TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

MSc European Economy and Business Law

Statistics Pre-course 2022

Topic 4: Statistical inference

Francesca Centofanti

francesca.centofanti@students.uniroma2.eu

Outline

- Fundamentals of Inference
- Central Limit Theorem for Estimation
- Hypothesis Testing
- Introduction to Estimation tools

Fundamentals of Inference: Inference vs Probability

- **Probability** starts from the *population*, which is described by the means of a probability distribution function, and predicts what happens in a *sample* extracted from it: given a probability law, which is the probability of a given event?
- **Inference** starts from a *sample* and describes the observed data with the aim of inferring relevant information on the *population*: given a sample, which are the parameters of the probability law that generated my sample?

Fundamentals of Inference: Definitions

- **Estimate**: recover some parameter explaining the *phenomenon* that generates the data
  1. *Point estimate*: a single number that is our best guess for the parameter
  2. *Interval estimate*: an interval of numbers that is believed to contain the actual value of the parameter
- **Hypothesis testing**: using data to validate certain *statements* or *predictions*

Fundamentals of Inference: Making Inference

- For making inference we assume that our data (collected in a sample) come from a *probability distribution*. The probability distribution is assumed to be known but its parameters are unknow
- *Examples*:
  1. We want to estimate the true proportion (p) of Americans who favour doctor-assisted suicide
  2. We want to estimate the true daily mean (μ) time spent driving their motor vehicles by the Americans

Formally speaking, when we record for each element of the sample the corresponding opinion, we are making a *Bernoulli experiment* for each observation $i = 1, 2, ..., n$:

$$X_i = \begin{cases} 0 & \textit{with probability } 1-p \\ 1 & \textit{with probability } p \end{cases}$$

Example 1

1. True proportion (p) of Americans who favour doctor-assisted suicide:

$$p = \frac{Americans\ \ who\ \ favour\ \ doctor\ \ assisted\ \ suicide}{Americans}$$

BE CAREFUL! The population of interest (Americans) is too large: we observe a random sample and estimate such proportion on the sample:

$$\widehat{p} = \frac{Americans\ \ who\ \ favour\ \ doctor\ \ assisted\ \ suicide\ \ in\ \ the\ \ SAMPLE}{Americans\ \ in\ \ the\ \ SAMPLE}$$

Suppose that, in a given sample of $n = 50$ we find 35 Americans that favour doctor assisted suicide. If, within the following days, we collect another sample, we expect to obtain a different result. The sample proportion has its own variability!

Example 2

1. True daily mean (μ) time spent driving their motor vehicles by the Americans:

$$\mu = \frac{\textit{total amount of time spent by Americans driving}}{\textit{total number of Americans who drive}}$$

BE CAREFUL! We always have to use an estimate obtained from a sample:

$$\overline{x} = \frac{\textit{total amount of time spent by Americans in the SAMPLE driving}}{\textit{total number of Americans who drive in the SAMPLE}}$$

Central Limit Theorem for Estimation: Point Estimates

- The *Central Limit Theorem* states:

Given n random variables $X_1, X_2, ..., X_n$ that are independent and that have the same distribution with mean $\mathbb{E}[X] = \mu$ and variance $\mathbb{V}[X] = \sigma^2$, the following holds

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ \textit{is distributed as} \ \overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Knowing the probability distribution associated to the results we obtained in a given sample allows us to solve problems like:
  1. In a survey it was reported that 33 percent of women believe in the existence of aliens. If 100 women are selected at random, what is the probability that more than 45 percent will say that they believe in aliens?
  2. A tire manufacturer claims that its tires last an average 60,000 miles with a standard deviation of 3,000 miles. 64 tires are placed on test. What is the probability that their failure miles will be more than 59,500 miles?

Problems' Solutions

1. We need to compute $P(\widehat{p} > 0.45)$ and we know that $\widehat{p} \sim N(0.33, 0.0022)$, since $\widehat{p} = 0.33$ and $p(1-p)/n = 0.0022$:

$$P(\widehat{p} > 0.45) = P\left(\frac{\widehat{p} - 0.33}{\sqrt{0.33(1-0.33)/100}} > \frac{0.45 - 0.33}{\sqrt{0.33(1-0.33)/100}}\right)$$
$$= P(Z > 2.55) = 0.0054$$

2. $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{59{,}500-60{,}000}{30{,}000/\sqrt{64}}$, thus:

$$P(\overline{x} > 59{,}500) = P(Z > -1.33) = 0.4082$$

Central Limit Theorem for Estimation: Confidence Interval

- So far we evaluated the *sampling distribution* of quantities tailored at estimating *population parameters* ($\widehat{p}$ for estimating population proportion p and sample mean $\overline{x}$ for estimating population mean μ). These are **point estimates**
- The sampling distribution allows us to construct *intervals* of plausible values associated to the estimate of a *population parameter*. These are called **confidence intervals**:

Given n random variables $X_1, X_2, ..., X_n$ and a parameter of interest θ, and interval $[L_1(X_1, X_2, ..., X_n), L_2(X_1, X_2, ..., X_n)]$ is Confidence Interval at $1 - \alpha$ confidence if it contains with probability $1 - \alpha$ the unknown θ parameter

Confidence Interval: Construction

- The Central Limit Theorem states:

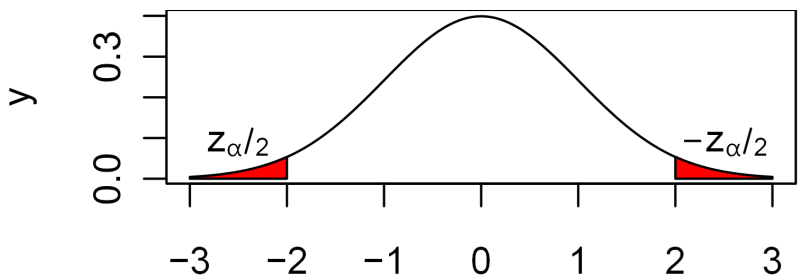
$$\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})/n}} = Z \sim N(0,1) \;\; thus \;\; 1-\alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$$

- Knowing that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P(-z_{\alpha/2} \leq \frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})/n}} \leq z_{\alpha/2}) = 1-\alpha$$

- We obtain:

$$P\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq \widehat{p} \leq \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = 1-\alpha$$



Exercise

- In a random sample composed by $n = 100$ persons, 77% of them declared that they regularly pray. Determine a 90% confidence interval for the true proportion of people that pray.
- So we have:
  1. $\widehat{p} = 0.77$
  2. $z_{\alpha/2} = 1.645$
  3. $\sqrt{\widehat{p}(1-\widehat{p})/n} = 0.042$
- Thus the interval that contains with probability 0.9 the true proportion of persons that pray is:

$$[0.77 - 1.645 \times 0.042; 0.77 + 1.645 \times 0.042]$$

Confidence Interval for the Sample mean

- The Central Limit Theorem states:

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = Z \sim N(0,1) \;\; thus \;\; 1-\alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$$

- Thus the following holds:

$$P(-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1-\alpha$$

- Thus the interval which contains the true μ with probability $1-\alpha$ is:

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

Exercise

- A random sample composed of $n = 100$ public school teachers has a mean salary of $31,578\$$ with a Standard deviation of $4,415\$$. Construct a 99% for the true mean salary.
- So we have:
  1. $z_{\alpha/2} = 2.575$
  2. $\overline{x} = 31,578$
  3. $n = 100$
  4. $\sigma = 4,415$
- Thus the true average salary lies within the following interval with probability 99%:

$$\left[31,578 - 2.575 \times \frac{4,415}{100}; 31,578 + 2.575 \times \frac{4,415}{100}\right]$$
$$= [31,578 - 1,136; 31,578 + 1,136]$$

Hypothesis Testing

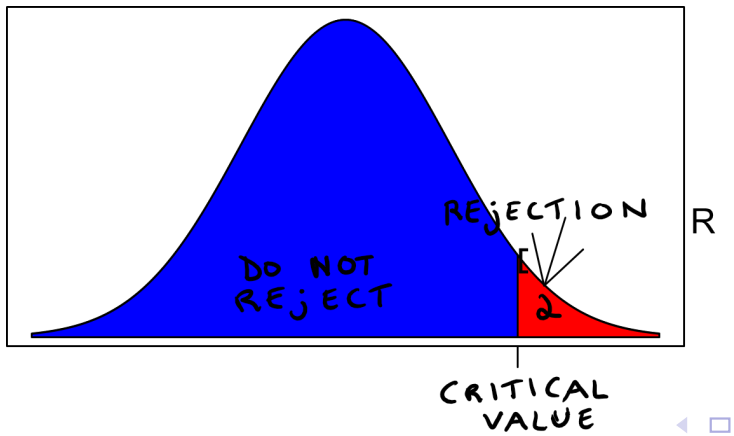
The main goal of *statistical testing* is to check whether the data support certain statements (hypothesis), usually expressed in terms of population parameters for variables measured in the study

- A **statistical hypothesis** is an opinion about a population parameter θ:
  1. $\theta = \theta_0 \Rightarrow$ punctual hypothesis
  2. $\theta \geq \theta_0$ or $\theta \leq \theta_0 \Rightarrow$ one-sided hypothesis
  3. $\theta \neq \theta_0 \Rightarrow$ two-sided hypothesis
- In an **hypothesis test** we compare two types of hypothesis:
  1. *Null hypothesis* (H_0): the hypothesis that is held to be true unless sufficient evidence to the contrary is obtained
  2. *Alternative Hypothesis* (H_1): the new theory we would like to test

Hypothesis Testing: Rejection and Non-Rejection region

A *statistical test* represents a **rule** that allows to discern the samples that lead to the *non-rejection* of the null hypothesis from those that lead instead to its *refusal*

- The test is based on the value assumed by a statistic, which is nothing more than a sample statistic whose distribution is known under the null hypothesis
- The set of values of the test statistic that allows one to accept (not reject) the null hypothesis is called the **non-rejection region**, while the set of values that leads to the rejection of the null hypothesis is called the **rejection region**



Example

- Your teacher claims that 60% of American males are married. You feel that such proportion is higher. In a random sample of $n = 100$ American males, 65 of them were married. Test the teacher's claim at 5% of significance

$$\begin{cases} H_0 : p \leq 0.6 \\ H_1 : p > 0.6 \end{cases}$$

REMEMBER! From the Central Limit Theorem we know that:

$$Z_p = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

Example: Testing Procedure

- So we have:
  1. Null Hypothesis ($H_0 : p \leq 0.6$)
  2. Alternative Hypothesis ($H_1 : p > 0.6$)
  3. Test Statistics (a function of the data whose distribution is known)
  4. A critical value ($\alpha = 0.05$)
  5. A decision rule (reject the null hypothesis if $z > z_\alpha$, with $z_\alpha = 1.645$)

REMEMBER! The value we are testing is $p_0 = 0.6$

- We compute the test statistics $z = 1.0201$
- According to the stated decision rule, we do not reject H_0 because $1.0201 < 1.645$

Elements of Hypothesis Testing: Test Statistics

The *Test Statistics* is a Numerical Summary of a dataset

- It is used because its sampling distribution is known (it can be calculated)
- The sampling distribution allows us to evaluate if the difference among the values observed within the sample and the value stated within the Null Hypothesis is statistically significant or is a consequence of the variability among the different samples

Elements of Hypothesis Testing: Critical Value

The *critical value* separates Rejection Region from Non-Rejection Region

- Whenever you do a test, you specify the level of significance (typical values are 1% or 5%)
- **Type I and Type II Errors**:
  1. *Type I*: committed by rejecting a true null hypothesis. The probability of committing a Type I error is called α (the level of significance)
  2. *Type II*: committed when a researcher fails to reject a false null hypothesis. The probability of committing a Type II error is called β

Hypothesis Testing: Alternative Method

P-value is another way to reach statistical conclusion in hypothesis testing

- p-value defines the smallest value of α for which the null hypothesis can be rejected:
  1. $p - value \leq \alpha \Rightarrow$ reject H_0
  2. $p - value > \alpha \Rightarrow$ do not reject H_0

Example: to solve a test $H_0 = \theta \leq \theta_0$ versus $H_1 = \theta > \theta_0$ (one-tailed test) we compute

$$p - value = P(Z > |z|)$$

where z is the value of the statistic Z in the sample.
BE CAREFUL! The test statistics and the p-value tests have to confirm each other

- **Basic concepts**
  1. *Parameter*: numerical characteristic of the population that we are trying to recover (hence typically unknown). *Example*: λ in a Poisson, μ in a Gaussian, etc.
  2. *Statistics*: numerical function of the sample that does not directly depend on any unknown parameter. *Example*: test statistic z
  3. *Estimator*: a statistic used to estimate the population parameter. *Example*: $\overline{X}$ is an estimator for μ
  4. *Estimate*: the value of an estimator corresponding to an observed sample. *Example*: $\overline{x}$ is an estimate corresponding to $\overline{X}$

Estimation Tools: Estimator

The aim of the *estimator* is to try to recover the distribution that generated the data

- The are several automatic ways to derive an estimator, depending on how to use the data to recover the generating distribution:
  1. Methods of Moments
  2. Maximum likelihood

Point Estimation: Methods of Moments Estimator

The core idea is to equate *sample moments* to *population moments*, i.e.

$$\begin{cases} \mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \mathbb{E}[X^2] = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ \mathbb{E}[X^3] = \frac{1}{n}\sum_{i=1}^{n} X_i^3 \\ \dots \end{cases}$$

Example: Consider a random sample $X_1, ..., X_n \sim Unif(0, \theta)$, for which $\mathbb{E}[X] = \frac{\theta}{2}$. The **MOM estimator** is found by equating $\mathbb{E}[X] = \frac{\theta}{2}$ to $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\frac{\theta}{2} = \overline{X} \Rightarrow \widehat{\theta}_{MOM} = 2\overline{X}$$

Point Estimators: Evaluation

- **Unbiasedness**: an estimator T for a parameter θ, is said to be unbiased if $\mathbb{E}[T] = \theta$
- **Efficiency**: an estimator T is precise if its variance $\mathbb{V}[T]$ is small

The *Mean Squared Error* (MSE) evaluates the performance of the estimator combining these two elements:

$$MSE(T) = \mathbb{V}[T] + Bias(T)^2$$

REMEMBER! A "good" estimator is on average close to the real value of the parameter of interest and is always on target. Of course, if $\mathbb{E}[T] = \theta$ the MSE reduces to its variance

Point Estimators: Asymptotic Property

- **Consistency**: the MSE, when $lim_{n\to\infty} MSE(T) = 0$, can be alternatively defined as

$$MSE(T) = \mathbb{E}[(T-\theta)^2]$$

so, we have that as n grows T becomes closer and closer to the real value of the parameter θ.
REMEMBER! Consistency reassures us that adding more observations improves the performances of the estimator

Example

- Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$. Consider the two estimators:
  1. $T_1(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$
  2. $T_2(X) = \frac{X_{(1)}+X_{(n)}}{2}$
- Let us compute the **bias**:
  1. $\mathbb{E}[T_1(X)] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{n\mu}{n} = \mu$
  2. $\mathbb{E}[T_2(X)] = \frac{\mathbb{E}[X_{(1)}]+\mathbb{E}[X_{(n)}]}{2} = \frac{\mu+\mu}{2} = \mu$
- The estimators are both unbiased. Which is the best?
- Let us look at the **efficiency**, computing the MSE (which is now equal to the variance):
  1. $\mathbb{V}[T_1(X)] = \frac{1}{n^2}\mathbb{V}[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$
  2. $\mathbb{V}[T_2(X)] = \frac{1}{2}(\mathbb{V}[X_1] + \mathbb{V}[X_n]) = \frac{1}{2}(\sigma^2 + \sigma^2) = \sigma^2$
- T_1 has a smaller MSE, so it is better than T_2

Point Estimation: Maximum Likelihood Estimator

- We start defining the **Likelihood function**:

Given i.i.d random variables $X_1, ..., X_n$ in a population, whose distribution depends on the parameter θ, and a sample of observed $x_1, x_2, ..., x_n$ extracted from the population, the *likelihood function* $L(\theta)$ indicates the probability of observing the sample as the parameter changes

- Since the observations (sample) are independent and identically distributed (i.i.d.), we can write the likelihood function as a *product of the probabilities* of the single sample observations:

$$L(\theta) = P(data; \theta) = P(X_1 = x_1; \theta) \times ... \times P(X_1 = x_1; \theta)$$
$$= \prod_{i=1}^{n} f(x_i; \theta)$$

where $f(x_i; \theta)$ is the probability function (for discrete r.v.) or density function (for continous r.v.)

Example (discrete population)

- Suppose a small municipality intends to build a new road in a green area. The parameter of interest is the proportion p of citizens who prefer to preserve the green area (the population is represented by a Bernoulli r.v.):

$$\begin{cases} x = 1 : \textit{the citizen prefers the green area} \\ x = 0 : \textit{otherwise} \end{cases}$$

- Suppose a random sample of $n = 8$ citizens was observed:

$$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0, x_6 = 1, x_7 = 1, x^8 = 1$$

- The random variables $x_1, x_2, ..., x_8$ representing the sample are i.i.d., with $P(X_i = 1) = p, P(X_i = 0) = 1 - p$.
- What is the probability of extracting the sample observed when the parameter varies?

$$P(X_1 = x_1; \theta) \times ... \times P(X_1 = x_1; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = p^6(1-p)^2$$

Maximum Likelihood Estimator

To *estimate* the parameter θ we take the most plausible value of θ given the observed sample, i.e. the value of θ which **maximizes** the likelihood function $L(\theta)$

- Since the logarithmic function is an increasing monotonic function, the maximum likelihood estimate of θ $(\widehat{\theta})$ is also the value that maximizes $log\ L(\theta)$, which is more convenient to calculate than $L(\theta)$
- The estimate of the maximum likelihood of θ is the value $\widehat{\theta}$ that maximizes $log\ L(\theta)$:

$$log\ L(\widehat{\theta}) = \sup_{\theta} logL(\theta)$$

Maximum Likelihood Estimator: Operation Steps

1. Compute the *derivative* of the log-likelihood and equate it to 0: $\partial logL(\theta)/\partial\theta = 0$
2. Isolate θ to find the candidate for the MLE (i.e. the *critical point*)
3. Check the sign of $\partial^2 logL(\theta)/\partial^2\theta$ in the candidate θ to verify that this is not a minimum or a saddle point

Example: in general, for a Bernoulli distribution, the maximum (log) likelihood is given by

$$\frac{\partial logL(\theta)}{\partial\theta} = \frac{\partial[log(p^{\sum_{i=1}^n x_i} \cdot (1-p)^{n-\sum_{i=1}^n x_i})]}{\partial p}$$

$$= \frac{\sum_{i=1}^n x_i}{p} - \frac{n-\sum_{i=1}^n x_i}{1-p} = 0$$

$$\widehat{p} = \overline{x}$$

Recap of Interval Estimates: Interval Estimator

- With a **point estimator** we provide an estimate of the unknown parameter
- We can also provide a set of plausible **intervals** of values for the unknown parameter
- Recall that the Central Limit Theorem (CLT) states that, for large n,

$$\frac{\frac{\sum_{i=1}^{n} X_i}{n} - \mu}{\frac{S}{\sqrt{n}}} \sim N(0,1)$$

where S is the square root of the sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- The quantity $\frac{S}{\sqrt{n}}$ is called **standard error**

Interval Estimator: Definitions

An *interval estimator* for a parameter θ is a random interval $[L(X_1, ..., X_n), U(X_1, ..., X_n)]$, containing the most believable values for the parameter

- Intuitively, it is very difficult to predict the exact value of the unknown parameter (if T is a continuous random variable, this is even impossible, as by definition $P(T = \theta) = 0$), hence is more reasonable to ask for a range of possible parameters
- In addition, a set of plausible values is more informative on the phenomenon than just a single guess

Interval Estimator: Confidence Interval

A *confidence interval* of level $1-\alpha$ is a random interval $[L, U]$, where L and U are two statistics, such that

$$P(\theta \in [L, U]) = 1 - \alpha$$

- The confidence level $1-\alpha$ is the probability that the interval contains the **true value** of the parameter θ, before the sample is observed (typically this value is chosen to be high (0.95 or 0.99))
- A **confidence interval** is built using the formula

$$T \pm err$$

where T is the point estimator for θ and err measures how accurate the point estimate is and depends on the level of confidence as well as $\mathbb{V}[T]$