Statistical Analysis

Interval estimation

Ana Cristina Costa

ccosta@novaims.unl.pt

NOVA IMS Information Management School

Mater in Statistics and Information Management 2021

Topics

- **LU5 – Interval estimation**
  - Introduction
  - Confidence intervals for the mean
  - Confidence intervals for the difference between means
  - Confidence intervals for the proportion
  - Confidence intervals for the difference between proportions
  - Confidence intervals for the variance
  - Sample size determination

Objectives

- **At the end of this learning unit students should be able to**
  - Define confidence interval and confidence level
  - Build and interpret confidence intervals for the mean
  - Build and interpret confidence intervals for the difference between means
  - Build and interpret confidence intervals for the proportion
  - Build and interpret confidence intervals for the difference between proportions
  - Build and interpret confidence intervals for the variance
  - Understand what factors affect the length of an interval
  - Determine the sampling error, or precision, of the point estimate
  - Calculate the sample size given the precision of the point estimate

Suggested reading

- Newbold, P., Carlson, W. L., Thorne, B. (2013). Statistics for Business and Economics. 8th Edition, Boston: Pearson, pages 291-309 (ch. 7), 315-319 (ch. 7), chapter 8.
- Mariappan, P. (2019). Statistics for Business. New York: Chapman and Hall/CRC, pages 208-221 (ch. 12).
- Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2013). *Introduction to Probability and Statistics*. 14th Edition, Boston: Brooks/Cole, Cengage Learning, pages 291-323.
- Holmes, A., Illowsky, B. & Dean, S. (2019) "VIII. Confidence Intervals". In Introductory Business Statistics. (accessed: July 2021)
- Oakley, J. (2021) "5. Interval estimates and confidence intervals". In MAS113 Part 2: Data Science, updated 2021-02-17. (accessed: July 2021)
- The Pennsylvania State University (2021) "Lesson 6: Sample Size". In STAT 415 Introduction to Mathematical Statistics. (accessed: July 2021)

Introduction

Concepts

- A confidence interval estimator for a population parameter θ is a rule for determining (based on sample information) an interval that is likely to include the parameter. The corresponding estimate is called a confidence interval estimate.
  - The interval $(\widehat{\Theta}_L, \widehat{\Theta}_U)$ is named confidence interval
- The confidence interval should
  - Contain the parameter with large probability
  - Have a small range
- The confidence interval allows evaluating the sampling error, but not considering any systematic errors associated to the process of obtaining the sample

Introduction

- **Definition**

A **(1 – α)×100% confidence interval** for the parameter θ is a random interval $(\widehat{\Theta}_L, \widehat{\Theta}_U)$, where the confidence limits $\widehat{\Theta}_L$ and $\widehat{\Theta}_U$ are two sample statistics so that

$$P\left(\hat{\Theta}_L < \theta < \hat{\Theta}_U\right) = 1 - \alpha$$

where **1–α** is the **confidence level** and

α∈]0, 1[is the **significance level**

Introduction

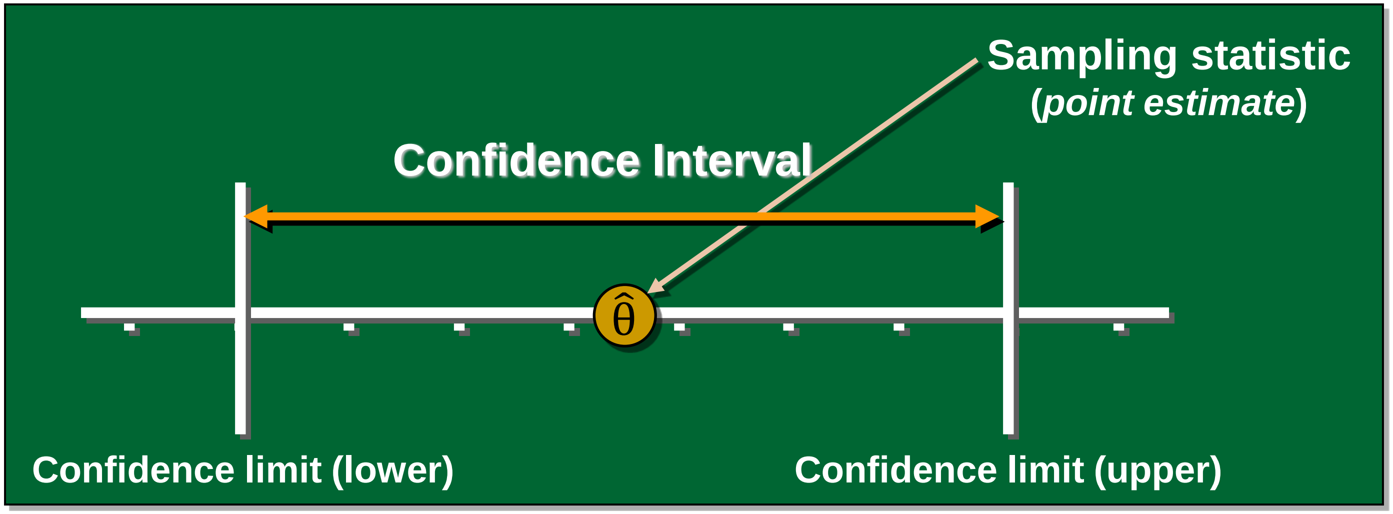
- **Different samples produce different interval estimates**
  - **1 – α** is
    - The probability that the CI contains the parameter θ (before sampling)
    - The proportion of times that the observed intervals contain the parameter value for all possible samples
  - For many estimation problems, a confidence interval estimate of the unknown parameter θ takes on the general form

$$\hat{\boldsymbol{\theta}} \mp \boldsymbol{ME}$$

where $\boldsymbol{ME}$, the margin of error, measures the **sampling error** or **precision of the estimate**.

Introduction

- **Sampling error** or **precision**
  - The (absolute) **sampling error** or (absolute) **precision** of the estimate of θ is given by **half range of the CI**
  - The relative sampling error or relative precision of the estimate of θ is given by the absolute value of the ratio between half range of the CI and $\hat{\theta}$ (in %)



Introduction

- **Fulcral variable**
  - **Definition**
    - ❑ The random variable $T=t(X_1, X_2,\ldots,X_n \mid \theta)$ is called a fulcral variable if its probability distribution does not depend on θ
  - **Example**
    - ❑ If $X_1, X_2,\ldots,X_n$ is an iid random sample drawn from a $N(\mu, \sigma)$ population then

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Z is a fulcral variable because its distribution does not depend on μ

Confidence intervals for the mean

Confidence interval for μ – case I

- Let $X_1, X_2, \ldots, X_n$ be a random sample of iid random variables from a Normal population with mean μ and known variance σ^2

- Normal population
- σ^2 known
- Any sample size

- Let $X_1, X_2, \ldots, X_n$ be a random sample of iid random variables with mean μ and known variance σ^2

- Any population
- σ^2 known
- Large sample size

$$\left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- $z_{1-\alpha/2}$ is the 1–α/2 quantile, or 100(1–α/2) percentile, from the N(0,1) distribution

- The level of confidence is exact for normal populations, but it is approximate for other populations (CLT- Central Limit Theorem)

Confidence intervals for the mean

- **Example 1**
  - The weekly wage of industrial workers has a standard deviation of σ = 40 Euros. In a sample of 81 workers, the average wage was 360 Euros. Determine 90%, 95% and 99% confidence intervals for the mean wage of the industrial workers. When the confidence level increases what happens to the precision of the sample mean?

    ✓ CI for μ: $\left[360 - z_{1-\frac{\alpha}{2}}\frac{40}{\sqrt{81}}, 360 + z_{1-\frac{\alpha}{2}}\frac{40}{\sqrt{81}}\right]$

    ✓ $z_{0.95}$=1.65, $z_{0.975}$=1.96 and $z_{0.995}$=2.58, thus

    ✓ 90% CI for μ: [352.69, 367.31]

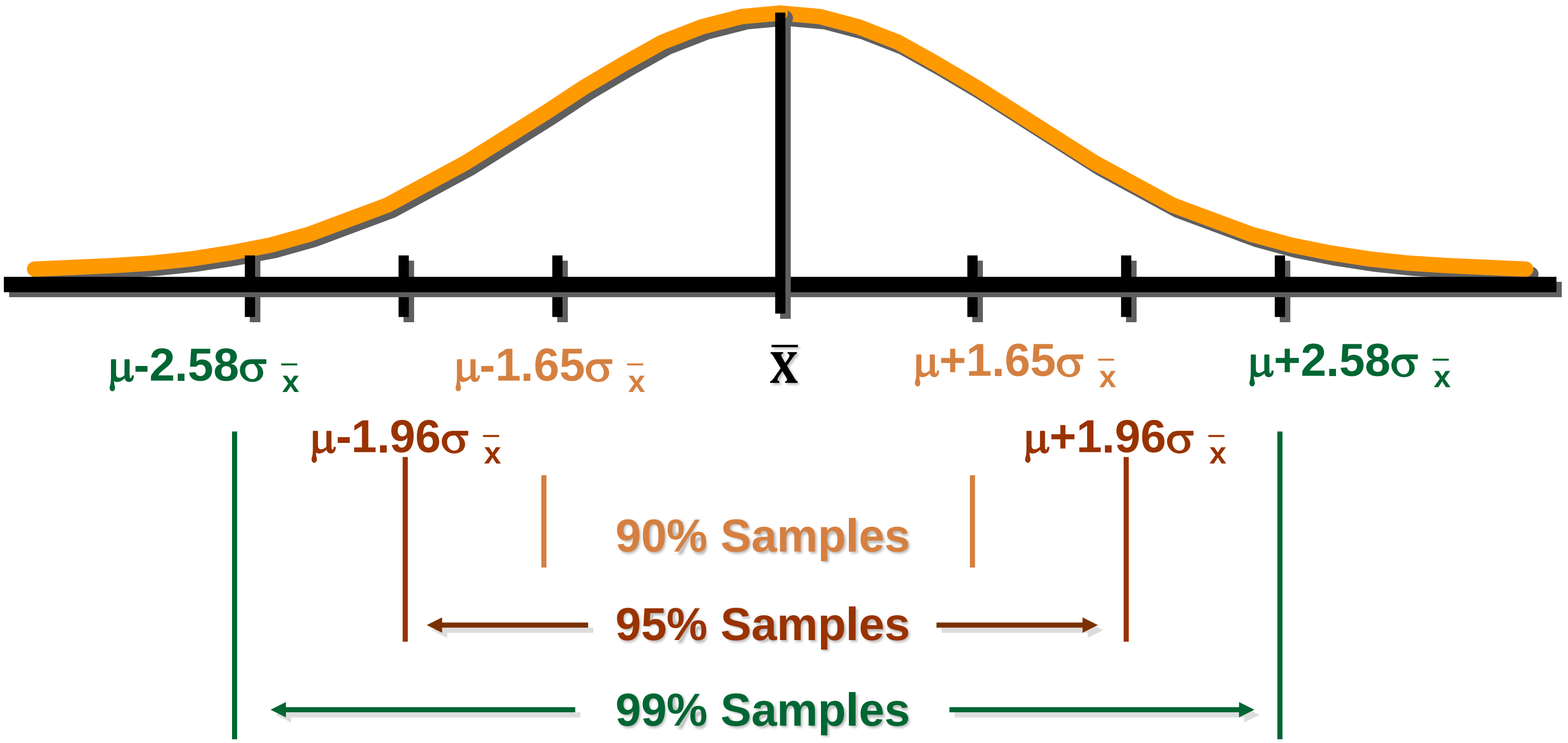
    ✓ 95% CI for μ: [351.29, 368.71]

    ✓ 99% CI for μ: [349.66, 370.34]

Solved in the **Example1** sheet of the LU5_Examples Excel file

Confidence intervals for the mean

- **The larger the confidence level**, the larger the range of the interval, **the smaller the precision of the sample mean**



$\mu-2.58\sigma_{\bar{x}}$
$\mu-1.65\sigma_{\bar{x}}$
$\bar{X}$
$\mu+1.65\sigma_{\bar{x}}$
$\mu+2.58\sigma_{\bar{x}}$
$\mu-1.96\sigma_{\bar{x}}$
$\mu+1.96\sigma_{\bar{x}}$
90% Samples
95% Samples
99% Samples

Confidence intervals for the mean

Confidence interval for μ – case II

- Let $X_1, X_2, \ldots, X_n$ be a random sample of iid random variables with mean μ and unknown variance σ^2

- Any population
- σ^2 unknown
- Large sample size

$$\overline{X} \pm z_{1-\alpha/2} \frac{S}{\sqrt{n}}$$

- $z_{1-\alpha/2}$ is the 1–α/2 quantile from the N(0,1) distribution
- The level of confidence is approximate (Slutsky theorem)

Confidence intervals for the mean

- **Example 2**
  - The director of a commercial company claims that vendors make, on average, less than 15 contacts per week. To verify this idea, 36 vendors were randomly selected and recorded the number of contacts made in a given week. The mean and variance of the sample were equal to 16 and 9 respectively. Is there evidence for the claim of the director?

✓ 95% CI for μ: $\left[16 - 1.96\sqrt{\frac{9}{36}}\,, 16 + 1.96\sqrt{\frac{9}{36}}\right] \equiv [15.02\,, 16.98]$

✓ The sample shows that the weekly average number of contacts of each vendor is between 15.02 and 16.98. Since values less than or equal to 15 are not within the confidence interval, there is evidence that the commercial director is wrong.

Solved in the **Example2** sheet of the LU5_Examples Excel file

Confidence intervals for the mean

Confidence interval for μ – case III

- Let $X_1, X_2, \ldots, X_n$ be a random sample of iid random variables from a Normal population with mean μ and unknown variance σ^2

- Normal population
- σ^2 unknown
- Any sample size

$$\overline{X} \pm t_{(n-1);1-\alpha/2} \frac{S}{\sqrt{n}}$$

- $t_{(n-1);\ 1-\alpha/2}$ is the 1–α/2 quantile from the $t_{(n-1)}$ distribution

Confidence intervals for the mean

- **Example 3**
  - The daily sales of milk in a supermarket follow a normal distribution. In a random sample of 16 days it was observed an average of 300 litres and a standard deviation of 20 litres. Obtain 90%, 95% and 99% confidence intervals for the mean daily sales in the supermarket.

✓ CI for μ: $\left[300 - t_{(15;1-\frac{\alpha}{2})}\frac{20}{\sqrt{16}}\ , 300 + t_{\left(15;1-\frac{\alpha}{2}\right)}\frac{20}{\sqrt{16}}\right]$

✓ $t_{(15;\ 0.95)}$=1.753, $t_{(15;\ 0.975)}$=2.131 and $t_{(15;\ 0.995)}$=2.947, thus

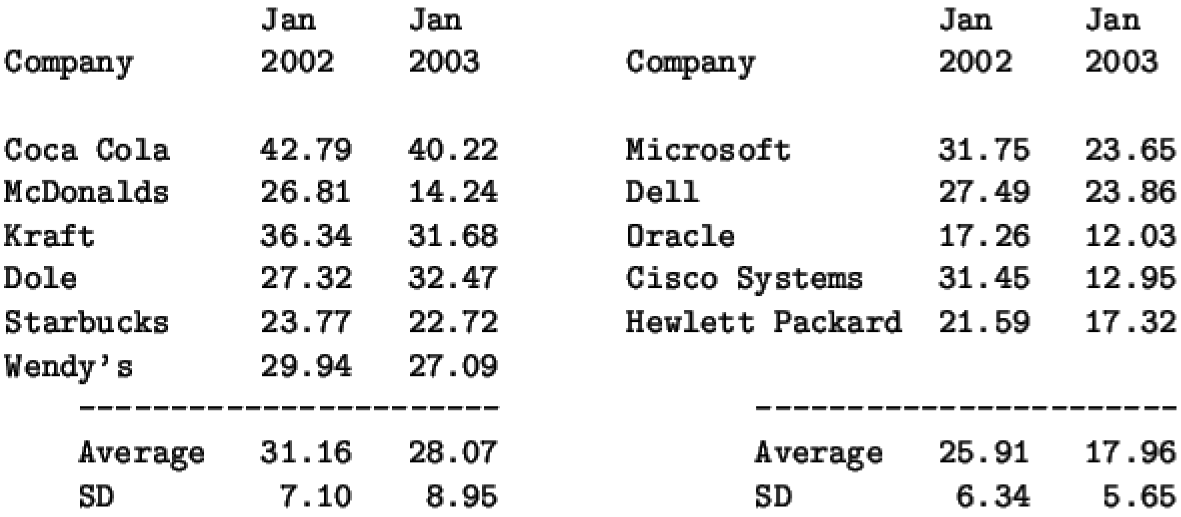
- ✓ 90% CI for μ: [291.23, 308.77]
- ✓ 95% CI for μ: [289.34, 310.66]
- ✓ 99% CI for μ: [285.27, 314.73]

Solved in the **Example3** sheet of the LU5_Examples Excel file

Confidence intervals for the difference between means

Independent *versus* paired samples

- Stock prices for 6 food service companies and 5 computer industry companies in January of 2002 and 2003



Company	Jan 2002	Jan 2003
Coca Cola	42.79	40.22
McDonalds	26.81	14.24
Kraft	36.34	31.68
Dole	27.32	32.47
Starbucks	23.77	22.72
Wendy's	29.94	27.09
Average	31.16	28.07
SD	7.10	8.95

Company	Jan 2002	Jan 2003
Microsoft	31.75	23.65
Dell	27.49	23.86
Oracle	17.26	12.03
Cisco Systems	31.45	12.95
Hewlett Packard	21.59	17.32
Average	25.91	17.96
SD	6.34	5.65

- **Independent Samples**: if we compare food company versus computer company prices

- **Paired Samples**: if we compare Jan 2002 prices versus Jan 2003 prices

Adapted from: http://www.stat.wmich.edu/s216/book/node82.html
(accessed 20 October 2017)

Confidence intervals for the difference between means

- **Independent *versus* paired samples**
  - The statistical treatment is different because paired samples involve sampling randomness just once. This gives paired-samples better statistical properties than independent samples. For independent samples, luck-of-the-draw occurs two times.
  - The best way to tell whether you have independent or paired samples is to ask: Do I have "two samples" or "one sample measured twice"?
    - <u>Repeated measurements</u>: two measurements taken on the same person or object (e.g., different treatments are applied to the same individual; making “before” and “after” measurements on a single individual or object)
    - <u>Clinical trials</u> with subjects grouped according to similar characteristics (e.g., individuals are matched by age, gender, weight, lifestyle, and other pertinent factors)
    - Studies with twins

Confidence intervals for the difference between means

Independent samples – case I: confidence interval for $\mu_1 - \mu_2$

- Let $\overline{X}_1$ and $\overline{X}_2$ be the means of two mutually independent random samples of sizes $\mathbf{n_1}$ and $\mathbf{n_2}$ drawn from two populations with means $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ and known variances $\boldsymbol{\sigma_1^2}$ and $\boldsymbol{\sigma_2^2}$, respectively.

- Normal populations
- σ_1 and σ_2 known
- Any sample sizes

- Any population
- σ_1 and σ_2 known
- Large sample sizes

$$(\overline{X}_1 - \overline{X}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^{\;2}}{n_1} + \frac{\sigma_2^{\;2}}{n_2}}$$

- $z_{1\text{-}\alpha/2}$ is the 1–α/2 quantile from the N(0,1) distribution
- The level of confidence is exact for Normal populations, but it is approximate for other populations (CLT- Central Limit Theorem)

Confidence intervals for the difference between means

- **Independent samples – case II: confidence interval for $\mu_1 – \mu_2$**
  - Let $\overline{X}_1$ and $\overline{X}_2$ be the means, and S_1^2 and S_2^2 the variances, of two mutually independent random samples of sizes $\mathbf{n_1}$ and $\mathbf{n_2}$ drawn from two populations $N(\boldsymbol{\mu_1},\boldsymbol{\sigma})$ and $N(\boldsymbol{\mu_2},\boldsymbol{\sigma})$, respectively, where σ is unknown.

- Normal populations
- $\sigma_1 = \sigma_2$ unknown
- Any sample sizes

$$(\overline{X}_1 - \overline{X}_2) \pm t_{(n_1+n_2-2);1-\alpha/2} S' \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- S'^2 is the combined estimator of σ^2: $S'^2 = \dfrac{(n_1 - 1)S_1^{\ 2} + (n_2 - 1)S_2^{\ 2}}{n_1 + n_2 - 2}$

- $t_{(n_1+n_2-2);1-\alpha/2}$ is the $1-\alpha/2$ quantile from the $t_{(n_1+n_2-2)}$ distribution

Confidence intervals for the difference between means

- **Independent samples – case III: confidence interval for $\mu_1 - \mu_2$**
  - Let $\overline{X}_1$ and $\overline{X}_2$ be the means, and S_1^2 and S_2^2 the variances, of two mutually independent random samples of sizes n_1 and n_2 drawn from two populations with means μ_1 and μ_2, respectively, and unknown variance σ^2.

❖ Any populations

❖ $\sigma_1 = \sigma_2$ unknown

❖ Large sample sizes

$$(\overline{X}_1 - \overline{X}_2) \pm z_{1-\alpha/2} S' \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- S'^2 is the combined estimator of σ^2: $S'^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

- $z_{1-\alpha/2}$ is the 1–α/2 quantile from the N(0,1) distribution (confidence level is approximate)

Confidence intervals for the difference between means

- **Independent samples – case IV: confidence interval for $\mu_1-\mu_2$**
  - Let $\overline{X}_1$ and $\overline{X}_2$ be the means, and S_1^2 and S_2^2 the variances, of two mutually independent random samples of sizes $\mathbf{n_1}$ and $\mathbf{n_2}$ drawn from two populations $N(\mu_1,\sigma_1)$ and $N(\mu_2,\sigma_2)$, respectively, where the variances are unknown.

❖ Normal populations

❖ $\sigma_1 \neq \sigma_2$ unknown

❖ Any sample sizes

$$(\overline{X}_1-\overline{X}_2)\pm t_{(r);1-\alpha/2}\sqrt{\frac{S_1^{\ 2}}{n_1}+\frac{S_2^{\ 2}}{n_2}}$$

- *r* is the **integer part** of: $r^*=\dfrac{\left(\dfrac{S_1^{\ 2}}{n_1}+\dfrac{S_2^{\ 2}}{n_2}\right)^2}{\dfrac{1}{n_1-1}\left(\dfrac{S_1^{\ 2}}{n_1}\right)^2+\dfrac{1}{n_2-1}\left(\dfrac{S_2^{\ 2}}{n_2}\right)^2}$

- $t_{(r);1-\alpha/2}$ is the $1-\alpha/2$ quantile from the $t_{(r)}$ distribution

Confidence intervals for the difference between means

- **Independent samples – case V: confidence interval for $\mu_1 – \mu_2$**
  - Let $\overline{X}_1$ and $\overline{X}_2$ be the means, and S_1^2 and S_2^2 the variances, of two mutually independent random samples of sizes $\mathbf{n_1}$ and $\mathbf{n_2}$ drawn from two populations with means $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ and unknown variances $\boldsymbol{\sigma_1^2}$ and $\boldsymbol{\sigma_2^2}$, respectively.

- ❖ Any populations
- ❖ $\sigma_1 \neq \sigma_2$ unknown
- ❖ Large sample sizes

$$(\overline{X}_1 - \overline{X}_2) \pm z_{1-\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

- $z_{1-\alpha/2}$ is the 1–α/2 quantile from the N(0,1) distribution (confidence level is approximate)

Confidence intervals for the difference between means

- **Example 4**
  - The East Electricity decided to estimate the evolution of the mean electricity consumption per household in the last year before making new investments. To do so, they selected two mutually independent random samples of domestic consumers, with sizes n_1=120 and n_2=150, from January of last year and the current year. They obtained $\bar{x}_1 = 550$ and $\bar{x}_2 = 567$ kilo-watt-hour, respectively. Assuming that the standard deviation of consumption per household in January of both years was σ_1= σ_2=110, determine a 95% confidence interval for the evolution of the mean electricity consumption (i.e., $\mu_2 - \mu_1$). Should the company make new investments?

    ✓ 95% CI for $\mu_2 - \mu_1$ (case I): [–9.41, 43.41]

    ✓ This result does not guarantee with 95% confidence that there has been a positive trend in consumption, since it admits negative values for the difference $\mu_2 - \mu_1$. Hence, before carrying out new investments, it is advisable to carry out a study with a larger sample to reduce the sampling error.

Solved in the **Example4** sheet of the LU5_Examples Excel file

Confidence intervals for the difference between means

Example 5

- A transport company decided to compare the quality of the tires of two brands, A and B, that equipped its fleet. They used the records of the past two years and found that the $n_A = 28$ tires of the A brand travelled on average 43400 km with a standard deviation of s_A=5100 km, whereas the n_B=32 tires of the B brand travelled on average 45100 km with a standard deviation of s_B=5900 km.

- Determine a 95% CI for the difference between the mean travelled distance (in km) of the tires of the two brands (i.e., $\mu_B - \mu_A$). Do brand B tires last longer than brand A tires?

  - ✓ 95% CI for $\mu_B - \mu_A$ (case V): [–1083, 4483]

  - ✓ Although one can not say with a degree of confidence of 95% that the B brand tires last longer than brand A tires, the result suggests that this may happen.

Confidence intervals for the difference between means

- **Example 6**
  - The residents of St. Paul, Minnesota, complain that traffic speeding fines given in their city are higher than the traffic speeding fines that are given in nearby Minneapolis. Independent random samples of the amounts paid by residents for speeding tickets in each of the two cities over the last 3 months were obtained. These data are in the **Example6** sheet of the LU5_Examples Excel file.

1. Compute descriptive statistics and investigate normality using the Shapiro-Wilk test for both samples in the Real Statistics Resource Pack (http://www.real-statistics.com/free-download/real-statistics-resource-pack/)

2. Assuming equal population variances, find a 95% confidence interval for the difference in the mean costs of speeding tickets in these two cities (St. Paul – Minneapolis)

   - ✓ 95% CI for difference (case II): [25.84, 52.76]
   - ✓ The CI only contains positive values (St. Paul – Minneapolis >0), thus there is evidence that the amounts paid for speeding tickets by residents from St. Paul are larger than those paid by residents from Minneapolis

Confidence intervals for the difference between means

Example 7

- An accounting firm conducts a random sample of the accounts payable for the east and the west offices of one of its clients. From these two independent samples, the company wants to estimate the difference between the population mean values of the payables. Assume normality and different population variances.

	East offices (X)	West offices (Y)
Sample size	16	11
Sample mean	$290	$250
Sample variance	225	2500

✓ 95% CI for $\mu_X - \mu_Y$ (case IV): $r^* = \dfrac{\left[\frac{225}{16}+\frac{2500}{11}\right]^2}{\frac{\left(\frac{225}{16}\right)^2}{15}+\frac{\left(\frac{2500}{11}\right)^2}{10}} \approx 11.25$

$$40 \pm 2.201\sqrt{\frac{225}{16}+\frac{2500}{11}} \qquad \Rightarrow \qquad [5.81, 74.19]$$

✓ The CI only contains positive values ($\mu_X - \mu_Y > 0$), thus we are 95% confident that East offices' payables are larger than those from West offices.

Solved in the **Example7** sheet of the LU5_Examples Excel file

Confidence intervals for the difference between means

- **Paired samples**
  - The elements are paired up so that measurements are *not independent*
  - **Examples**
    - Duplicate (double) measurements on the same sample: when data are collected twice from the same element or individual (meant to account for within-subject variability)
    - Sequential measurements (pre-test/post-test): when data from the same set of elements are collected both before and after a time period passes or before and after an intervention
    - Cross-over trials: individuals are randomly assigned to one of two treatments and then afterward assigned to the second treatment
    - Matched samples: the participants share every characteristic except for the one under investigation (e.g., individuals are matched on similar characteristics, such as age and sex, and then one individual is assigned to a treatment group and another to a control group)

Confidence intervals for the difference between means

Paired samples

- Let X_1 and X_2 be two populations with means μ_1 and μ_2

- Lets consider the **population of differences** $\mathbf{D = X_1 - X_2}$
  - The mean of population D is $\mu_D = \mu_1 - \mu_2$
  - The standard deviation of population D is σ_D
  - If the populations X_1 and X_2 are Normal then D is also Normal

- Lets consider the **sample of differences** of **size n**: $D_i = X_{1i} - X_{2i}$, i=1,2,…,n,
  - The sample mean is $\overline{X}_D = \overline{D} = \overline{X}_1 - \overline{X}_2$
  - The sample standard deviation is $S_D{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$

Confidence intervals for the difference between means

- **Paired samples & normality: confidence interval for $\mu_1 – \mu_2$**

- Normal population of differences
- σ_D known
- Any sample size

$$\overline{D} \pm z_{1-\alpha/2} \frac{\sigma_D}{\sqrt{n}}$$

- $z_{1-\alpha/2}$ is the 1–α/2 quantile from the N(0,1) distribution

- Normal population of differences
- σ_D unknown
- Any sample size

$$\overline{D} \pm t_{(n-1);1-\alpha/2} \frac{S_D}{\sqrt{n}}$$

- $t_{(n-1);1-\alpha/2}$ is the 1–α/2 quantile from the $t_{(n-1)}$ distribution

Confidence intervals for the difference between means

- **Paired samples & non-normality: confidence interval for $\mu_1 – \mu_2$**
  - Any population of differences
  - σ_D unknown
  - Large sample size

$$\overline{D} \pm z_{1-\alpha/2} \frac{S_D}{\sqrt{n}}$$

- $z_{1-\alpha/2}$ is the 1–α/2 quantile from the N(0,1) distribution (confidence level is approximate)

Confidence intervals for the difference between means

Example 8

- A random sample of six salespeople who attended a motivational course on sales techniques was monitored 3 months before and 3 months after the course. The data in the **Example8** sheet of the LU5_Examples Excel file shows the values of sales (in thousands of dollars) generated by these six salespeople in the two periods. Assume that the population distributions are normal. Find an 80% confidence interval for the difference between the two population means.

✓ $\bar{d} = -7.5$ $s_D = 18.7697$ $t_{(5;\,0.90)} = 1.4759$

✓ 80% CI for $\mu_{before} - \mu_{after}$: $-7.5 \pm 1.4759\frac{18.7697}{\sqrt{6}} \equiv [-18.81, 3.81]$

✓ The CI contains the value zero ($\mu_{before} - \mu_{after} = 0$), thus there is not enough evidence to say that the motivational improved the sales amounts with 80% confidence. However, considering that the CI has more negative than positive values, it is advisable to carry out a study with larger samples to reduce the sampling error of 11.31thousands of dollars.

Confidence intervals for the proportion

Confidence interval for *p*

- Let $\widehat{P}$ be the proportion of successes in an iid random sample of size *n*

- Bernoulli population
- np ≥ 5 and nq ≥ 5

$$\hat{P} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

- $z_{1-\alpha/2}$ is the 1–α/2 quantile from the N(0,1) distribution
- The confidence level is approximate (Central Limit Theorem)

Confidence intervals for the difference between proportions

Confidence interval for $p_1 - p_2$

- Let $\widehat{P}_1$ and $\widehat{P}_2$ be the proportions of successes in two independent random samples with *large* sizes n_1 and n_2

- Bernoulli populations
- $n_1p_1 \geq 5$ and $n_1q_1 \geq 5$
- $n_2p_2 \geq 5$ and $n_2q_2 \geq 5$

$$(\hat{P}_1 - \hat{P}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

- $z_{1-\alpha/2}$ is the 1–α/2 quantile from the N(0,1) distribution

- The confidence level is approximate

Confidence intervals for the difference between proportions

- **Example 9**
  - A bond proposal for school construction will be submitted to the voters at the next municipal election. A major portion of the money derived from this bond issue will be used to build schools in a rapidly developing section of the city, and the remainder will be used to renovate and update school buildings in the rest of the city. To assess the viability of the bond proposal, a random sample of n_1=50 residents in the developing section and n_2=100 residents from the other parts of the city were asked whether they plan to vote for the proposal: 76% of the residents in the developing section were in favor, and 65% of the other residents were in favor. Estimate the difference in the true proportions favoring the bond proposal with a 99% confidence interval.

✓ 99% CI for $p_1 - p_2$: $0.11 \pm 2.58\sqrt{\frac{0{,}76 \times 0.24}{50} + \frac{0.65 \times 0.35}{100}}$ $\Rightarrow$ [–0.09, 0.31]

✓ Since this CI contains the value zero ($p_1 - p_2 = 0$), there may be no difference in the proportions favoring the bond issue in the two sections of the city

Solved in the **Example9** sheet of the LU5_Examples Excel file

Confidence intervals for the variance

Confidence interval for σ^2

- Let S^2 be the variance of an iid random sample of size *n* drawn from a Normal population with variance σ^2

- Normal population

$$\left[\frac{(n-1)S^2}{\chi^2_{(n-1);1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{(n-1);\alpha/2}} \right]$$

- $\chi^2_{(n-1);1-\alpha/2}$ and $\chi^2_{(n-1);\alpha/2}$ are the $1-\alpha/2$ and $\alpha/2$ quantiles, respectively, from the $\chi^2_{(n-1)}$ distribution

- The validity of the interval estimator for σ^2 depends far more critically on the assumption of normality than does that of the interval estimator for μ

- Notice this CI does not have the usual form: sample point estimator ± margin of error.

Confidence intervals for the variance

- **Example 10**
  - The manager of Northern Steel, Inc., wants to assess the temperature variation in the firm’s new electric furnace. It is known that temperatures are normally distributed. A random sample of 25 temperatures over a 1-week period is obtained, and the sample variance is found to be $s^2 = 100$. Find a 95% confidence interval for the population variance temperature.

✓ $\chi^2_{(24);0.975} = 39.364$; $\chi^2_{(24);\alpha/2} = 12.401$

✓ $\left[\frac{(25-1)100}{39.364}, \frac{(25-1)100}{12.401}\right] \equiv [60.97, 193.53]$

Solved in the **Example10** sheet of the LU5_Examples Excel file

Sample size determination

- **Based on a confidence interval for μ**

  - Normal population
  - σ² known
  - Any sample size

  - Any population
  - σ² known
  - Large sample size

  - What should the sample size be so that the absolute sampling error is less than ε for a given (1–α)% confidence level?

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

$$Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \varepsilon \quad \Rightarrow \quad n \geq \left(\frac{Z_{1-\alpha/2}\sigma}{\varepsilon}\right)^2$$

! If the number *n* resulting from the sample-size formula is not an integer, then round up to the next integer value in order to guarantee that the confidence interval does not exceed the required sampling error

Sample size determination

- **Based on confidence interval for μ**
  - How can we estimate an *unknown* σ^2 ?
    - ❑ One way is to use a relatively small-scale pilot study from which the sample standard deviation *S* is used as a point estimate of the population standard deviation
    - ❑ A second approach is to estimate it by using the results of a similar study done at some time in the past
    - ❑ A third method is to estimate σ as 1/6 the approximate range of data values

Sample size determination

- **Example 11**
  - A state politician would like to determine the average amount earned during summer employment by state teenagers during the past summer's vacation period. She wants to have 95% confidence that the sample mean is within $50 of the actual population mean. Based on past studies, she has estimated the population standard deviation to be σ = $400. What sample size should she consider?

✓ $Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \varepsilon \quad \Leftrightarrow \quad 1.96\frac{400}{\sqrt{n}} \leq 50 \ \Rightarrow n \geq \left(\frac{1.96\times 400}{50}\right)^2 = 245.8624$

✓ Since we can't include a fraction of a person in the sample, we round up to n=246 to ensure 95% confidence in being within $50 of the population mean

Sample size determination

- **Based on confidence interval for *p***
  - What should the sample size be so that the absolute sampling error is less than ε for a given $(1–\alpha)$% confidence level?

$$z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \varepsilon \;\Rightarrow\; n \ge \frac{(z_{1-\alpha/2})^2\,\hat{p}(1-\hat{p})}{\varepsilon^2}$$

- The value of $\hat{p}$ that maximizes *n* is $\hat{p} = 0.5$

  (value obtained when we equal to zero the 1st derivative; the 2nd derivative is negative)

- A conservative estimate of *n* that can be used when *p* is totally unknown is

$$n \ge \left(\frac{z_{1-\alpha/2}}{2\varepsilon}\right)^2$$

Sample size determination

- **Example 12**
  - Suppose a company wants to launch a new product in the market. To this end, a telephone survey will be carried out to estimate the proportion of potential customers for the new product. The company wishes to be 95% confident that the sampling error will be no more than 0.07 (i.e., 7% points). What should the sample size be?

✓ $$n \geq \left(\frac{z_{1-\alpha/2}}{2\varepsilon}\right)^2 = \left(\frac{1.96}{2 \times 0.07}\right)^2 = 196$$

Sample size determination

- **Example 13**
  - The **SampleSize** sheet of the LU5_Examples Excel file shows how we can use Excel to determine the necessary sample size for estimating a population mean or proportion
  - With these procedures, it is very easy to examine “what-if” scenarios and instantly see how changes in confidence level or specified maximum sampling error will affect the required sample size

Interval estimation

Do the homework*!*

Ana Cristina Costa, ccosta@novaims.unl.pt