Fundamentals of Inference

Carlo Cavicchia

carlo.cavicchia@uniroma1.it

Inference vs Probability

from population to data and back

- **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.

- **Inference** starts from a sample and describes the observed data with the aim of inferring relevant information on the population.

What is Inference?

a general introduction

- **Estimate**: recover some parameter explaining the phenomenon that generates the data

  point estimate: a *single number* that is our best guess for the parameter.

  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

Random sample

A **random sample** is a collection of random variables $X_1, \dots, X_n \sim f_{X_1,\dots,X_n}$, that are:

- *independent*

$$f_{X_1,\dots,X_n} = \prod_{i=1}^{n} f_{X_i}(x_i)$$

- *identically distributed*

$$f_{X_i}(x_i) = f_X(x_i) \quad \forall i$$

As a consequence

$$f_{X_1,\dots,X_n} = \prod_{i=1}^{n} f_X(x_i)$$

An **observed sample** $(x_1, \dots, x_n)$ is a realization of the random sample.

Toy Example

how to compute the sample distribution

Let $X_1, \dots, X_n$ i.i.d. (**i**ndependente **i**dentically **d**istributed) from a Poisson(λ).

The **sampling distribution** $f_{X_1,\dots,X_n}$ can be derived as follows:

$$
\begin{aligned}
f_{X_1,\dots,X_n}(x_1, \dots, x_n) &= \prod_{i=1}^{n} f_X(x_i) \\
&= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
&= \frac{1}{\prod_{i=1}^{n} x_i!} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}
\end{aligned}
$$

Basic Concepts

short glossary of estimation tools

- **Parameter**: numerical characteristic of the population that we are trying to recover (hence typically unknown)

  Examples: λ in a Poisson

- **Statistics**: numerical function of the sample that does not directly depend on any unknown parameter

  Example: $S(X_1, \dots, X_n) = X_{(n)} - X_{(1)}$

- **Estimator**: a statistic used to estimate the population parameter

  Example: $T(X_1, \dots, X_n) = \bar{X}$ is an estimator for μ

- **Estimate**: the value of an estimator corresponding to an *observed* sample:

  Example: $T(x_1, \dots, x_n) = \bar{x}$ is an estimate corresponding to $\bar{X}$

Variability of Estimators

walking our way through it with an example

In order to assess the IQ of Torvergata students, we interview 10 people, and we use the sample mean $\bar{X}$ as an estimator of the population mean μ.

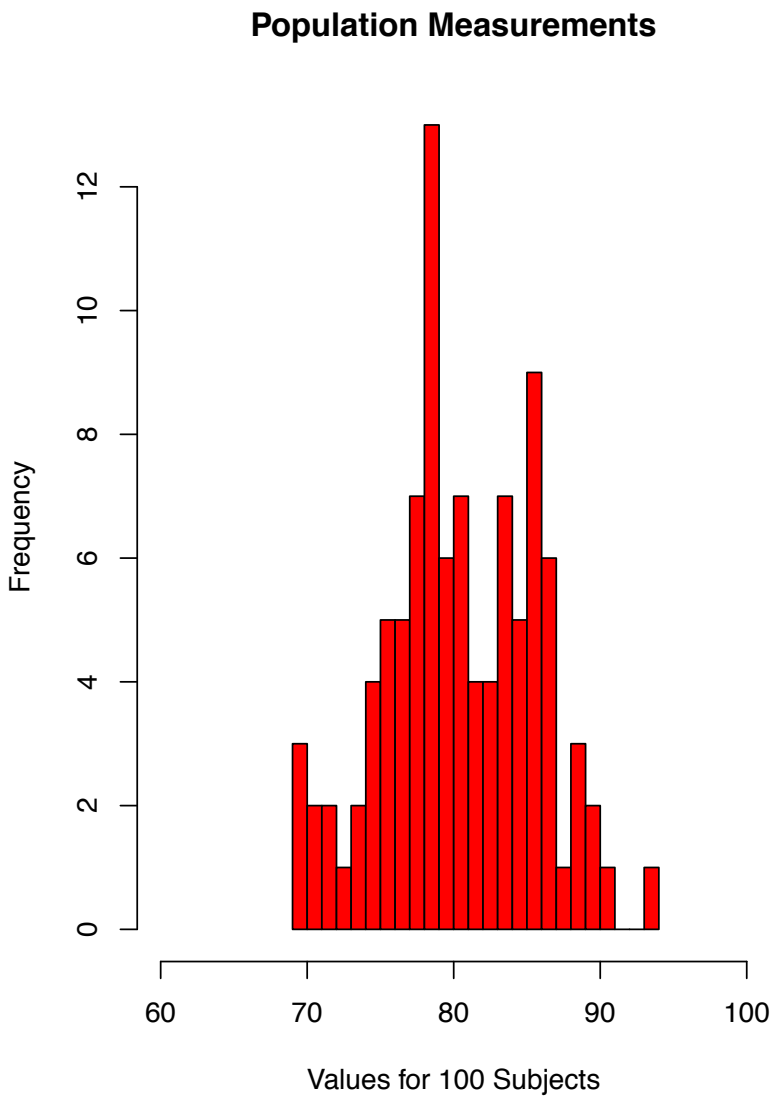
- observed sample: $x = (x_1 = 95, x_2 = 104, x_3 = 104, x_4 = 95, x_5 = 88, x_6 = 126, x_7 = 77, x_8 = 112, x_9 = 111, x_{10} = 105)$
- estimate: $T(x_1, \dots, x_n) = \bar{x} = 101.7$

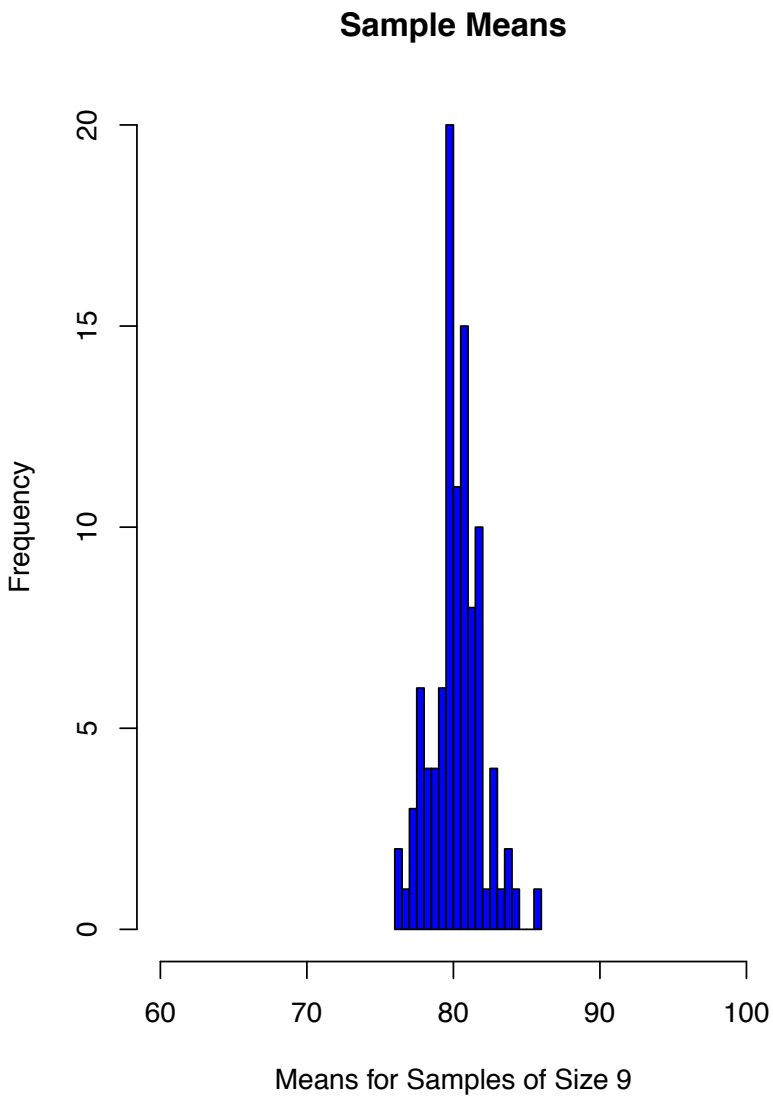
CAVEAT: if we draw another sample from the same population, we will observe different results:

- 2-nd observed sample: $x' = (123, 119, 94, 116, 106, 91, 88, 107, 91, 103)$
- estimate: $T(x_1, \dots, x_n) = \bar{x'} = 103.8$

Since it is a function of a random object, an **estimator** is a *random variable*, and the **estimates** are its *realizations*.

Toy example


Population Measurements
Frequency
0
2
4
6
8
10
12
60
70
80
90
100
Values for 100 Subjects


Sample Means
Frequency
0
5
10
15
20
60
70
80
90
100
Means for Samples of Size 9

Exercises

- Let X be the random variable *score of statistics' exam* and let x be a observed sample equal to $(x_1 = 22, x_2 = 18, x_3 = 18, x_4 = 20, x_5 = 22, x_6 = 26, x_7 = 28, x_8 = 30, x_9 = 30, x_{10} = 22)$.

1. Try to compute the sample mean of each sample with $n = 1, n = 3$ and $n = 10$.
2. Try to compute the expected value of the sample mean of samples with $n = 1, n = 3$ and $n = 10$.
3. Comment the results.

Comments on estimators

not all estimators are good

There is no “universal estimator”, but it must be chosen according to:

- the distribution of the data

  we wouldn’t try to estimate the max of a discrete variable with a continuous value

- the parameter of interest

  we wouldn’t try to estimate the mean and the variance of a Normal distribution with the same estimator

Example:

- parameter of interest: mean of a Normal population
- estimator: $T(X_1, \dots, X_n) = X_{(n)}$

How do we define an estimator

There are several *automatic* ways to derive an estimator, based on different

- **Methods of Moments**:
- **Maximum Likelihood**:
- **Least squares**:

Methods of Moments

for point estimation

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 $X_1, \dots, X_n \sim \text{Unif}(0, \theta)$, for which $\mathbb{E}[X] = \theta/2$.

The MOM estimator is found by equating $\mathbb{E}[X] = \theta/2$ with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\theta/2 = \bar{X} \qquad \Rightarrow \hat{\theta}_{MOM} = 2\bar{X}$$

Exercise:

Let $X_1, \dots, X_n \sim \text{Unif}(a, b)$, compute the MOM estimator for a and b.

Remember that

$$X \sim \text{Unif}(a, b) \qquad \Rightarrow \mathbb{E}[X] = \frac{b+a}{2} \qquad \mathbb{V}[X] = \frac{(b-a)^2}{12}$$

The Likelihood Function

the basic intuition

Let $X \sim$ Binomial(n, p), the probability mass function gives us the probability of observing a value x, once we know p and n.

Assume that $n = 10$ and we observe $x = 8$

- if $p = 0.5$, $P(X = 8) = \binom{10}{8}(0.5)^8(0.5)^2 = 0.043$
- if $p = 0.7$, $P(X = 8) = \binom{10}{8}(0.7)^8(0.3)^2 = 0.233$

For $x = 8$, the parameter $p = 0.7$ seems to be more likely than $p = 0.5$.

When we fix the realization x and we consider it a function of the parameter p, the p.m.f gives us a measure of **how compatible** x is with the value p.

This tells us how **plausible** a value of the parameter is, but it does not measure its **probability**.

The Likelihood Functiom

a little more formally

- $X_1, \dots, X_n$ i.i.d random variables from a discrete distribution with parameter θ, and let $x_1, \dots, x_n$ be an observed drawn from it. The **Likelihood function** $L(\theta; x_1, \dots, x_n)$ corresponds to the **Probability Mass function** when taken to be a function of the parameter θ for a fixed value of $x_1, \dots, x_n$.

- $X_1, \dots, X_n$ i.i.d random variables from a continuous distribution with parameter θ, and let $x_1, \dots, x_n$ be an observed drawn from it. The **Likelihood function** $L(\theta; x_1, \dots, x_n)$ corresponds to the **Probability Density function** when taken to be a function of the parameter θ for a fixed value of $x_1, \dots, x_n$.

The **log-likelihood function**, denoted by $l(\theta; x_1, \dots, x_n)$ is the *logarithm* of the Likelihood function.

Maximum Likelihood Estimator

The **Maximum Likelihood Estimator (MLE)** is the value of the parameter that maximises the Likelihood:

$$\hat{\theta}_{MLE} = \arg\max L(\theta; x_1, \dots, x_n) = \arg\max l(\theta; x_1, \dots, x_n)$$

Operationally the steps to find the **MLE** are:

1. **Compute the derivative** of the log-likelihood and equate it to 0: $dl(\theta; x_1 \dots, x_n)/d\theta = 0$
2. **Isolate** θ to find the candidate for the **MLE** (i.e. the critical point)
3. **Check the sign** of $d^2 l(\theta; x_1 \dots, x_n)/d\theta^2$ in the candidate θ to verify that this is not a min or a saddle

Example

Maximum Likelihood for the parameter λ of a Poisson:

Remember that if $X_1, \dots, X_n$ random sample, with $X_i \sim \text{Poisson}(\lambda)$ then:

- joint distribution

$$p_{X_1,\dots,X_n}(x_1, \dots, x_n; \lambda) = \frac{1}{\prod_{i=1}^{n} x_i!} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}$$

- Likelihood

$$L(\lambda; x_1, \dots, x_n) = \frac{1}{\prod_{i=1}^{n} x_i!} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}$$

- log-Likelihood

$$l(\lambda; x_1, \dots, x_n) = \log\left(\frac{1}{\prod_{i=1}^{n} x_i!}\right) - n\lambda + \sum_{i=1}^{n} x_i \log(\lambda)$$

Example

Maximum Likelihood for the parameter λ of a Poisson:

1. Compute the derivative of $l(\lambda; x_1, \dots, x_n)$ and equate it to 0:

$$\frac{dl(\lambda; x_1, \dots, x_n)}{d\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

2. Isolate λ to get the MLE estimate:

$$-n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \quad \iff \quad \hat{\lambda}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}_n$$

CAVEAT Even if $p_{X_1,\dots,X_n}(x_1, \dots, x_n; \lambda)$ denotes a discrete distribution, it is **a continuous function in** λ, hence we can compute derivatives to find the max.

Core of the Likelihood

The multiplicative factor **depending on the data** but **not on the parameter** $\frac{1}{\prod_{i=1}^{n} x_i!}$ disappeared when we computed the derivative. This is always true:

- if $L(\lambda; x) = h(x)g(x, \theta)$, then $l(\lambda; x) = \log(h(x)) + \log(g(x, \theta))$
- the derivative of $\log(h(x))$ does not depend on θ

$$\frac{dl(\theta; x)}{d\theta} = \frac{d\log(h(x))}{d\theta} + \frac{d\log(g(x, \theta))}{d\theta} = \frac{d\log(g(x, \theta))}{d\theta}$$

The function $g(x, \theta)$ is called the **core** of the likelihood and it contains all the information we need from the data.

Since **we can replace** L **with** g **without loss of information**, when we talk about *Likelihood* we actually talk about its *core*.

Exercise

Let $X_1, \dots, X_n$ be a random sample (i.i.d.), where each X_i has the following density function

$$f_X(x;\theta) = (\theta+1)x^\theta \qquad\qquad x \in (0,1),\ \theta > -1$$

- Compute the joint distribution $f_{X_1,\dots,X_n}(x_1, \dots, x_n)$
- Find the likelihood distribution
- Determine the Maximum Likelihood estimator for θ

Evaluating Point estimators

- An estimator T for a parameter θ, is said to be **unbiased** if $\mathbb{E}[T] = \theta$.
  a "good" estimator is on average close to the real value of the parameter of interest

- An estimator T is **precise** if its variance $\mathbb{V}(T)$ is small.
  a "good" estimator is *always* on target

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

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

MSE

- if $\mathbb{E}[T] = \theta$ we say that the estimator is **unbiased** and the MSE reduces to its variance

Consistency

- the MSE can be alternatively defined as

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

- when

$$\lim_{n \to \infty} MSE(T) = 0$$

we have that as n grows T becomes closer and closer to real value of the parameter θ. This important property is called **consistency**, and reassures us that adding more observations improves the performances of the estimator

Exercise

- Let $X_1, X_2, \ldots, X_n$ be iid Poisson (λ) random variables. Let consider the following estimators:

$$T_1 = \sum_{i=1}^{n} X_i \quad T_2 = \sum_{i=1}^{n} iX_i \quad T_3 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Compute bias and MSE for each estimator
- Which one is the best?
- Which one is the worst?

Interval Estimates

A **interval estimator** for a parameter θ is a random interval $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$, 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.

The ingredients

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 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).

Typically 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 $\mathbb{V}[T]$.

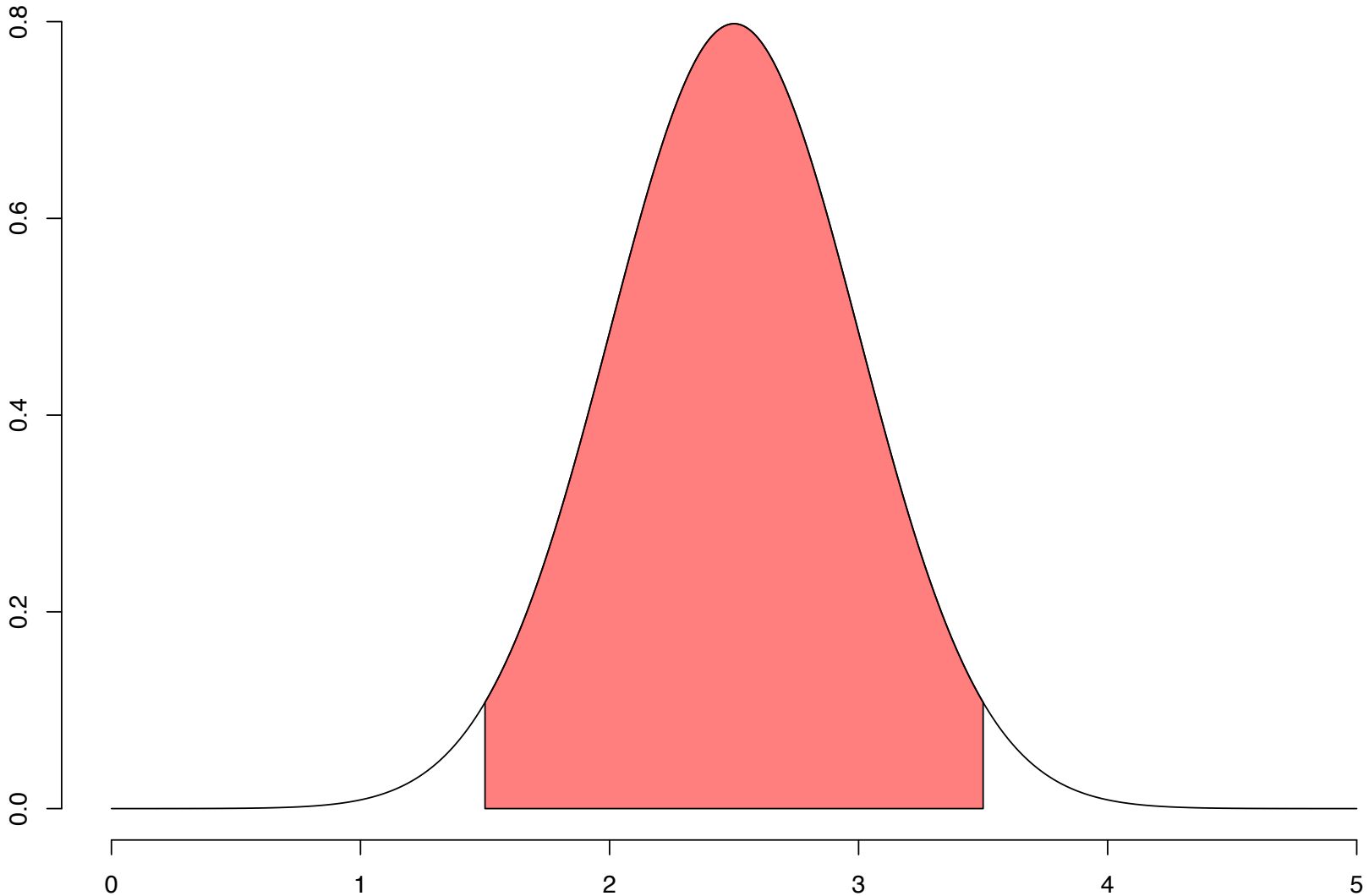
Confidence

a word of caution

BE CAREFUL: once we observe the sample, and we have an *estimate* of the confidence interval $[l, u]$, the probability that the parameter lies in this interval is either 0 or 1.

However, remembering the definition of probability as the limit of the relative frequency of an event, we can be **confident** that if we build a large number of confidence intervals, the parameter will be contained in the 95% of them.

Toy example



Exercise

- When a General Social Survey asked 1326 subjects, “Do you believe in science?”, the proportion who answered yes was 0.82.

1. Compute the standard error of this estimator.
2. Construct the 95% confidence interval. Interpret it in context.
3. How does the result in (2) change if you construct a 99% confidence interval?
4. Another source claims, ”75% of people believe in science.” Does the confidence intervals support this claim?
5. Describe the effect of the sample size on the confidence interval.

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.

Usually, an *hypothesis* on the parameter θ is formalized as follows:

- $\theta = \theta_0$ *punctual* hypothesis
- $\theta \geq \theta_0$ or $\theta \leq \theta_0$ *one-sided* hypothesis
- $\theta \neq \theta_0$ *two-sided* hypothesis

Hypothesis

In a **hypothesis test** we compare two alternative hypothesis H_0 and H_1:

- The **Null Hypothesis** (H_0) is the hypothesis that is held to be true unless sufficient evidence to the contrary is obtained.
- The **Alternative Hypothesis** (H_1) represent the new theory we would like to test.

Example: We want to test whether an astrologer can correctly predict which of 3 personalities charts applies to a person.

- $H_0 : p = 1/3$

  the astrologer doesn't have any predictive power (the probability of guessing the personality is $1/3$)
- $H_1 : p \geq 1/3$

  the astrologer does have predictive power

Test logic

Innocent until proven guilty

	H_0 is true	H_0 is false
Accept H_0	✌	Type II Error
Reject H_0	Type I Error	✌

- If we want to completely avoid Type II Error we should **always Reject** H_0
- If we want to completely avoid Type I Error we should **always Accept** H_0

It is impossible to simultaneously avoid both: which one is more important?

As H_0 represent the current condition, we would like to subvert it only when the data provide strong evidence against it

Testing procedure:

How to solve a test $H_0 = \theta \leq \theta_0$ versus $H_1 = \theta > \theta_0$:

1. Choose a level α of significance (i.e. the probability of Type I Error), typically $\alpha = 0.05$
2. Choose a test statistic T, i.e. a statistic that describes how far that point estimate falls from the parameter value given in the null hypothesis
3. Given an observed sample $(x_1, \dots, x_n)$, compute the $t = T(x_1, \dots, x_n)$
4. Compute the p-value, $P(T > t|H_0) = p$, a measure of how compatibles the data are with H_0
5. If $p \leq \alpha$, reject H_0, otherwise do not reject it

Toy Example

- A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15.

  Step 1: State the Null hypothesis. The accepted fact is that the population mean is 100, so: $H_0 : \mu = 100$
  Step 2:State the alternative hypothesis. The claim is that the students have above average IQ scores, so: $H_1 : \mu > 100$
  Step 3: Find the rejection region area (given by your α level equal to 0.05) from the z-table. An area of 0.05 is equal to a z-score of 1.645.
  Step 4:Find the test statistic using this formula: $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = 4.56$
  Step 5:The value of Z is greater than z_α (4.56 > 1.645), so you can reject the null.

Exercise

In a sample of 402 Tor Vergata first-year students, 174 are enrolled into Statistics course.

1. Find the sample proportion.
2. Is the proportion of students enrolled into Statistics course in the population of all Tor Vergata first-year students different from 0.50 at the significance level $\alpha = 0.05$?