BEMACC

A.Y. 2025/2026

B.LAB

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.

21 **Statistics** module 1 – Describing Data

- statistical units
- Population → size N
- Sample → size n
  ↓
  representative of the population

PARAMETER → numerical summary at population level
STATISTIC → numerical summary at sample level

INFERENTIAL process: to make inference on population parameters based on sample statistics
DESCRIPTIVE statistics: describe sample data through sample statistics

Question ⇒ statistical VARIABLE: result of measurment process

TYPES:
- categorical
  - nominal → no ranking
  - ordinal → ranking
- Numerical
  - discrete → counting process
  - continuos

not a priori: same characteristic → ≠ measurments process
↓
Kind of variable → statistical tool

- GSS WAVE 12

FREQUENCY DISTRIBUTION TABLES
1st col: distinct values
2nd col: Ab. freq. / counts
relative freq.

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

22 **Histogram**

several distinct values
- group observations into INTERVALS → 1st col.
- (Absolute/Relative) Frequency → 2nd col
- horizontal axis → intervals
  Bar → AREA = interval's Relative Frequency
- vertical axis → DENSITY (RF/interval lenght)

Central tendency → condense data in one value

- Categorical var → typical level
- Numerical var → typical value or average

1) MODE → all variables (few values)
2) MEDIAN → ordinal or numerical
3) MEAN → numerical

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

2) **median** central value (sample split in 2 halves)
↳ based on freq.
RAW DATA: rank the obs.
n odd → value with pos: $\frac{(n+1)}{2}$
n even → 2 middle obs.

cumulative relative frequency
RF of units with values ≤ to it
ordinal and numerical

median: 1st value CF > 0.5 or = 0.5 any value in the interval with the value as lower bound

3) **mean**: arithmetic avg of all values
numerical
Raw data: $\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$ ↱ value observed on the 1st sample unit
x_1 = value · Ab. freq.
$\bar{X}$ = value · RF

DEVIATION → value − mean

↳ fulcrum

→ affected by OUTLIERS

Balancing point sum of deviations = 0

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

mean from Freq. Distribution: weighted avg of the values
↓
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

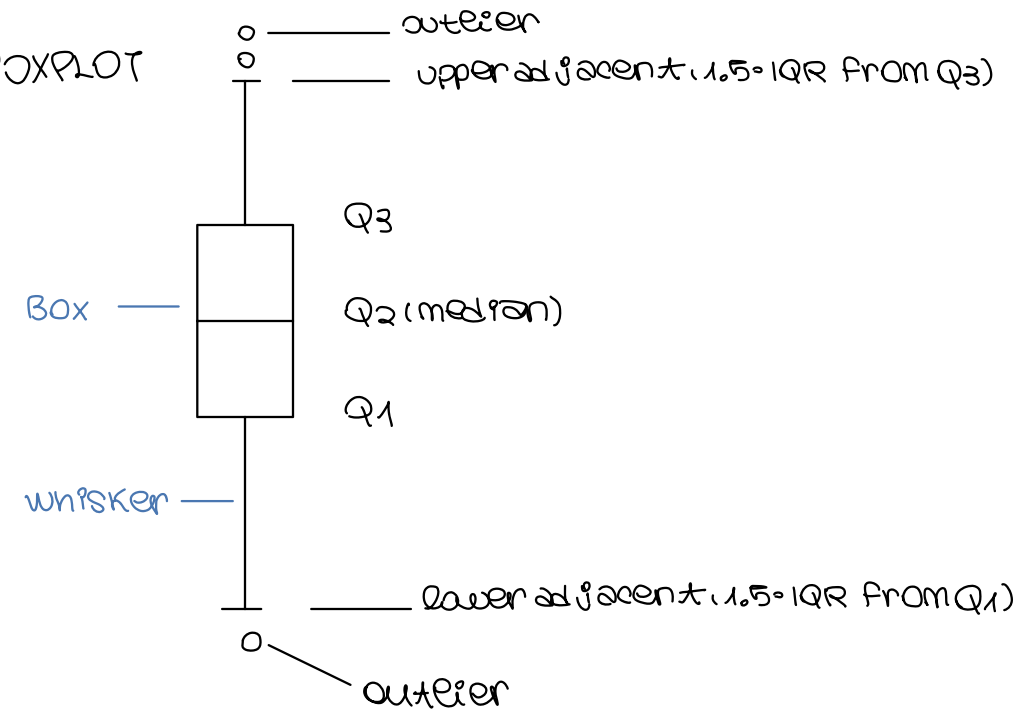
- Rank obs + divide in quarters
  - Q1 (lower) → 1st value where CF ≥ 0.25
  - Q3 (upper) → 1st value where CF ≥ 0.75

PERCENTILES

- p^{th} percentile → value such that p% of obs are below

FIVE numbers SUMMARY ⇒ min, Q1, Q2 (median), Q3, max

BOXPLOT



measuring the variability
numerical

variables with:
- = mean
- ≠ means or unit of measurment

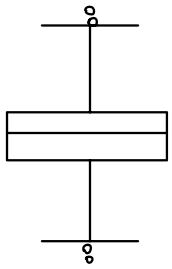
- range = max − min → affected by outliers
- IQR = Q3 − Q1 → variability of 50% central part
- variance, $s^2 = \frac{(x_1-\bar{x})^2+\dots+(x_n-\bar{x})^2}{n-1}$ variability about the mean
- standard deviation, $s = \sqrt{\frac{(x_1-\bar{x})^2+\dots+(x_n-\bar{x})^2}{n-1}}$
- coefficent of variation, $CV = \frac{s}{|\bar{x}|}$ → SD expressed as % of the mean

Alice Elliston

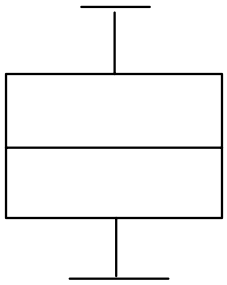
Shape numerical

Histogram

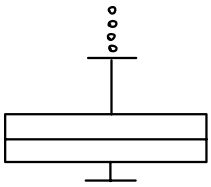
- SYMMETRIC distribution
  - BELL-shaped
    median − Q1 = Q3 − median
    mean = median
    Q1 − min = max − Q3
    median − Q1 < Q1 − min



  - Not BELL-shaped: U-SHAPED
    median − Q1 = Q3 − median
    mean = median
    Q1 − min = max − Q3
    median − Q1 > Q1 − min

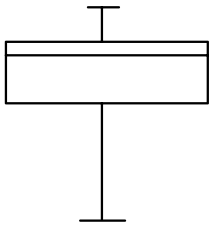


- asymmetric
  - on right (tail rg):



SKEWED-RIGHT distribution
median − Q1 < Q3 − median
Q1 − min < max − Q3
mean > median

  - on left (tail lf): SKEWED-LEFT distribution
    median − Q1 > Q3 − median
    Q1 − min > max − Q3
    mean < median



L4 Concentration analysis → Descriptive technique
numerical positive variable

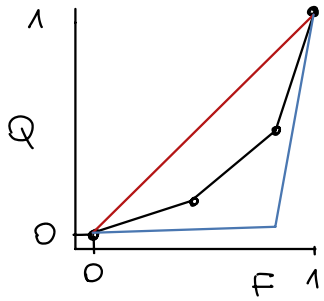
EXTREME SCENARIOS:
- perfect equality: all households same income
- maximum concentration: 1 household total amount

transferable characteristics → economic variables
↓
income, revenue, wages

How to measure concentration:
- CONCENTRATION CURVE

variable x → values $x_1, x_2 \dots x_n$ ranked in ascending order $x_1 \le x_2 \le \dots \le x_n$
for each unit $i = 1 \dots n$ $F_i = \frac{i}{n}$ $Q_i = \frac{x_1 + x_2 + \dots + x_i}{x_1 + x_2 + \dots + x_n}$
$F_0 = Q_0 = 0$
$F_n = Q_n = 1$

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



perfect equality

OBSERVED situation

maximum concentration

$(Q_i \times 100)\%$ percentage or Q_i proportion of total amount held by bottom F_i prop. of the sample $(F_i \times 100)\%$

x-axis

The closer concentration curve to bisector ↓ the LOWER the concentration HIGHER

the closer to 0: the lower concentration ↑ 1 higher

0 CONCENTRATION INDEXES $0 < R < 1$

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 → all Q_i's $= 0$ but $Q_n = 1 \rightarrow R = 1$

⇒ the greater R, the greater the concentration

PIETRA'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 asymmetric perspective

→ assessing DEPENDENCY

- Dependent variable → measuring what we want to EXPLAIN
- Independent variable → measuring what we propose as the EXPLANATION

→ selection techniques depends on variable's type

CASES

1) Both Dependent and Independent var → categorical

contingency tables

cross table // cross-tab cells: joint counts / prop.

→ marginal distributions

dep var

indep var

Food type	Pesticide Status		
	Present	Not present	Total
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

↳ RF distribution of Dep var in each group

CONDITIONAL frequency distributions

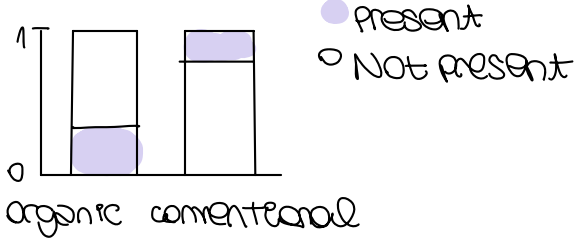
→ comparison: behaviour Dep var across groups defined by the indep var ↓ splits sample into sub-samples

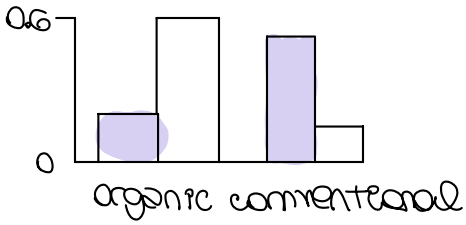
conditional freq. distributions SIMILAR → no association

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

16 – 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 r* value of x observed on unit n
- covariance: $Cov(X,Y) = \frac{(x_1-\bar{x})(y_1-\bar{y}) + \dots + (x_n-\bar{x})(y_n-\bar{y})}{n-1}$

II	I
III	IV

- I and III positive value of cov
  overall trend → line with POSITIVE slope
- II and IV negative value of the value
  overall trend → line with NEGATIVE slope

cov = 0 → no linear association

- Pearson's correlation index $r = \frac{Cov(X,Y)}{S_X S_Y}$ $\boxed{-1 < r < 1}$
  - $r=0$ 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
- strenght linear association

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

---

using R:

```
tab1 = table(dataframe$variable)
prop.table: RF distribution
mean( )
median( )
summary( )      (five numbers summary + mean)
assess variability: sd(      )/mean( )
Histogram with density y-axis: hist(   ,freq=FALSE)
tapply(dep.var, indep.var, mean)
tapply(dep.var, indep.var, summary)
boxplot (dep.var~indep.var)
tapply( dep.var, indep.var, sd)/tapply( dep.var, indep.var, mean)
cor( dep.var, companies$qr)
cor( dep.var, companies$qr) → correlation index
tab2=table( var1, var2) → table with JOINT-COUNTS
prop.table( tab2,2)
                ↓
            indep.var.
```

Module 2 – Elements of probability

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

– probability distribution PROPERTIES
$p_i = Pr(X=x_i)$
- $0 \le p_i \le 1$ for $i=1,2,\dots,K$
- $p_1+p_2+\dots+p_K=1$

EXPECTED VALUE: $E(X)=\mu=x_1p_1+x_2p_2+\dots+x_Kp_K$ = mean
weighted (weights = probabilities) average → summarize probabilistic behaviour
VARIANCE: $var(X)=\sigma^2=(x_1-\mu)^2p_1+\dots+(x_K-\mu)^2p_K$: dispersion w.r.t. expected value

2) **Continuos** 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

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

Linear transformation → for any random variable X
Y → linear transformation of X
- for any two numbers a and b $Y=a+bX$

- $E(Y) = E(a+bX) = a + bE(X)$
- $var(Y) = var(a+bX) = var(bX) = b^2 var(X)$

STANDARDIZATION

X random variable with EV μ and variance σ^2

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

- $E(Z) = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$
- $var(Z) = \frac{1}{\sigma^2} var(X) = \frac{\sigma^2}{\sigma^2} = 1$

numbers

28

Bernoulli distribution

Bernoulli random variable

- Experiment with two outcomes → success or failure

p → success probability

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

Probability distribution distribution parametrized by p

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

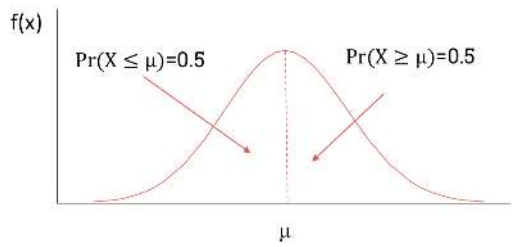
- $E(X) = 0\cdot(1-p) + 1\cdot p = p$
- $var(X) = (0-p)^2(1-p) + (1-p)^2 = p(1-p)$

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



- μ → expected value and median
- σ^2 → variance

STANDARD NORMAL DISTRIBUTION

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

STANDARDIZATION of normal random variable

$X \sim N(\mu, \sigma^2) \rightarrow \boxed{Z = \frac{X-\mu}{\sigma}}$ 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

$Pr_X(X = x_i)$ for $i = 1, 2, \dots, K$

$Pr_Y(Y = y_j)$ for $j = 1, 2, \dots, h$

Joint probability function → Joint probability behaviour

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

-cross-tab

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

$Pr(X=0, Y=1) = 0.2$ $\qquad$ $Pr(X=1, Y=3) = 0.1$

X: 1 student
0 otherwise

Y: 1 unsatisfied
2 not satisfied nor unsatisfied
3 satisfied

X	Y			
	1	2	3	Marginal
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 distribution of X

↓

marginal probability distribution 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	0.3·0.3	0.3·0.7	0.3
1	0.3·0.7	0.7·0.7	0.7
	0.3	0.7	

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

Perfect independence

Linear association

Cov and Pearson's correlation coefficients → def and interpretation as in Descriptive statistics

Pearson's correlation index $\boxed{\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}}$

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

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

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

BIVARIATE vector

- X → expected value μ_X and variance σ^2_X
- Y → expected value μ_Y and variance σ^2_Y

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

$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) = \sigma^2_X + \sigma^2_Y + 2Cov(X,Y)$

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

$Var(X+Y) = Var(X) + Var(Y) = \sigma^2_X + \sigma^2_Y$

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

$Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y) = \sigma^2_X + \sigma^2_Y - 2Cov(X,Y)$

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

$Var(X+Y) = Var(X) + Var(Y) = \sigma^2_X + \sigma^2_Y$

I.I.D Random variables

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

$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n) = n\mu$

$var(X_1 + X_2 + \dots + X_n) = var(X_1) + var(X_2) + \dots + var(X_n) = n\sigma^2$

} sum of i.i.d. r.v.

arithmetic average of i.i.d. r.v.: $\boxed{\overline{X} = \frac{X_1 + X_2 + \dots + X_n}{n}}$

- $E(\overline{X}) = E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n} E(X_1 + X_2 + \dots + X_n) = \frac{1}{\cancel{n}} \cdot \cancel{n}\mu = \mu$
- $var(\overline{X}) = var\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n^2} var(\underbrace{X_1 + X_2 + \dots + X_n}_{= T}) = \frac{\cancel{n}\sigma^2}{n^{\cancel{2}}} = \frac{\sigma^2}{n}$

$E(T) = n\mu$

$var(T) = n\sigma^2$

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 $n\sigma^2$ $\boxed{T \approx N(n\mu, n\sigma^2)}$
- 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 $\frac{\sigma^2}{n}$

$E(\overline{X}) = \mu \quad var(\overline{X}) = \frac{\sigma^2}{n}$

$\boxed{\overline{X} \approx N(\mu, \frac{\sigma^2}{n})}$

Sum of Bernoulli r.v.

$X_1, \dots, X_n$ i.i.d. $\sim Ber(p)$

$E(X_i) = p \quad var(X_i) = p(1-p)$

$(X_1 + \dots + X_n)$ has distribution that for n large can be approximated by a normal distribution with E: np and var $np(1-p)$

EXAMPLE

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

⇒ work out probability that more than 35 students out of 150 randomly drawn are admitted to the internship program

$X_i = 1$ if student selected

$X_i = 0$ otherwise

$X_1, \dots, X_{150}$ i.i.d $\sim Ber(0.3)$

⇒ $Pr((X_1 + \dots + X_n)) > 35)$

$E(X_1 + \dots + X_n) = np = 150 \cdot 0.3 = 45$

$var(X_1 + \dots + X_n) = np(1-p) = 150 \cdot 0.3 \cdot 0.7 = 31.5$

Alice Elliston

n large → Normal approximation

$Pr((X_2+\dots+X_{150})>35) = 1-Pr((X_2+\dots+X_{150})<35) \approx 1-pnorm(35,45,sqrt(31.5)) = 0.9626$

ARITHMETIC average of Bernoulli r.v.

- $X_1,\dots,X_n$ i.i.d $\sim Ber(p)$

$\boxed{\bar{X} = \frac{X_1+X_2+\dots+X_n}{n}}$

- 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:

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

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

⇒ 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 $\sim Ber(0.3)$

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

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

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

$Pr(\hat{p}<0.25)$ ⇒ Normal approximation μ $sd = \sqrt{var}$

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

↓

returns probability of observing a value ≤ to

Module 3 - Inferential Statistics: ESTIMATION PROBLEM

29 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 → cannot stop at description

INFERENTIAL STATISTICS:

- collection of techniques that let draw conclusions on the entire population (inferences)

→ from the analysis of a portion of the population, a sample drawn from it

SAMPLE → representative of the population

random sampling

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

- same prob. to be drawn ⇒ no BIASES
- samples of given size n, same prob. to be drawn

- Description of any inferential problem within a PROBABILISTIC FRAMEWORK

→ 1st step in building the probabilistic framework

Quantify 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 ($n^o \geq 0$)

Alice Elliston

EXPERIMENT - RADOM variable $X \rightarrow$ "n° 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 n° of hours spent watching TV per day by ALL adults in USA in 2012
Parameter = unknown n°

Population variance $\sigma^2 \rightarrow$ variance of the n° of hours spent watching TV per day by ALL adults in USA in 2012
Parameter = unknown n°

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

INFERENTIAL PROBLEMS:
- Point estimation
- confidence interval estimation
- (Hypotesis testing)

STATISTICS: inference based on sample data

SAMPLE STATISTICS:
- function of the sample
- random variable

sample mean: $\bar{X} = \frac{X_1 + \dots + X_n}{n}$
- random variable
- observed value $\rightarrow \bar{x}$

$(X_1, \dots, X_n)$ i.i.d $E(X_i) = \mu$

μ $var(X_i) = \sigma^2$

$\bar{X}$ $\bar{x} = 20000$

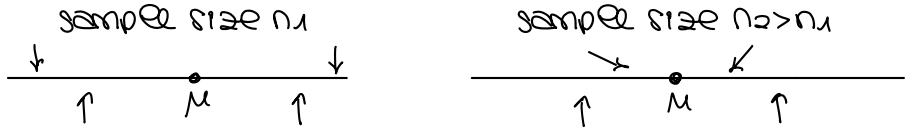
PROPERTIES OF SAMPLE MEAN: for an i.i.d sample, the sample mean has:

1) 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 - Is the sample mean a good estimator?
UNBIASED estimator $\rightarrow$ no systematic error

2) $var(\bar{X}) = \frac{\sigma^2}{n}$
   - observed 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 → the closer to the population mean
- sample mean → CONSISTENT estimator

STANDARD ERROR of the mean

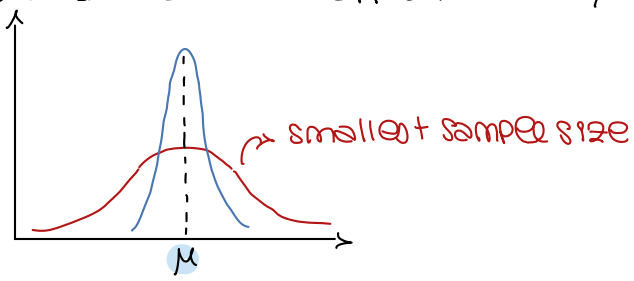
- standard deviation of the sample mean $\frac{\sigma}{\sqrt{n}}$ → standard error of the mean

↳ unknow parameter

$\frac{\sigma}{\sqrt{n}}$ standard error → 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 $\boxed{\bar{X} \approx N(\mu, \frac{\sigma^2}{n})}$



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 → population variance σ^2

σ^2 = variance of our characteristic of interest ≈ 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 σ^2 → S^2 estimator of σ^2

n−1: so it is UNBIASED $\boxed{E(S^2) = \sigma^2}$

- S estimator of σ

$$S = \sqrt{S^2} = \sqrt{\frac{(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + \dots + (X_n+\bar{X})^2}{n-1}}$$

estimating σ^2 and suggest sample variance S^2

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

UNBIASNESS concerns expected value

2.5783: estimate of σ

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

σ → spread of my char. of interest

enters in the expression of standard error

Alice Elliston

ESTIMATION OF THE STANDARD ERROR OF the mean

SE standard deviation of $\bar{X}$

in case of sample mean the standard error is $\boxed{\frac{\sigma}{\sqrt{n}}}$ parameter ↘ when $n \uparrow$

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

r.v. delivering estimates $\boxed{\frac{S}{\sqrt{n}}}$ → estimator of the population standard deviation

→ ESTIMATOR of the parameter $\frac{\sigma}{\sqrt{n}}$

estimator of the SE of the mean

number never known

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

$s = 2.5783$ observed value of SD

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

standard error of the mean → it is variable

$\bar{x} = 2.5783$

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

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

μ → 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)

↳ expected value $E(\bar{X})$

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

has a standard normal distribution

↳ SD(SE) regardless value of μ → standard normal distribution that is known

CONFIDENCE interval estimation: μ

- confidence level → $(1-\alpha)\cdot 100\%$ of a given level of confidence (it is fixed)

– typical values for α → 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 90% confident that unknown parameters μ belongs to it

length: measure of accuracy

↳ obj that matters

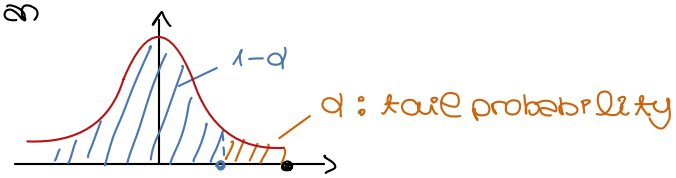
a confidence interval for an

↳ standard normal

$Z \sim N(0,1)$

quantile of order $(1-\alpha)$

$0 < \alpha < 1$



quantile named after the right tail probability

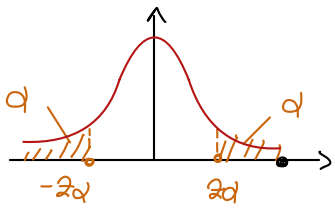
$Pr(Z \leq z_\alpha) = 1-\alpha$ quantile is number s.t. Pr of $1-\alpha$ to observe a smaller value

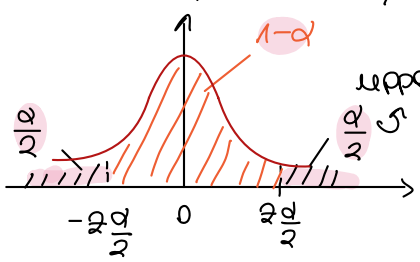
quantile of order $1-\alpha$ leaves above mass $= \alpha$

$z_{0.75}$ leave 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

QUARTILE of a standard normal

z_α → quantile of order $1-\alpha$ of a standard normal distribution

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

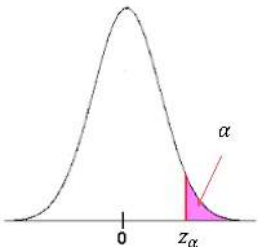
$\frac{\alpha}{2}$ → divide 2 tails symmetric w.r.t. y-axis

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

$$\Downarrow$$

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

$$\Downarrow$$



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

$\Downarrow$

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

number contains μ with probability $1-\alpha$

take estimator (r.v.)

and identify 2 quantities: $\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{n}$ $\bar{X}$ - number $\rightarrow$ r.v.

r.v.

<u>The RANDOM interval</u>

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

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 $\alpha \rightarrow$ interval of values

90% confident that my <u>unknown pop. mean</u> 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:

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

<u>LENGHT and MARGIN OF ERROR</u>

$\bar{x} \pm z_{\frac{\alpha}{2}}\boxed{\frac{\sigma}{\sqrt{n}}}$ $\rightarrow$ S.E. (standard error)

SD of estimator (depends on n and σ) $\Rightarrow$ measure of accuracy and UNCERTAINTY of estimation of μ

$\downarrow$

level of confidence

- $z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ = margin of error

taking estimate and stretching it to account for the uncertainty

- mid point confidence interval tells ESTIMATE
- lenght gives measure of error $\rightarrow$ twice the margin of error

$\hookrightarrow$ LENGHT of the confidence interval:

$\ell = \boxed{2\, z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}} = 2 \cdot ME$ (margin of error)

$\rightarrow$ measure of accuracy

FIX confidence level (i.e. 80%) fixing $(1-\alpha) \rightarrow$ fixed α

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

SD $\downarrow$

Alice Elliston

for given confidence level
and given σ^2, the greater n →
- the lower the SE
- the lower the margin of error
- the lower is the LENGHT of the interval

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

The greater the confidence level $(1-\alpha)$ →
- the lower α
- the greater $z_{\frac{\alpha}{2}}$
- 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 → estimate it

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

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

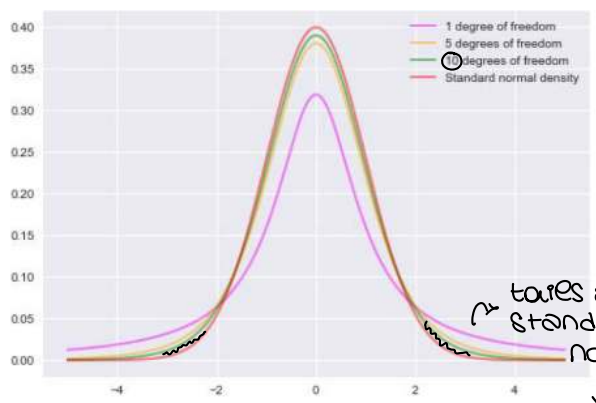
$(X_1, \dots, X_n)$ i.i.d ~ $N(\mu, \sigma^2)$

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

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

$\frac{S}{\sqrt{n}}$ → 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

tails above standard normal
↓
has thicker TAILS

Probability student's is > than standard normal when $n\uparrow$ → tails closer to standard normal (see green very close)

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

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

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

Alice Elliston

$$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

confidence interval:

center: estimate

Quantile from a Student's t distribution

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

In R:

Student's t distribution with n degrees of freedom

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

qt(1-α, n) returns the quantile of order (1-α), the value q such that $Pr(X \leq q) = 1-\alpha$

for X with Student's t distribution with 10 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-α)

→ returns the (1-α)100% confidence interval on the pop. mean of the considered value: t.test(variable_name) returns by default the 95% confidence level

Alice Elliston

FOR DOUBTS OR SUGGESTIONS ON THE HANDOUTS

ALICE ELLISTON

alice.elliston@studbocconi.it
@alice.elliston_
+39 3442849243

FOR INFO ON THE TEACHING DIVISION

NICOLA COMBINI

nicola.combini@studbocconi.it
@nicolacombini
+39 3661052675

MARTINA PARMEGGIANI

martina.parmegiani@studbocconi.it
@martina_parmegiani05
+39 3445120057

MARK OLANO

mark.olano@studbocconi.it
@mark_olano._
+39 3713723943

TEACHING DIVISION

OUR PARTNERS

Wall Street English®

TEGAMINO'S

ETHAN SUSTAINABILITY

700+ CLUB

LA PIADINERIA

LA PIADINERIA dal 1994