Confidence Intervals Estimation

Sorana D. Bolboacă

OBJECTIVES

- Point estimators
- Confidence interval for mean
- Confidence interval for proportion

INFERENTIAL STATISTICS

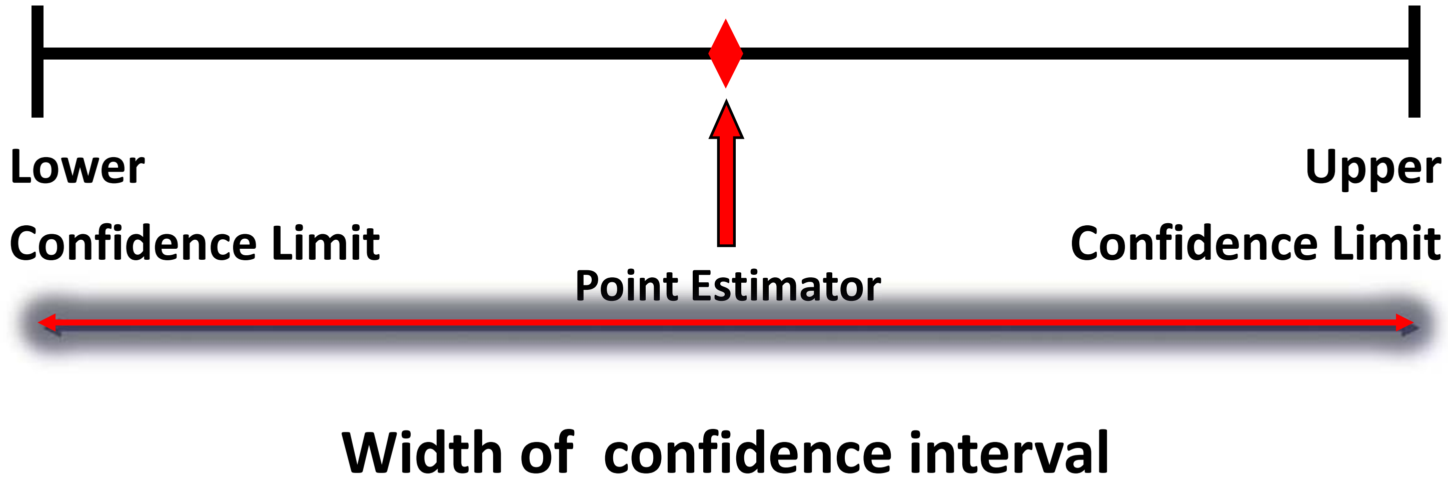
- Inferential statistics = the process of making guesses about the truth on the population by examining a sample extracted from the population
- Sample statistics = summary measures calculated from data belonging to a sample (e.g. mean, proportion, ratio, correlation coefficient, etc.)
- Population statistics = true value in the population of interest

POINT ESTIMATOR VS. INTERVAL ESTIMATION

- It is recommended to estimate a theoretical parameter by using a range of value not a single value
  - It is called confidence intervals
  - The estimated parameter belong to the confidence intervals with a high probability.

Point Estimator 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 characteristics 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 1 sample
  - Gives information about closeness to unknown population parameters
  - Stated in terms of level of confidence. (Can never be 100% confident)
- **The general formula for all confidence intervals is equal to:**

$$\text{Point Estimator} \pm \underbrace{(\text{Critical Value})(\text{Standard Error})}_{\textbf{Margin or error}}$$

Table value

INTERVAL ESTIMATION

$$\text{Point Estimator} \pm \underbrace{(\text{Critical Value})(\text{Standard Error})}_{\textbf{Margin or error}}$$

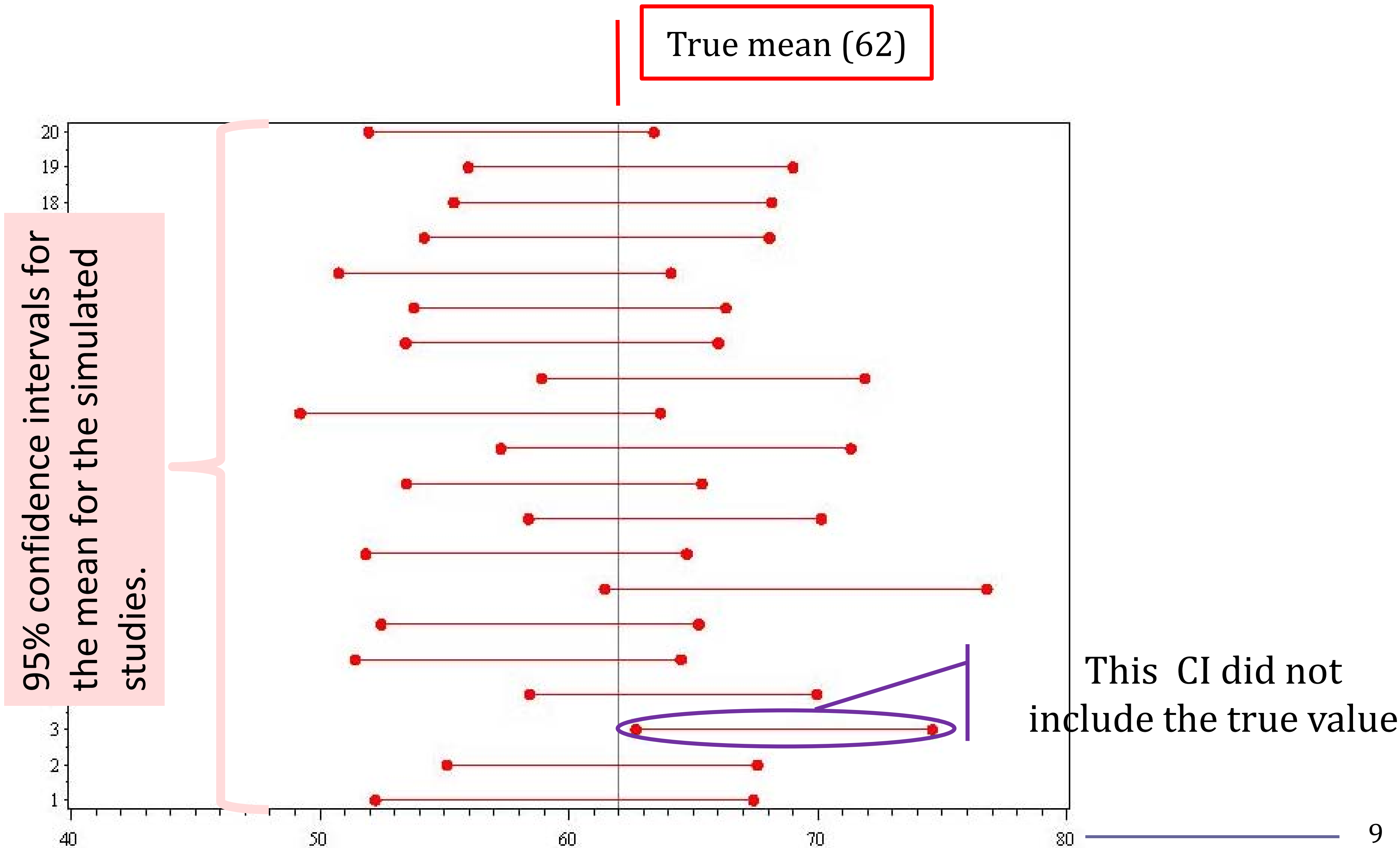
- The margin of error, and hence the width of the interval, gets smaller the as the sample size increases.
- The margin of error, and hence the width of the interval, increases and decreases with the confidence.

INTERVAL ESTIMATION

- Significance level α = 5% $\longrightarrow$ 95% confidence interval (CI)
- CI = $(1 - \alpha) = 0.95$

- 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 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



CONFIDENCE INTERVALS

- Provides:
  - A plausible range of values for a population parameter.
  - The precision of an 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)

- 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
  - Degree of confidence (standard values)

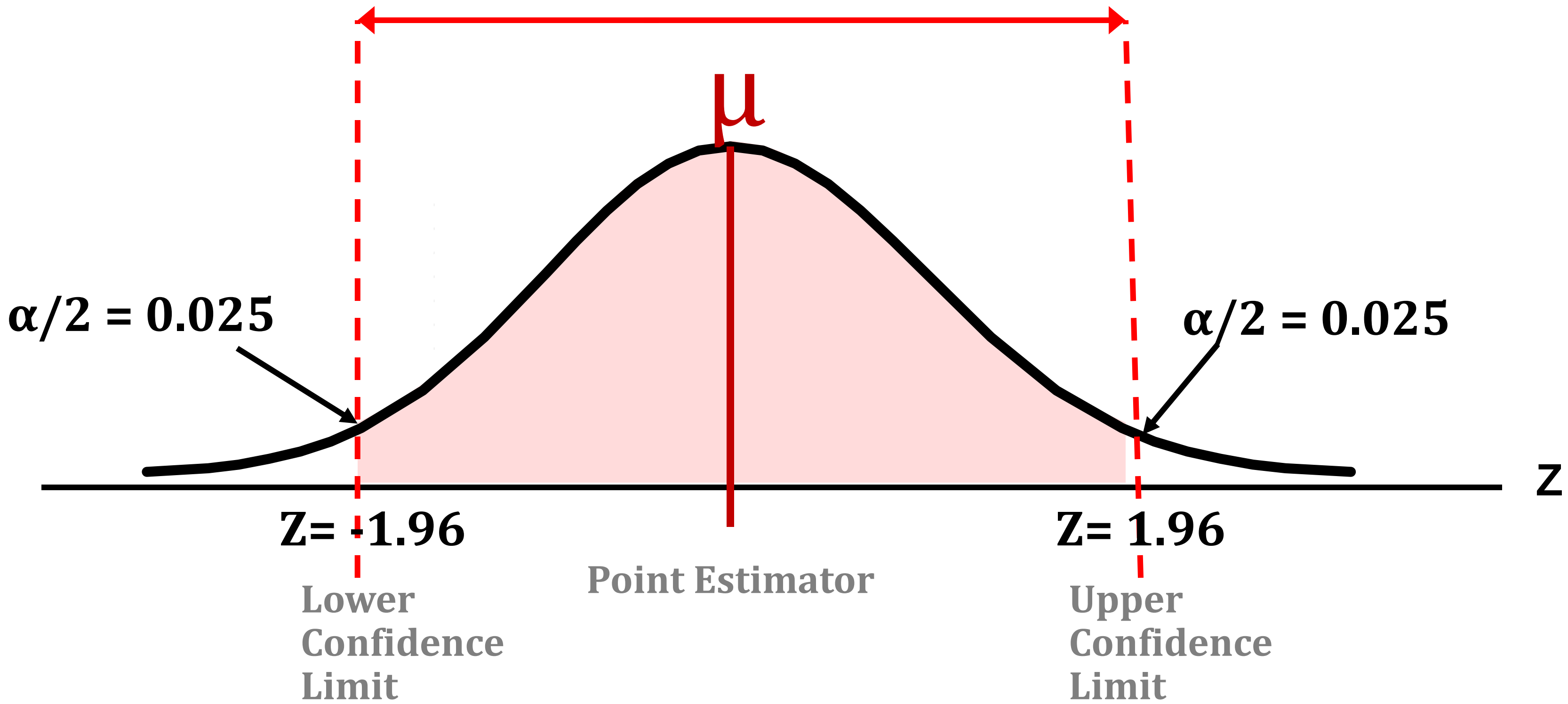
CONFIDENCE INTERVALS FOR MEANS

- Assumptions
  - Population standard deviation σ is known
  - Population is normally distributed
  - If population is not normal, use large sample

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

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

- Consider a 95% confidence interval:
- $1-\alpha = 0.95$ & $\alpha = 0.05$ & $\alpha/2 = 0.025$



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, suppose that 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 Interval by Examples

- Interpretation of the results from
  - Sample A: We are **95%** confident that the true mean price is between €4.47 and €5.33
  - Sample B: We are **95%** confident that the true mean price is between \$4.82 and \$5.58
- 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 true mean.

CONFIDENCE INTERVALS

- Interpretation of CI for difference between two means
  - If 0 is contains by the confidence intervals for the distance between an observed and theoretical mean, the difference between the two investigated means is 0.
  - If 0 is NOT contains by the confidence intervals for the distance between an observed and theoretical mean, the difference between the two investigated means is NOT 0.

Comparing Means using Confidence Intervals

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

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

CONFIDENCE INTERVAL FOR FREQUENCY

- Could be computed if:
  - n*f > 10, where n = sample size, f = frequency

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

CONFIDENCE INTERVAL FOR FREQUENCY

- We are interested in estimating the frequency of breast cancer in women between 50 and 54 years with positive family history. In a randomized trial involving 10,000 women with positive history of breast cancer were found 400 women diagnosed with breast cancer.
- What is the 95% confidence interval associated frequently observed?

- f = 400/10000 = 0.04

$$\left[0.04-1.96\sqrt{\frac{0.04\cdot 0.96}{10000}};\, 0.04+1.96\sqrt{\frac{0.04\cdot 0.96}{10000}}\right]$$

- [0.04-0.004; 0.04+0.004]
- [0.036; 0.044]

COMPARING ORS USING CONFIDENCE INTERVALS

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

Remember!

- Correct estimation of a statistical parameter is done with confidence intervals.
- Confidence intervals depend by the sample, size and standard error.
- The confidence intervals is larger for:
  - High value of standard error
  - Small sample sizes