



A.Y. 2025/2026



HANDOUTS

STATISTICS -FORMULAS-

WRITTEN BY

ALICE ELLISTON



TEACHING DIVISION

“

This handout is written by students and is not intended to replace university materials.

It is a useful tool for studying the subject, but it does not guarantee the same comprehensive and complete preparation for passing the exam as the materials recommended by the University.

The content may contain errors and has not been reviewed or approved in any way by the instructors. It is recommended that it be used as a supplementary resource, in conjunction with the official sources and materials indicated in the exam syllabus.



1.1

Statistics

module 1 - Describing Data



- statistical units
- Population $\rightarrow$ size N PARAMETER $\rightarrow$ numerical summary at population level
- Sample $\rightarrow$ size n STATISTIC $\rightarrow$ numerical summary at sample level



representative of the population

INFERENTIAL process: to make inference

on population parameters based on sample statistics

DESCRIPTIVE statistics: describe sample data through sample statistics

Question $\Rightarrow$ statistical VARIABLE: result of measurement process

TYPES:

- categorical
 - nominal $\rightarrow$ no ranking
 - ordinal $\rightarrow$ ranking
- Numerical
 - discrete $\rightarrow$ counting process
 - continuous

not a priori: same characteristic $\Rightarrow$ $\neq$ measurements process



- GSS WARE 12

kind of variable $\rightarrow$ statistical tool

FREQUENCY DISTRIBUTION TABLES

1st col: distinct values

2nd col: Abs. Freq. / counts
relative freq.

GRAPHS

- Pie chart (circle = total, slice = share)
- Bar chart (value, bar = height equal to frequency)

1.2

Histogram

several distinct values

- group observations into INTERVALS $\rightarrow$ 1st col.
- (Absolute, Relative) Frequency $\rightarrow$ 2nd col.
- horizontal axis $\rightarrow$ intervals
Bar $\rightarrow$ AREA = interval's Relative Frequency
- vertical axis $\rightarrow$ DENSITY (RF / interval length)

Central tendency $\rightarrow$ condense data in one value

- categorical var $\rightarrow$ typical level
- numerical var $\rightarrow$ typical value or average



- 1) MODE $\rightarrow$ all variables (few values)
- 2) MEDIAN $\rightarrow$ ordinal or numerical
- 3) MEAN $\rightarrow$ numerical

1) **mode** level with highest frequency
 unique synthetic measure for **categorical nominal**

2) **median** central value (sample split in 2 halves)
 $\hookrightarrow$ based on freq.

RAW DATA: rank the obs.

n odd $\rightarrow$ value with pos: $\frac{(n+1)}{2}$
 n even $\rightarrow$ 2 middle obs.

cumulative relative frequency
 RF of units with values $\leq$ to it
ordinal and numerical

median: 1st value $CF > 0.5$ or $= 0.5$ **only value in the interval with the value**
 $\hookrightarrow$ **lower bound**

3) **mean**: arithmetic avg of **all values** **numerical**
 $\hookrightarrow$ value observed on the 1st sample unit

RAW data: $\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$

$x_1 = \text{value} \cdot \text{Ab. Freq.}$

$\bar{X} = \text{value} \cdot \text{RF}$

DEVIATION $\rightarrow$ value - mean



$\rightarrow$ affected by OUTLIERS

$\hookrightarrow$ fulcrum

Balancing point sum of deviations = 0

$$(x_1 - a)^2 + (x_2 - a)^2 + \dots + (x_n - a)^2 \Rightarrow \text{minimized at } \bar{x}$$

mean from Freq. Distribution: weighted avg of the values

$\downarrow$
RF

value	absolute frequency	relative frequencies
0	60	0.30
1	40	0.20
2	60	0.30
3	20	0.10
4	20	0.10

$$\bar{x} = \frac{0 \cdot 60 + 1 \cdot 40 + 2 \cdot 60 + 3 \cdot 20 + 4 \cdot 20}{200} =$$

$$= 0 \cdot 0.30 + 1 \cdot 0.20 + 2 \cdot 0.30 + 3 \cdot 0.10 + 4 \cdot 0.1 = 1.5$$

23 measuring the location in the sequence of ordered observations

- ordinal + numerical

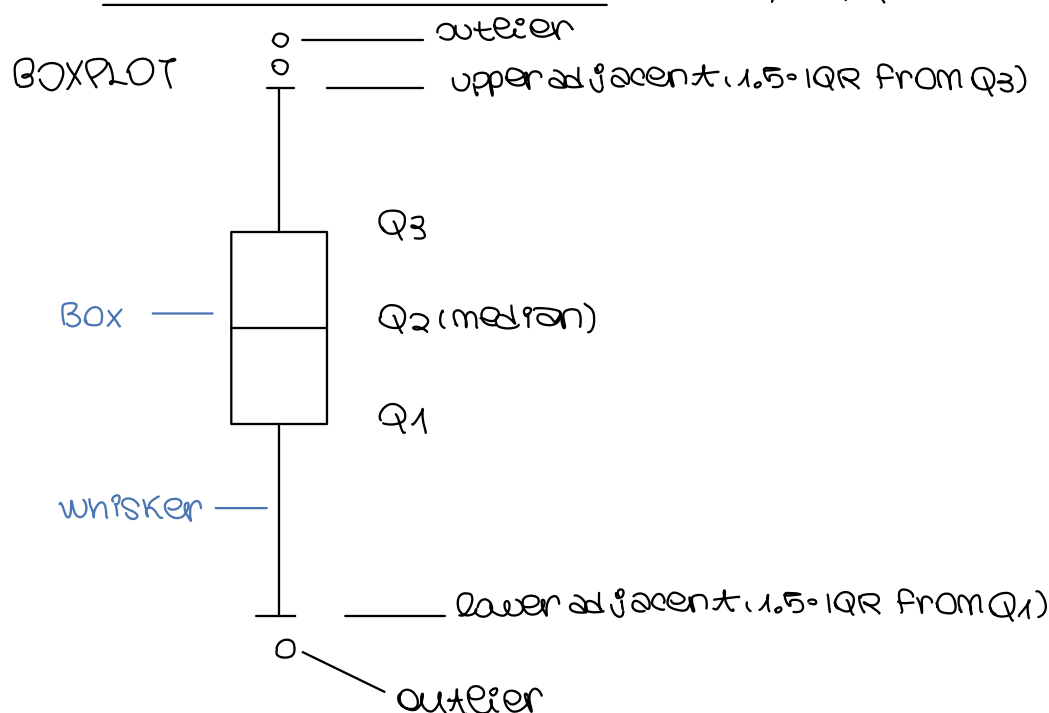
QUARTILES

- RANK obs + divide in quarters
 - Q_1 (lower) $\rightarrow$ 1st value where $CF \geq 0.25$
 - Q_3 (upper) $\rightarrow$ 1st value where $CF \geq 0.75$

PERCENTILES

- p^{th} percentile $\rightarrow$ value such that $p\%$ of obs are below

FIVE numbers SUMMARY $\Rightarrow$ min, Q_1 , Q_2 (median), Q_3 , max



measuring the variability numerical

variables with:

- = mean
- $\neq$ means or unit of measurement

- range = max - min $\rightarrow$ affected by outliers
- IQR = $Q_3 - Q_1 \rightarrow$ variability of 50% central part

- variance, $s^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$ variability about the mean

- standard deviation, $s = \sqrt{\frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$

- coefficient of variation, $CV = \frac{s}{|\bar{x}|} \rightarrow$ SD expressed as % of the mean

Histogram

◦ SYMMETRIC distribution

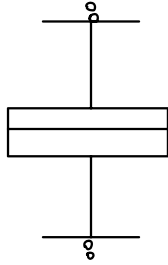
- BELL-shaped

$$\text{median} - Q_1 = Q_3 - \text{median}$$

$$\text{mean} = \text{median}$$

$$Q_1 - \min = \max - Q_3$$

$$\text{median} - Q_1 < Q_1 - \min$$



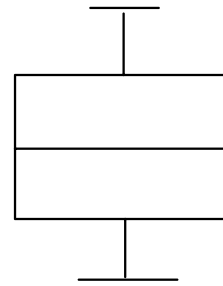
- NOT BELL-shaped: U-SHAPED

$$\text{median} - Q_1 = Q_3 - \text{median}$$

$$\text{mean} = \text{median}$$

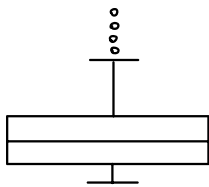
$$Q_1 - \min = \max - Q_3$$

$$\text{median} - Q_1 > Q_1 - \min$$



◦ asymmetric

- on right (tail right):



SKewed - RIGHT distribution

$$\text{median} - Q_1 < Q_3 - \text{median}$$

$$Q_1 - \min < \max - Q_3$$

$$\text{mean} > \text{median}$$

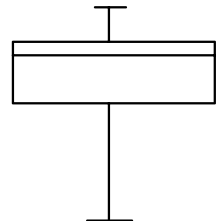
- on left (tail left):

SKewed - LEFT distribution

$$\text{median} - Q_1 > Q_3 - \text{median}$$

$$Q_1 - \min > \max - Q_3$$

$$\text{mean} < \text{median}$$



L4 Concentration analysis → Descriptive technique numerical positive variables

EXTREME SCENARIOS:

- perfect equality: all households same income
- maximum concentration: 1 household total amount

transferable characteristics → economic variables



income, revenue, wages

How to measure concentration:

◦ CONCENTRATION CURVE

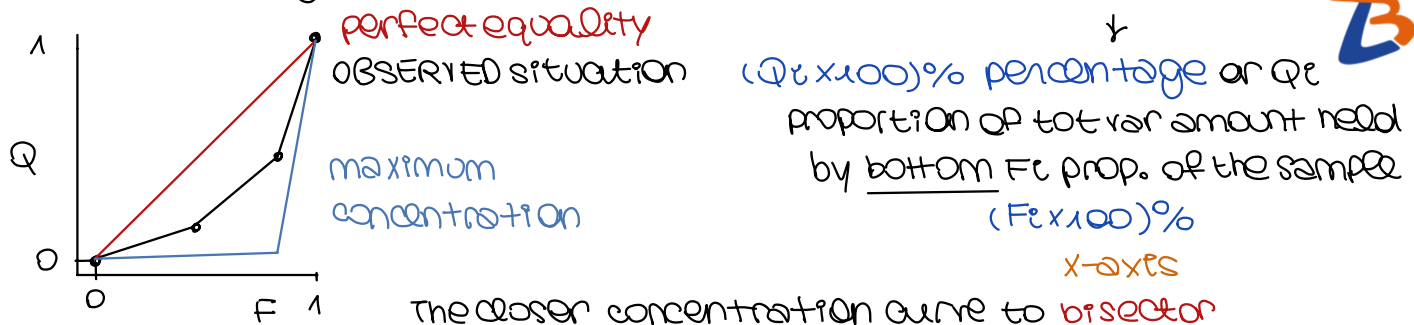
variable $x \rightarrow$ values $x_1, x_2, \dots, x_n$ ranked in ascending order $x_1 \leq x_2 \leq \dots \leq x_n$

$$\text{for each unit } i=1, \dots, n \quad F_i = \frac{i}{n} \quad Q_i = \frac{x_1 + x_2 + \dots + x_i}{x_1 + x_2 + \dots + x_n}$$

$$F_0 = Q_0 = 0$$

$$F_n = Q_n = 1$$

curve joining the $n+1$ points with coordinates (F_i, Q_i)



the closer to 0: the lower concentration

↑ 1 higher
0 CONCENTRATION INDEXES $0 < R < 1$

↓ the LOWER the concentration
HIGHER

GINI INDEX:
$$R = \frac{(F_1 - Q_1) + (F_2 - Q_2) + \dots + (F_{n-1} - Q_{n-1})}{F_1 + F_2 + \dots + F_{n-1}}$$

- perfect equality ($F_i = Q_i$) $\rightarrow R = 0$
- max concentration $\rightarrow$ all $Q_i = 0$ but $Q_n = 1 \rightarrow R = 1$
 $\Rightarrow$ the greater R , the greater the concentration

PIERRE'S INDEX:
$$P = \frac{n}{n-1} \max_{i=1, \dots, n-1} (F_i - Q_i)$$

- perfect equality $P = 0$
- max concentration $P = 1$

L5 Aim of BIVARIATE descriptions

joint behaviour of 2 variables

assessment of association as symmetric perspective
 $\rightarrow$ assessing DEPENDENCY

- **Dependent** variable $\rightarrow$ measuring what we want to EXPLAIN
- **Independent** variable $\rightarrow$ measuring what we propose as the EXPLANATION
 $\rightarrow$ selection techniques depends on variable's type

CASES

- 1) Both dependent and independent var $\rightarrow$ **categorical**

contingency tables

cross table // cross-tab

dep var

cells: joint counts / prop.

$\rightarrow$ marginal distributions

	Food type	Pesticide Status		Total
		Present	Not present	
indep var $\leftarrow$	Organic	0.001	0.004	0.005
	Conventional	0.73	0.265	0.995
	Total	0.731	0.269	1

Marginal univariate distribution of Food Type

Marginal univariate distribution of Pesticide Status

$\Rightarrow$ RF distribution of Dep var in each group

CONDITIONAL frequency distributions

$\rightarrow$ comparison: behaviour Dep var across groups defined by the indep var

conditional freq. distributions SIMILAR $\rightarrow$ no association

↓

splits sample into sub-samples



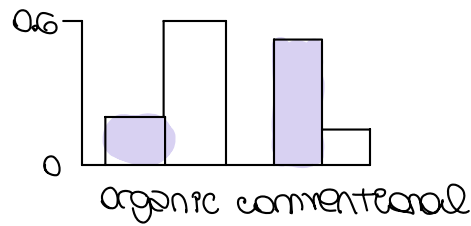
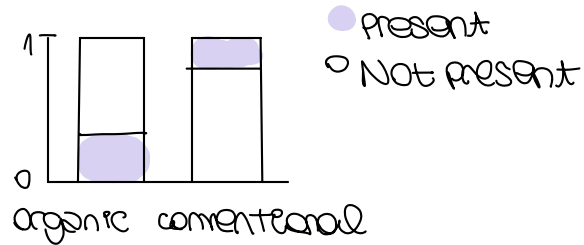
RF distribution of Pesticide status separately for each of the 2 groups:

- conditional freq. Distribution of Pesticide status given food type = organic
- conditional freq. Distribution of Pesticide status given food type = conventional

⇒ calculated in the same way: Joint RF / marginal RF of indep var

CHARTS of conditional distributions

- **Stacked** bar plot
- **Side to side** bar plot



BIVARIATE association

- 2) Dep **numerical**
Indep **categorical**

Compare behaviour num var in the groups determined by cat var

Behaviour → synthetic measures

1st step: mean comparison

- indep var: splits sample in sub-samples
- calculate means of dep var in each group
- Compare the means

very close: No association ←

- 3) two **numerical** var

⇒ cr// 5 numbers summary

LG - SCATTERPLOT

x axis: values indep var

y axis: values dep var

LINEAR association ⇒ correlation

- **positive linear**: high (low) values of one variable tend to occur with high (low) values of the other
- **negative linear**: high (low) values of one variable tend to occur with low (high) values of the other

measures LINEAR association

x_i value of x observed on unit i

• covariance:
$$\text{cov}(X, Y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{n-1}$$

I	II
III	IV

- I and III positive value of cor
Overall trend → line with **positive slope**
- II and IV negative value of the value
Overall trend → line with **negative slope**

Cor = 0 → no **linear** association



o Pearson's correlation index

$$r = \frac{\text{Cov}(X,Y)}{S_X S_Y}$$

$$-1 < r < 1$$

- $r=0$ no correlation
 - $r>0$ positive linear association
 - $r<0$ negative linear association
- The closer to -1 or +1 the STRONGER the association

measures:

- direction linear association
- strength linear association

$\Rightarrow \text{sign } r = \text{sign } \text{cov}(X,Y)$ since $S > 0$

using R:

```
tab1 = table(dataframe$variable)
```

```
prop.table: R distribution
```

```
mean( )
```

```
median( )
```

```
summary( ) (five numbers summary + mean)
```

```
assess variability: sd( ) / mean( )
```

```
Histogram with density y-axis: hist( , freq=FALSE)
```

```
tapply(dep.var, indep.var, mean)
```

```
tapply(dep.var, indep.var, summary)
```

```
boxplot(dep.var~indep.var)
```

```
tapply( dep.var, indep.var, sd) / tapply( dep.var, indep.var, mean)
```

```
Cor( dep.var, companies$qr)
```

```
cor( dep.var, companies$qr) → correlation index
```

```
tab2 = table( var1, var2) → table with JOINT-COUNTS
```

```
prop.table( tab2, 2)
```

↓

```
indep.var.
```



PROBABILITY → way of quantifying uncertainty

INFERENCE process → affected by UNCERTAINTY

mathematical-probabilistic framework

EXPERIMENT → process with uncertain outcome

→ unknown outcome before

Random **EXPERIMENT** in statistics:

- draw at random a unit from population
- measure the concept or characteristic of interest

Random variable

Function that associates with an experiment's outcome a number

- Range/support set → set of all possible values

Types:

- Support set finite/countable: **DISCRETE** random variable
- Support set infinite/not countable: **CONTINUOUS** random variable

1) **Discrete** random variable

Few values/n° of values that we can count



come from a counting process

PROBABILITY FUNCTION

↗ associates to each possible value

- X discrete with values $x_1, x_2, \dots, x_K$ the probability to observe it
- **PROBABILITY FUNCTION**, for each value x_i , returns: $\Pr(X = x_i) = p_i$ for $i = 1, 2, \dots, K$

- probability distribution **PROPERTIES**

$$p_i = \Pr(X = x_i)$$

- $0 \leq p_i \leq 1$ for $i = 1, 2, \dots, K$
- $p_1 + p_2 + \dots + p_K = 1$

EXPECTED VALUE: $E(X) = \mu = x_1 p_1 + x_2 p_2 + \dots + x_K p_K$ = mean

weighted (weights = probabilities) average → summarize probabilistic behaviour

VARIANCE: $\text{var}(X) = \sigma^2 = (x_1 - \mu)^2 p_1 + \dots + (x_K - \mu)^2 p_K$: dispersion w.r.t. expected value

2) **Continuous** random variable

- takes any value in an interval

DENSITY FUNCTION

- not negative

- area of the region delimited by the graph of the function and the x-axis equal 1

$$\Rightarrow \Pr(0.6 \leq X \leq 0.8) = \Pr(0.6 \leq X < 0.8) = \Pr(0.6 < X \leq 0.8) = \Pr(0.6 < X < 0.8)$$

Linear transformation → for any random variable X

Y → linear transformation of X

- for any two numbers a and b $Y = a + bX$



$$- E(Y) = E(a+bx) = a+bE(X)$$

$$- \text{var}(Y) = \text{var}(a+bx) = \text{var}(bx) = b^2 \text{var}(X)$$

STANDARDIZATION

X random variable with $E(X) = \mu$ and variance σ^2

standardized variable $z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma} X - \frac{\mu}{\sigma}$

$$\bullet E(z) = \frac{1}{\sigma} E(X) - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$$

$$\bullet \text{var}(z) = \frac{1}{\sigma^2} \text{var}(X) = \frac{\sigma^2}{\sigma^2} = 1$$

numbers

2.8

Bernoulli distribution

Bernoulli random variable

• Experiment with two outcomes $\rightarrow$ success or failure

$p \rightarrow$ success probability

X: "N° of times we observe a success" $X \sim \text{Ber}(p)$

Probability distribution distribution parametrized by p

X	$P_r(X=x)$
0	$1-p$
1	p

$$\bullet E(X) = 0 \cdot (1-p) + 1 \cdot p = p$$

$$\bullet \text{var}(X) = (0-p)^2(1-p) + (1-p)^2 = p(1-p)$$

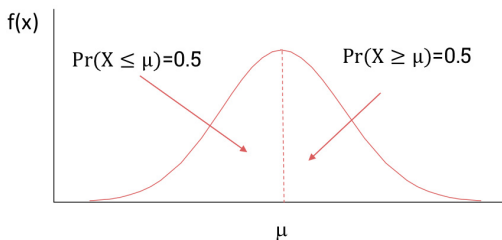
Normal (Gaussian) Distribution

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

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

• μ real number

• σ^2 positive real number



• $\mu \rightarrow$ expected value and median

• $\sigma^2 \rightarrow$ variance

STANDARD NORMAL DISTRIBUTION

$$\mu=0, \sigma^2=1 \quad Z \sim N(0,1)$$

STANDARDIZATION of normal random variable

$$X \sim N(\mu, \sigma^2) \rightarrow Z = \frac{X-\mu}{\sigma} \text{ has a standard normal distribution}$$

Random vector

Joint probabilistic behaviour (of several random variables)

DISCRETE bivariate vector

X discrete r.v. taking values $x_1, x_2, \dots, x_K$

Y discrete r.v. taking values $y_1, y_2, \dots, y_h$

marginal probability functions

$$P_{r_X}(X=x_i) \text{ for } i=1, 2, \dots, K$$

$$P_{r_Y}(Y=y_j) \text{ for } j=1, 2, \dots, h$$

Joint probability function $\rightarrow$ joint probability behaviour

For each pair (x_i, y_j) returns $P_r(X=x_i \text{ and } Y=y_j) = P_r(X=x_i, Y=y_j) = p(x_i, y_j)$

-cross-tab



X	Y		
	1	2	3
0	0.2	0.05	0.15
1	0.15	0.35	0.1

$$\Pr(X=0, Y=1) = 0.2$$

$$\Pr(X=1, Y=3) = 0.1$$

X: 1 student

0 otherwise

Y: 1 unsatisfied

2 not satisfied nor unsatisfied

3 satisfied

X	Y			Marginal
	1	2	3	
0	0.2	0.05	0.15	0.4
1	0.15	0.35	0.1	0.6
Marginal	0.35	0.4	0.25	1

marginal probability distributi of x

marginal probability distributi of Y

INDEPENDENCE

two DISCRETE r.v. are statistically independent

$$\Pr(X=x_i, Y=y_j) = \Pr(X=x_i)\Pr(Y=y_j)$$

for each $i=1, 2, \dots, K$ and each $j=1, 2, \dots, h$

	-2	2	
0	$0.3 \cdot 0.3$	$0.3 \cdot 0.7$	0.3
1	$0.3 \cdot 0.7$	$0.7 \cdot 0.7$	0.7
	0.3	0.7	

$$\Pr(X=0, Y=-2) = \Pr(X=0)\Pr(Y=-2)$$

Perfect independence

Linear association

Cor and Pearson's correlation coefficients $\rightarrow$ def and interpretation as in Descriptive statistics

Pearson's correlation index

$$\rho(X, Y) = \frac{\text{Cor}(X, Y)}{\sigma_X \sigma_Y}$$

$\rho(X, Y) = 0$ no linear association $\rightarrow$ X and Y independent

$\rho(X, Y) > 0$ Positive linear association

$\rho(X, Y) < 0$ Negative linear association

BIVARIATE vector

• X $\rightarrow$ expected value μ_X and variance σ_X^2

• Y $\rightarrow$ expected value μ_Y and variance σ_Y^2

Sum of two r.v.: $E(X+Y) = E(X) + E(Y) = \mu_X + \mu_Y$

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cor}(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2\text{Cor}(X, Y)$$

X and Y not correlated $E(X+Y) = E(X) + E(Y) = \mu_X + \mu_Y$

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) = \sigma_X^2 + \sigma_Y^2$$

Difference two r.v.: $E(X-Y) = E(X) - E(Y) = \mu_X - \mu_Y$

$$\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cor}(X, Y) = \sigma_X^2 + \sigma_Y^2 - 2\text{Cor}(X, Y)$$

X and Y not correlated $E(X-Y) = E(X) - E(Y) = \mu_X - \mu_Y$

$$\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) = \sigma_X^2 + \sigma_Y^2$$

I.I.D Random variables



- n random variables $x_1, \dots, x_n$
- independent
- Identically distributed $\rightarrow$
 - same expected value μ
 - same variance σ^2

$$\begin{aligned} E(x_1 + x_2 + \dots + x_n) &= E(x_1) + E(x_2) + \dots + E(x_n) = n\mu \\ \text{Var}(x_1 + x_2 + \dots + x_n) &= \text{Var}(x_1) + \text{Var}(x_2) + \dots + \text{Var}(x_n) = n\sigma^2 \end{aligned} \quad \left. \vphantom{\begin{aligned} E(x_1 + x_2 + \dots + x_n) \\ \text{Var}(x_1 + x_2 + \dots + x_n) \end{aligned}} \right\} \text{SUM of I.I.D. n.r.}$$

arithmetic average of i.i.d. r.v.s: $\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$

$$\circ E(\bar{X}) = E\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) = \frac{1}{n} E(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \cdot n\mu = \mu$$

$$\circ \text{Var}(\bar{X}) = \text{Var}\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) = \frac{1}{n^2} \text{Var}(x_1 + x_2 + \dots + x_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$\Downarrow$
 $= T$

$$E(T) = n\mu$$

$$\text{Var}(T) = n\sigma^2$$

Central Limit theorem (CLT)

- The sum of n r.v. independent and identically distributed with same mean μ and same variance σ^2 has distribution that for a large n, can be approximated by a NORMAL distribution with mean $n\mu$ and variance $n\sigma^2$ $T \approx N(n\mu, n\sigma^2)$

- The arithmetic average of n r.v. independent and identically distributed with same mean μ and same variance σ^2

has distribution that for a large n, can be approximated by a NORMAL distribution with mean μ and variance $\frac{\sigma^2}{n}$ $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$

Sum of Bernoulli r.v.

$$x_1, \dots, x_n \text{ i.i.d. } \sim \text{Ber}(p)$$

$$E(x_i) = p \quad \text{Var}(x_i) = p(1-p)$$

$(x_1 + \dots + x_n)$ has distribution that for n large can be approximated by a normal distribution with Er np and var $np(1-p)$

EXAMPLE

30% of probability for a student to be selected to an internship program

$\Rightarrow$ work out probability that more than 35 students out of 150 randomly drawn are

$x_i = 1$ if student selected

admitted to the internship program

$x_i = 0$ otherwise

$$x_1, \dots, x_{150} \text{ i.i.d. } \sim \text{Ber}(0.3)$$

$$\Rightarrow P(x_1 + \dots + x_n) > 35$$

$$E(x_1 + \dots + x_n) = np = 150 \cdot 0.3 = 45$$

$$\text{Var}(x_1 + \dots + x_n) = np(1-p) = 150 \cdot 0.3 \cdot 0.7 = 31.5$$



n large $\rightarrow$ Normal approximation

$$Pr((X_1 + \dots + X_{150}) > 35) = 1 - Pr((X_1 + \dots + X_{150}) < 35) \approx 1 - \text{pnorm}(35, 45, \sqrt{31.5}) = 0.9626$$

ARITHMETIC average of Bernoulli r.v.

- $X_1, \dots, X_n$ i.i.d $\sim \text{BER}(p)$ $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$
- proportion of successes in n trials
- change of notation: $\bar{X} = \hat{p}$

$\hat{p}(\bar{X})$ has distribution that for a large n can be approximated by a normal distribution with:

$$E(\hat{p}) = E(\bar{X}) = p$$

$$\text{var}(\hat{p}) = \text{var}(\bar{X}) = \frac{p(1-p)}{n}$$

$\Rightarrow$ work out probability that out of 200 students randomly drawn, less than 25% are admitted to the internship program

$X_1, \dots, X_{200}$ i.i.d $\sim \text{BER}(0.3)$

$\hat{p}(\bar{X}) \rightarrow$ proportion of students selected out of 200

$$E(\hat{p}) = E(\bar{X}) = 0.3$$

$$\text{var}(\hat{p}) = \text{var}(\bar{X}) = \frac{0.3 \cdot 0.7}{200} = 0.001$$

$$Pr(\hat{p} < 0.25) \Rightarrow \text{Normal approximation} \quad \mu \quad \text{sd} = \sqrt{\text{var}}$$

$$Pr(\bar{X} < 0.25) \approx \text{pnorm}(0.25, 0.3, \sqrt{0.001}) = 0.0658$$

$\downarrow$

returns probability of observing a value $\leq$ to

Module 3 - Inferential Statistics: ESTIMATION PROBLEM

2.9 population: set of all statistical units on which we might measure a characteristic or concept of interest

sample: any subset of the population

Data analysis make predictions $\rightarrow$ cannot stop at description

INFERENTIAL STATISTICS:

- collection of techniques that let draw conclusions on the entire population (inferences)
- $\rightarrow$ from the analysis of a portion of the population, a sample drawn from it

SAMPLE $\rightarrow$ representative of the population

random sampling

RANDOM SAMPLE: units drawn at RANDOM from pop. on at time

- same prob. to be drawn $\Rightarrow$ no BIASES
- samples of given size n , same prob. to be drawn

- Description of any inferential problem within a PROBABILISTIC FRAMEWORK

$\rightarrow$ 1st step in building the probabilistic framework

Quantity and control for the probability of making errors

OUR EXPERIMENTS

- Draw at random one at time from a population
- Measure a characteristic of interest on them

RANDOM variable: mathematical description of the outcome of each draw
Before experiment: not know answer ($n \geq 0$)

EXPERIMENT-RANDOM variable

$X \rightarrow$ "no of hours spent watching TV by a respondent ~~drawn~~ at random"



- $\mu \rightarrow$ mean (Expected value) of X
- $\sigma^2 \rightarrow$ variance of X

Population mean $\mu \rightarrow$ arithmetic avg of the no of hours spent watching TV
Parameter = unknown no per day by All adults in USA in 2012

Population variance $\sigma^2 \rightarrow$ variance of the no of hours spent watching TV
Parameter = unknown no per day by All adults in USA in 2012

Random Sampling

Draw at random n units from the population

n random variables $\rightarrow (X_1, \dots, X_n) \rightarrow$ same distribution

$\downarrow$
r.v. of 1st drawn

- same mean μ
- same variance σ^2
- independent

Random sample $\rightarrow (X_1, \dots, X_n)$ i.i.d.

sample realization $\rightarrow (x_1, \dots, x_n)$

Inference on the population mean

INFERENCEAL PROBLEMS:

- Point estimation
- confidence interval estimation
- (Hypothesis testing)

STATISTICS: inference based on sample data

SAMPLE statistics:

- function of the sample
- random variable

sample mean: $\bar{X} = \frac{X_1 + \dots + X_n}{n}$ $\circ$ random variable
 $(X_1, \dots, X_n)$ i.i.d. $E(\bar{X}) = \mu$ $\circ$ observed value $\rightarrow \bar{x}$
 μ $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
 $\bar{X}$ $\bar{x} = 20000$

PROPERTIES OF SAMPLE MEAN: for an i.i.d sample, the sample mean has:

- MEAN: $E(\bar{X}) = \mu$: sample mean has expected value = population mean μ
 - observed value of sample mean $\rightarrow$ one of the several values
 - arithmetic average of all possible values $\rightarrow$ population mean



Sample mean $\rightarrow$ ESTIMATOR of the population mean (μ)

Observed value $\rightarrow$ estimate $\circ$ Is the sample mean a good estimator?

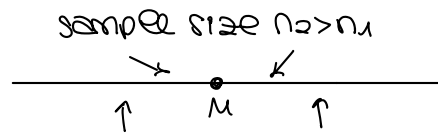
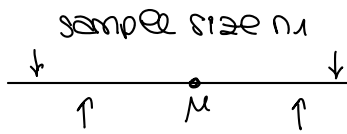
UNBIASED estimator $\rightarrow$ no systematic error

2) $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

- observed n value of sample mean $\rightarrow$ one of the several values
- variance of all possible values $\rightarrow \frac{\sigma^2}{n}$



- decreases with sample size n



- sample mean might over/underestimate population mean
- the greater the sample size $\rightarrow$ the closer to the population mean
- sample mean $\rightarrow$ **CONSISTENT** estimator

STANDARD ERROR of the mean

- standard deviation of the sample mean $\frac{\sigma}{\sqrt{n}} \rightarrow$ **standard error** of the mean

$\rightarrow$ unknown parameter

$$\frac{\sigma}{\sqrt{n}}$$

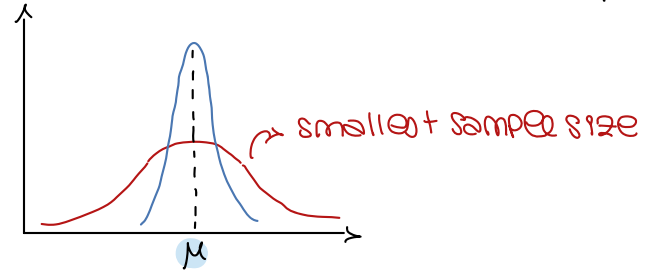
standard error $\rightarrow$ parameter consistent

SD of all possible values of the sample mean observed on all possible samples

- c) For **Central Limit Theorem**, for a large n distribution that can be approximated by a **normal** distribution $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$

sample variance

$$S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$



sample standard deviation $S = \sqrt{S^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$

sample variance as an UNBIASED estimator

- mean (expected value) of the sample variance $\rightarrow$ population variance σ^2

$\sigma^2 =$ variance of our characteristic of interest $\approx$ using sample variance

estimate median using μ

estimate σ^2 using our sample variance

sample mean is a random variable

- S^2 has mean/expected value of $\sigma^2 \rightarrow S^2$ estimator of σ^2

$n-1$: so it is UNBIASED

$$E(S^2) = \sigma^2$$

- S estimator of σ

estimating σ^2 and suggest sample variance S^2

this is the mean of all possible values of S^2

UNBIASEDNESS concerns expected value

2.5783: estimate of σ

$$S = \sqrt{S^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$$

number { observed value of sample standard deviation s $\uparrow$
observed value of sample variance s^2

$\sigma \rightarrow$ spread of my char. of interest

enters in the expression of standard error



ESTIMATION OF THE STANDARD ERROR OF THE MEAN

SE standard deviation of $\bar{X}$

parameter

in case of sample mean the standard error is $\frac{\sigma}{\sqrt{n}}$ when $n \neq$

Replace σ with its estimate: $\frac{2.5783}{\sqrt{709}}$ observed value of the sample SD

r.v. delivering estimates $\frac{S}{\sqrt{n}}$ estimator of the population standard deviation
 ESTIMATOR of the parameter $\frac{\sigma}{\sqrt{n}}$
 estimator of the SE of the mean number never known

$\frac{1}{\sqrt{n}}$ → estimate of $\frac{\sigma}{\sqrt{n}}$ ⇒ number

$n = 2.5783$ observed value of SD

it is approximating σ (so estimate of σ) → using a function that depends on set of values

Standard error of the mean → it is variable

$\pi = 2.5783$

0.0963 estimate of the SE, the estimate of the variability of all possible estimators

problem var and SD → Not RELATIVE index

lower bound is 0

but my SE is the 1st way through which I assess accuracy → but I cannot read it

- sample mean → good estimator of the pop. mean
- assessment of accuracy
- How close to the parameter value the estimate is likely to fall

STANDARD ERROR

- first assessment of accuracy
- The lower → the more accurate the estimation

⇒ never stop at point estimator (SD not enough!) estimate could be wrong
 210 (part 2)

share same normal distribution → symmetric

Normal population → assumption: it is known

$(X_1, \dots, X_n)$ iid $\sim N(\mu, \sigma^2)$

unknown

satisfaction of an individual drawn at random with $\sigma^2 = 4$

μ is our object of inference → start from point estimation

- $\bar{X} \sim$ Normal with mean μ and known variance $\frac{\sigma^2}{n}$

↓ $\bar{X}$ normal r.v.

arithmetic average = sample mean

variance is known

standard error (SE) of $\bar{X}$ = SD is known: $\frac{\sigma}{\sqrt{n}}$

→ linear transformation

Standardization of the sample mean (standardization of the estimator)

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

↗ expected value $E(\bar{X})$

has a standard normal distribution

↳ SD (SE) regardless value of $\mu \rightarrow$ standard normal distribution that is known

CONFIDENCE interval estimation: μ

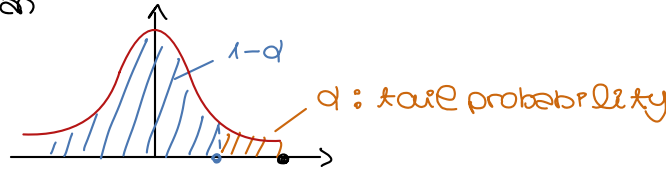
- confidence level $\rightarrow (1-\alpha) \cdot 100\%$ of a given level of confidence (it is fixed)
- typical values for $\alpha \rightarrow 0.01, 0.05, 0.1$
- typical levels of confidence 95%, 90%, 99% expressed in %
- confidence interval is an interval of numbers for which we are 95% confident that unknown parameters μ belongs to it
- length measure of accuracy
- ↳ obj that matters

2 confidence interval for an

→ standard normal

$$Z \sim N(0,1)$$

quantile of order $(1-\alpha)$
 $0 < \alpha < 1$



quantile named after the right tail probability

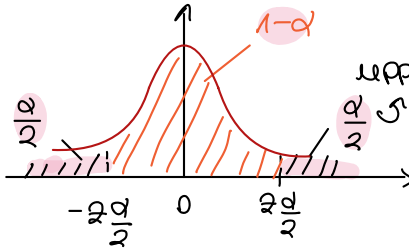
$Pr(Z \leq z_{1-\alpha}) = 1-\alpha$ quantile is number s.t. Pr of $1-\alpha$ to observe a smaller value

quantile of order $1-\alpha$ leaves above mass $= \alpha$

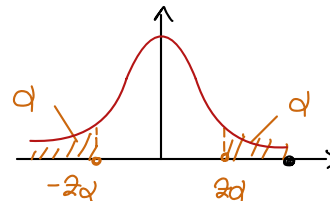
$z_{0.75}$ leaves above mass 0.75 and leaves below 0.25

$z_{0.05}$ 5% probability above

→ axis of symmetry



upper tail probability



Probability is the area of a region of a plane

QUANTILE of a standard normal

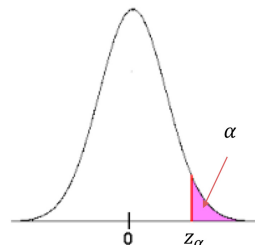
$z_{\alpha} \rightarrow$ quantile of order $1-\alpha$ of a standard normal distribution

$$Pr(Z \leq z_{\alpha}) = 1-\alpha$$

$\frac{\alpha}{2} \rightarrow$ divide 2 tails symmetric wrt to y-axis

$$Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$Pr\left(-z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$





$$P_r \left(-\bar{X} - 2 \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + 2 \frac{\sigma}{\sqrt{n}} \right) = 1-\alpha$$

probabilistic result

$$P_r \left(\bar{X} - 2 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2 \frac{\sigma}{\sqrt{n}} \right) = 1-\alpha \Rightarrow \left(\bar{X} - 2 \frac{\sigma}{\sqrt{n}}, \bar{X} + 2 \frac{\sigma}{\sqrt{n}} \right)$$

number contains μ with probability $1-\alpha$

take estimator (r.v.)

and identify 2 quantities: $\bar{X} - 2 \frac{\sigma}{\sqrt{n}}$

$\bar{X}$ - number $\rightarrow$ r.v.
 $\downarrow$

r.v.

The random interval

as lower bound $\bar{X} - 2 \frac{\sigma}{\sqrt{n}}$ and upper bound $\bar{X} + 2 \frac{\sigma}{\sqrt{n}}$

contains μ with probability $1-\alpha$

with $\alpha = 0.05 \rightarrow 95\%$ probability

95% confidence interval is an interval of numbers and

having as lower bound I compute once I am given the data
and the same with upper one.

I am given $d \rightarrow$ interval of values

90% confident that my unknown pop. mean belong to this interval
 μ

sample mean splits in 2 the confidence interval

Regardless level of confidence: mid point is its estimate: $\bar{X}$

CONFIDENCE INTERVAL $100(1-\alpha)\%$ for μ

- the confidence interval of level $(1-\alpha)100\%$ for the unknown population mean μ mean is:

$$\left(\bar{x} - 2 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \frac{\sigma}{\sqrt{n}} \right) \quad \text{-----} \quad \bar{x}$$

LENGTH and MARGIN OF ERROR

$\rightarrow$ S.E. (standard error)

$$\bar{x} \pm 2 \frac{\sigma}{\sqrt{n}}$$

SD of estimator
(depends on n and σ)

$\Rightarrow$ measure of accuracy and UNCERTAINTY of
estimation of μ

level of confidence

$$2 \frac{\sigma}{\sqrt{n}} = \text{margin of error}$$

taking estimate and stretching it to account for the uncertainty

- mid point confidence interval tells ESTIMATE

- length gives measure of error $\rightarrow$ twice the margin of error

LENGTH of the confidence interval:

$$L = 2 \frac{\sigma}{\sqrt{n}} = 2 \cdot \text{ME (margin of error)}$$

$\rightarrow$ measure of accuracy

Fix confidence level (i.e. 90%) fixing $(1-\alpha) \rightarrow$ fixed α

when n increases ($n \rightarrow +\infty$) length $\downarrow$

SD $\downarrow$

for given confidence level

and given σ^2 , the greater $n \rightarrow$

- the lower the SE
- the lower the margin of error
- the lower is the length of the interval

Confidence level	α	$\frac{z\alpha}{2}$
90%	0.1	1.645
95%	0.05	1.96
99%	0.001	2.576

The greater the confidence level $(1-\alpha) \rightarrow$

- the lower α
- the greater $\frac{z\alpha}{2}$
- the longer the confidence interval

FREQUENTIST INTERPRETATION

- Draw repeatedly samples of a given size n from pop.
- Compute for each of them the confidence interval of level $(1-\alpha)100\%$ for a given parameter
- we expect that $(1-\alpha)100\%$ of the delivered intervals will contain the unknown parameter

NORMAL POPULATION - unknown variance

2) σ unknown $\rightarrow$ estimate it

$(X_1, \dots, X_n)$ i.i.d. $\sim$ Normal, μ and σ^2 unknown

- $\bar{X} \rightarrow \mu$ $(X_1, \dots, X_n)$ i.i.d. $\sim N(\mu, \sigma^2)$
- $S \rightarrow \sigma$ $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

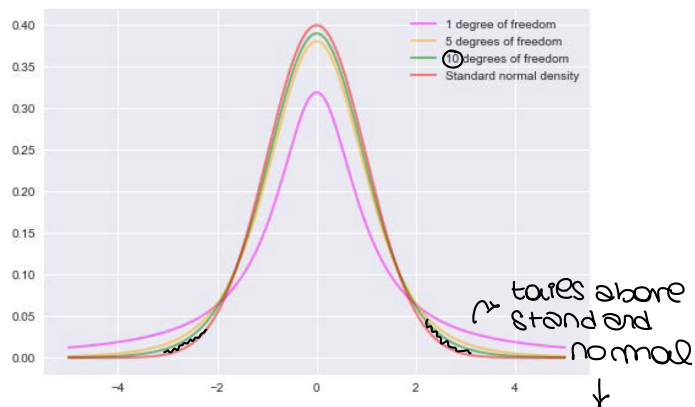
$$\frac{S}{\sqrt{n}} \rightarrow \frac{\sigma}{\sqrt{n}}$$

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

$\frac{S}{\sqrt{n}} \rightarrow$ estimator of SE (not SD)

STUDENT'S T DISTRIBUTION: kind of standardization

has a Student's t distribution with $(n-1)$ degrees of freedom



looks like standard normal
w.r.t. the tails

Probability student's t is $>$ than standard normal
when $n \uparrow \rightarrow$ tails closer to standard normal
(see green very close)

has thicker tails

$100(1-\alpha)\%$ confidence interval for the mean of a normal population, with unknown variance

$\rightarrow$ The confidence interval of level $(1-\alpha)100\%$ for the unknown population mean μ mean is

$$\left(\bar{X} - t_{(n-1), \frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + t_{(n-1), \frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right)$$



$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$$

confidence interval:

center: estimate

Quantiles from a student's t distribution

$$t_{(n-1), \frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad \text{margin of error}$$

In R:

Student's t distribution with n degrees of freedom

`pt(x, n)` returns $\Pr(X \leq x)$

`qt(1- $\alpha$ , n)` returns the quantile of order $(1-\alpha)$, the value q such that $\Pr(X \leq q) = 1-\alpha$ for X with Student's t distribution with n degrees of freedom

The quantile of order 0.95 is `qnorm(0.95, 10) = 1.812461`

is the value such that $\Pr(X \leq 1.812461) = 0.95$

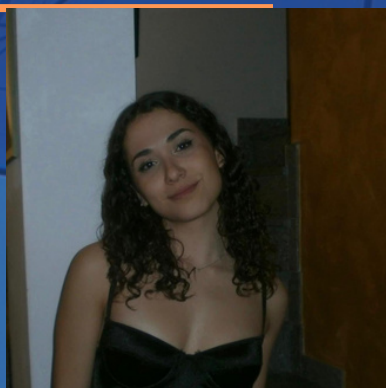
Confidence intervals on the pop. mean

`t.test(variable_name, conf.level = 1- $\alpha$ )`

→ returns the $(1-\alpha)100\%$ confidence interval on the pop. mean of the considered value: `t.test(variable_name)`

returns by default the 95% confidence level

FOR DOUBTS OR SUGGESTIONS ON THE HANDOUTS



ALICE ELLISTON

alice.elliston@studbocconi.it

[@alice.elliston_](https://www.instagram.com/alice.elliston_)

+39 3442849243

FOR INFO ON THE TEACHING DIVISION



NICOLA COMBINI

nicola.combini@studbocconi.it

[@nicolacombini](https://www.instagram.com/nicolacombini)

+39 3661052675



MARTINA PARMEGGIANI

martina.parmeggiani@studbocconi.it

[@martina_parmeggiani05](https://www.instagram.com/martina_parmeggiani05)

+39 3445120057



MARK OLANO

mark.olano@studbocconi.it

[@mark_olano_](https://www.instagram.com/mark_olano_)

+39 3713723943



TEACHING DIVISION



OUR PARTNERS



ETHAN
SUSTAINABILITY

700+
CLUB

LA PIADINERIA

