Applied Statistics

Rita Gaio[1,2]

argaio@fc.up.pt

[1]Faculdade de Ciências da Universidade do Porto
[2]Centro de Matemática da Universidade do Porto

CENTRO DE MATEMÁTICA UNIVERSIDADE DO PORTO

U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Confidence Intervals

Two possibilities for the estimation of a parameter θ (belonging to the population) from values observed in a random sample:

- estimate the *most probable value* of θ, given the observations in the sample - **point estimation**. This theory leads to **hypothesis tests**.
- estimate a region that, somehow, contains information about θ - **interval estimation**. This theory leads to **confidence intervals**.

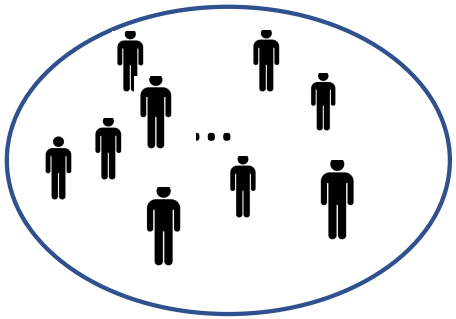
Confidence Intervals

In this section, we will **review confidence intervals**:

- skipping the theory about point estimation
- concentrating on the interpretation and application of confidence intervals.

Confidence Intervals

What is the mean height μ of portuguese men?



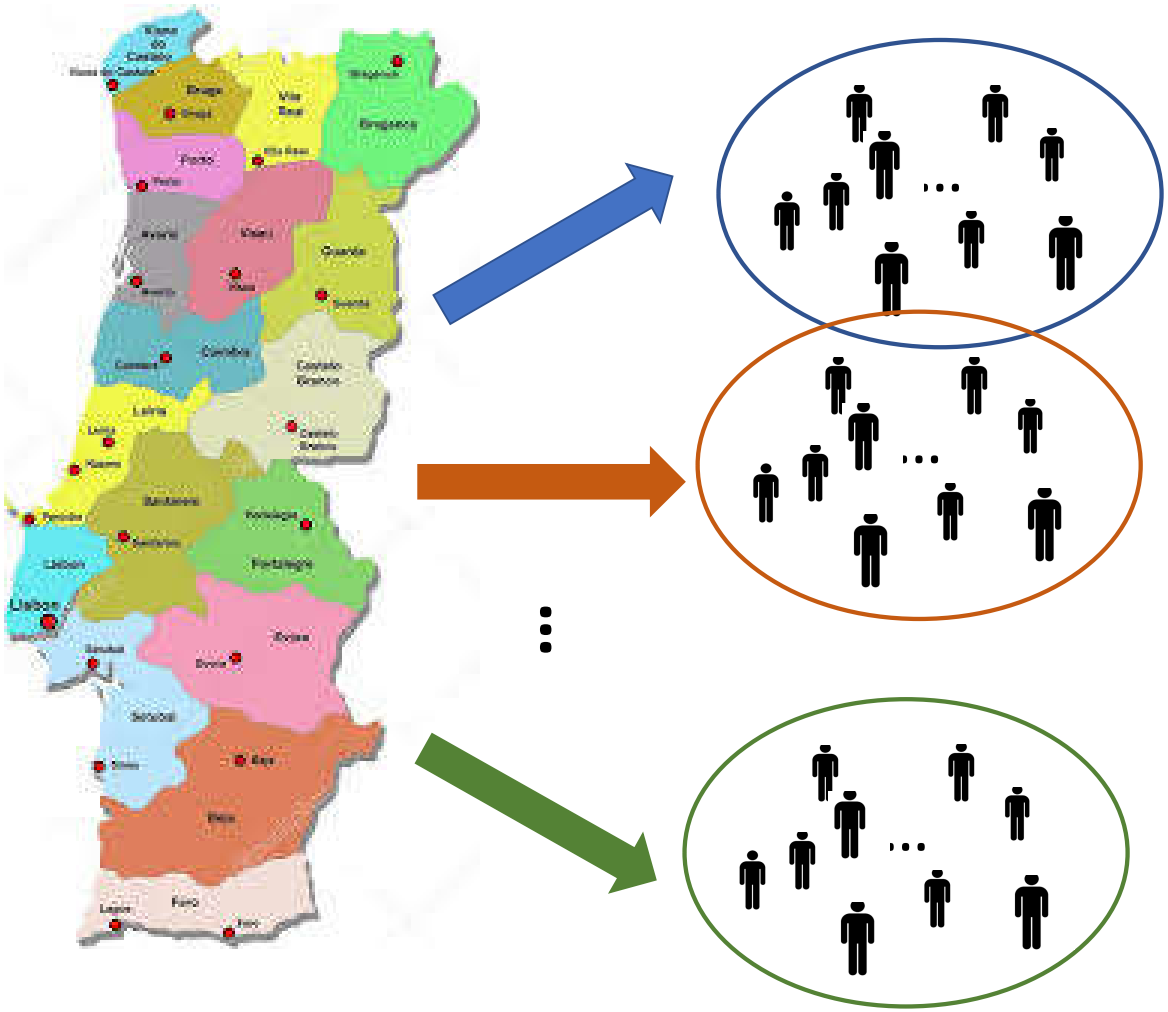
n=198

$\bar{x} = 173\ cm$

$s = 46\ cm$

Confidence Intervals

Changing sample leads to other estimates of μ.



n=198
$\bar{x} = 173\ cm$
$s = 46\ cm$

n=198
$\bar{x} = 172\ cm$
$s = 51\ cm$

n=198
$\bar{x} = 173\ cm$
$s = 48\ cm$

Confidence Intervals

Questions:

- How do the values of $\overline{x}$ change?
- $\overline{X}$: r.v. representing all sample means from random samples of size n What is the distribution of $\overline{X}$?

Recall:

- $\overline{x}$, s^2, s are numbers, computed in the sample
- $\overline{X}$, S^2, S are random variables, computed in the population

Confidence Intervals

Solution: **Central Limit Theorem**
Regardless of the distribution of X,

$$\overline{X} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

Equivalently,

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

Prop: If $X \sim N(\mu, \sigma^2)$ then $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Confidence Intervals

Suppose $X_1, \ldots, X_n$ are independent samples from a normal distribution with unknown mean μ, and known variance σ^2.

Then a (symmetric) $100(1-\alpha)\%$ **confidence interval for** μ is the interval

$$\left(\overline{x} - \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2}, \overline{x} + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2}\right)$$

where $Z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard $N(0,1)$-distribution.

For instance, for a 95% confidence interval, $Z_{0.975} = 1.96$ and, for a 99% confidence interval, $Z_{0.995} = 2.6$.

Confidence Intervals

θ: unknown parameter (fixed quantity), to be estimated

Situation 1: Given a random sample $X = (X_1, ..., X_n)$, a $100(1-\alpha)\%$ confidence interval for θ is the **random** interval

$$(A(X), B(X))$$

such that

$$P(A(X) \leq \mu \leq B(X)) = \alpha.$$

Situation 2: Given a realization of a random sample $x = (x_1, ..., x_n)$, a $100(1-\alpha)\%$ confidence interval for θ is the **real** interval

$$(a(x), b(x)).$$

In this case, nothing can be said about the probability of μ being in $(a(x), b(x))$.

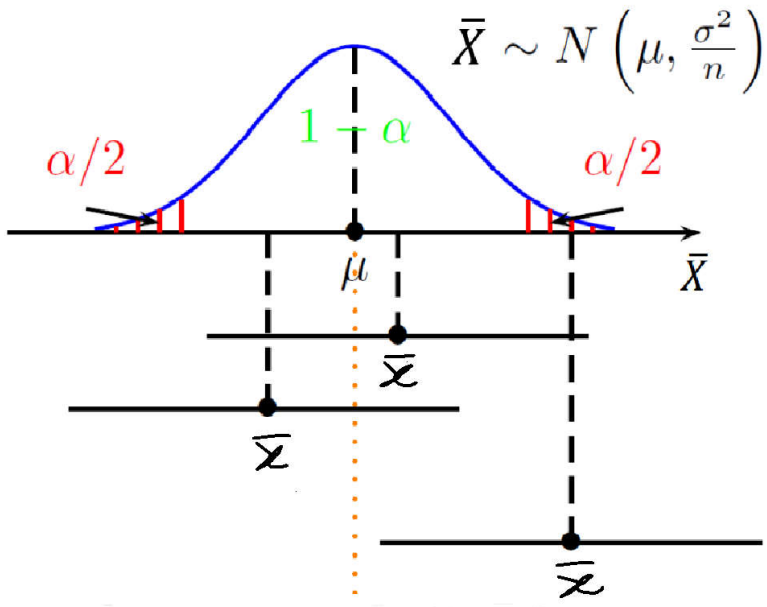
Confidence Intervals

For example, for $\alpha = 0.05$ and $\theta = \mu$:

- 95% of the time, the *random interval* $(a(X), b(X))$ covers μ,
  i.e.,
  95% of all real intervals $(a(x), b(x))$ contain μ (and 5% of them don't)
- by a frequentist argument, among N confidence intervals for μ, approximately $0.95N$ contain μ
- the confidence interval is for μ and not $\overline{x}$. Indeed, $\overline{x}$ is completely known once we have a random sample.

Confidence Intervals

- given a realization of a r.s., $(a(x), b(x))$ may or may not contain μ



- given a realization of a r.s., it is not true that $(a(x), b(x))$ contains μ with probability $1-\alpha$. That probability is either 0 or 1.

Confidence Intervals - Normal Distributions

- **Parameter**: Mean μ (σ^2 known)

  Requirements: $X \sim N(\mu, \sigma^2)$

  $(1-\alpha)100\%$ **CI**: $\overline{x} \pm \frac{\sigma}{\sqrt{n}} N_{1-\frac{\alpha}{2}}(0,1)$

- **Parameter**: Mean μ (σ^2 unknown)

  Requirements: $X \sim N(\mu, \sigma^2)$

  $(1-\alpha)100\%$ **CI**: $\overline{x} \pm \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)$,[1]

- **Parameter**: Variance σ^2

  Requirements: $X \sim N(\mu, \sigma^2)$

  $(1-\alpha)100\%$ **CI**:

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right)$$

[1] $\left(1-\frac{\alpha}{2}\right)$-quantile of the $t(n-1)$ distribution, also denoted by $t_{n-1,1-\frac{\alpha}{2}}$.

Confidence Intervals - Normal Distributions

- **Parameter**: Difference of means $\mu_1 - \mu_2$ (independent pop; σ_1^2, σ_2^2 known)

  Requirements: $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$

  $(1-\alpha)100\%$ **CI**: $\overline{x}_1 - \overline{x}_2 \pm \sigma^* N_{1-\frac{\alpha}{2}}(0,1)$, $\quad \sigma^* = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- **Parameter**: Difference of means $\mu_1 - \mu_2$ (independent pop.; σ_1^2, σ_2^2 unknown)

  Requirements: $X_1 \sim N(\mu_1, \sigma_1^2)$, $\quad X_2 \sim N(\mu_2, \sigma_2^2)$

  $(1-\alpha)100\%$ **CI**: $\overline{x}_1 - \overline{x}_2 \pm s^* t_{1-\frac{\alpha}{2}}(\nu)$, $\quad s^* = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and ν is the largest integer not exceeding

$$\nu' = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}$$

Confidence Intervals - Normal Distributions

- **Parameter**: Ratio of Variances $\frac{\sigma_2^2}{\sigma_1^2}$

  Requirements: $X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2)$

  $(1-\alpha)100\%$ **CI**:

$$\left(\frac{s_2^2}{s_1^2} F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1), \frac{s_2^2}{s_1^2} F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \right)$$

- **Parameter**: Difference of means $\mu_1 - \mu_2$ (paired samples)

  Requirements: $D = X_1 - X_2 \sim N(\mu_D, \sigma_D^2)$

  $(1-\alpha)100\%$ **CI**:

$$\overline{x}_1 - \overline{x}_2 \pm \frac{s_D}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)$$

Confidence Intervals - Large Samples

- **Parameter**: Mean μ (σ^2 known)

  $(1-\alpha)100\%$ **CI**:

$$\overline{x} \pm \frac{\sigma}{\sqrt{n}} N_{1-\frac{\alpha}{2}}(0,1),$$

- **Parameter**: Mean μ (σ^2 unknown)

  $(1-\alpha)100\%$ **CI**:

$$\overline{x} \pm \frac{s}{\sqrt{n}} N_{1-\frac{\alpha}{2}}(0,1),$$

- **Parameter**: Proportion π

  Requirements:[2] $n \geq 30$, $np > 5$ e $n(1-p) > 5$

  $(1-\alpha)100\%$ **CI**:

$$p \pm \sqrt{\frac{p(1-p)}{n}} N_{1-\frac{\alpha}{2}}$$

[2] Otherwise, use the binomial distributions

Confidence Intervals - Large Samples

- **Parameter**: Difference of means $\mu_1 - \mu_2$
  (independent pop.; σ_1^2, σ_2^2 finite)

  $(1-\alpha)100\%$ **CI**: $\overline{x}_1 - \overline{x}_2 \pm s^* N_{1-\frac{\alpha}{2}}(0,1), \quad s^* = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$

- **Parameter**: Difference of means $\mu_1 - \mu_2$ (paired samples)

  $(1-\alpha)100\%$ **CI**: $\overline{x}_1 - \overline{x}_2 \pm \frac{s_D}{\sqrt{n}} N_{1-\frac{\alpha}{2}}(0,1), \quad D = X_1 - X_2$

- **Parameter**: Difference of proportions $\pi_1 - \pi_2$

  Requirements: $n_1, n_2 \geq 30$; $n_1 p_1, n_2 p_2 > 5$ e
  $n_1(1-p_1), n_2(1-p_2) > 5$ [3]

  $(1-\alpha)100\%$ **CI**:

  $$p_1 - p_2 \pm p^* N_{1-\frac{\alpha}{2}}(0,1)$$

  onde $p^* = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

[3] Otherwise, use the Fisher exact test, with the hypergeometric distributions.

Confidence Intervals - Question

Question: What will happen to the **margin of error** in a confidence interval as (keeping all other values fixed):

- the confidence level is increased?
- the sample size is increased?
- the variability in the sample is increased?

Confidence Intervals - Examples

Question: Suppose we take a random sample of 400 households in Porto. We find that they have an average income of 34 700 eur with an SD of 15 400. What can we infer about the average income of all households in Porto, with a 99% confidence level?

Answer: Although the distribution of incomes is not known to be normal, the average of 400 incomes is approximatively normally distributed, once 400 is *large* - CLT. A 99% confidence interval is

$$34700 \pm \frac{15400}{\sqrt{400}} N_{0.995}(0,1)$$

which gives aprox. $(32717, 36683)$ eur.

In **R**, $N_{0.995(0,1)}$ is given by `qnorm(0.995)`.
Note that, by default, mean=0 and sd=1.

Confidence Intervals - Examples

Question: Assume that the r.v. height follows a normal distribution in the whole population of portuguese male adults. Data from a random sample of 28 men provided a sample mean of 173 cm and a std deviation of 46 cm.

- What can be said about the mean height of all portuguese male adults, with 95% confidence level?
- What is the margin of error in the above confidence interval?

Answer: Once we have normality for the r.v. height, the variance of height in the population is unknown and $\alpha = 0.05$, the formula to be applied is

$$\overline{x} \pm \frac{s}{\sqrt{n}} N_{0.975}(0,1)$$

which gives $173 \pm \frac{46}{\sqrt{28}} 1.96$, that is, aproximately $(156, 190)$ cm.
The associated margin of error is $\frac{46}{\sqrt{28}} 1.96 \approx 17$ cm.

Exercise: repeat the question without assuming normality in the population. Does the 95% confidence interval remain the same?

Confidence Intervals - Exercises

Exercise: The Gallup organisation carried out a poll in October, 2005, of Americans' attitudes about guns. They surveyed 1012 Americans, chosen at random. 30% said they personally owned a gun. If they'd picked different people, purely by chance, they would have gotten a somewhat different percentage. What does this survey tell us about the true proportion of Americans who own guns, and what is the associated margin of error? Consider a 90% confidence level.

Solution: $(27.6, 32.4)\%$ thus the margin of error is of 2.4%.

Confidence Intervals - Exercises

Exercise: The Gallup organisation wants to carry out a poll again next year in Americans' attitudes about guns. How many Americans, chosen at random, should they pick if they are willing to accept a margin of error of 3% for the true proportion of Americans who own guns? Consider the usual confidence level.

Solution: $n \geq 1068$

Confidence Intervals - Exercises

A researcher would like to estimate the mean difference in weight following a specific diet using a two-sided 95% confidence interval. The standard deviation estimate, based on the range of paired differences, is 18.3 Kg. The researcher would like the interval to be no wider than 20 Kg. What is the required sample size?

Extension: To be solved in **R**, with an appropriate function.
The confidence level is set at 0.95, but include 0.99 for comparative purposes. The researcher would like the interval to be no wider than 20 Kg but will also examine widths of 12, 16, 24 and 28 Kg. The goal is to determine the necessary sample size for each situation.

Confidence Intervals - Exercises

3539 participants attending the 7th examination of the Offspring cohort in the Framingham Heart Study. In 1623 men, mean (std deviation) blood pressure was 128.2 (17.5) mmHg while in 1911 women, the values were 126.5 (20.1) mmHg.
In the population, are there significant differences in mean systolic blood pressure between men and women, with a 95% confidence level?

Solution: $(0.44, 2.96)$ mmHg; with 95% confidence level, men have a significantly higher mean blood pressure than women.