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?

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.

Example

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

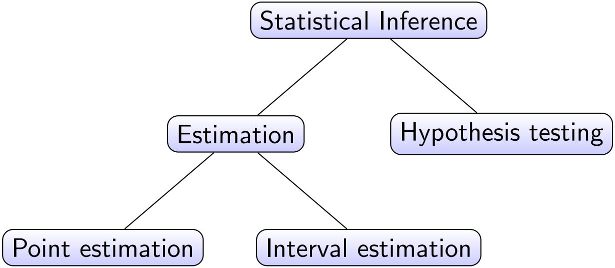
1. You measure the height of each university student in Spain and then average the results.
2. You measure the height of a **sample** of university student in Spain and then average the results.
    - How to choose this sample? How many samples would you need?
    - How close would our **estimation** be to the real value?
    - How likely would our **estimate** be within a certain range of values?

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

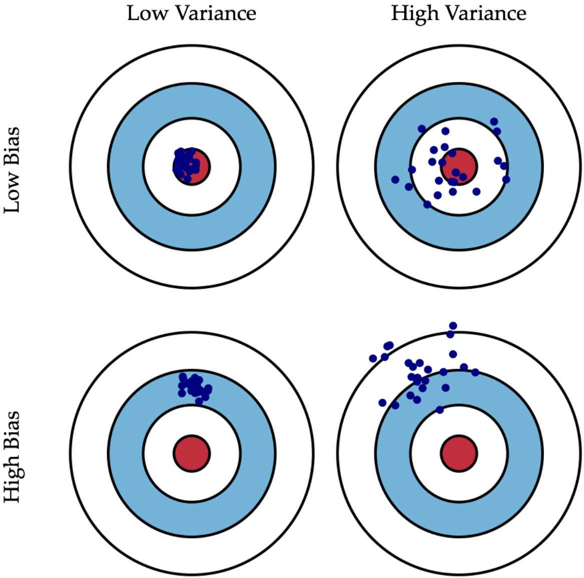
$$b = E[\hat{\theta}] - \theta$$

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 $\hat{\mu}_2$

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

Interval Estimation

- Interval estimation: specify a range within which the parameter values are likely to fall

Interval estimation of the mean

- From our previous discussion we know that

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

- In the standardized form,

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

- Hence, $95\,\%$ of the z values from repeated samples of size n will fall within the interval $[-1.96, +1.96]$

$$P(z_{0.025} < Z < z_{0.975}) = P(-1.96 < Z < 1.96) = 0.95$$

- We would then have a $95\,\%$ certainty that $\bar{X}$ would fall in the interval

$$[\mu - 1.96 \cdot \sigma/\sqrt{n}, \mu + 1.96 \cdot \sigma/\sqrt{n}]$$

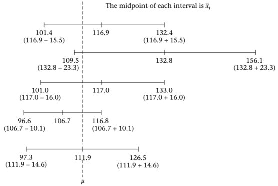
Confidence interval (CI) of the mean

- Or equivalently, the probability that these limits contain μ is $95\,\%$
- The quantity:

$$\boxed{\overline{X} \pm 1.96 \cdot \sigma/\sqrt{n}}$$

  is called a $95\,\%$ **interval** for μ

- Notice that the **interval is different for each sample**



Over the collection of all $95\,\%$ CIs that could be constructed from repeated random samples of size n, $95\,\%$ will contain the parameter μ

Confidence interval (CI) of the mean

- More generally, we can write $95\,\% = 100\,\%(1-\alpha)$, so that

$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1-\alpha$$

  - $1-\alpha$ is called the **confidence level**
  - α is called the **significance level**

- And then[1]

$$\boxed{\overline{X} \pm z_{1-\alpha/2} \cdot \sigma/\sqrt{n}}$$

is the confidence interval of μ with a confidence level $100\,\%(1-\alpha)$

[1] Using that $z_{1-\alpha/2} = -z_{\alpha/2}$

$t-$distribution

- In practice $\boldsymbol{\sigma}$ **is rarely known**.
- Thus, it is reasonable to estimate σ by the sample standard deviation s_*
- However, the quantity

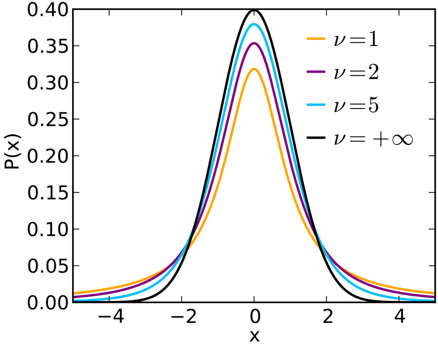
$$t = \frac{\bar{X} - \mu}{s_*/\sqrt{n}}$$

  is **no longer normally distributed**. It distributes as a **Student's** t-**distribution**[2]
    - The shape of this distribution depends on the sample size n.
    - Thus, the t-distribution is not a unique distribution but is instead a **family of distributions** indexed by a parameter referred to as the **degrees of freedom (df)** of the distribution

[2] This problem was first solved in 1908 by a statistician named William Gossett. For his entire professional life, Gossett worked for the Guinness Brewery in Ireland. He chose to identify himself by the pseudonym "Student"

$t-$distribution

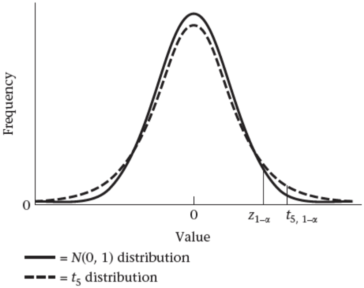
If $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$ and are independent, then $\frac{\bar{X}-\mu}{s_*/\sqrt{n}}$ is distributed as a $t-$distribution with $d = (n-1)$ df, which is sometimes referred to as the t_d-distribution.



$t-$distribution percentiles

- We denote the $u-$th percentile of the t_d distribution (d degrees of freedom) as $t_{d,u}$ so that

$$P(t_d < t_{d,u}) = u$$



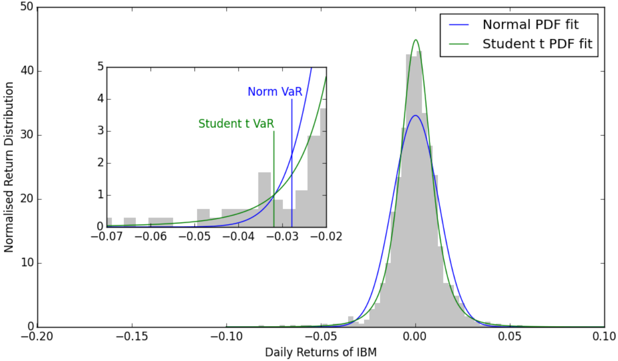
Python code

```
from scipy.stats import norm, t
print(norm.ppf(0.95))
print(t(df=5).ppf(0.95))
```

- $>>$ 1.6448536269514722
- $>>$ 2.015048372669157

$t-$distribution example in the real world

- Asset portfolio optimization Value at Risk metric calculation[3]



[3]This figure was extracted from here

CI of the mean (unknown variance)

- Remember that the confidence interval when the variance is know, can be calculated as

$$\overline{X} \pm z_{1-\alpha/2} \cdot \sigma/\sqrt{n}$$

- And now, if the **variance is unknown**, we have

$$\boxed{\overline{X} \pm t_{n-1,1-\alpha/2} \cdot s_*/\sqrt{n}}$$

which is the confidence interval of μ with a confidence level $100\,\%(1-\alpha)$

  - if $n > 200$ then $t_{n-1} \sim \mathcal{N}(0,1)$ and in this case

$$\overline{X} \pm z_{1-\alpha/2} \cdot s_*/\sqrt{n}$$

Factors affecting the length of a CI

$$\boxed{\overline{X} \pm t_{n-1,1-\alpha/2} \cdot s_*/\sqrt{n}}$$

- n: if $n \uparrow \Rightarrow$ length of CI $\downarrow$
- s_*: if $s_* \uparrow \Rightarrow$ length of CI $\uparrow$
- α: if $\alpha \uparrow \Rightarrow$ length of CI $\downarrow$

Interval estimation of the variance

- To obtain an interval estimate for σ^2, we need a **new family of distributions** called chi-square (χ^2) distributions

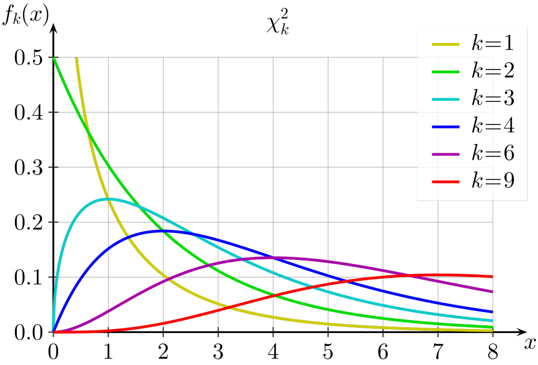
Let $X_1, X_2, \ldots, X_n$ be **independent** r.v.'s following a $\mathcal{N}(0,1)$ distribution. Then,

$$G = \sum_{i=1}^{n} X_i^2$$

is said to follow a chi-square distribution with n degrees of freedom (df), which is denoted by $\boldsymbol{\chi_n^2}$

- Distribution parameters:
  - $E[\chi_n^2] = n$
  - $\mathrm{Var}[\chi_n^2] = 2n$

Chi-square distribution

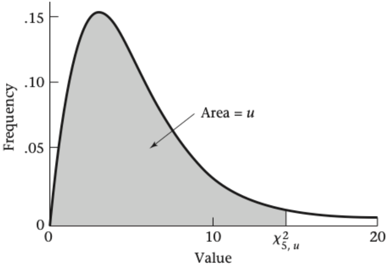


- Only takes on positive values and is always skewed to the right
- The skewness diminishes as n increases

Chi-square percentiles

- We denote the $u-$th percentile of the χ^2_d distribution (d degrees of freedom) as $\chi^2_{d,u}$ so that

$$P(\chi^2_d < \chi^2_{d,u}) = u$$



Python code

```
from scipy.stats import chi2
print(chi2(df=5).ppf(0.95))
```

- $>> 11.070497693516351$

Interval estimation

- Let Z_i be a standard normal. Then, by definition

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

- Since $Z_i = \frac{X_i - \mu}{\sigma}$, we can write

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2 \Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \sim \chi_n^2$$

- If we estimate μ by $\overline{X}$ (we usually don't know μ), then we lose 1 df.

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 \sim \chi_{n-1}^2$$

Interval estimation

- Then, by using the relationship $s_*^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ it results in

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(n-1)s_*^2}{\sigma^2} \sim \chi^2_{n-1}$$

- So we can obtain that

$$s_*^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

- And then

$$P\left(\frac{\sigma^2}{n-1}\chi^2_{n-1,\alpha/2} < s_*^2 < \frac{\sigma^2}{n-1}\chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha$$

Confidence interval of the variance

- Thus, the interval

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

is a $100\,\%(1-\alpha)$ CI for σ^2