Biomedical Engineering Degree

2. Estimation

Felipe Alonso Atienza
✉felipe.alonso@urjc.es
@FelipeURJC

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Rey Juan Carlos

References

1. R. Bernard. *Fundamentals of Biostatistics*. Ed.: Thompson. Chapter 6
2. B. Caffo. *Statistical Inference for Data Science*. Leanpub. Chapter 7
3. D. Díez, M Cetinkaya-Rundel and CD Barr. *OpenIntro Statistics*. Chapter 5.

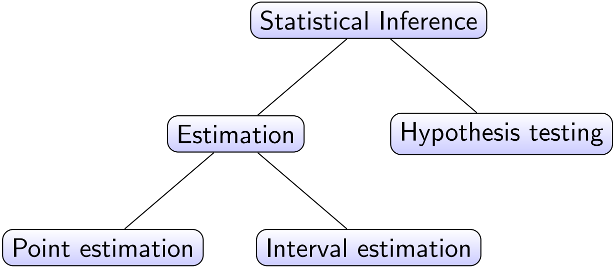
Outline

Example

We want to measure the average height of the university students **population** in Spain. Who would you do it?

1. You measure the height of each university student in Spain and then average the results.
2. You measure the height of a **sample** of university student in Spain and then average the results.
   - How to choose this sample? How many samples would you need?
   - How close would our **estimation** be to the real value?
   - How likely would our **estimate** be within a certain range of values?
3. You assume that the height of university student in Spain follows a Normal distribution with mean value μ and variance σ^2
   - Does this assumption help? Is this a valid assumption?
   - How can we estimate μ? and σ^2?
   - Under this assumption, can we **compare** the height of students from Valencia versus students from Bilbao?

Mind map



- Statistical inference: is the process and result of drawing conclusions about a population from **one or more samples**
- Point estimation: estimating the values of specific population parameters
- Interval estimation: specify a range within which the parameter values are likely to fall
- Hypothesis testing: is concerned with testing whether the value of a population parameter is equal to some specific value.

Random sample vs population

- **Population**, reference, or target refer the group we want to study.
- From the population, a sample is drawn at random (**random sample**) to select some members of the population such that **each member is independently chosen**.
- If we can take action on the sampling process, we must consider:
  1. Building a sample big enough to have reliable data
  2. Building a representative sample of the population
     - Example: randomized clinical trials

Outline

Point estimation

- We will study two estimators for different conditions and distributions:
  1. Estimation of the mean
  2. Estimation of the variance

Given a specific random sample $x_1, x_2, \ldots, x_n$, how can we estimate μ and σ^2?

- We will not study how to mathematically derived (robust) estimators using different criteria like
  1. Maximum likelihood, maximum a posteriori
  2. Method of moments
  3. Least squares

Estimation of the mean

Given a specific random sample $x_1, x_2, \ldots, x_n$, how can we estimate μ?

- Answer: use the **sample mean**

$$\boxed{\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

- *But, why?* Let's examine its properties ...
- ... *OK, but, how can I do it?* Use the **sampling distribution**

Sampling distribution

We must forget about our particular sample for the moment and consider the set of all possible samples of size n that could have been selected from the population

SAMPLING DISTRIBUTIONS ARE NEVER OBSERVED, BUT WE KEEP THEM IN MIND

Example

- Sorry, but I do not believe you, my estimator is better than yours:
  a. Mine: $\hat{\mu}_1 = \frac{1}{n}\sum_{n=i}^{n} x_i$
  b. Yours: $\hat{\mu}_2 = x_1$

Exercise 1: Let's run some simulations

- Represent the sampling distribution of both estimators. To do so, consider:
  1. The population follows a Normal distribution with $\mu = 2$ and $\sigma^2 = 2$
  2. Use $n = 10$

Exercise 1 (cont.): Let's do some thinking (it is free!)

- Which is the best estimator? and why?
- What if we increase/decrease n, how does it affect to our results?

Properties of an estimator

Take-home message

The estimator $\hat{\theta}$ of a distribution parameter θ is always a random variable

- Thus, properties of an estimator have to be assessed statistically:
  - Analytically, through its pdf
  - Computationally, through computer simulations (Monte Carlo methods)

Bias

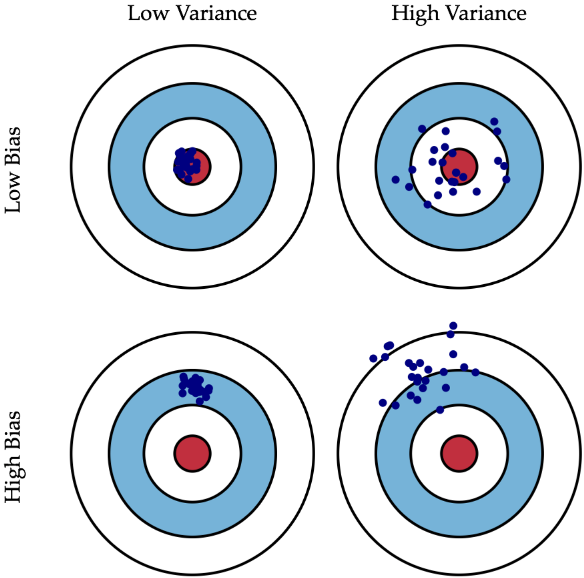
$$b = E[\hat{\theta}] - \theta$$

where b is the **bias**. If $b = 0$ we say that $\hat{\theta}$ is **unbiased**

Variance

$$\mathrm{Var}(\hat{\theta}) = E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right]$$

Bias vs Variance



Example

Calculate the bias and variance of our estimators

a. Mine: $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$

b. Yours: $\hat{\mu}_2 = x_1$

Bias (example solution)

- As for $\hat{\mu}_1$

$$E[\hat{\mu}_1] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{n=1}^{n} \mu = \mu$$

  Thus, $\hat{\mu}_1$ is **unbiased**.

- The estimator $\hat{\mu}_2$

$$E[\hat{\mu}_2] = E[x_1] = \mu$$

  is also **unbiased**

- In terms of bias, both estimators are equally good.
- If both are unbiased, which one should I choose?

Variance (example solution)

- Variance for $\hat{\mu}_1$ is

$$\mathrm{Var}(\hat{\mu}_1) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(x_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

- And for $\hat{\mu}_2$

$$\mathrm{Var}(\hat{\mu}_2) = \mathrm{Var}(x_1) = \sigma^2$$

- So, $\hat{\mu}_1$ is better than $\hat{\mu}_2$

Standard error (se) of the mean

$$\mathrm{se} = \frac{\sigma}{\sqrt{n}}$$

Exercise 2

Let X be a r.v. that follows a $\mathcal{N}(\mu, \sigma)$ with $\mu = 100$ and $\sigma = 15$. What's the sampling distribution of $\bar{X}$ for different values of n? Check your analytical solution with computer simulations.

Exercise 3

Let X be a r.v. that follows a uniform $\mathcal{U}(a, b)$ with $a = 150$ and $b = 190$. What's the sampling distribution of $\bar{X}$ for different values of n? Check your analytical solution with computer simulations.

Central-Limit Theorem

- Let $X_1, X_2, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2.

For large n ($n > 30$), $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ even if the underlying distribution of individual observations in the population is not normal.

- If we standardized the sampling distribution then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows a $\mathcal{N}(0, 1)$

Estimation of the variance

Given a specific random sample $x_1, x_2, \ldots, x_n$, how can we estimate σ^2?

- Answer: use the (corrected) **sample variance**

$$\boxed{\hat{\sigma}^2 = s_*^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean

- This constitutes an **unbiased** estimator of the variance. Proofs here and here.
- In Python, we can use `np.std(x,ddof=1)` for calculating the (corrected) sample standard deviation s_*

Outline