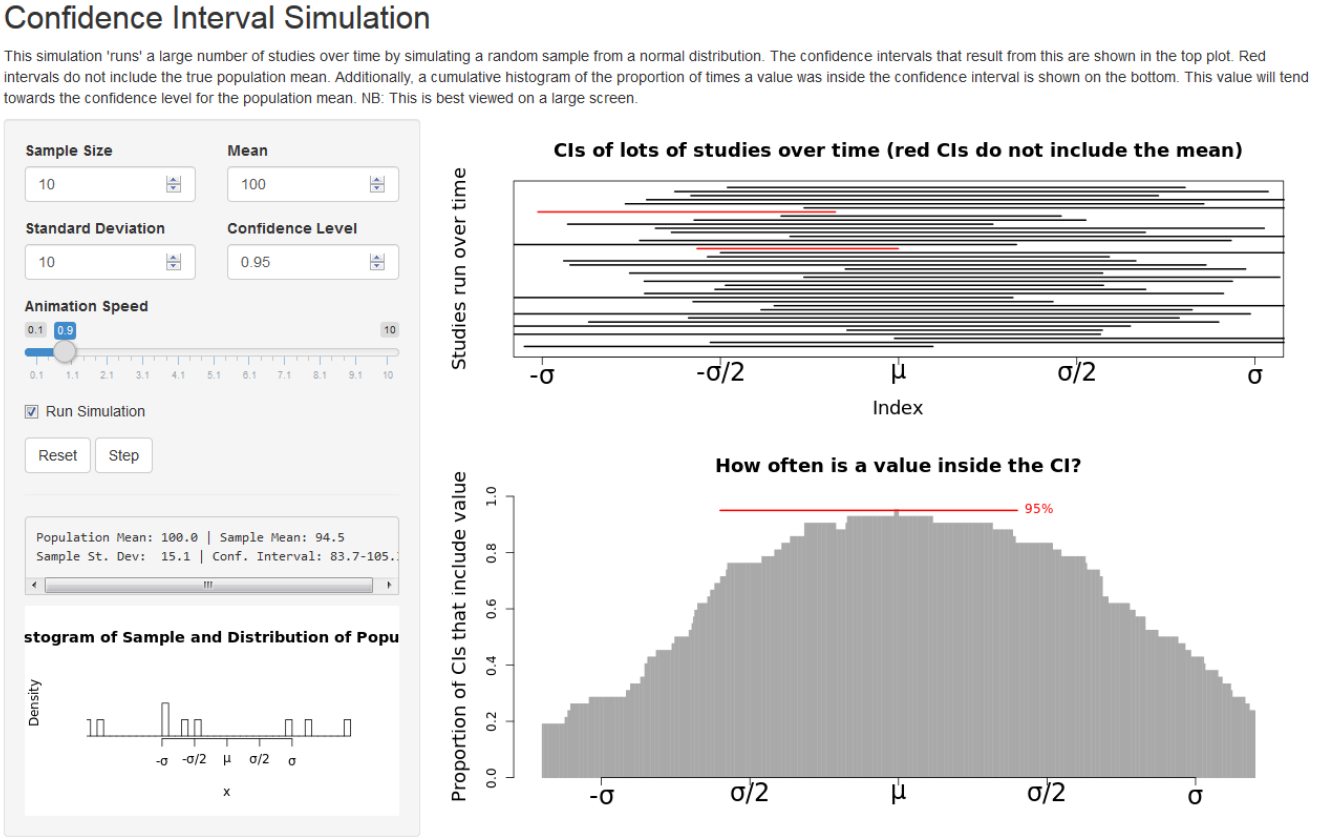
- The sample or population size
- The type of investigated variable (qualitative OR quantitative)

The formula of calculus comprised two parts:

- One estimator of the quality of sample based on which the population estimator was computed (standard error)
  - Standard error: is a measure of how good our best guess is.
    - Standard error: the bigger the sample, the smaller the standard error.
    - Standard error: is always smaller than the standard deviation
- The degree of confidence (standard values)

Confidence intervals

- Can be calculated for any point estimator
- Levels of confidence usually used:
  - 90% (α=10%)
  - 95% (α=5%)
  - 98% (α=2%)
  - 99% (α=1%)



Confidence intervals

Confidence interval: the likelihood that a specified interval will contain the population parameter.

95% confidence interval: there is a 0.95 probability that a specified interval DOES contain the population mean.

- In other words, there are 5 chances out of 100 (or 1 chance out of 20) that the interval DOES NOT contain the population parameter.

99% confidence interval: there is 1 chance out of 100 that the interval DOES NOT contain the population parameter.

Confidence intervals

Continuous	Categorical
One sample: known variance	dichotomous outcome
One sample: unknown variance	categorical/ ordinal outcome
Matched samples	

Confidence intervals for one sample

CONTINUOUS OUTCOME

DICHOTOMOUS OUTCOME

Confidence intervals

Assumptions:

- Population standard deviation (σ) is known by replacing s with σ or if n>200
- The population is normally distributed

$$\bar{X} \pm Z_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

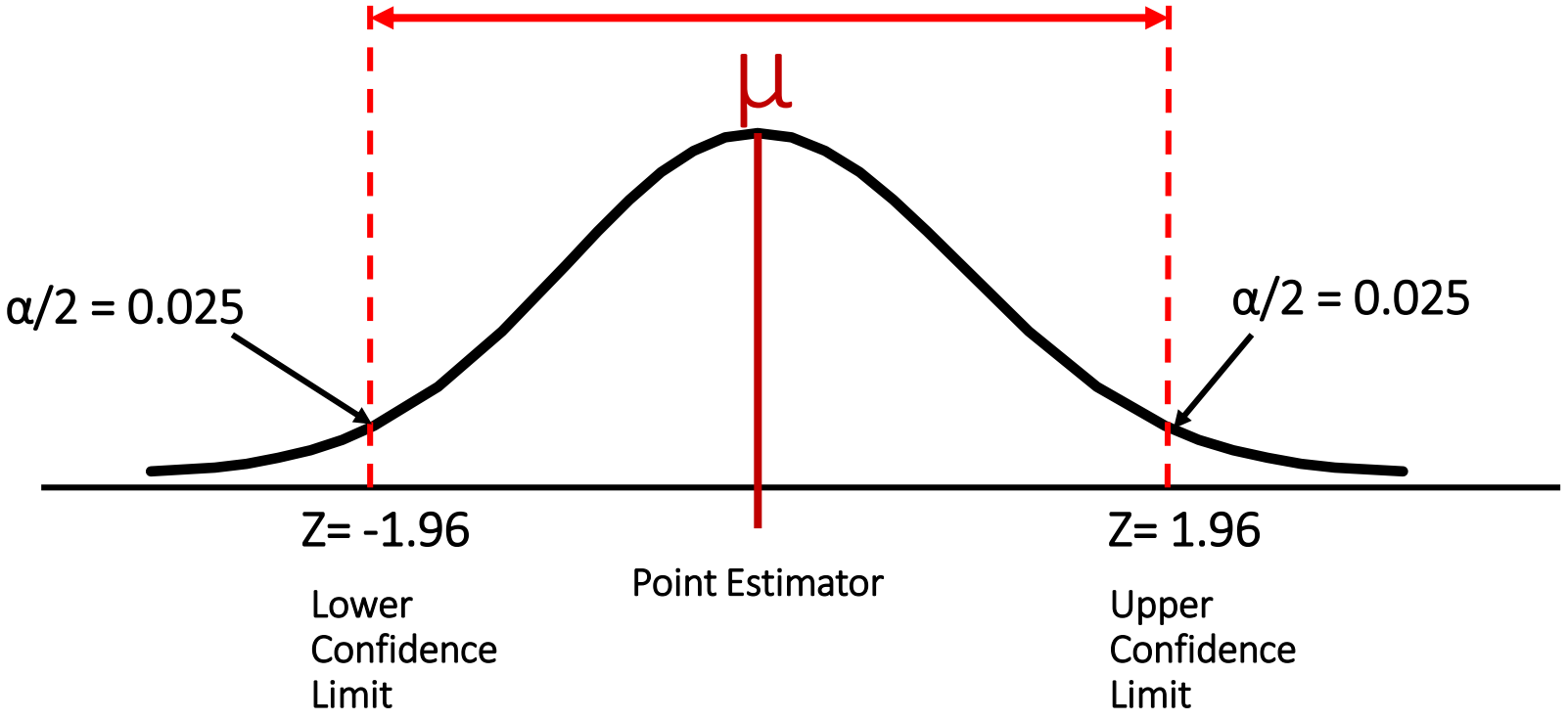
where Z is the normal distribution's critical value for a probability of α/2 in each tail

For large sample sizes (n > 30), σ can be estimated from the sample standard deviation (s) based on the Central Limit Theorem.

Confidence interval

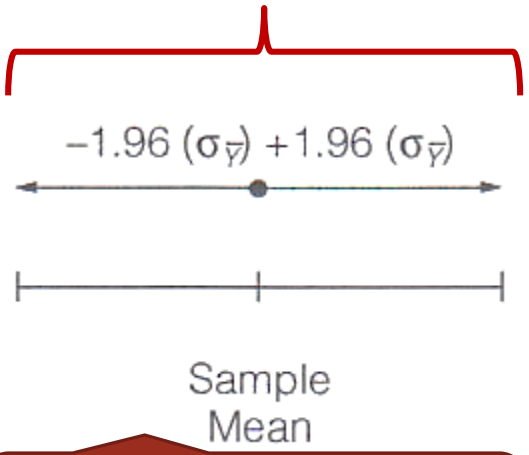
Consider a 95% confidence interval:

$1-\alpha = 0.95$ & $\alpha = 0.05$ & $\alpha/2 = 0.025$



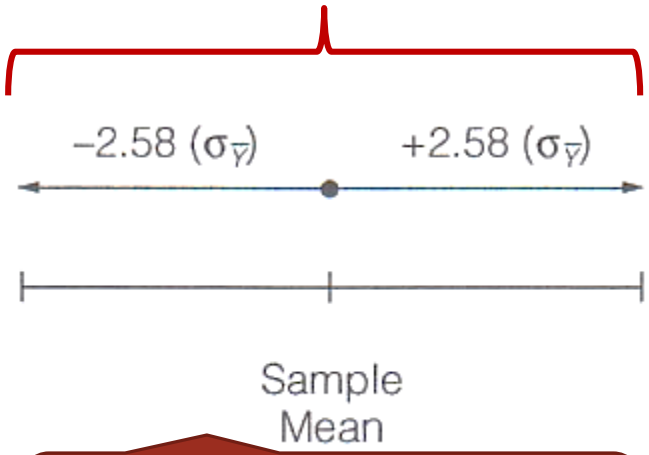
Confidence interval

95% confidence interval



More precise
Less confident

99% confidence interval



Less precise
More confident

Increasing our confidence interval from 95% to 99% means we are less willing to draw the wrong conclusion – we take a 1% risk (rather than a 5%) that the specified interval does not contain the true population mean.

Confidence level and corresponding Z values

Confidence level	Z value
90%	1.65
95%	1.96
99%	2.58

Larger samples result in smaller standard errors, and therefore, in sampling distributions that are more clustered around the population mean. A more closely clustered sampling distribution indicates that our confidence intervals will be narrower and more precise.

Smaller sample standard deviations result in smaller, more precise confidence intervals.

Confidence interval

Student t-distribution with n-1 degree of freedom will be used

The unknown population mean (μ) & unknown population standard deviation (σ)

Small sample size (n < 30)

$$\bar{X} \pm t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

- *t* values depend on *n*
- small samples have larger *t* value (less precision)
- values are indexed by degrees of freedom (*df = n-1*)

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

Confidence interval

A sample of 20 female students gave a mean weight of 60kg and a standard deviation of 8 kg. Assuming normality, find the 90%, 95%, and 99% confidence intervals for the population mean weight.

$$90\%CI = \left[60 - 1.73\frac{8}{\sqrt{20}}, 60 + 1.73\frac{8}{\sqrt{20}}\right] = [56.91, 63.09]$$

$$95\%CI = \left[60 - 2.09\frac{8}{\sqrt{20}}, 60 + 2.09\frac{8}{\sqrt{20}}\right] = [56.26, 63.74]$$

$$99\%CI = \left[60 - 2.86\frac{8}{\sqrt{20}}, 60 + 2.86\frac{8}{\sqrt{20}}\right] = [54.88, 65.12]$$

Consider the distribution of serum cholesterol levels for all female in Romania who are hypertensive and overweight. This population has an unknown mean (μ) and a known standard deviation (σ) equal to 30 mg/dl. We extracted from this population a sample of 20 subjects and we found a mean of serum cholesterol level ($\bar{X}$) equal with 220 mg/dl.

$\bar{X} = 220\, mg/dl$ is the point estimator of the unknown mean serum cholesterol level (μ) in this population

It is important to construct the interval able to take into account the sampling variability:

$$95\%CI = \left[220 - 2.09 \cdot \frac{30}{\sqrt{20}}\ to\ 220 + 2.09 \cdot \frac{30}{\sqrt{20}}\ t\right] = [206\ to\ 234]$$

Length = 234-206 = 28

$$99\%CI = \left[220 - 2.86 \cdot \frac{30}{\sqrt{20}}\ to\ 22 + 2.86 \cdot \frac{30}{\sqrt{20}}\ t\right] = [201\ to\ 239]$$

Length = 239-201 = 38

Continuous outcome

Factors affecting the length of a confidence interval

$$\overline{X} \pm Z_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

$$\bar{X} \pm t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}}$$

- As the sample size (n) increases, the length of the confidence interval decreases
- As the standard deviation (s, which reflect the variability of the distribution of individual observations) increases, the length of the confidence interval increases
- As the confidence desire increases (significance level α decreases), the length of the confidence interval increases

Confidence interval by examples

Let us suppose that there are 65 country and imported beer brands in the Romanian market. We have collected 2 different samples of 20 brands and gathered information about the price of a 6-pack, the calories, and the percent of alcohol content for each brand. Further, we know the population standard deviation (σ) of price is €1.15. Here are the samples' information:

Sample A: m_A= €4.90, s_A= €1.09

Sample B: m_B= €5.20, s_B= €0.98

Provide 95% confidence interval **estimates of population mean price** using the two samples.

Confidence intervals by example

Interpretation of the results from

- A: 95%CI = [4.90±2.09 ·(1.09/sqrt(20)] = [4.39 to 5.41] → We are **95%** confident that the true mean price is between €4.39 and €5.41
- B: 95%CI = [5.20±2.09 ·(0.98/sqrt(20)] = [4.74 to 5.66] → We are **95%** confident that the true mean price is between €4.39 and €5.41

After the fact, I am informing you know that the population mean was €4.50. Which one of the results hold?

- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the real mean.

Dichotomous outcome

Confidence interval for frequencies

Is proper to computed if:

- $n \times f > 10$ and $n(1-f) > 10$, where n = sample size, f = frequency

Estimating the standard error of a proportion

- based on the Central Limit Theorem, a sampling distribution of proportions is approximately normal, with a mean equal to the population proportion, π, and with a standard error of proportions equal to:

$$\sigma_f = \sqrt{\frac{\pi \cdot (1-\pi)}{N}}$$

- Since the standard error of proportions is generally not known, we usually work with the estimated standard error:

$$s_f = \sqrt{\frac{f \cdot (1-f)}{n}}$$

Confidence interval for frequencies

Formula:

$$f \pm Z_{\alpha/2} s_f$$

where

- f = observed sample frequency (estimate the population frequency π)
- Z = Z critical value
- s_f = estimated standard error of the frequency

Dichotomous outcome

$$f \pm Z_{\alpha/2} \cdot s_f$$

Confidence interval for frequencies

Parameter: proportion of Romanian population treated for hypertension

Sample size: n=200

Frequency of hypertension treatment: 0.25

Confidence level: 95% (Z = 1.96)

Confidence interval for f:

$$0.25 \pm 1.96 \cdot \sqrt{\frac{0.25 \cdot (1 - 0.25)}{200}}$$

$$0.25 \pm 1.96 \times 0.03$$

$$0.25 \pm 0.06$$

$$0.19 \leq \pi \leq 0.31$$

From the sample, we estimate the proportion of persons treated for hypertension to be 0.25, and we are 95% confident that the true proportion lies between the interval of 0.19 to 0.31

Two samples

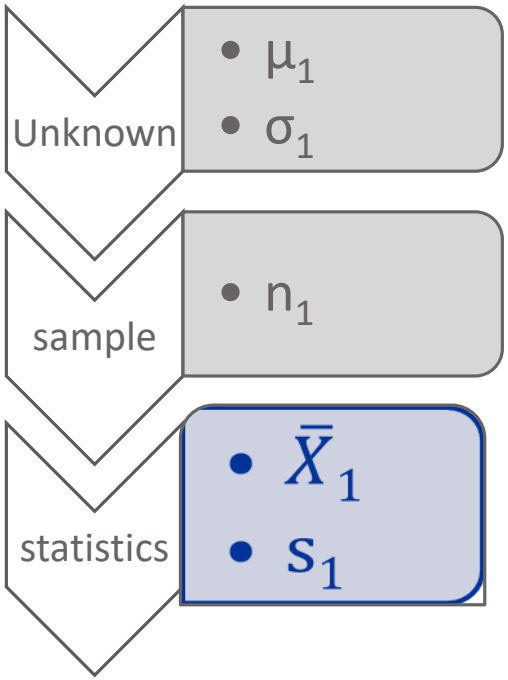
CONTINUOUS OUTCOME

DICHOTOMOUS OUTCOME

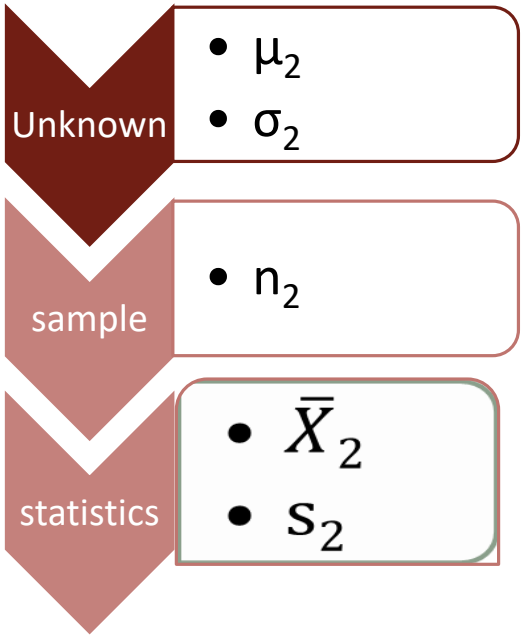
Continuous outcome

Confidence interval: Independent samples

Population 1



Population 2



Interpretation:

If 0 is in the confidence interval ➔ the difference between mean is not statistically different by zero

If 0 is not in the confidence interval ➔ the difference between mean is statistically different by zero

Estimate $(\mu_1 - \mu_2)$ with $\bar{X}_1 - \bar{X}_2$

Confidence interval for difference between means

$$(\bar{X}_1 - \bar{X}_1) \pm t_{df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_{12}}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Group 1	7	7	8	8	8	6	9	6	5
Group 2	8	10	9	6	10	8	9	7	8

	Group 1	Group 2
Mean	7.11	8.33
s	1.27	1.32
s^2	1.61	1.75

df=15.97

for $\alpha = 0.05 \longrightarrow t_{15.97} = 2.13$

$-1.22 \pm 2.13 \cdot 0.61$

$-1.22 \pm 1.30 \longrightarrow$ [-2.52 to 0.08]

Since 0 is in the confidence interval we can conclude that no significant differences between marks on Group 1 and 2 exists.

Continuous outcome

Confidence interval Paired samples

- E.g., a pre- and post-measurement design (e.g. before and after treatment)
- The goal is to compare the mean score before and after the intervention (to see if the treatment works)
- Because the sample is matched (same persons completing pre- and post measurements), cannot use aggregate means
- Parameter of interest is the mean difference, denoted μ_d
- Parameter of interest is SD of the difference scores, denoted s_d

$$\bar{X}_d \pm Z_{1-\alpha/2}\frac{s_d}{\sqrt{n}}$$

Continuous outcome

Confidence interval for frequencies

Difference between the two frequencies

$$(f_1-f_2)\pm Z_{critical}\cdot\sqrt{(f_1\cdot(1-f_1)/n_1)+(f_2\cdot(1-f_2)/n_2)}$$

Confidence interval for RR

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

	Disease+	Disease-
Exposure +	a	b
Exposure -	c	d

$$ln(\widehat{RR}) \pm Z_{crit}\sqrt{\frac{b/a}{a+b} + \frac{d/c}{c+d}} \rightarrow$$

$$\left[\exp\left(\ln(\widehat{RR}) - Z_{crit}\sqrt{\frac{\frac{b}{a}}{a+b} + \frac{\frac{d}{c}}{c+d}}\right); \exp\left(\ln(\widehat{RR}) + Z_{crit}\sqrt{\frac{\frac{b}{a}}{a+b} + \frac{\frac{d}{c}}{c+d}}\right)\right]$$

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

Confidence interval for RR

	Lung cancer+	Lung cancer -
Smoking+	30	20
Smoking-	15	35

$$\mathrm{RR} = \frac{30/(30+20)}{15/(15+35)} = 2$$

ln(RR) = 0,69

Z_{critic}= 1.96

ES = sqrt(0,06) = 0,24

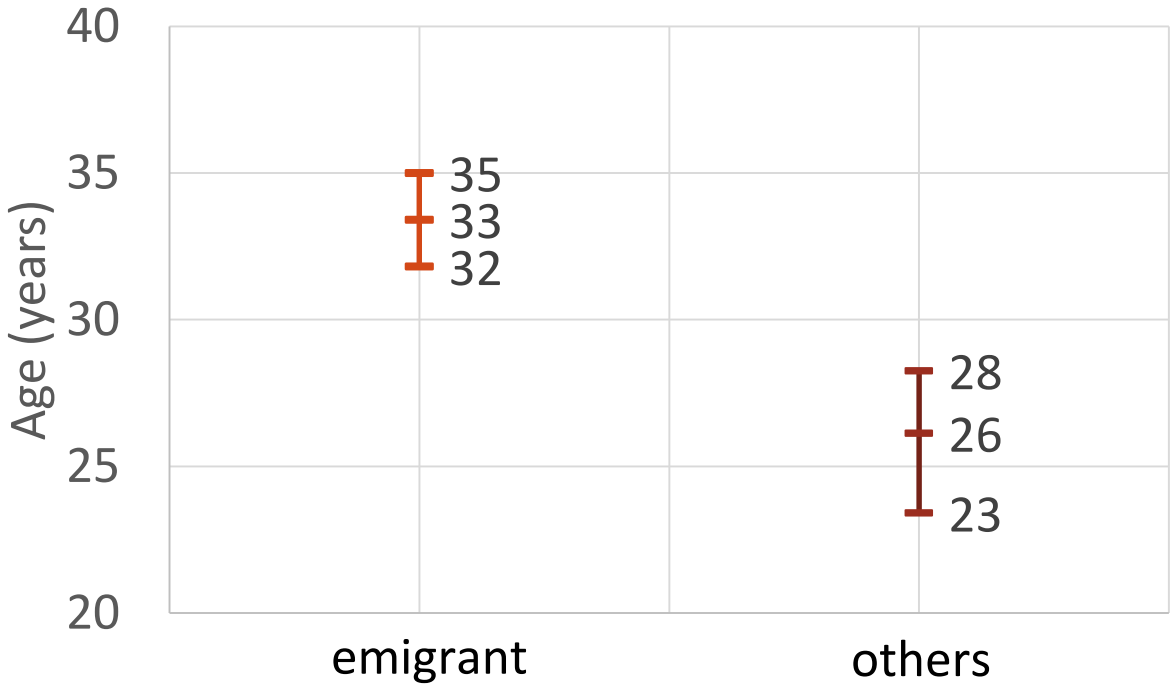
[exp(0.69-1.96*0.24) to exp(0.69+1.96*0.24)] → [1.24 to 3.23]

Comparing samples with confidence intervals

Table 1 Living conditions of the MS-MV and the immigrant population (CASEN survey 2006)

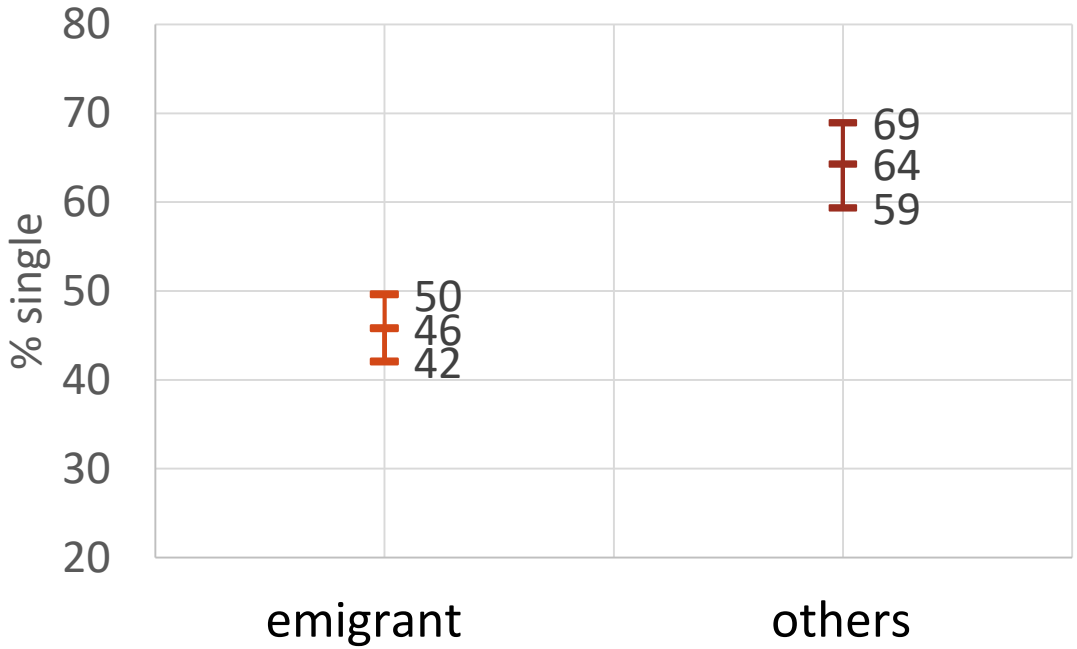
	IMMIGRANT POPULATION 1% total sample, n = 154 431 weighted population (1877 real observations)		**MS-MV GROUP 0.67% total sample, n = 108 599 weighted population (1477 real observations)**	
	% or mean	**95% CI**	**% or mean**	**95% CI**
DEMOGRAPHICS				
Mean age**	*X* = 33.41	31.81–35.00	X = 26.13	23.41–28.26
Age categories:				
<16 years old**	13.60	11.29–16.28	45.25	39.53–51.10
16-65 years old**	79.08	75.92–81.93	47.26	41.64–52.94
>65 years old	7.32	5.33–9.97	7.49	5.31–10.46
Sex (female = 1)	45.21	41.74–48.72	51.27	47.99–55.41
Marital status:				
Single**	45.81	42.06–49.62	64.30	59.36–68.95
Married**	45.49	41.66–49.36	29.39	25.09–34.10

Comparing samples with confidence intervals



Since the confidence intervals for age are not overlap on each other, we can conclude that the age of emigrant is significantly higher compared to the age others.

Comparing samples with confidence intervals



Since the confidence intervals for % declared as singles are not overlap on each other we can conclude that the others are in a significantly higher % single than emigrants.

Confidence intervals for OR

http://www.biomedcentral.com/content/pdf/1471-2458-12-1013.pdf

Table 3 Odds Ratio (OR) of presenting **any disability and any chronic condition or cancer**, adjusted by different sets of factors separately (CASEN survey 2006)

	ANY DISABILITY				**ANY CHRONIC CONDITION OR CANCER**			
	International immigrants		**MS-MV**		**International immigrants**		**MS-MV**	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
DEMOGRAPHICS								
Age	**1.04***	**1.02-1.06**	**1.04***	**1.02-1.06**	**1.05***	**1.02-1.08**	**1.02***	**1.01-1.04**
Sex (female = 1)	0.56	0.25-1.25	**0.39***	**0.20-0.75**	**2.78****	**1.26-6.71**	1.05	0.46-2.36

Determine sample size

FOR MEAN

FOR PROPORTION (FREQUENCY)

Sampling error

The required sample size can be found to reach a desired margin of error with a specified level of confidence $(1 - \alpha)$

The margin of error is also called sampling error

- the amount of imprecision in the estimate of the population parameter
- the amount added and subtracted to the point estimate to form the confidence interval

Determining sample size

Sampling error (margin of error ME)

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$ME = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$n = \frac{Z_{\alpha/2}^2\sigma^2}{ME^2}$$

Determining sample size: mean

To determine the required sample size for the mean, you must know:

- The desired level of confidence (1 - α), which determines the critical value, $Z_{\alpha/2}$
- The acceptable sampling error, ME
- The standard deviation, σ

Required sample size example

If $\sigma = 45$, what sample size is needed to estimate the mean within ± 5 with 90% confidence?

$$n = \frac{Z^2 \sigma^2}{ME^2} = \frac{(1.645)^2(45)^2}{5^2} = 219.19$$

So the required sample size is n = 220 (Always round up)

When σ is not known, select a pilot sample and estimate with the sample standard deviation, s

Determining sample size: proportion

$$ME = Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \qquad n = \frac{Z^2_{\alpha/2}\pi(1-\pi)}{ME^2}$$

To determine the required sample size for the proportion, you must know:

- The desired level of confidence (1 - α), which determines the critical value, $Z_{\alpha/2}$
- The acceptable sampling error, e
- The true proportion of events of interest, π
  - π can be estimated with a pilot sample if necessary (or conservatively use 0.5 as an estimate of π)

Required sample size example

How large a sample would be necessary to estimate the true proportion of defectives in a large population within ±3%, with 95% confidence?
(Assume a pilot sample yields p = 0.12)

$$n = \frac{Z_{\alpha/2}^2 \pi(1-\pi)}{ME^2} = \frac{1.96^2 \cdot 0.12 \cdot (1-0.12)}{0.03^2} = 450.74$$

So use n = 451

Ethical Issues

A confidence interval estimate (reflecting sampling error) should always be included when reporting a point estimate.

The level of confidence should always be reported .

The sample size should be reported.

An interpretation of the confidence interval estimate should also be provided.

Formulas

If standard deviation of the population is known or n>200

$$\bar{X} \pm t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}} \qquad \bar{X} \pm Z_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

$$(\bar{X}_1 - \bar{X}_2) \pm t_{critical} \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

$$f \pm Z_{critical} \sqrt{\frac{f(1-f)}{n}}$$

$$(f_1 - f_2) \pm Z_{critical} \cdot \sqrt{(f_1 \cdot (1-f_1)/n_1) + (f_2 \cdot (1-f_2)/n_2)}$$

Recall

Correct estimation of a population parameter is done with confidence intervals.

Confidence intervals depend by the sample size and standard error.

The confidence intervals are larger for:

- The high value of the standard error
- Small sample sizes

The sample size necessary to develop a confidence interval for the population mean or population proportion could be calculated.

Thank you for your attention!