

Introduction - and motivation



- Why studying Statistics in the biomedical field
 - Role of Statistics in research, education, decision making
 - Evidence-Based Medicine / Nursing / Prevention

Sources of knowledge in Medicine

How do we know about ...

- features of different diseases
 - risk / prognostic factors
 - efficacy of treatments
- etc?

Personal experience of MD with patients

Lab research (studying tissues, cells, molecules ...)

... are not sufficient!!

- "Multiple sclerosis is a disease of young people, mainly women ..."
- "Smoking causes cancer"
- "In breast cancer, negativity of lymph nodes is associated with good prognosis"
- "In a clinical trial for relapsed myeloma, Velcade has proven to be superior to the standard treatment based on Dexamethasone for the survival of the patients"

Sources of knowledge in Medicine

How do we know about ...

- features of different diseases
 - risk / prognostic factors
 - efficacy of treatments
- etc?

Personal experience of MD with patients

Lab research (studying tissues, cells, molecules ...)
... are not sufficient!!

Answering questions about clinical facts needs the **observation** of the phenomena of interest in (a large number of) human beings

- select a set of individuals reasonably "general" to represent the reality we are interested in without distortion
- collecting information
- processing data to get explicative summaries of issues of interest
- communicating results with proper interpretation

The role of Statistics in Medicine

- Statistics is a
- collection of methodological approaches and mathematical methods
- for observing **collective phenomena** characterized by **variability**
- to gather quantitative knowledge
- useful for research and decision making

Answering questions about clinical facts needs the **observation** of the phenomena of interest in (a large number of) human beings

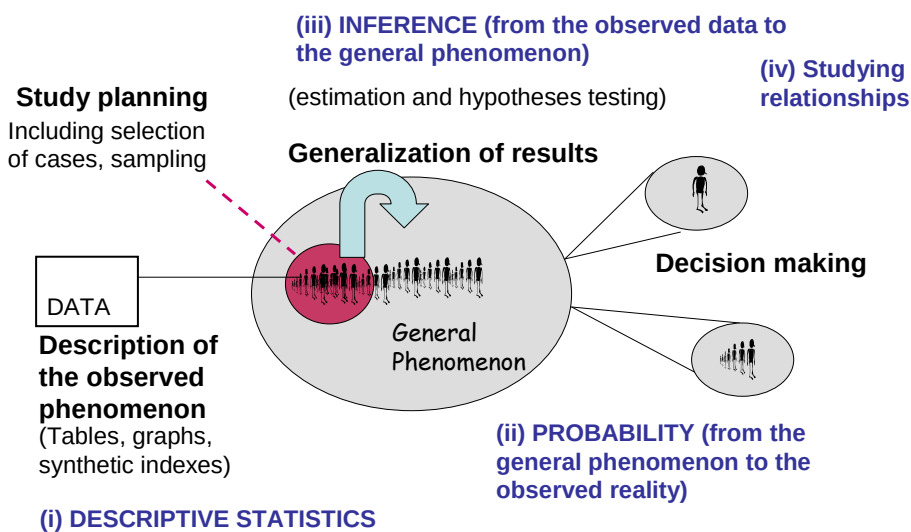
- select a set of individuals reasonably "general" to represent the reality we are interested in without distortion
- collecting information
- processing data to get explicative summaries of issues of interest
- communicating results with proper interpretation

Errors in Medicine, ethics, EBM

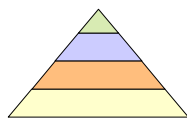
- It is estimated that 15% of errors in clinical practice is of cognitive type, being a consequence of:
 - Bad information
 - Bad reasoning, due to the overlooking or bad usage of «good» information, and/or systematical use of wrong methods
- ETHICS imposes to USE COGNITIVE RESOURCES AT THE BEST
- Decisions and interventions, both for single individuals or for the community, are more and more based on data:
 - Evidence-Based Medicine
 - Evidence-Based Nursing
 - Evidence-Based Prevention

Statistics and phases of research

Our course



Basic terminology and first concepts



- Population and sample
- Types of data. Classification of variables

Objects of interest

It's an object of interest any phenomenon that concerns a group of individuals and that shows to have **variability** from one individual to another.

	Italians with residence in Italy on 9/10/2011	..., home property, possibility of web connection, job conditions, time distance from workplace, etc ...
	Breast cancer patients	Age, menopause, tumor stage, tumor dimension, treatment, survival duration...
	Counseling assisted family	Number of components, education of family master, income, presence of members needing assistance...
	Hospitals in the territory	Number of employees, number of beds, presence of an intensive care unit...

Population

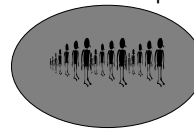
→ Considering a phenomenon of interest, we can imagine that it exists a **POPULATION** of individuals that, if entirely observed, allows us to know every aspect of interest of the phenomenon

→ It's called **TARGET POPULATION**

→ It can be a **real** population which is potentially fully observable (eg survey among Italians), or a **hypothetical** and **not identifiable** population

Example: studying the outcomes of the novel influenza

Target population: all individuals previously exposed to the contagion, individuals that will be exposed to it, patients already ill, and those exposed and ill in the past and subsequently dead.



We can represent this population like a set

Notice: the term 'individual' is used for **STATISTICAL UNIT**, the smallest element of the population (patient, family, hospital, cell etc:)

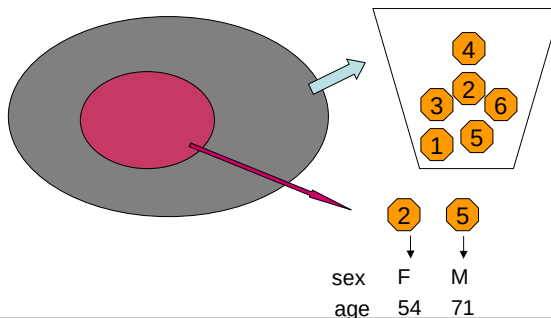
Sample

The set of individuals for which we actually observe the phenomenon is called observed population, or **SAMPLE**.

→ The sample is also represented as a set, and being a part of population (subset) it is completely included in the population set

Ideally, these **individuals** have been **extracted from target population, like bullets from a container.**

→ This intuitive concept is the reason why Statistics needs probability theory (a branch of Mathematics) to link the observed data to the target population



The word «sample» can refer to the units (individuals) extracted but also to the observed data

→ Observed data can be considered as numbers drawn at random from a truffle base (container)

Variables

We are interested in patients screened with colonoscopy.

For a sample of 12 patients, we observe some "characteristic", or variables

unit	Gender	AGE	N. polyps	Education	Chronic pathology
1	M	55	0	University	IBD
2	F	51	2	Secondary	None
3	F	44	2	University	Obesity
4	M	62	3	Secondary	IBD
5	M	48	0	University	None
6	F	51	0	Secondary	None
7	M	69	1	Primary	Obesity
8	F	58	0	Primary	None
9	F	72	0	Secondary	None
10	M	50	0	Secondary	Obesity
11	F	78	1	Primary	Other
12	F	46	0	Secondary	IBD

The characteristics we are interested have a different nature, which is relevant for describing them and/or manipulating the information (summarizing, establishing relationships ...)

Variables

- Statistical unit's characteristics of interest are called **VARIABLES**.
- Variables present (they are expressed through) **MODALITIES** or **VALUES**, each one different from unit to unit.
- Variables are classified according to the following terminology, which allows to define the types of possible operations on modalities, to manipulate, compare and summarize them.

QUALITATIVE

NOT ORDERED

sex → M,F
disease → ulcer, gastric tumor, intestinal tumor...

ORDERED

Education → none, primary school, secondary school, degree
Disease stage → I,II,III, IV

QUANTITATIVE

DISCRETE

Members (of a family) → 1,2,3,4, ...
Previous pregnancies → 0, 1, 2, 3, ...

CONTINUOUS

weight (kg) → 56.4, 78.2, ...
WBC ($\times 10^3/\text{ml}$) → 3.4, 2.8, ...

Qualitative variables

- They present modalities corresponding to attributes and features that could be described through «words» (or through numbers expressing conventions, not truly derived from counting or measurements)
- We cannot apply any mathematical operation to the modalities!!
 - **NOT ORDERED**: there is no natural or typical or conventionally decided sorting
 - It is only possible to establish if two units are equal or different (if they have the same or different modalities)
 - **ORDERED**: a natural or typical sorting exists
 - It is possible to establish superiority/inferiority relations between two units;
 - it is anyway not possible (it's a non-sense) to calculate differences to estimate «distance» between two units
 - (Do not get confused by numerical coding!!)

Dichotomous Variables

- A particular kind of not ordered qualitative variable is the BINARY or DICHOTOMOUS one, that can take only two modalities.
- It can be usually considered like an indicator of the presence/absence of a certain feature.
- We often use the numeric coding 0/1 (0=no=absence, 1=yes=presence)

Examples

- Smoker: yes/no
- Responding to therapy: yes/no
- Sex = M/F, or:
- Male patient: yes/no

Manipulating qualitative variables

Recoding

DISEASE
a – gastric tumor
b – gastric ulcer
c – intestinal tumor

DISEASE
ulcer (b)
tumor (a, c)

Concerning qualitative variables we can lump together modalities

DISEASE
gastric (a, b)
Intestinal (c)

For not ordered qualitative variables, this can be done following various criteria.

TUMOR STAGE
I
II
III
IV

TUMOR STAGE
I – early stage
II-III – advanced stage
IV – terminal stage

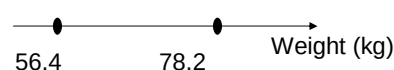
For ordered qualitative variables, it's necessary to respect modalities order

Quantitative Variables

- They present numeric modalities, obtained through counting or measurements; it is possible to apply mathematical operations on modalities; two modalities can be compared using the ratio or the difference
- **DISCRETE**: modalities can be enumerated: values between two modalities can NOT be modalities too.
- Generally obtained by counting



- **CONTINUOUS**: modalities cannot be enumerated; values between two modalities are ALWAYS modalities too
- Generally obtained by measurement



→ The measuring instrument inaccuracy produces an approximation or ROUNDING, but the variable is continuous in nature

→ A discrete variable with a very high number of modalities, ex. inhabitants of a municipality or age expressed in years – can be treated as a continuous variable

Manipulating quantitative variables

Recoding in classes; mathematical transforms

Age	
- 25	
25 - 45	$25 < \text{Age} \leq 45$ (25, 45]
45 - 65	
65 -	Age > 65 (open class)

Quantitative variables can be reduced in CLASSES, by lumping together modalities. In this way they look like ordered qualitative variables.

WBC		log(WBC)
2.2		0.788
3.2		1.160
1.8		0.588
2.1		0.742

Quantitative modalities can be transformed through mathematical operations.

Variables Hierarchy (1)

Variable	Possible manipulation of its modalities
Not ordered qualitative	Comparison: to establish equality or difference (= or ≠) Manipulation: lumping, according to several criteria
Ordered qualitative	Comparison: to establish superiority/inferiority relations Manipulation: lumping keeping the order
Quantitative	Comparison : difference or ratio (-, /) Manipulation: creating classes; mathematical operations or transforms (+, -, ·, /, log, ...)

Description of one variable: tables, graphics and synthetic indexes



- Qualitative variables
 - Relative Frequencies and percentages; cumulative frequencies. Mode.
- Quantitative variables
 - Concept of Frequency Density, histogram; boxplot
 - Statistical summaries: Position indexes (mean, median, quantiles) and Variability indexes (standard deviation, variance, variation coefficient)
 - Normal Curve and other shapes of the distribution

Frequency tables

unit	Gender	AGE
1	M	55
2	F	51
3	F	44
4	M	62
5	M	48
6	F	51
7	M	69
8	F	58
9	F	72
10	M	50
11	F	78
12	F	46

Gender	n
M	5
F	7
tot	12

AGE	n
- 50	4
50 - 65	5
65 -	3
tot	12

- The first useful operation to summarize data is **counting**: to each modality (or class, ie range of values) corresponds a **frequency** that is the number of units presenting that modality (or falling into that class).
- A table is a **synthesis** of the original series, in which we lose a part of initial information (concerning the reference to each single unit), but gain in terms of a quick and general overview about the phenomenon

Original data set:

All sample units with the modality observed for each variable

Relative frequencies and Percentages

Hospitalized patients distribution according to the TYPE OF DISEASE.

Disease	n	f	p (%)
Renal failure	454	0.244	24.4
Diabetes	1227	0.659	65.9
Other organic disease	153	0.082	8.2
Psychiatric disease	27	0.015	1.5
	1861	1.000	100.0

Absolute frequency

Relative freq.

Percentage freq.(%)

Eg. for the second modality:

$$\frac{1227}{1861} = 0.659; \quad 0.659 \times 100 = 65.9$$

These quantities express the same fraction

This is the concept of **proportion**

$$1227 : 1861 = 0.659 : 1 = 65.9 : 100$$

Sample total
number

Total
=1

total=
100

Usage and interpretation of Percentages

Treatment response	%
No	42.9
Yes	57.1
tot	100.0

freq.
6
8
14

freq.
600
800
1400

(*) In these 2 scenarios, different absolute frequencies are consistent with the same percentages.

Do the two studies have the same reliability?

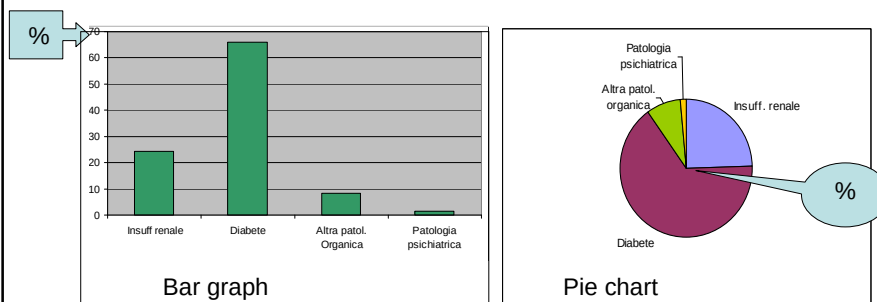
Which one is more reliable?

- Response **percentages could be interpreted** as the probability of response to treatment for a generic patient
 - If 20 (new) patients are treated, we expect that 11 of them (57.1%) respond to treatment: $0.571 \times 20 = 11.42$
 - Although useful, percentages are not exhaustive of a distribution: we miss the information about the **total number**, that is **necessary to assess data reliability/precision**
- When you have percentages, always report / look at / the total sample size!! (*)

Graphs from qualitative variables tables

Hospitalized patients distribution according to the TYPE OF DISEASE.

Disease	n	p (%)
Renal failure	454	24.4
Diabetes	1227	65.9
Other organic diseases	153	8.2
Psychiatric disease	27	1.5
	1861	100.0



A synthesis of the whole table: the Mode

Hospitalized patients distribution according to the TYPE OF DISEASE.

Disease	n	p (%)
Renal failure	454	24.4
Diabetes	1227	65.9
Other organic diseases	153	8.2
Psychiatric disease	27	1.5
	1861	100.0

The most representative modality of a qualitative variable is the one with the highest frequency: this index is called **MODE**

Here, the Mode is the modality «Diabetes».

We can say that the «typical» hospitalized patient in this hospital unit is a diabetic patient. Or, in this hospital unit the largest proportion of patients suffers from diabetes.

Cumulative frequencies

This is a manipulation of frequencies only suitable for ordered variables (or continuous variables represented in classes)

Number of children	freq	p (%)	cum	% cum
0	25	40%	25	40%
1	21	33%	46	73%
2	12	19%	58	92%
3	4	6%	62	98%
4 +	1	2%	63	100%
total	63	100%		

The cumulative frequencies (absolute ones or percentage ones) simply represent the partial sums of the frequencies up to the current modality.

For example, looking at the last column, you can quickly see that:

→ 73% (3 out of 4 women) have 1 child at most;

→ 92% of women have 2 children at most, thus only 8% have more than two children
etc

Practicality

Basic synthesis of frequency tables

Number of prior hospitalization	n_i	%
0	4	$4/12 \times 100 = 33.3$
1	5	$5/12 \times 100 = 41.7$
2	3	$3/12 \times 100 = 25.0$
tot	12	

Mode = ?

Number of prior hospitalization	n_i	N (cumulative)
0	4	4 (33.3%)
1	5	9 (75.0%)
2	3	12 (100%)
tot	12	

Which is the percentage of patients who had prior hospitalizations?

33% has never been hospitalized before, thus 77% had at least one prior hospitalization



Do not confuse modalities (Hospitalizations = 0, 1, 2) with the frequencies
The MODE is "1", not "5"!

The sum of percentages must be equal to 100

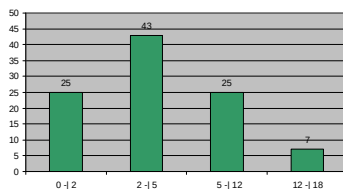
Compute the cumulative freq. only for ORDERED variables

Graphs from tables for continuous variables

Distribution of 56 pediatric patients according to their age

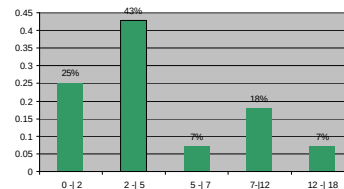
Age	freq.	%
0 - 2	14	25
2 - 5	24	43
5 - 12	14	25
12 - 18	4	7
	56	100

For example: classes 0-|2 and 5-|12 have the same frequency, thus in a bar chart they are represented as if they had the same importance:



If classes don't have the same width the simple representation of percent frequencies of the classes on the vertical axis of a bar chart returns a **distorted representation** of the phenomenon

Let's try to split the range 5-|12 in two classes: we have 4 patients from 5 to 7 years, and 10 patients from 7 to 12 years: now they both look «less important» than the 0-|2 class !!



Concept of frequency density

Age	freq.	%
0 - 2	14	25.0
2 - 5	24	42.9
5 - 12	14	25.0
12 - 18	4	7.1
	56	100

The same frequency (14 units) of the first class and of the third one refers to classes with different width, respectively 2 years (2-0) and 7 years (12-5);

If we now consider small intervals of 1 year width (0-1 year; 1-2 years; 2-3 years; etc) we have:

The 14 cases of class 0-|2 → about $14 / 2 = 7$ cases in each small interval

• the 14 cases of class 5-|12 → about $14 / 7 = 2$ cases in each small interval

To properly evaluate the 'importance' of the class, it is necessary to relate the frequency to the width of the class, considering the frequency density, a value which **represents how many units are contained in each small class of width 1**

$$\text{density} = \frac{\text{frequency}}{\text{width}}$$

$$\Leftrightarrow \text{frequency} = \text{width} \times \text{density}$$

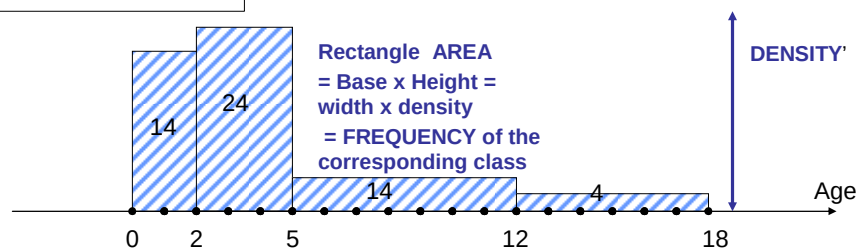
Histogram: a graph for the density

Age	freq.	%
0 - 2	14	25.0
2 - 5	24	42.9
5 - 12	14	25.0
12 - 18	4	7.1
	56	100

Classes are represented on the X-axis

width	density
$2 - 0 = 2$	$14 / 2 = 7.0$
$5 - 2 = 3$	$24 / 3 = 8.0$
$12 - 5 = 7$	$14 / 7 = 2.0$
$18 - 12 = 6$	$4 / 6 = 0.7$

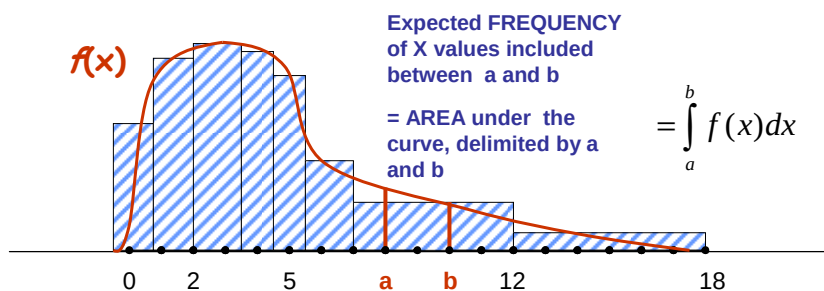
Density is represented on the Y-axis



Theoretical density curves

If we imagine an histogram using very small classes, and we connect the columns' top midpoints, we get the plot of a continuous curve.

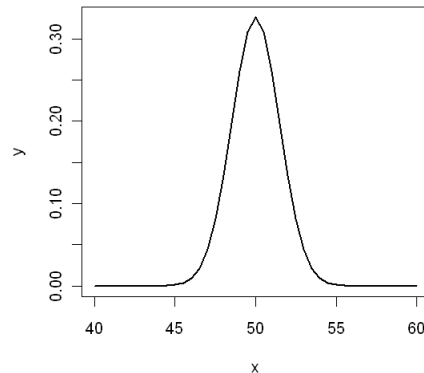
Mathematics provides equations of continuous curves that can be interpreted like theoretical density curves, which corresponds to «ideal» distributions of quantitative phenomena of interest X.



The Normal Curve (i)

A model for biological variability / errors

The main **theoretical density curve** is the Normal (Gaussian), which describes phenomena that can be measured as continuous variables and depend on «chance», like measurement errors. It's symmetric and it is bell-shaped



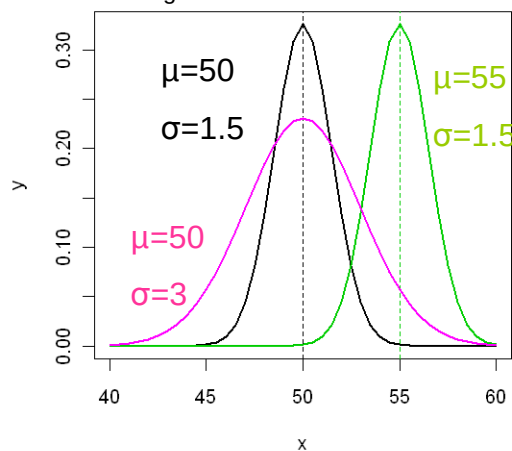
$$y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Eg: distribution of repeated weight measurements of a patient who weighs 50 kg

The Normal Curve (ii)

A model for biological variability / errors

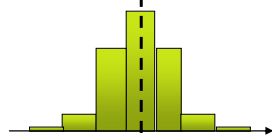
The formula describing the Curve contains two parameters μ and σ ; the first one indicates the curve position along to the horizontal axis and the second one is related to how large the bell is.



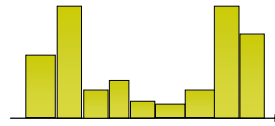
$$y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Various distribution shapes

SYMMETRIC distributions: the density mass is distributed «equally» on two sides of an imaginary axis (symmetry axis)

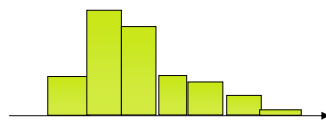


The **bell-shape** is typical of phenomena considered as being the result of chance variation, for example individual height

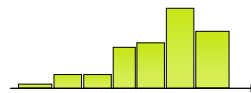


Bimodal distribution, that is with density concentrated in two masses.

It's often considered as an indication of presence of different sub-populations: e.g. height in females and males



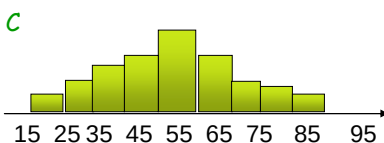
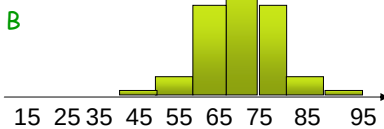
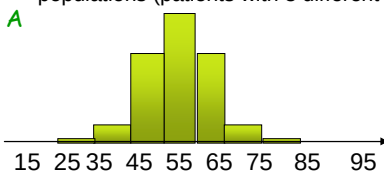
Asymmetric (skewed) distribution on the right is typical of several biological phenomena, e.g. variables with positive values that can take very high values, but not very low ones, as the body weight, WBC values, etc



In the asymmetric distribution on the left, according to an hypothetical symmetry axis, there's a density mass on the left, on low values

Synthesis of quantitative variables

Distribution of the age at diagnosis for 3 different populations (patients with 3 different diseases)



Frequency tables and graphs return a complete representation of data.

Statistical indexes can give us synthesis of some distribution aspects.

The two main features are:

The **position** of the variable on the axis; ideally we identify one single value to represent all other values.

The **variability** of the variable, that is if observations are homogeneous i.e. similar, or heterogeneous, i.e. scattered along the axis.

Arithmetic mean

Arithmetic mean is one of the most important **position indexes**.

The mean is the total amount (sum of all observations) divided in equal parts

→ If the mean replaces each observation, we have the same total amount

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

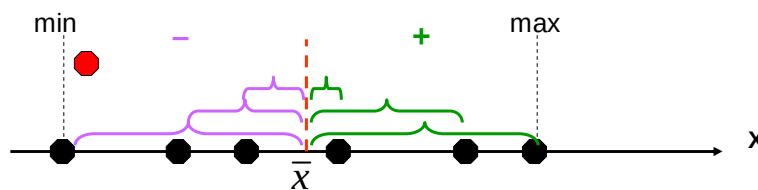
$$\bar{x} = \frac{\sum x_i}{n} \Leftrightarrow \sum x_i = n\bar{x}$$

Mark	
26	24
18	24
28	24
Sum = 72	72

Mean = 72 / 3 = 24

➤ Arithmetic mean is like a **center of gravity**, as shown by several of its properties: it's in the center of observations, that's why it represents them, it's an useful synthesis

Main properties of the mean



$$\min(x_i) \leq \bar{x} \leq \max(x_i)$$

The mean falls within the range i.e. it is included between the smallest and the largest observed value

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

The differences of each observed value and the mean are called 'errors'. The sum of all errors is equal to zero: in other words the mean is the "centre" of the observed value, balancing positive and negative errors

$$\text{dist} = \sqrt{\sum_{i=1}^n (x_i - C)^2}$$

Measuring the distance of each observation and a value C and using this as global measurement of distance, the minimum value is obtained with C equal to the mean. In other words the mean is the value globally nearest to the observations

(There are other means besides the arithmetic one (quadratic; geometric; harmonic) that have other properties, they are anyway less useful and we don't consider them)

Mean of a quantitative discrete variable from a frequency table

Sample of 8 women giving birth, distributed by the number of prior deliveries:

Deliveries (x_i)	freq. (n_i)	$x_i \cdot n_i$
0	4	0
1	3	3
2	1	2
tot	8	5

Total number of deliveries =
 $(0+0+0+0)+(1+1+1)+(2)$
 $= 0 \cdot 4 + 1 \cdot 3 + 2 \cdot 1$

Mean = $5 / 8 = 0.6$



Do not confuse modalities (Deliveries = 0, 1, 2) with the frequencies

There are $n=8$ units, and 3 modalities.
 We need to compute the total of the variable and then divide it by the number of units (or sample size)

The total of the variable for each modality is the product of the modality by the frequency

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_i}{n}$$

Mean of a quantitative continuous variable from a frequency table (data in classes)

Es: Body weight for a sample of 64 patients

weight (kg)	freq. (n_i)	x_i	$x_i \cdot n_i$
- 50	4	45	180
50 - 60	17	55	935
60 - 70	24	65	1560
70 - 80	11	75	825
80 -	8	85	680
	64		4180

Mean = $4180 / 64 = 65.3$



The idea is again that we first need the total per class, then divide the sum by n . The problem is that we don't have a single value but a class as modality.

Solution: to assign to each class a representative value – usually the central value

$$x_i = \frac{l_{i-1} + l_i}{2}$$

For open classes we choose a “likely” value. The mean depends on what we choose.

Mean of two (or more) groups

A paper reports the mean cholesterol level separately for a group of 40 men (198 mg/dl) and 16 women (190 mg/dl). What was the overall mean in the sample?

	mean $\bar{x}$	nr of cases n	total
M	198	40	198×40=7920
F	190	16	190×16=3040
		56	10960



mean $\neq (198+190)/2=194$

We must as usual divide the total amount by the total sample size

Following the logic, we derive the formula of the WEIGHTED AVERAGE

mean = 10960 / 56 = 195.7

$$\bar{x} = \frac{\sum x_i}{n} \Leftrightarrow \sum x_i = n\bar{x}$$

$$\bar{x} = \frac{\sum_{all} x_i}{n} = \frac{\sum_{gr1} x_i + \sum_{gr2} x_i}{n_1 + n_2} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Weighted average

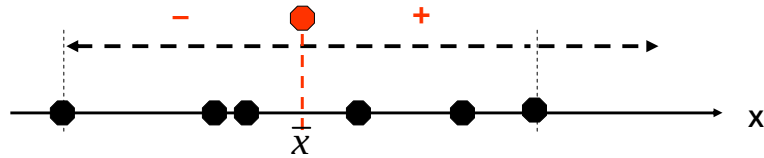
General formula for the mean of a series of K means $\bar{x}_j$, each computed in a subgroup with n_j units

$$\bar{x} = \frac{\sum_{j=1}^K \bar{x}_j n_j}{\sum_{j=1}^K n_j}$$

The idea can be generalized: we can obtain the mean of K objects x_j giving to each one a different «weight» p_j

$$\bar{x}_p = \frac{\sum_{j=1}^K x_j p_j}{\sum_{j=1}^K p_j}$$

Limitations of arithmetic mean

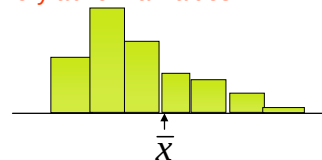


Since the mean is set at the centre of the observed values, BALANCING positive and negative deviations, it is influenced by very high or very low values

(think what happens if the large values become larger: the mean "follows" them: it is attracted by EXTREME VALUES)

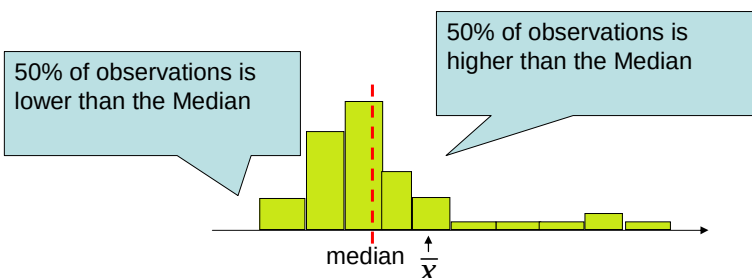
→ Arithmetic mean is an unsatisfactory synthesis of the distribution

- When we have one or more extreme and very abnormal values
- When we have a skewed distribution



The Median

- The arithmetic mean is an unsatisfactory synthesis of the distribution:
 - When we have one or more extreme and very abnormal values
 - When we have a skewed distribution
- In these cases the median is more representative: it is a value x such that a half of the observations are lower than x and a half are higher than x

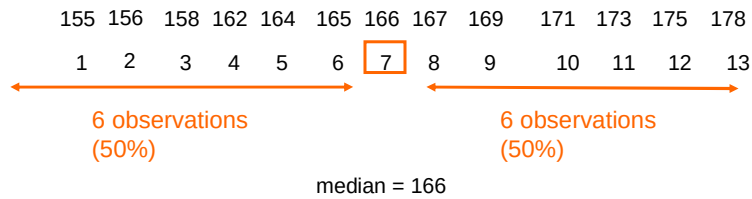


The Median

Example: in a sample of 13 subjects we study the variable Height (cm)

173 155 162 165 167 175 171 169 164 178 156 158 166

→ we sort observations in ascending order, giving them a position in the list
(RANK)



Odd number of obs → median = modality with rank $(n+1)/2$

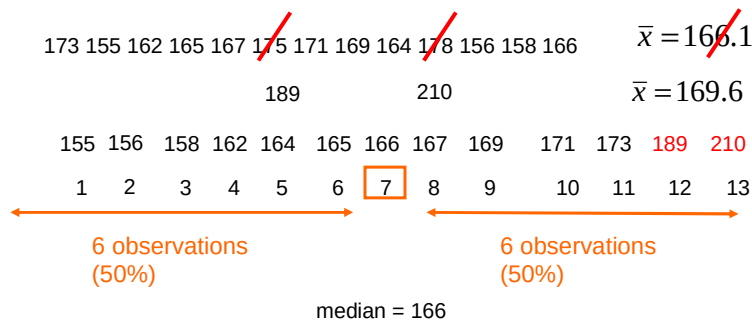
Even number of obs → median = intermediate modality between the two modalities with rank $n/2$ and $n/2+1$

(for example, if $n=6$, it's the central modality between the 3rd and the 4th)

Median robustness

The median is not affected or is slightly affected by extreme values (e.g. values much higher than the other ones) (it's «robust»)

look what happens if in the previous sample the two taller subjects are even more tall:



→ The median doesn't change because the ranking doesn't change (instead the mean changes because the total amount changes)

Mean and Median of a quantitative variable

n = 6 units

id	Weight (kg)	Rank
p1	54	1
p2	65	4
p3	71	6
p4	55	2
p5	58	3
p6	68	5
	$\Sigma = 371$	

Ranking

Rank	1	2	3	4	5	6
Weight	54	55	58	65	68	71

$$n=6 \rightarrow n/2=3 \quad n/2+1=4$$

$$\text{median}=(58+65)/2 = 61.5$$

Mean = sum / n

$$= 371 / 6 = 61.8$$

Median for n = even number

Identify the modalities ranked n/2 and n/2+1

Median = sum of the two modalities / 2

Median (class) of a quantitative continuous variable from a frequency table

Sample of 8 women giving birth, distributed by the number of prior deliveries:

Prior deliveries	n_i	N_i
0	6	6
1	5	11
2	3	14
3	3	17
4	1	18
tot	18	

$n/2=9 \rightarrow$ the 9th unity has modality "1"

In fact, at modality "0" we have only the first 6 patients, moving to "1" we get up to the 11th, and thus we have included the 9th

$\rightarrow$ Median =1



The idea is as usual to identify the modality ranked n/2 (and the subsequent).

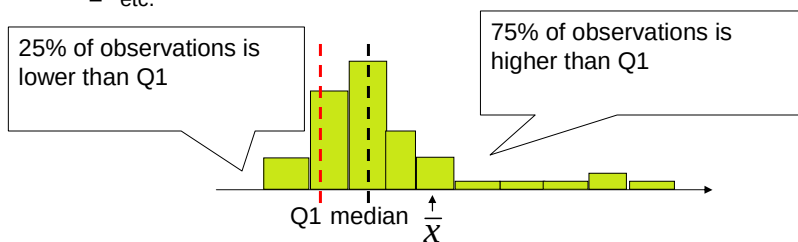
That's why we look at the cumulative frequencies

With a continuous variable, we can identify in the same way the median class (and we would identify the modal class).

(There is a formula to identify a single median point, we can disregard it)

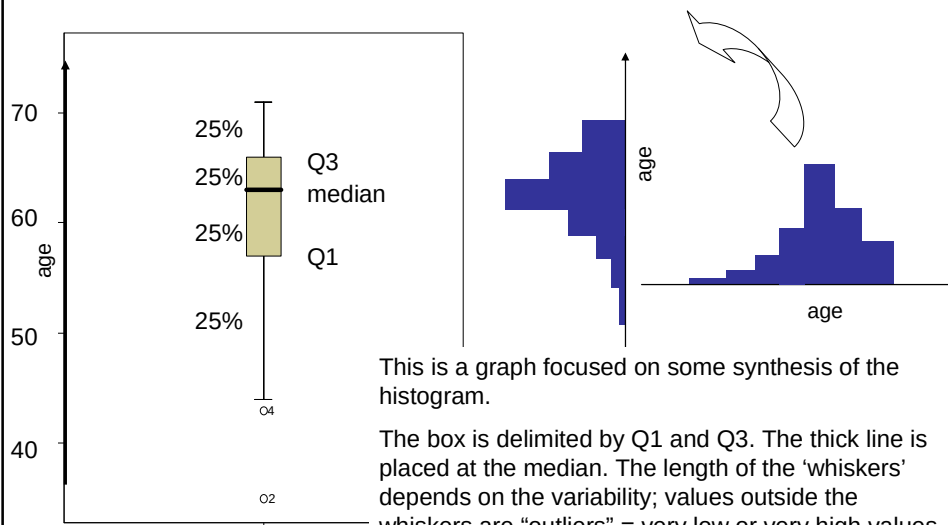
Generalization of the median : other quantiles

- The median divides the distribution into 2 parts, each one including 50% of observations
- Quantiles separate the distribution in other percent fractions, for example:
 - 1st quartile (Q1): divides the first 25% from the remaining 75%
 - 3rd quartile (Q3): divides the first 75% from the remaining 25%
 - 1st decile: divides the first 10% from the remaining 90%
 - 95th percentile: only 5% of values are higher than it
 - etc.



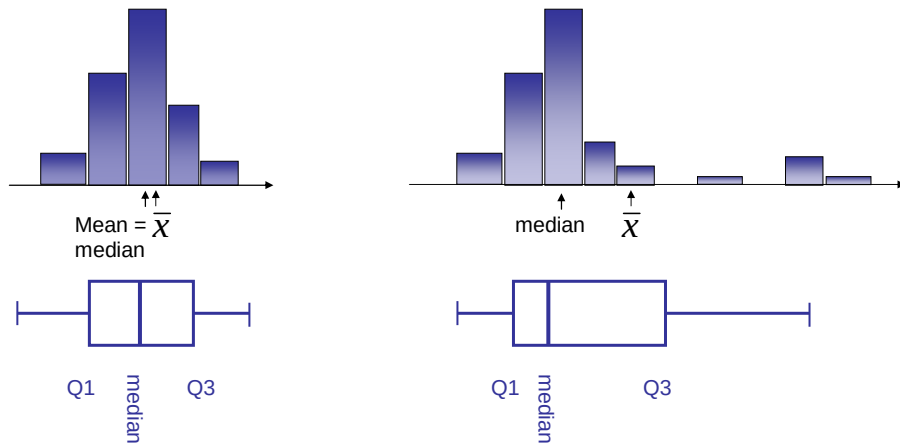
Note: we can apply the concepts of median and quantile to **ordered qualitative** variables too

Boxplot: a quartile-based graph

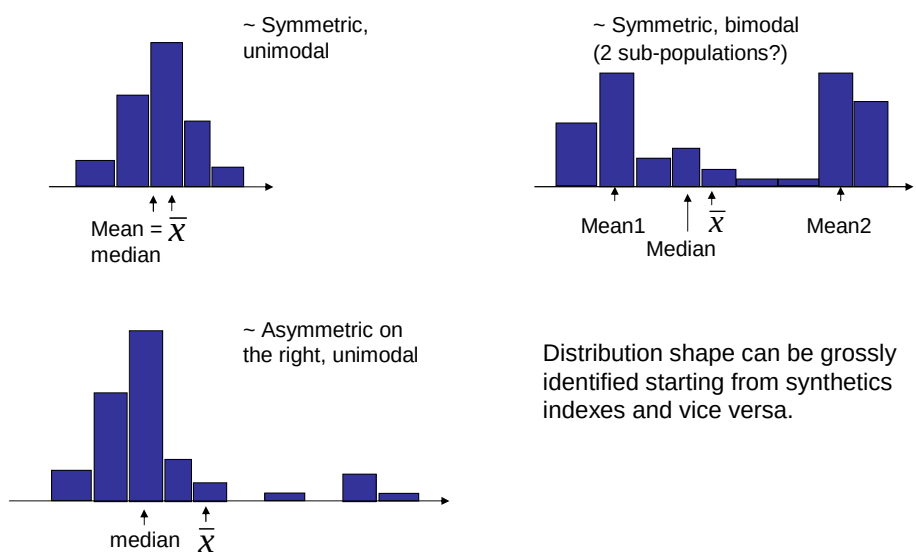


Distribution shapes and indexes (1)

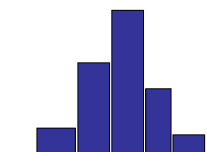
We consider some representations through histogram and boxplot: distribution shape can be grossly identified starting from synthetics indexes and vice versa.



Distribution shapes and indexes (2)

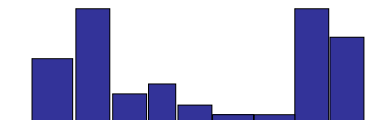


Appropriateness of the indexes



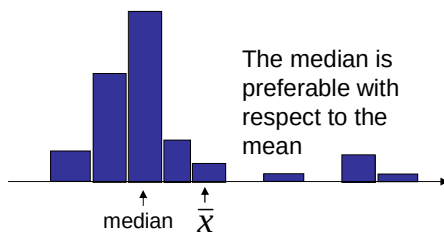
Mean = $\bar{X}$
median

The mean is a satisfactory synthesis, it tends to coincide with median and with modal class



Mean1 Median Mean2

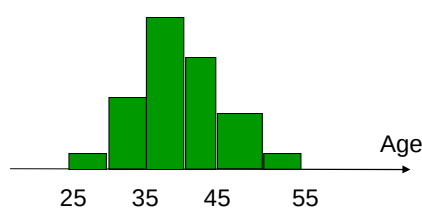
It's good to emphasize the bimodality aspect: neither the mean nor the median are good syntheses.



median $\bar{X}$

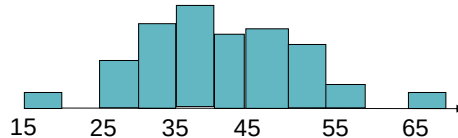
The median is preferable with respect to the mean

Measuring variability using the distances from mean $(x_i - \bar{x})$

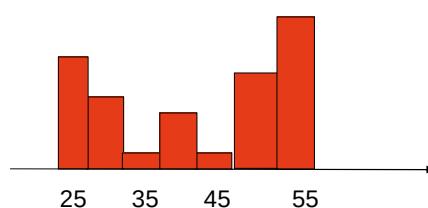


These 3 distributions are symmetric, they have the same mean = median = 38 years

Here, the largest part of observations is close to the mean, there are few 20-yrs-old and there isn't any elderly



Here we have a lot of subjects in each class, someone is very young or very old: a lot of observations are far from mean



Here there are few subjects in the central class and more in the younger or older classes: the largest part of observations is far from mean

Standard Deviation

- The Standard Deviation is an index of variability as it represents the **average (mean) distance between the observations and the arithmetic mean**

$$std = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Taking all distances of the observations from the mean ("errors"), we compute their *quadratic* mean

Notice: in the denominator we use (n-1) instead of n to have a property that we will see in the part dedicated to inference

- The standard deviation is a sort of "**relevant unit of measurement**" of the observed phenomenon
 - Ex. X = patient's weight, std = 4.5kg: this is the "relevant distance" between two patients (1kg is not relevant to describe the variation of the variable)
 - See the information related to the std in a Normal distribution (intervals around the mean)
- The quantity below the squared root (i.e. the std squared) is called **VARIANCE** and it is itself a measure of variability

Practicality

Computing the standard deviation

Age for a sample of 7 patients

AGE x_i	$x_i - m$	$(x_i - m)^2$
65	9.29	86.22
35	-20.71	429.08
44	-11.71	137.22
43	-12.71	161.65
71	15.29	233.65
63	7.29	53.08
69	13.29	176.51
	0	1277.43

mean $m=55.7$

For example at the second row:

$$(35-55.7) = -20.71 ; (-20.71)^2 = 161.65$$

$$\text{Variance} = 1277.43 / 6 = 212.90$$

$$\text{std} = \sqrt{212.90} = 14.59$$

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



Watch out to follow the correct order:

First compute the errors, $x_i - \text{mean}$;

Then rise to power 2 the errors;

Then sum the squared errors;

divide by (n-1), getting the VARIANCE;

Extract the squared root



Computing the standard deviation faster

Age for a sample of 7 patients

AGE x_i	$(x_i)^2$
65	4225
35	1225
44	1936
43	1849
71	5041
63	3969
69	4761
	23006

mean $m=55.7$

To compute the variance:

$$23006 / 7 = 3286.57$$

$$3286.57 - (55.7)^2 = 182.49$$

$$\text{Variance} = 182.49 \times 7/6 = 212.90$$

$$\text{std} = \sqrt{212.90} = 14.59$$



The VARIANCE is obtained faster with the following formula:

$$\text{var} = \left(\frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 \right) \cdot \frac{n}{n-1}$$

Remind to extract the squared root in order to get the std!!

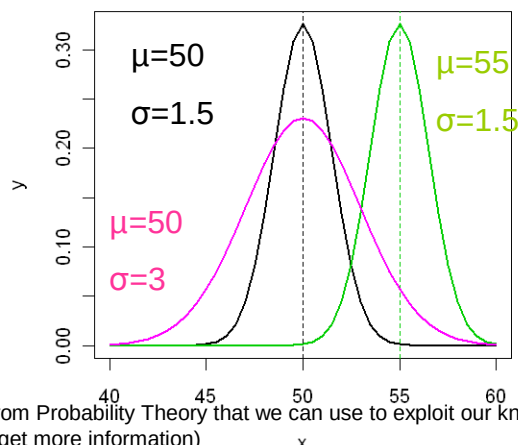
(possible discrepancies can be due to rounding)

The Normal Curve (ii)

parameters μ and σ

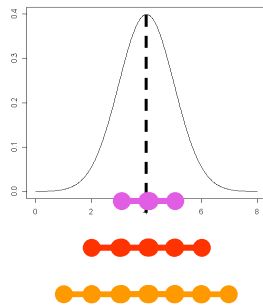
→ μ , it places the symmetry axis and it can be interpreted as the arithmetic mean (**average value**)

→ σ , it determines the bell width, that is variability or **dispersion** and it coincides with the **Standard Deviation**



(This is the first tool from Probability Theory that we can use to exploit our knowledge on a phenomenon to get more information)

Properties of the Normal



The **area** included under the curve in the following intervals is = the **frequency** of values x included in those intervals. Approximately (*):

$$(\mu - \sigma, \mu + \sigma)$$

68%

$$(\mu - 2\sigma, \mu + 2\sigma)$$

95%

Precisely: $\mu \pm 1.96\sigma$

$$(\mu - 3\sigma, \mu + 3\sigma)$$

99.7%

Median=Mean= μ . The two quartiles Q_1 and Q_3 are far 0.67σ from mean:

$$Q_1 = \mu - 0.67 \cdot \sigma$$

$$Q_3 = \mu + 0.67 \cdot \sigma$$

(*) vd. See also probability

Coefficient of variation

- CV is a relative measurement of variability: it expresses the variability with reference to the average size of the variable: moreover it hasn't got any unit of measure
- So it's a suitable way to compare variability between two different populations and between different variables too.

$$CV = \frac{std}{\bar{x}} \cdot 100$$

Ratio between the Standard Deviation and the arithmetic mean (in %)

X = baby weight: mean= 3.2 kg, std = 0.5 kg

Y = mother weight: mean= 60 kg, std = 4.5 kg

Z = baby height: mean= 51 cm, std = 2.5 cm

→ Is weight more variable in babies or in mother?

→ Are babies more variable with respect to height or to weight?

X : CV = (0.5 kg / 3.2 kg) · 100 = 15.6

Y : CV = (4.5 kg / 60 kg) = 7.5

Z : CV = (2.5 cm / 51 cm) = 4.9

→ Babies are more variable with respect to weight than to height (3 times more), and concerning weight, they are twice as much variable than mothers

Indexes interpretation

Ex: For a patients group at diagnosis of multiple sclerosis:

Mean= 36 St. Dev.. = 7

Q1=29 Median=35 Q3=41

→ Was the distribution symmetric or asymmetric?

Symmetric, around the mean value of 36 years

→ Could it have a bell-shape?

Yes: in the Normal Curve Q1 and Q3 are $36 \pm 0.67 \cdot 7 \approx 31$ e 41

→ If yes, which was age range?

Range: $36 \pm 3 \cdot 7 = (15, 57)$

→ A half of patients was younger than..

35 years

→ One out of four was younger than...

29 years

→ One out of four was older than...

41 years

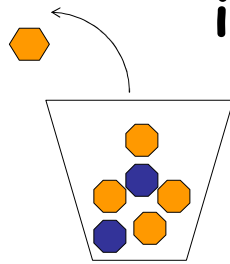
→ Was the age highly variable?

no: the Standard Deviation was 20% of mean (coefficient of variation= $7/36 \cdot 100 = 19.4$)

Variables hierarchy (2)

Variable	Possible synthesis
Not ordered qualitative	Mode
Ordered qualitative	Mode Median
Quantitative	If in classes: Modal Class and Median Class Mean Median (and other quantiles) Standard Deviation and Coefficient of Variation

Elements of probability theory and their application in medicine

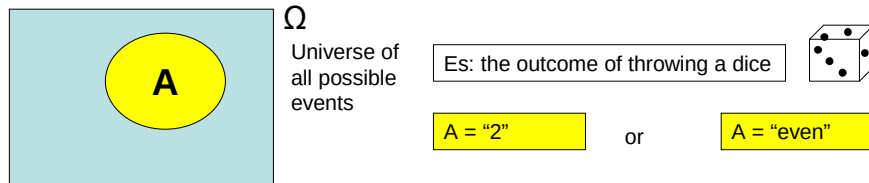


- Events and Probability: basic rules
 - Concept of conditioning and of probabilistic (in)dependence
 - Bayes rule
- Main probability distributions
 - Binomial, Poisson, Normal
- Application of prob. to diagnostic tests

Events and Probability

- “Event” and “probability” are intuitive concepts.
- An event is a circumstance that can occur or not, depending on chance and/or on a multiplicity of conditions.
- Its probability represents the expectation that the event will occur, and it is usually expressed in percent (or, by a number between 0 and 1)
- We can consider an event as the outcome of a trial – such as drawing lots; we can consider the probability as a bet on the occurrence of the event.
- No limitation to future time (i.e. we can bet on events occurred in the past provided there is still uncertainty about their occurrence).
 - Ex. of objects of interest in biomedicine: having an infection during hospitalization, having a chromosome abnormality, having been exposed to a pollutant, having less than 1,000 WBC etc.
- In the scientific context there are several different “philosophical” settings where the probability is defined more precisely; mathematics offers a common framework – we limit our course to the intuition and the basic mathematics of the probability theory.

Events - Sets - Probability



The probability of the event A is a number:

$$0 < p(A) < 1$$

Ω is the certain event: $p(\Omega)=1$

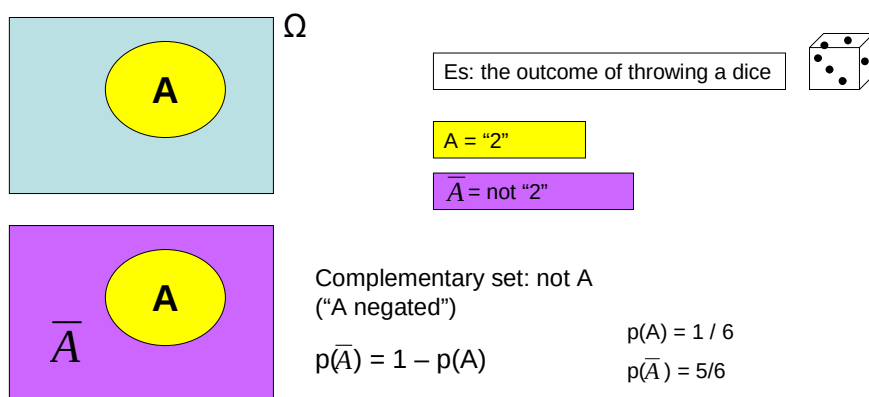
Respectively:

In simple situations where the trial has a finite number of possible outcomes and all are equally probable $p(A) = \text{number of favorable cases} / \text{number of possible cases}$.

$$p(A) = 1 / 6$$

$$p(A) = 3 / 6 = 1 / 2$$

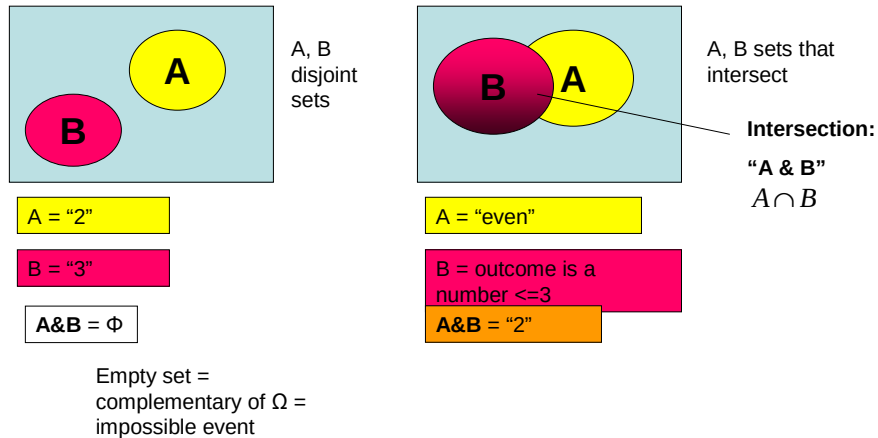
Complementary Event



The event complement to A is simply the event that includes all cases when A is not true, when A does not occur

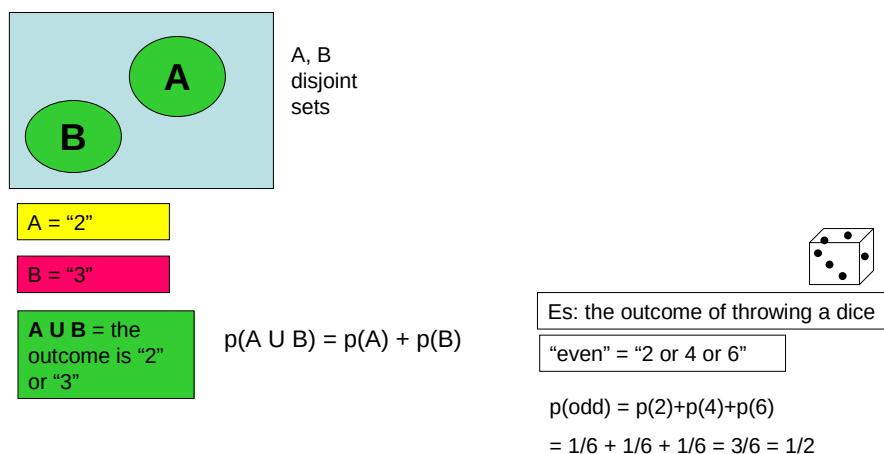
$$\begin{aligned} p(\text{infection}) &= 0.7 \\ p(\text{no infection}) &= 0.3 \\ P(\text{dead}) &= 0.25 \\ P(\text{alive}) &= 0.75 \end{aligned}$$

Intersection $\rightarrow$ "and" (&)

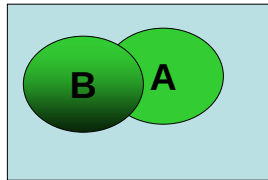


The intersection of two event A and B includes all cases when both A and B are true, they both occur; it can be empty, when this is "impossible"

Union $\rightarrow$ "Or" (either ... or) (1)



Union $\rightarrow$ "Or" (either ... or) (2) "or both"



A, B sets that intersect

The union of A and B includes all cases when either A or B or both are true ("both" = the intersection: the intersection is included in the union)

A = "even"

B = outcome is a number ≤ 3

A & B = "2"

A $\cup$ B = outcome is "1" or "2" or "3" or "4" or "6"

$$p(A \cup B) = p(A) + p(B) - p(A \& B)$$

Notice: this general case includes the previous one with disjoint events.

Also the formula is a more general version which returns the previous formula when events are disjoint, since $p(A \& B) = 0$

Conditional probability

The probability of an event can change depending on the information we have

Es: $p("2") = 1/6$

But: if I know that the outcome is "even" $\rightarrow$ prob. That outcome is "2" increases to $1/3$: $p("2" | \text{"even"}) = 1/3$

We thus introduce the concept of conditional probability:

$p(A|B)$ = prob. of A conditional on B

- > "given B"
- > "if B is true B"
- > "knowing that B is true"
- > "restricting to the cases when B is true"

Es: In the general population the probability of death due to stroke is 5%; among obese it is 10%.

$p(\text{"death due to stroke"}) = 0.05$

$p(\text{"death due to stroke"} | \text{"obese"}) = 0.10$

Dependent and independent events

When the probability of an event DOES NOT CHANGE given a conditioning event, they are called **independent**

$$p(A|B) = p(A)$$

- The expectation about A does not change knowing that B occurs
- The conditioning on B does not have any effect on “betting” on event A
- Notice: this also implies that B is independent of A i.e. $P(B|A)=P(B)$.
- This also implies the independence of any couple of events among A, B, not-A and not-B (not-A i.e. the complementary of A)

Similarly, A and B are called **dependent** if:

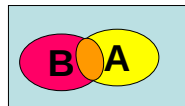
$$p(A|B) \neq p(A)$$

Formulas for conditional probability and intersections

$$p(A | B) = \frac{p(A \cap B)}{p(B)}$$

Prob. of A conditional on B

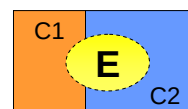
$$\begin{aligned} p(A \cap B) &= p(A | B) \cdot p(B) \\ &= p(B | A) \cdot p(A) \end{aligned}$$



$$p(A \cap B) = p(A) \cdot p(B)$$

Specific case for independent events A and B

$$p(C1 | E) = \frac{p(E | C1) \cdot p(C1)}{p(E | C1) \cdot p(C1) + p(E | C2) \cdot p(C2)}$$



Bayes formula: to compute posterior probability of cause C1 given evidence E

Exercise: using probabilities to answer to a clinical question

In a certain population:

- 20% of deliveries has complications;
- Half of them needs a cesarean;
- In general, 30% of deliveries are cesarean.

What is the probability of having a cesarean delivery or complications?

Events: A = complications; B = cesarean

Information: $p(A)=0.2$ $p(B) = 0.3$ $P(B|A) = 0.5$

Question: $P(A \cup B)$

$$P(A \cup B) = p(A) + p(B) - p(A \cap B)$$

$$p(A \cap B) = p(B|A) \cdot p(A) = 0.5 \cdot 0.2 = 0.1 \text{ (prob. cesarean with complications)}$$

$$P(A \cup B) = 0.2 + 0.3 - 0.1 = 0.4$$

Using Bayes: framework $\approx$ diagnosis

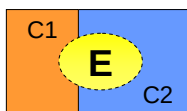
A typical problem in medicine is making a diagnosis based on the observation:

- of symptoms or other possible consequences of a condition (disease)
- of the result of a diagnostic procedure

The more general framework is: there is an EVIDENCE observed, and there are 2 (or more) possible CAUSES or hypotheses that can explain the evidence.

For cause C1 we want to know its probability GIVEN that E was observed:

$$P(C1 | E)$$



We know:

$P(C1)$: the prob of cause C1 before observing any evidence

$P(E|C1)$: the prob of event E if cause=C1

$P(E|C2)$: the prob of event E if cause= not C1

Updating prior probabilities based on observation

WHAT IS THE PROBABILITY THAT IT IS A STROKE??

A patient goes to his GP having pain in his left arm, he is afraid it is a stroke.

The GP thinks that:

- IF it is a stroke, patients have pain in 80% of the cases
- IF instead it is an inflammation, patients have this kind of pain only 30% of the times

On the other hand:

- This man is young, fit, practices sport: the probability that he gets a stroke is only 5%
- Instead given his sport practice the prob. of inflammation is 95%

E = pain

C1 = stroke

C2 = inflammation

$$p(E|C1) = 0.8$$

$$p(E|C2) = 0.3$$

LIKELIHOOD of hypotheses C1 e C2 given E

$$p(C1) = 0.05$$

$$p(C2) = 0.95$$

Prior prob. of hypotheses C1 and C2

The Bayes formula

WHAT IS THE PROBABILITY THAT IT IS A STROKE??

In order to make a diagnosis, the GP needs to appraise all elements and consider the probability of each possible cause GIVEN the observation of the symptoms (the "evidence").

Clinical question: $P(C1|E)$. Bayes formula:

$$= \frac{p(E|C1) \cdot p(C1)}{p(E|C1) \cdot p(C1) + p(E|C2) \cdot p(C2)}$$

$$= \frac{0.8 \cdot 0.05}{0.8 \cdot 0.05 + 0.3 \cdot 0.95} = 0.12$$

$p(C1|E)$: Posterior prob. of the hypothesis C1

E = pain

C1 = stroke

C2 = inflammation

$$p(E|C1) = 0.8$$

$$p(E|C2) = 0.3$$

LIKELIHOOD of hypotheses C1 e C2 given E

$$p(C1) = 0.05$$

$$p(C2) = 0.95$$

Prior prob. Of hypotheses C1 and C2

The diagnostic tests

- A diagnostic test is a tool for detecting the presence of a certain condition (e.g. a genetic abnormality or simply a disease) which is used in clinics and in prevention. A “positive” test indicates the presence of that condition (“disease”).
- Most diagnostic tests don’t provide correct results with certainty: the disease is not always detected when present, on the contrary some subjects are erroneously indicated as having the disease.
 - Thus we have respectively **FALSE NEGATIVE and FALSE POSITIVE** outcomes
 - Nonetheless the test are useful when the method for correct detection of the disease is too expensive – invasive – dangerous - etc

The features of a diagnostic test are expressed by two parameters:

SENSITIVITY: the capacity of detecting disease when present

SPECIFICITY: the capacity of recognizing subjects without that particular disease

Issues about the diagnostic test

Example: it is known that 10% of people of a certain risk group has the HIV virus (for simplicity, let’s say “is diseased”).

Suppose a person of this group has a diagnostic test for HIV; the test has sensitivity = 90% and specificity = 80%.

- The questions of the operator are:
 - For how many diseased people will we miss to detect disease?
 - How many sane people will be erroneously given further tests?
 - How many diagnostic errors will we do in total?
- The questions of the patients are for example:
 - Does the test give a certain result?
 - If I am sick will I have a positive result?
 - If I get a positive result, does this mean I have HIV?

The probability in the diagnostic test (1)

		Result of the diagnostic test	
Status of the subject (unknown)		Test +	Test –
	Diseased	ok	False negative
	Not Diseased	False positive	ok

The features of the diagnostic test are expressed by two parameters:

SENSITIVITY: the capacity of detecting disease when present (giving positive result)

SPECIFICITY: the capacity of recognizing subjects without disease (giving negative result)

SENSITIVITY: $p(\text{Test +} \mid \text{Diseased})$

SPECIFICITY: $p(\text{Test –} \mid \text{Not Diseased})$

Errors: Prob. of errors given the status

False negative $\rightarrow p(\text{Test –} \mid \text{Diseased})$

$= 1 - \text{SENSITIVITY}$

False positive $\rightarrow p(\text{Test +} \mid \text{Not Diseased})$

$= 1 - \text{SPECIFICITY}$

The probability in the diagnostic test (2)

		Result of the diagnostic test	
Status of the subject (unknown)		Test +	Test –
	Diseased	ok	False negative
	Not Diseased	False positive	ok

If the test is positive, is the subject diseased?

The result ("positive") is known, the cause ("Diseased") is unknown. We need to compute a "posterior" prob. $\rightarrow$ Bayes formula

$$\begin{aligned}
 p(D \mid T+) &= \frac{p(T+ \mid D) \cdot p(D)}{p(T+ \mid D) \cdot p(D) + \underbrace{p(T+ \mid \text{not } D)}_{1 - p(T- \mid \text{not } D)} \cdot \underbrace{p(\text{not } D)}_{1 - p(D)}} \\
 &= \frac{\text{sens} \cdot \text{prev}}{\text{sens} \cdot \text{prev} + (1 - \text{spec}) \cdot (1 - \text{prev})}
 \end{aligned}$$

Predictive values of the test:

$p(\text{Diseased} \mid \text{Test +})$

$p(\text{Not Diseased} \mid \text{Test –})$

we need to know the "prior", unconditional prob. of having the disease [PREVALENCE]

$P(\text{Disease})$

Example: diagnostic test

It is estimated that a disease affects 1 person out of 50. Detecting this disease is invasive. A test based on a blood sample can detect diseased people with sensitivity = 70% and specificity = 90%. We want to compute the probability that a person with positive Test is diseased.

Events: PT= positive test ; D = disease

Information: $p(\text{PT}|\text{D})=0.7$ $p(\text{not PT}|\text{not D}) = 0.9$ $P(\text{D}) = 1/50=0.02$

Clinical question: $P(\text{D} | \text{PT})$

We apply the Bayes formula:

$$\begin{aligned}
 & \frac{p(\text{PT} | \text{D}) \cdot p(\text{D})}{p(\text{PT} | \text{D}) \cdot p(\text{D}) + \underbrace{p(\text{PT} | \text{not D})}_{1 - p(\text{not PT} | \text{not D})} \cdot \underbrace{p(\text{not D})}_{1 - p(\text{D})}} \\
 &= \frac{0.7 \cdot 0.02}{0.7 \cdot 0.02 + 0.1 \cdot 0.98} = 0.125
 \end{aligned}$$

Exercise: the total probability of error of a diagnostic test

Develop a formula for the total probability of error of a diagnostic test, then apply it to compute how many errors you will make when a test with sensitivity = 70% and specificity = 90% is given to 40 people among whom the prevalence of the disease is 10%.

Hint: Error is: False Negative for the Diseased, False Positive for the Non-diseased

$$\begin{aligned}
 \text{Pr(Error)} &= \text{Pr(Error \& Diseased)} + \text{Pr(Error \& Not Diseased)} = \\
 &= \text{Pr(Error} | \text{Diseased)} P(\text{Diseased}) + \text{Pr(Error} | \text{Not Diseased)} \text{Pr(Not Diseased)} = \\
 &= \text{Pr(Test -} | \text{Diseased)} P(\text{Diseased}) + \text{Pr(Test +} | \text{Not Diseased)} (1 - \text{Pr(Diseased)}) = \\
 &= (1 - \text{sensitivity}) \text{Pr(Diseased)} + (1 - \text{specificity}) (1 - \text{Pr(Diseased)})
 \end{aligned}$$

Application:

$$\text{Pr(Error)} = (1 - 0.7) \cdot 0.1 + (1 - 0.9) \cdot (1 - 0.1) = 0.093$$

$$\text{Expected number of Errors} = 0.093 \cdot 40 = 3.72$$

Estimation of Sensitivity and Specificity

To estimate sensitivity and specificity of a new diagnostic technique alternative of another with certain results but more invasive / expensive, we consider a sample of subjects, that we could classify into "diseased" and "not diseased" using the old technique. The subjects are given the new test, results are:

	Test +	Test –	
Diseased	tp	fn	d
Not Diseased	fp	tn	n-d
	tp+fp	fn+tn	n

n subjects among whom d diseased, the other (n-d) not diseased.

Among diseased, tp who got a positive test are the "true positive" and fn who got a negative test are "false negative". etc

$$\text{sensitivity} = p(T+ | D) = \frac{\text{est. } tp}{d} = \frac{tp}{tp + fn}$$

$$\text{specificity} = p(T- | nD) = \frac{\text{est. } tn}{n-d} = \frac{tn}{tn + fp}$$

Estimation of the Positive Predictive Value of the diagnostic test

Apparently the positive predictive value $P(\text{Diseased} | \text{Test} +)$ can be estimated from the table in a straightforward way:

	Test +	Test –	
Diseased	tp	fn	d
Not Diseased	fp	tn	n-d
	tp+fp	fn+tn	n

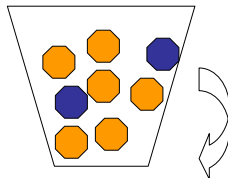
$$p(D | T+) = \frac{\text{est. } tp}{tp + fp}$$

This is in general NOT CORRECT! In fact to estimate the predictive value we need to know (an estimate of) the prevalence of the disease. This is normally NOT the one that we get from the sample, because the number of Diseased and Not Diseased is fixed by the investigators, so it does not necessarily reproduce the prevalence in the population.

The formula above is correct only IF the prevalence in the experiment ($=d/n$) is equal to the true prevalence in the population. (In this case, using the formula above or the Bayes formula will return the same result)

Random variables: random numbers

We associate a number to the event which is a result of a trial. This number is not deterministic, but varies depending on chance: it is called a *random variable*.



 $X=1$ $\Pr(X=1)=0.25=p$
 $X=0$ $\Pr(X=0)=0.75=1-p$

Ex: X Response to treatment
= 0 (no) or 1 (yes)

Ex: X white blood count

X could be any value between ...

eg 1,000 to 5,000

Urn with infinite balls and infinite colours

A random number (random variable) is described by the **PROBABILITY DISTRIBUTION**: the "list" of all possible values for X, with the probability of each possible value

Example: X counts the number of Heads when tossing a coin some times



Ex: tossing a coin 3 times and computing X = the number of outcomes = "Heads". $P(H)=0.5$. X is a random variable (r.v.)

Possible results

T	T	T	T	T	H	T	H	H	H	H	H
			T	H	T	H	T	H			
			H	T	T	H	H	T			

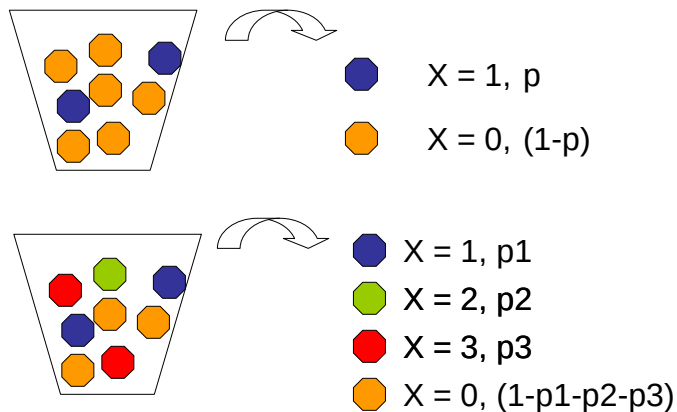
Probability distribution of the random variable X when $p(H)=0.5$

X = 0	X = 1	X = 2	X = 3
$p(X=0) = 1/8$	$p(X=1) = 3/8$	$p(X=2) = 3/8$	$p(X=3) = 1/8$

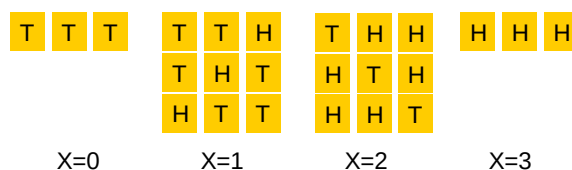
Notice: the sum of the probabilities of all possible results for X is equal to 1. Thus we need to specify 3 of the 4 probabilities to describe the whole distribution.

Discrete probability distributions

Being X a discrete random variable with K possible values, we imagine a container with bullets in K colours. We describe the container with $K-1$ numbers, representing the probabilities, or percentages of bullets of each colour.



Ex: X counting n. of Heads if $p(\text{Head})=0.3$



Random Variable: X = number of Heads in $N=3$ trials (tosses of coin)

Information: $p(H)=0.3$ $p(T)=0.7$

$$P(X=0) = P(T, T, T) = 0.7 \times 0.7 \times 0.7 = 0.343$$

$$P(X=1) = 3 \times P(T, T, H) = 3 \times 0.7 \times 0.7 \times 0.3 = 0.441$$

$$P(X=2) = 3 \times P(T, H, H) = 3 \times 0.7 \times 0.3 \times 0.3 = 0.189$$

$$P(X=3) = P(H, H, H) = 0.3 \times 0.3 \times 0.3 = 0.027$$

Notice: same situation for X =number of responders among N patients when $p(\text{Response})=p$ (here, $N=3$, $p=0.3$)

Expected value of n. of Heads if $p(H)=0.3$

T T T	T T H	T H H	H H H
	T H T	H T H	
	H T T	H H T	
X=0	X=1	X=2	X=3
P=0.343	P=0.441	P=0.189	P=0.027

Expected value:

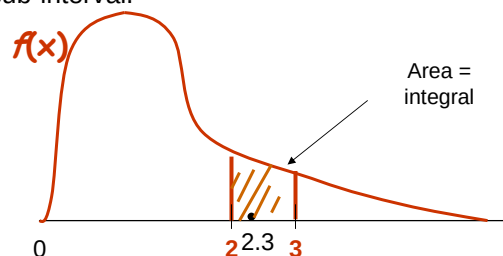
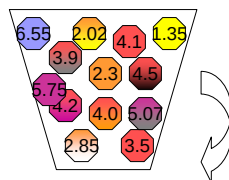
$$E(X) = \sum_i x_i \cdot p_i$$

$$= 0 \cdot 0.343 + 1 \cdot 0.441 + 2 \cdot 0.189 + 3 \cdot 0.027 = 0.9$$

Continuous probability distributions

When the r.v. X is continuous (e.g. age at onset of disease; WBC after treatment) it is like if the container includes bullets of infinite colors: how shall we describe the container?

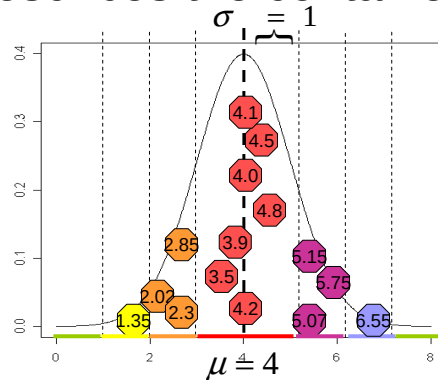
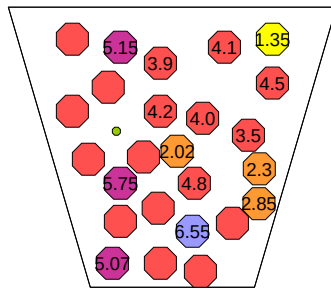
The possible values of X are all values of an interval (possibly of the axis or semi-axis of real numbers). We use a **density function** such that for each sub-interval of values, the corresponding **area under the curve returns the probability** that X belongs to that sub-interval.



$$\text{Prob}(X=2.3) = 0$$

$$\text{Prob}(X \text{ in } (2,3)) = \int_2^3 f(x) dx = \text{Area}$$

The Normal curve is a theoretical density function that describes the container



The variable X with Normal density with $\mu=4$, $\sigma=1$ has values:

- Very "dense" around 4; 68% within less than 1, higher or lower than 4 i.e. between 3 e 5
- Slightly less dense between 2 and 3 (about 14%) or between 5 and 6 (another 14%)
- Only 2% of values are between 1 and 2, or between 6 and 7
- Almost no value is <1 or >7 : overall, only 3 out of 1000

Computing probabilities for a Normal

For the "standard" Normal (mean=0, std=1) computers and tables return for each $z > 0$ the value of the area below the curve until z , representing the probability of getting one of the values on the abscissa $< z$.

Let's call the area until z as $\Phi(z)$.

For any other Normal(μ, σ), the area until x is found as the area $\Phi(z)$ of a Standard Normal until the transformed value:

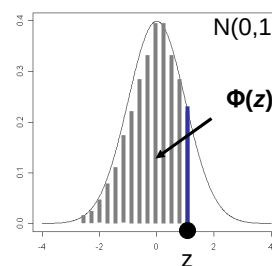
$$z = \frac{x - \mu}{\sigma} \quad (\text{Standardization})$$

To compute areas with another shape the shape has to be composed or decomposed in pieces of the shape of $\Phi(z)$, using the symmetry around the vertical axis in μ , thus:

$$\text{Area}(Z < -z) = \text{Area}(Z > z)$$

$$\Phi(-z) = 1 - \Phi(z)$$

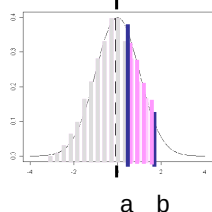
$$\Phi(0) = 0.5 \quad \Phi(+\infty) = 1$$



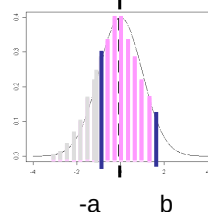
Computing probabilities for a Normal

Using tables that return $\Phi(z) = \text{Area}(-\infty, z)$ for $z > 0$:

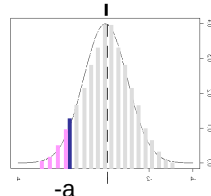
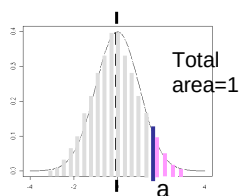
$$\text{Area}(a, b) = \Phi(b) - \Phi(a)$$



$$\text{Area}(-a, b) = \Phi(b) - (1 - \Phi(a))$$



$$\text{Area}(a, +\infty) = 1 - \Phi(a) = 1 - \Phi(a)$$



Two values of Φ to learn:

$$\begin{aligned} \Pr(Z > 1.96) &= \Pr(Z < -1.96) = 2.5\% \\ \Pr(Z > 1.64) &= \Pr(Z < -1.64) = 5\% \end{aligned}$$

Example: using the Normal distribution

In a population of teenager girls the Body Mass Index (BMI) is distributed according a Normal with mean 23 and variance 7. If we define “underweight” the girls with BMI<18, which is the probability of being underweight? How many girls will be underweight in a group of 60?

Random Variable: X = value of the BMI

Information: $\mu=23$ $\sigma^2=7$

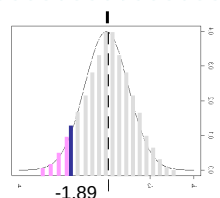
Clinical question: $P(X < 18)$

Let's standardize the value $x=18$: $z = \frac{18 - 23}{\sqrt{7}} = -1.89$

(it is negative!) $\Phi(-1.89) = 1 - \Phi(1.89)$

$= 1 - 0.971 = 0.029 \approx 3\%$. This is the probability that a girl is underweight.

Out of 60 girls, around 3% i.e. $0.029 \cdot 60 = 1.74$, thus about 2 girls will be underweight

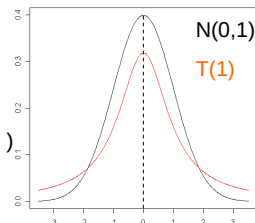


Other continuous distributions

Student T DISTRIBUTION $T(g)$

The density curve is bell-shaped, symmetric around the vertical axis placed in $\mu=0$, with tails heavier (= more dense) than the $N(0,1)$

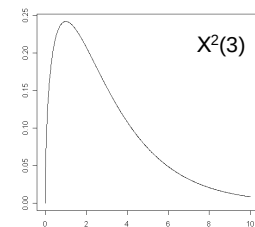
(with large parameter g (>30) it is very similar to $N(0,1)$)



Chi Squared DISTRIBUTION $\chi^2(g)$

Positive continuous variable $X \geq 0$

The shape of the density $f(x)$ is different depending on the parameter g



The parameter g is called for both the number of “degrees of freedom”. These distributions are used in statistical inference

The expected value

Given a random variable X , we call “expected value” the mean of the values that it can assume (“weighted” with their probabilities).

The meaning is equivalent to the meaning of the mean in statistics.

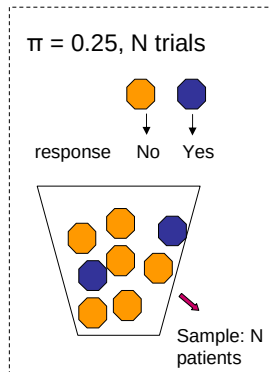
It is the value we expect “on average” to observe.

$$E(X) = \sum_i x_i \cdot p_i \quad \text{Discrete random variable}$$

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx \quad \text{Continuous random variable}$$

The Binomial probability law

The context is the repeated observation of an event with probability of occurrence ("success") equal to π in a number N of similar trials, the random variable we are interested in is the total number of successes (X).



- The expected value (mean number of successes) is $N \cdot \pi$
- The probability of observing exactly x events is:

$$p(X = x) = \binom{N}{x} \pi^x (1 - \pi)^{N-x}$$

where: $\binom{N}{k} = \frac{N!}{(N-k)! \cdot k!}$

$$k! = k \cdot (k-1) \cdot (k-2) \cdot \dots \cdot 2 \cdot 1$$

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

$$0! = 1$$

Exercise: using the Binomial for n. of Heads if $p(H)=0.3$

T T T	T T H	T H H	H H H
	T H T	H T H	
	H T T	H H T	
X=0	X=1	X=2	X=3
P=0.343	P=0.441	P=0.189	P=0.027

$X = \text{n. of Heads in 3 trials}$ follows a Binomial distribution with $N=3$ and $\pi=0.3$

$$p(X = x) = \binom{N}{x} \pi^x (1 - \pi)^{N-x}$$

$$P(X = 0) = \binom{3}{0} 0.3^0 (1 - 0.3)^{3-0} = \frac{3!}{(3-0)! \cdot 0!} \cdot 1 \cdot 0.7^3 = \frac{3!}{3! \cdot 1} \cdot 0.7^3$$

Expected value: $E(X) = n \cdot \pi = 3 \cdot 0.3 = 0.9$

Example: using the Binomial distribution

Response to treatment has probability 25%. 20 patients are treated. How many patients are expected to respond? Which is the prob. of having 10 responders?

Random Variable: X = number of responders

Information: $p(\text{response})=0.25$ number of trials: 20

Clinical questions: $E(X)$ and $P(X=10)$

We use the Binomial with $\pi=0.25$ and $N=20$

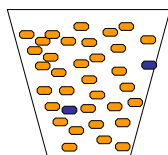
$$E(X) = N \cdot \pi = 20 \cdot 0.25 = 5$$

$$p(X = 10) = \binom{20}{10} 0.25^{10} (1-0.25)^{20-10} = 0.0099$$

The Poisson probability law

The context is the observation of a RARE event (small π) in an INFINITE series of trials or during an interval of time (or over an area, and similar). The random variable is the total number of events observed (X)

The parameter λ (called RATE) represents the expected mean number of events in the infinite trials (or in the interval, etc)



- The mean is equal to λ
- The probability of observing x events is:

$$p(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

where: $e^{-\lambda} = \frac{1}{e^{\lambda}}$

$$k! = k \cdot (k-1) \cdot (k-2) \cdot \dots \cdot 2 \cdot 1$$

$$k^0 = 1$$

The number X of events in N trials with $\pi = P(\text{Event}) \approx 0$ and $N \gg 0$ is distributed as a Poisson with $\lambda = N \cdot \pi$

Example: using the Poisson distribution

A certain disease occurs only in 1 child out of 1000. Which is the probability that in a sample of 50 children we observe at least 1 child with the disease?

Random Variable: X = number of children with the disease

Information: $p(\text{disease}) = \pi = 0.001$ number of trials: $N = 50$

Clinical question: $P(X=1 \text{ or } X=2 \text{ or } \dots) = P(X > 0) = P(\text{not } X=0) = 1 - P(X=0)$

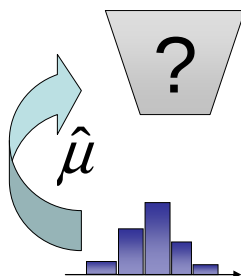
We use the Poisson with $\lambda = N \cdot \pi = 50 \cdot 0.001 = 0.05$

$P(X > 0) = 1 - P(X = 0)$

$$p(X = 0) = \frac{e^{-0.05} 0.05^0}{0!} = e^{-0.05} = 0.95$$

$$P(X > 0) = 1 - 0.95 = 0.05 = 5\%$$

Statistical inference: from the sample to the population.

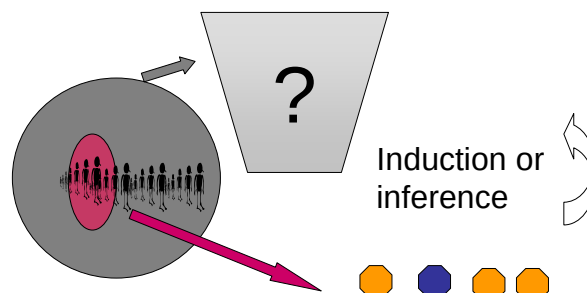


- Point and interval estimation
 - Unbiased estimators
 - Confidence Intervals (CI)
- Hypothesis testing
 - Significance
 - Formal elements
- Relationship between CI and Test
 - Remarks about significance and relevance

From the Sample to the Population

We are interested in a general phenomenon, in a target population. We use a mathematical (probability) model to represent it. In research, we ignore some elements of it: **we ignore some parameters that describe the composition of the container.**

We thus use the **DATA** from a sample to infer about the target population. They are like **bullets drawn from the container, thus they give information on the composition of the container.**



Different principles for Inference

- Statistical theory proposes different approaches to draw information on the population from the sample i.e. for making statistical inference. The main ones are:

- the frequentist approach
- the likelihood-based inference
- the Bayesian approach

- This brief course is limited to the **frequentist** approach, based on the **repeated sampling principle**. The core idea is:

- “Good rules” for inference are identified and they are used whatever the data of the sample are
- A “rule” is “good” if it has a **good behavior in the case that the procedure – drawing a sample and using that rule for inference – is repeated an infinite number of times**

Distribution of the sample mean

“Behavior”, i.e.: what are the values of the arithmetic mean of an infinite number of samples drawn from the same population

The container represents a population with mean μ and variance σ^2 .

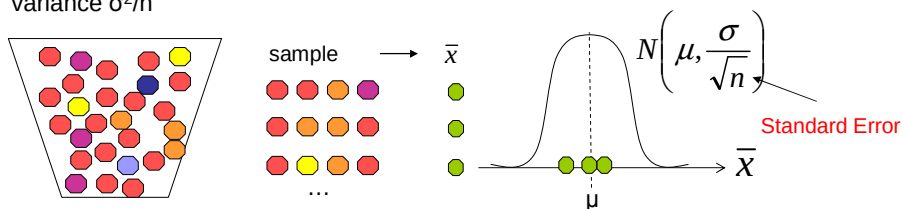
We repeatedly draw samples with size n (n observations) and we compute the average for each of them: we get a different value of the sample mean for each sample.

How are these values distributed?

Their distribution is a Normal with mean μ and variance σ^2/n

Notice: this result holds **if** at least one of these conditions hold:

- The population is Normal
- The sample size (n) is large (≥ 30)



Properties of the sample mean

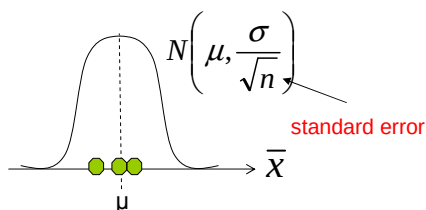
Is it useful to look at the sample mean to have information on the population mean?

Let us assess the behavior of the sample mean in repeated sampling:

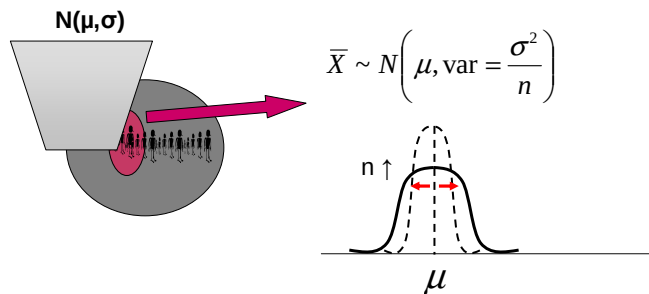
- $E(\bar{X}) = \mu$; the distribution of $\bar{X}$ is symmetric and centered on μ ; thus in repeated samples the cases when $\bar{X}$ is higher (overestimation) and the cases when $\bar{X}$ is lower than μ (underestimation) will be **balanced**.
- In 95% of the samples the sample mean belongs to the interval $\mu \pm 1.96 \cdot \sigma/\sqrt{n}$
- If the standard error is small, i.e. if the Normal is “thin”, with mass concentrated around its mean μ , **the sample mean will be close to μ** :

- This happens if the variance σ^2 of the population is small and/or if the sample size n is large:

- the larger the sample, the closer the sample mean $\bar{X}$ to the population mean μ .



Sample variation

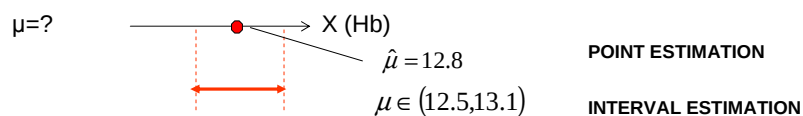


In general and not only for the sample mean, but for every sample index computed on the observed data, the variance of the sample distribution can be small (i.e. the sample information is “precise”) if:

- the variance of the population σ^2 is small $\rightarrow$ controlled experimental conditions, homogeneity of the subjects in the study (e.g. clinical trial)
- the sample size is large

Point vs. interval estimation

- Object of interest: the value (unknown) of parameter that describes the population (e.g. mean for continuous variable; prob. for $X=0,1$).
- Point estimation returns one value i.e. a single guess for the unknown parameter (“estimate”)
- Interval estimation returns an interval of possible values for the unknown parameter
 - This takes implicitly into account that the estimation procedure is affected by variability since it is based on one sample out of infinite possible samples from the population we are interested in.
 - Interval estimate is more cautious than point estimate, and thus more reliable, but point estimates are necessary in many practical situations
 - A “good” interval estimate needs to guarantee in some way that it includes the true (unknown) parameter but also to be “small”, for practical purposes



Point estimation: Estimators

In point estimation, the “rule” to apply to compute an estimate from the data that will be observed is a function of the data. Its result is a random variable. It is called **ESTIMATOR**.

The main properties of a “good rule” are:

→ **Correctness** or **unbiased** i.e. balanced i.e. not giving systematically overestimation or underestimation

→ **Efficiency**, i.e. being **precise**: the values obtained on different samples are similar to each other

→ **Consistency**: the **larger** the **sample**, the **closer** the **estimation** to the parameter

exemplification: estimator = weighing machine:

A balanced machine might give the wrong measurement in one single weigh, but repeating the procedure, on average it returns the true measure. The error is not systematic

Choosing between two weighing machines, I prefer the one that gives less extreme values and is more precise

Using the machine a larger number of times, the average weight should become more close to the true weight

Unbiased estimators

The sample mean, variance and relative frequency are unbiased estimators of the population mean, variance and probability respectively.

For a continuous variable with mean μ and variance σ^2 :

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

(If in s^2 we divide by n instead of $(n-1)$ the estimator is biased)

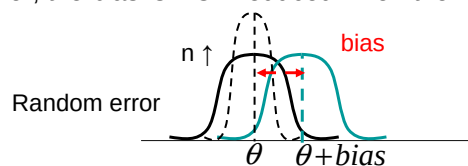
For a dichotomous variable with success ($X=1$) probability π the estimator is the relative (or percent) frequency of successes observed:

$$\hat{\pi} = f = \frac{\text{n.r cases with } (x=1)}{n}$$

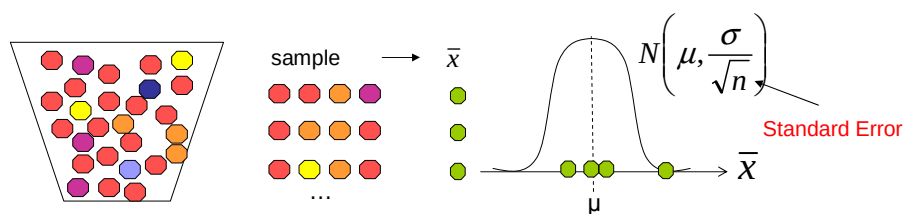
The concept of Bias

- Bias means “distortion”, i.e. SYSTEMATIC ERROR.
- The term can be used for a statistical tool, e.g. an estimator
 - Es. If we compute the variance of a sample dividing the SSE by n instead of by $n-1$, we have a biased estimator of the population variance
- The term can be used for a study (or experiment, or clinical trial)
- In the latter case there is an error in planning, e.g. in the way the sample is selected or the way variables are measured
 - Es. In an clinical trial to compare a new drug A with an old one B, we assess the incidence of side effects. It can happen that assessments are more careful in patients given the new drugs, thus side effects in A might be overestimated by design
- Differently from the random error, the bias is NOT reduced when the sample size increases

Exemplification of the random error and the systematic error in the estimation of θ , infinite repeated sampling



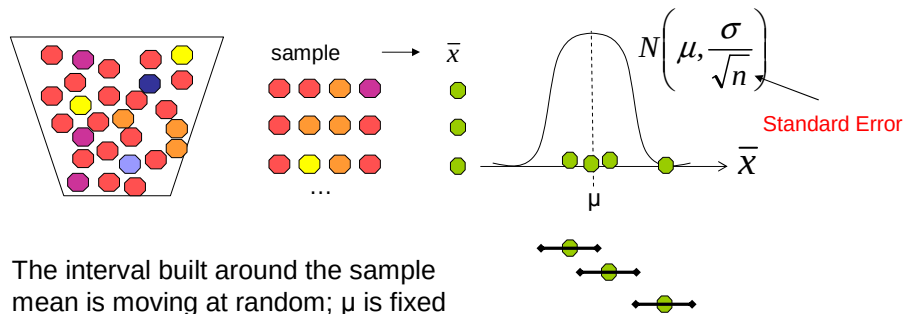
A rule for interval estimation of μ



$$\Pr\left(\underbrace{\mu - 1.96 \cdot \frac{\sigma}{\sqrt{n}}}_{\mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}} < \bar{X} < \underbrace{\mu + 1.96 \cdot \frac{\sigma}{\sqrt{n}}}_{\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu} \mid \mu\right) = 95\% \quad \text{Property of the Normal}$$

$$\Pr\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \mid \mu\right) = 95\%$$

95% Confidence interval for μ



The interval built around the sample mean is moving at random; μ is fixed (and unknown); the interval might or might not include μ

$$\Pr\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \mid \mu\right) = 95\%$$

Confidence Intervals

- A rule for computing interval estimates is “good” when it produces small and “reliable” intervals (that include the unknown parameter)
- The most common method is with “confidence intervals”. The method fixes a value $1-\alpha$ called “**confidence level**” usually equal to 0.95 or 0.90. So the full name of the estimator is “confidence interval of level $1-\alpha$ ” or “ $(1-\alpha)\%$ -Confidence Interval” (CI)
- The formula (“rule”) for computing a **confidence interval at 95% confidence level guarantees that for 95 samples out of 100 the interval will include the parameter and that the interval is the smallest with this property.**
- In general, a confidence level $1-\alpha$ means that the CI includes the unknown parameter except in $\alpha\%$ of the possible samples drawn from the population
- Thus the parameter α represents a degree of confidence in the computation rule – how much we rely on it to include the parameter. Lower α means larger confidence, at the expense of a larger interval.

Confidence Intervals at 95% level

(90%)

For a **continuous variable** with **mean μ** and **known variance σ^2** , IF:

- The variable has a Normal distribution in the population

or

- The sample is large

(the formula is more complicated if the population variance is not known – not done in our course)

$$\left(\underset{(1.64)}{\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}}, \underset{(1.64)}{\bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}} \right)$$

For a **dichotomous variable** with success ($X=1$) **probability π** , ONLY IF the sample is large:

$$\left(\underset{(1.64)}{\hat{\pi} - 1.96 \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}, \underset{(1.64)}{\hat{\pi} + 1.96 \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \right)$$

Example: 95% CI for the mean

We want to estimate the mean value of blood pressure among patients given a certain pharma therapy. We observe 130 patients, with sample mean equal to 160 and standard deviation equal to 25.

Since the sample size is large we can use our formula; we choose 95% as confidence level

$$\bar{x} = 160$$

$$s = 25, n = 130 \Rightarrow s.e.(\bar{X}) = 25 / \sqrt{130} = 2.19$$

$$z_{\frac{\alpha}{2}} = 1.96$$

$$\begin{aligned} 95\%CI &= \left(\bar{x} - 1.96 \cdot \sigma / \sqrt{n}, \bar{x} + 1.96 \cdot \sigma / \sqrt{n} \right) \\ &= (160 - 1.96 \cdot 2.19, 160 + 1.96 \cdot 2.19) \\ &= (155.7, 164.3) \end{aligned}$$

Hypothesis testing

Clinical questions can be approached in terms of statistical testing of hypotheses on unknown parameters of the population.

In a statistical test there are two hypotheses on the parameter of interest:
H0 ["null" hypothesis] : it is the reference situation, based on "neutral" knowledge

H1 [alternative hypothesis] : it is the experimental hypothesis, representing a new piece of knowledge that we wish to achieve

Examples:

- | | | | |
|----|---|---------------------|------------------------|
| 1. | Is the average value of Hb low i.e. <12 after chemotherapy? | H0: $\mu \geq 12$ | H1: $\mu < 12$ |
| 2. | Does the value of Hb depend on which type of chemo is given? | H0: $\mu_A = \mu_B$ | H1: $\mu_A \neq \mu_B$ |
| 3. | Is the probability of response higher with chemo A than with chemo B? | H0: $\pi_A = \pi_B$ | H1: $\pi_A \neq \pi_B$ |

The role of the hypotheses

In a statistical test there are two hypotheses on the parameter of interest:
H0 ["null" hypothesis] : it is the reference situation, based on "neutral" knowledge

H1 [alternative hypothesis] : it is the experimental hypothesis, representing a new piece of knowledge that we wish to achieve

The role of the two hypotheses is not the same. **THE DATA ARE USED TO TRY TO DISPROVE (REJECT) H0 IN FAVOUR OF THE ALTERNATIVE HYPOTHESIS.**

We need a strong evidence from the data in order to reject H0, which otherwise remains our reference assumption on the parameter.

The core idea of statistical test

The null hypothesis H_0 is our reference assumption for the unknown parameter. **We assume H_0 is true.**

What do we expect to observe in a sample drawn from a population when H_0 is true? **We use a probability model to say what we expect**

- 1) If what we observe is similar to what we expect, we conclude that H_0 is true.
- 2) **If what we observe is different than what we expect and it is consistent with the alternative hypothesis, we might reject H_0 in favor of the alternative hypothesis.**

Our data are the “clues” that can confirm H_0 (1) or can point out that H_0 might be false (2)

So we need to establish some criteria (or, a strict rule) to make a decision:

→ Accept H_0 (1) or Reject H_0 (2)

The core idea of statistical test

EXEMPLIFICATION

HYPOTHESES

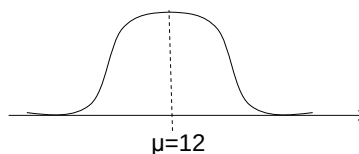
- $H_0: \mu \leq \mu_0 = 12$
- $H_1: \mu < \mu_0$

POPULATION:
 $N(\mu=12, \sigma=3,)$

We assume H_0 is true.

That's why we fix $\mu=12$.

The variance of the population is known, equal to 9

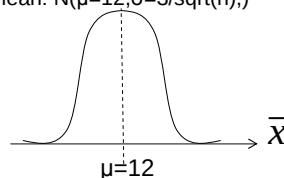


SAMPLE

Distribution of the sample mean: $N(\mu=12, \sigma=3/\sqrt{n},)$

Prob. theory tells us what to expect for the sample mean

$$\bar{X} \approx N\left(\mu, \frac{3}{\sqrt{8}}\right)$$



The core idea of statistical test

EXEMPLIFICATION

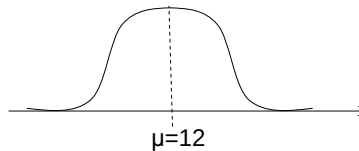
HYPOTHESES

- $H_0: \mu \leq \mu_0 = 12$
- $H_1: \mu < \mu_0$

POPULATION:
 $N(\mu=12, \sigma=3,)$

We assume H_0 is true.
That's why we fix $\mu=12$.

The variance of the
population is known,
equal to 9



SAMPLE

Distribution of the sample
mean: $N(\mu=12, \sigma=3/\sqrt{8},)$

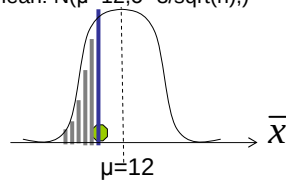
Prob. theory tells us what to
expect for the sample mean

$$\bar{X} \approx N\left(\mu, \frac{3}{\sqrt{8}}\right)$$

- Observed
sample mean

There is a high probability of
observing this value or a lower
one in the sample

ACCEPT H_0



The core idea of statistical test

EXEMPLIFICATION

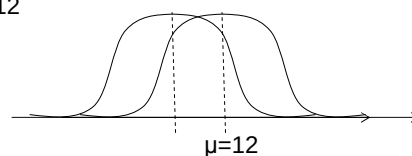
HYPOTHESES

- $H_0: \mu = \mu_0 = 12$
- $H_1: \mu < \mu_0$

POPULATION:
 $N(\mu=12, \sigma=3,)$

We assume H_0 is true.
That's why we fix $\mu=12$.

The variance of the
population is known,
equal to 9



SAMPLE

Distribution of the sample
mean: $N(\mu=12, \sigma=3/\sqrt{8},)$

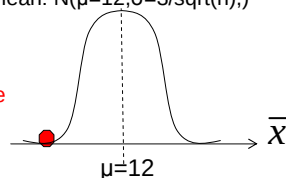
Prob. theory tells us what to
expect for the sample mean

$$\bar{X} \approx N\left(\mu, \frac{3}{\sqrt{8}}\right)$$

- Observed
sample mean

There is a very low probability of
observing this value or a lower one
in the sample; it is instead likely
under the alternative hypothesis

REJECT H_0



The core idea of statistical test

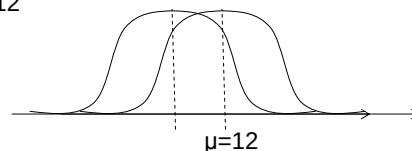
EXEMPLIFICATION

HYPOTHESES

- $H_0: \mu = \mu_0 = 12$
- $H_1: \mu < \mu_0$

POPULATION:
 $N(\mu=12, \sigma=3,)$

We assume H_0 is true.
That's why we fix $\mu=12$.
The variance of the population is known, equal to 9



SAMPLE

Distribution of the sample mean:
 $N(\mu=12, \sigma=3/\sqrt{n},)$

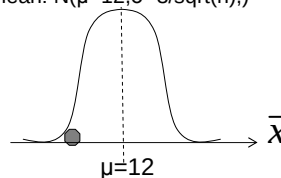
Prob. theory tells us what to expect for the sample mean

$$\bar{X} \approx N\left(\mu, \frac{3}{\sqrt{8}}\right)$$

- Observed sample mean

It is not very likely nor very unlikely that the observed value (or a lower one) comes from the null hypothesis.

?? We need a rule



Significance

P-value

The probability that under the null hypothesis we observe data (a test statistic) like the data actually observed or even more contradictory of the null hypothesis and in favor of H_1 is called **P-VALUE**.

Very small P-value (e.g. 0.001, 0.01) is a good evidence against H_0 , and we can reject H_0 in favor of H_1 . We use to say that "the data are significant".

Large P-value (e.g. 0.3 or higher) indicate a substantial absence of evidence against H_0 : we can hold H_0 . Accept H_0 , "the data are not significant".

In intermediate cases the decision needs expertise and might depend on the situation.

However often a rigid criterion is used, fixing a threshold to say what is "significant" and what is not significant. This threshold α , called "**significance level**", is often equal to **0.05** or 0.01. α is actually the probability of Type I error (next slide)

The adoption of a rigid threshold is critical: any $p < 0.05$ considered "significant" means that the information we get from $p=0.003$ and 0.049 is the same!!

Type I and II errors (1)

	Reject H_0	Accept H_0
H_0 true	Type I Error	ok
H_0 false	ok	Type II Error

Mind not to confuse this part (statistical test of hypothesis) with the diagnostic tests used in clinic!!

- Type I error is “more serious” as it means erroneously modifying correct knowledge on the population, acquiring false new knowledge.
- Type II error is “less serious” since new evidence (new data) can be acquired to disprove H_0 , while our knowledge is not modified from the experiment.
- However, also type II error is important since it means that our research failed to acquire new correct knowledge

$$p(\text{type I error}) = \Pr(\text{Reject } H_0 \mid H_0 \text{ true}) = \alpha \quad \text{significance level}$$

$$p(\text{type II error}) = \beta$$

$$p(\text{reject } H_0 \mid H_0 \text{ false}) = 1 - \beta \quad \text{power}$$

Type I and II errors (2)

	Reject H_0	Accept H_0
H_0 true	Type I Error	ok
H_0 false	ok	Type II Error

$$p(\text{type I error}) = \alpha \quad \text{significance level}$$

$$p(\text{type II error}) = \beta$$

$$p(\text{reject } H_0 \mid H_0 \text{ false}) = 1 - \beta \quad \text{power}$$

- Fixing the threshold to assess significance of the p-value equal e.g. to 0.05 corresponds to fixing the probability of Type I Error
 - This $\alpha = 0.05$ might not be appropriate depending on the circumstances. We might need to be more cautious depending on the consequences of the error. **Can we always afford to make the main mistake in 5% of the cases?**
 - Significance does not exhaust the problem. When the sample is large any test tends to give a highly significant p-value. On the contrary, in small samples we might miss significance even if H_0 is false. See “significance vs. relevance”
- Any research plan should have a sufficient power to reject a false H_0 .
 - In clinical trials usually it is requested to have 80% or 90% power
 - **Power is higher with low random variability in the experiment and using large samples**

Testing in practice (1)

Computing the P-value

The p-value is computed by statistics software, but it can also be computed using specific tables (in this course we will use tables to compute the p-value when the distribution of the test statistics is Normal)

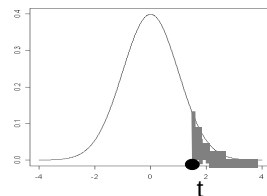
P = area under the density curve based on H_0 corresponding to the values in the tail(s) delimited by the observed test statistic.

Reject H_0 if p is small.

$H_0: \mu_A = \mu_B$ i.e. $\mu_A - \mu_B = 0$

i) $H_1: \mu_A > \mu_B$ i.e. $\mu_A - \mu_B > 0$ One-tail test ($p = \text{area}$)

ii) $H_1: \mu_A \neq \mu_B$ i.e. $\mu_A - \mu_B \neq 0$ Two-tail test ($p = 2 \times \text{area}$)



T is the test statistic, t is its observed value. The curve is the density of T if H_0 is true. The grey area is delimited at t. It is the p-value if the test is one-sided, it is p-value/2 if the test is 2-sided.

Testing in practice (2)

Inverse method: rejection region

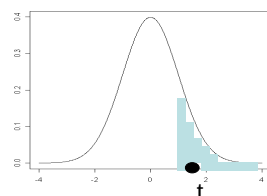
This method is equivalent to using the p-value with a rigid criterion that $p < \alpha$ leads to rejection of H_0 . It can be applied using specific tables – we will use this approach for example for Chi-Squared tests.

Rejection region = interval in the tail(s) for which the area under the density curve based on H_0 is equal to α . If the observed test statistic falls in the region, H_0 is rejected. This implies that a true H_0 is rejected with probability α .

$H_0: \mu_A = \mu_B$ i.e. $\mu_A - \mu_B = 0$

i) $H_1: \mu_A > \mu_B$ i.e. $\mu_A - \mu_B > 0$ One-tail test (rejection region = one tail with area= α)

ii) $H_1: \mu_A \neq \mu_B$ i.e. $\mu_A - \mu_B \neq 0$ Two-tail test (rejection region = two tails, each with area= $\alpha/2$)



T is the test statistic, t is its observed value. The curve is the density of T if H_0 is true. The blue area is equal to α if the test is one-sided, and equal to $\alpha/2$ if the test is 2-sided. In this example the test statistics falls into the rejection region.

Z- or T-test on one mean

$$H_0: \mu = \mu_0$$

Hypotheses: in the population the phenomenon is distributed according a Normal with mean μ (unknown) and variance σ^2 known

The formula is valid also if the population is not Normal or the variance of the population σ^2 is not known* **IF** the sample is large ($n \geq 30$)

*(replace σ^2 with the variance estimated from the sample)

Test statistic:

the sample mean, standardized:

$$t = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

Compute the p-value on the table of $N(0,1)$

Or

Fix α and compare t with the threshold(s) of the rejection region delimited by the quantile(s):

$$\begin{array}{ll} H_1 & z \\ \mu > \mu_0 (\mu < \mu_0) & z_\alpha (-z_\alpha) \\ \mu \neq \mu_0 & \pm z_{\frac{\alpha}{2}} \end{array}$$

Notice: when the variance σ^2 is not known and the sample is small the quantiles are taken from the Student T distribution with $(n-1)$ degrees of freedom

Ex: test for one mean – p-value

We know that in a population of sane subjects blood pressure is distributed as a Normal with mean 150 and standard deviation equal to 15, and we observe a sample of 20 subjects with mean equal to 160. We test the hypothesis that these subjects are drawn from that population versus the hypothesis that they have hypertension.

Hypotheses: $H_0: \mu = 150$ vs $H_1: \mu > 150$

Checking test assumptions: the population is Normal with known variance: ok.

Computing the test statistic: $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{160 - 150}{3.35} = 2.98$

Computing the p-value: $\Phi(2.98) = 0.999$ (using tables)

$$p\text{-value} = (1 - 0.999) = 0.001$$

If test 2-sided:
p=0.002

Conclusion: The observed value (160) is significantly different from the reference value 150: we can conclude that the subjects observed have hypertension.

Ex: test for one mean – rejection region

We know that in a population of sane subjects blood pressure is distributed as a Normal with mean 150 and standard deviation equal to 15, and we observe a sample of 20 subjects with mean equal to 160. We test the hypothesis that these subjects are drawn from that population versus the hypothesis that they have hypertension.

Hypotheses: $H_0: \mu = 150$ vs $H_1: \mu > 150$

Checking test assumptions: the population is Normal with known variance: ok.

Computing the test statistic: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{160 - 150}{3.35} = 2.98$

Method of rejection regions:

Let us fix $\alpha=0.05$; since the test is one-sided, the rejection region is only in the right tail, and due to a property of the Normal it is delimited by 1.64 (1.96 for a 2-sided test with level 5% or for a one-sided test with level 2.5%):

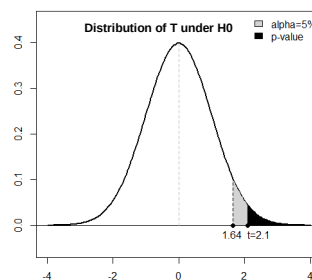
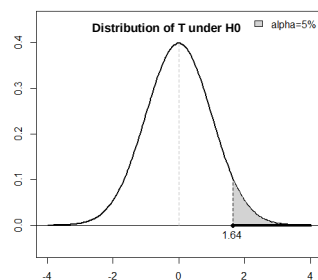
The test statistics is equal to 2.98, thus beyond 1.64, thus H_0 is rejected.

*Alpha and p-value

Example: $H_0: \mu \leq 160$ vs. $H_1: \mu > 160$. Known $\sigma=9$. Fix $\alpha=0.05$

$n=40$ $\bar{x} = 163$

$$t = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad \bar{x} = 163 \rightarrow t = 2.1$$

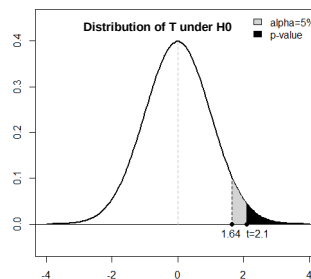
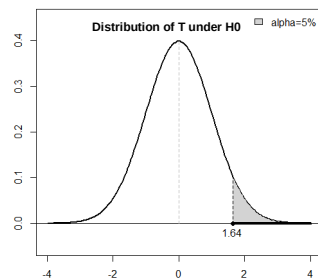


*Rejection region

Example: $H_0: \mu \leq 160$ vs. $H_1: \mu > 160$. Known $\sigma=9$. Fix $\alpha=0.05$
 $n=40$ $\bar{x} = 163$

We can also work on the original measurement scale: we apply the inverse transform to $z_\alpha=1.64$ to identify the threshold to compare to $\bar{x}$

$$z_\alpha = \frac{X^{thresh} - \mu_0}{\sigma/\sqrt{n}} \rightarrow X^{thresh} = \mu_0 + z_\alpha \cdot \sigma/\sqrt{n} = 162.3$$

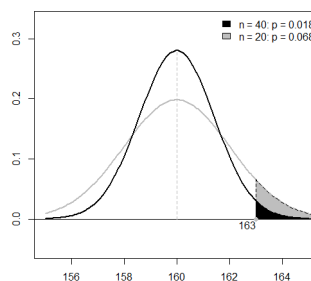
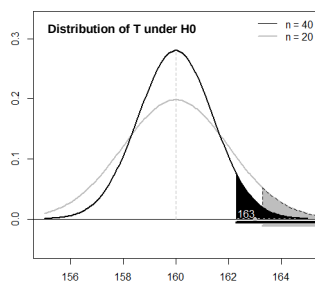


*Role of sample size

Example: $H_0: \mu \leq 160$ vs. $H_1: \mu > 160$. Known $\sigma=9$. Fix $\alpha=0.05$
 $\bar{x} = 163$

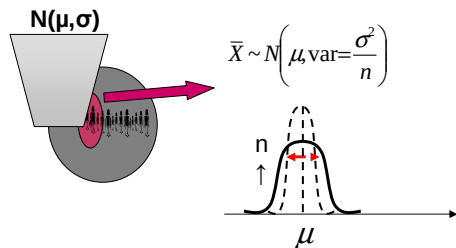
with $n = 40$: threshold for 5% significance: 162.3

with $n = 20$: threshold 163.3



Sample variation

RELOADED

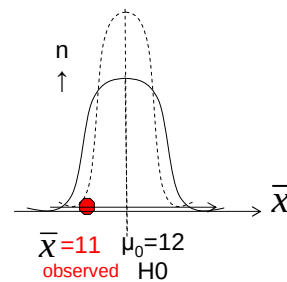


We learned that the variance of the sample distribution is small when the sample size is large.

We learned that the p-value is (2 times) the area in the tail of the curve that represents the sample distribution under H_0 , delimited by the observed value.

Keeping the observed value fixed, this area will be smaller with larger sample size. Thus relevance is the same, significance increases with n .

Significance depends on the sample size



*Test and CI

Example: $H_0: \mu = 160$ vs. $H_1: \mu \neq 160$. Fix α

$$\pm z_{\frac{\alpha}{2}} = \frac{X_{1,2}^{thresh} - \mu_0}{\sigma/\sqrt{n}} \rightarrow X_{1,2}^{thresh} = \mu_0 \pm z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}$$

Two-sided test. We reject H_0 if and only if:

$$X > \mu_0 + z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n} \quad \text{or} \quad X < \mu_0 - z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}$$

$$\bar{X} - z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n} > \mu_0 \quad \text{or} \quad \bar{X} + z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n} < \mu_0$$

We reject H_0 at α significance level if and only if μ_0 does not belong to the $(1-\alpha)$ confidence interval

Relationship between CI and test

This holds for any problem of estimation and testing of an unknown population parameter. Here explained for the mean μ .

The population mean is μ , unknown.

- Consider a 95% CI for estimating μ (more generally: a CI at $1-\alpha$ level)
- Consider a two-sided test to reject $H_0: \mu=\mu_0$ at 5% (α) significance level

- It is always true that the following conditions are equivalent:

→ μ_0 belongs to the CI $\leftrightarrow$ the test accepts H_0

→ μ_0 does not belong to the CI $\leftrightarrow$ the test rejects H_0



Thus a Confidence Interval gives ALSO an information on significance but in addition gives information on the clinical relevance of a parameter.

(1- α) % level CI and test for the prob. π

Hypothesis: the sample is large ($n \geq 30$)

Formula (1- α)% CI:

$$\left(\hat{\pi} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right)$$

Quantile: $N(0,1)$

$1-\alpha$	$z_{\frac{\alpha}{2}}$
90%	1.64
95%	1.96
99%	2.58

Test statistic for $H_0: \pi=\pi_0$: the sample proportion, standardized:

$$t = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}$$

Compute the p-value on the table of $N(0,1)$
Or:

Fix α and compare t with the threshold of the rejection region delimited by the quantile(s) from $N(0,1)$

$$\begin{array}{ll} H_1 & z \\ \pi > \pi_0 \ (\pi < \pi_0) & z_{\alpha} \ (-z_{\alpha}) \\ \pi \neq \pi_0 & \pm z_{\frac{\alpha}{2}} \end{array}$$

Ex: test for one proportion

We observe a sample of 160 myeloid acute leukemia patients; 110 (69%) get complete remission after a new chemotherapy. Can we conclude that the new chemo is more efficacious than the standard one, which gives 60% probability of complete remission?

Hypotheses: $H_0 : \pi = 0.6$ vs $H_1 : \pi > 0.6$; test di tipo T

To use the T-test we need large sample size: ok (n=160)

Computing the test statistic:

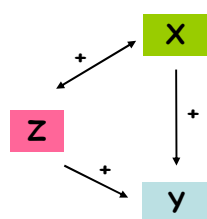
$$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{0.69 \cdot 0.31}{160}} = \sqrt{0.0013} = 0.036 \quad z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} = \frac{0.69 - 0.6}{0.036} = 2.50$$

Method of p-value: $\Phi(2.50) = 0.994$ (from the tables)

$$p\text{-value} = (1 - 0.994) = 0.006$$

Conclusion: the observed value 69% is statistically significantly higher than the reference historical value 60%: there is evidence that the new chemo is more effective than the usual one.

Studying relationships between two variables



- Association between two qualitative variables
 - Cross-tabulations and the Chi-Squared index & test
- General concepts seen in an example: relevance, significance and confounding
- Association of one quantitative and one qualitative variable
 - T-test for independent samples, other tests
- Association between two quantitative variables
 - Covariance, correlation coefficient, linear regression

Relationships in Medicine

Much useful knowledge in Medicine regards relationships between different variables

- risk factors for disease onset
 - Age→MS
 - Gender→MS
 - Smoke→Cancer
 - prognostic factors for disease outcome
 - Lymph nodes status→Relapse risk
 - Treatments for disease onset
- “Multiple sclerosis is a disease of young people, mainly women ...”
 - “Smoking causes cancer”
 - “In breast cancer, negativity of lymph nodes is associated with good prognosis”
 - “In a clinical trial for relapsed myeloma, Velcade has proven to be superior to the standard treatment based on Dexamethasone for the survival of the patients”

We need tools to describe and measure associations and interpret them

Contents

- Describe the relationship
- Assess the significance of the relationship
- Two qualitative variables
- One qualitative dichotomous variable and one quantitative continuous variable
- Two quantitative continuous variables
- Many methods available, we see only few of them
- Some notion about the rest: other test; notion of non-parametric method; notion of paired samples

Two qualitative variables: cross-tabulation (contingency table)

X = Gender	Y = smoke		total	
	no	yes		
M	24	42	66	46%
F	45	31	76	54%
total	69	73	142	

$69/142=49\%$

$73/142=51\%$

Central cells report the frequencies of the two variables' combinations

Ex. Women smokers are $22\% = 31/142$ of the total

The bottom line represents the distribution of Y, without considering X

The last column represents the distribution of X, without considering Y

← Marginal Distributions →

Conditional distributions

X = Gender	Y =smoke		total
	no	yes	
M	24	42	66
F	45	31	76
totale	69	73	142

Gender	Y = smoke		total
	no	yes	
M	$24/66 = 36\%$	$42/66 = 64\%$	
F	$45/76 = 59\%$	$31/76 = 41\%$	

The «row profile» (distribution of Y, given X), returns the distribution of Y when X is fixed equal to M / equal to F

[Notice: Conditioning = restriction]

Comparing smoke (dichotomous) in two groups

$$P(\text{smoke} \mid M) = 0.64$$

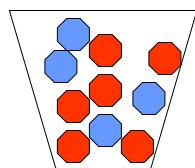
$$P(\text{smoke} \mid F) = 0.41$$

Comparison: better by the ratio: $RR = 0.64 / 0.41 = 1.56$

Gender	Y = smoke		total
	no	yes	
M	24/66 = 36%	42/66 = 64%	
F	45/76 = 59%	31/76 = 41%	

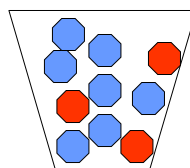
- **RR** usual tool to compare the occurrence of an event in two groups
- Several types of test to assess significance of the difference (measured either as $p_M - p_F$ or as $RR = p_M / p_F$)
- We will use only the **Chi-Squared test**

Comparing probabilities: Risk Ratio



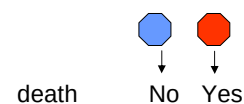
Population (container) of EXPOSED subjects, e.g. patients given surgery of type A

$$\Pr(\text{Death} \mid A) = 0.60$$



Population (container) of subjects NOT EXPOSED, e.g. patients given surgery of type B

$$\Pr(\text{Death} \mid B) = 0.30$$



The ratio of the two probabilities is usually used in the biomedical field to compare two probabilities (preferred to using the difference)

$$\text{RISK RATIO} \quad RR = \frac{\pi_E}{\pi_{NE}} = \frac{0.60}{0.30} = 2$$

Interpretation of the Risk Ratio

RISK RATIO $RR = \frac{\pi_E}{\pi_{NE}}$

= 1 : $\pi_E = \pi_{NE}$: there is no relation between Exposure and the occurrence of the event

> 1 : $\pi_E > \pi_{NE}$: the Exposure is a risk factor for the event

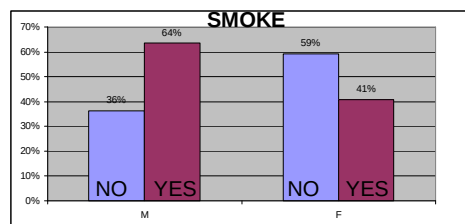
Between 0 and 1 : $\pi_E < \pi_{NE}$: the Exposure is a protective factor for the event

= 1.6 : the exposure increases the risk of event by 60%

= 2: the increase is by 100%

= 0.7: the exposure decreases the risk of event by 30% (in fact it is only 70% with respect to the risk of Non-Exposed)

Association or Independence



The distribution of smoking habits changes according to gender. This means that there's a certain **degree of statistical association** between the two variables

Gender	Y = smoke		total
	no	yes	
M	36%	64%	
F	59%	41%	
total	49%	51%	

As in probability: **independence = the probability / percentage of cases with one feature (e.g. Smokers) does not change when we condition / restrict to one subgroup or another (Male or Female)**. If Smoking and Gender were perfectly independent, we would observe 51% smokers both in M and F.

The table IF perfect independence holds

X = Gender	Y = Smoking		total
	no	yes	
M	32.07	33.93	66
F	36.93	39.07	76
total	69	73	142

The frequencies expected in the case of independence are such that there is the same proportion of smokers in each row as in the marginal row

For example the expected number E of Male Smokers would be such that:

$$\frac{E}{\text{males}(66)} = \frac{\text{smokers}(73)}{\text{all}(142)} = 51\% \Leftrightarrow E = \frac{\text{male} \cdot \text{smokers}}{\text{all}} = \frac{66 \cdot 73}{142} = 33.93$$

$$E = \frac{\text{tot row} \cdot \text{tot column}}{\text{general total}}$$

General formula of expected frequencies in each cell under the null hypothesis (H0) that the two variables X and Y are independent

The Chi-Squared index and test

The degree of association between two qualitative variables can be measured in terms of the distance between the observed frequencies and the frequencies expected in the case of independence

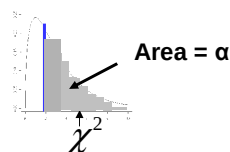
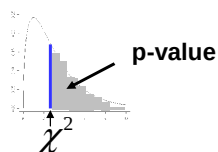
$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Under the hypothesis H0 of independence, the index χ^2 equals 0; it increases as the degree of association increases.

H₀: $\chi^2=0$: independence

H₁: $\chi^2>0$: association

We can then assess if the distance from 0 is due to pure chance or not by computing the p-value or by comparing the value of χ^2 with the threshold of a rejection region.



Computing the expected frequencies in a contingency table under the hypothesis of independence (for χ^2)

	pain			
gender	mode rate	strong	very strong	tot
M	2	5	3	10
F	5	4	3	12
tot	7	9	6	22

Observed
(absolute)
frequencies

Step 1 for the Chi-Squared index / test:
compute the expected frequency for each cell of the contingency table

	pain			
gender	mode rate	strong	very strong	tot
M	3.18	4.09	2.73	10
F	3.82	4.91	3.27	12
tot	7	9	6	22

Expected
(absolute)
frequencies

$$E = \frac{\text{tot row} \cdot \text{tot column}}{\text{general total}}$$

$$= 12 * 9 / 22$$

Computing the χ^2 index of association

	pain			
gender	mode rate	strong	very strong	tot
M	2	5	3	10
F	5	4	3	12
tot	7	9	6	22

Step 2 for the Chi-Squared index / test:
compute the distance between each observed freq. and the corresponding expected freq. using the formula:

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

0.44	0.20	0.03
0.37	0.17	0.02

$$= [(4 - 4.91)^2] / 4.91$$

	pain			
gender	mode rate	strong	very strong	tot
M	3.18	4.09	2.73	10
F	3.82	4.91	3.27	12
tot	7	9	6	22

Step 3: sum all distances

$$\chi^2 = 0.44 + 0.20 + \dots = 1.225$$

Practicality

χ^2 Test

H_0 : independence vs H_1 : association

We have computed the Chi-Squared index [test statistic] on a 2x2 table for two qualitative variables with respectively r and c modalities. To assess its significance, we use the chi-Squared distribution with g degrees of freedom, being $g = (r-1) \cdot (c-1)$.

One approach, if we have a computer, is to compute the p-value as the area in the tail, deciding if it is "sufficiently small" to reject the hypothesis of independence

The other approach is to identify a rejection region using the tables, fixing α (e.g. equal to 0.05) – and for g degrees of freedom

	pain			
gender	moderate	strong	very strong	tot
M	2	5	3	10
F	5	4	3	12
tot	7	9	6	22

$$\chi^2 = 0.44 + 0.20 + \dots = 1.225$$

In a table 2x3 $g=2$, and the threshold value for $\alpha=5\%$ is 5.99

1.225 belongs to the acceptance region

The χ^2 is non-significant: the data do not contradict the hypothesis of independence

(1.225 $\rightarrow$ P-value = 0.542)

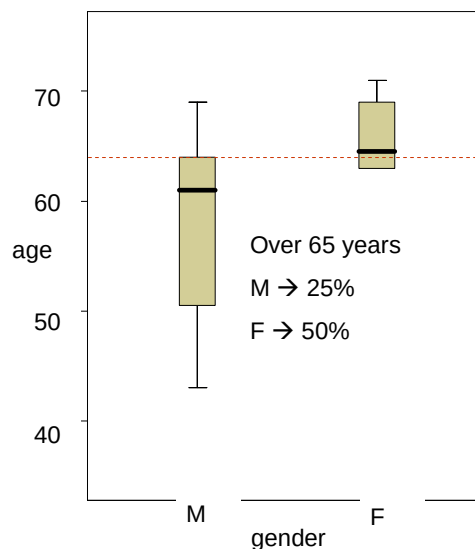
Association of a continuous and a categorical variable: description

Description is done by reporting descriptive indexes of the continuous variable Y by subgroups defined by the modalities of the categorical variable X

[= looking at the conditional distributions of Y given X]

Male:
Mean=55.7
Median=60.5

Female:
Mean=66.8
Median=64.5



Z- or T-test to compare 2 means (indep. samples)

$$H_0: \mu_1 = \mu_2 \text{ i.e. } H_0: \delta = \mu_1 - \mu_2 = 0$$

Hypotheses: the two groups come from independent populations with mean respectively μ_1 e μ_2 (unknown) and the same unknown variance σ^2 ; both samples are large ($n_1, n_2 > 30$)

$$\text{Compute: } s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Test statistic : the difference of the two sample means, standardized:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Compute the p-value on the table of $N(0,1)$
Or:

Fix α and compare t with the threshold(s) of the rejection region:

$$\begin{array}{ll} H_1 & z \\ \delta > 0 (\delta < 0) & z_\alpha (-z_\alpha) \\ \delta \neq 0 & \pm z_{\frac{\alpha}{2}} \end{array}$$

Notice: when the samples are small the quantiles are taken from the Student T distribution with $(n_1 + n_2 - 1)$ degrees of freedom

Ex: T-test to compare two means

Two groups of patients get respectively treatment A or B. We measure the value of blood pressure, assuming it is Normally distributed with the same variance in the two groups. Data are in the table. Verify the hypothesis of no difference vs. the hypothesis that blood pressure depends on treatment.

	group	
	A	B
n	32	36
$\bar{y}$	94	92
s^2	18	16

Hypotheses: $H_0: \mu_A = \mu_B$ vs $H_1: \mu_A \neq \mu_B$

Checking test assumptions: the population is Normal with equal variances: ok.

Computing the test statistic:

$$s = \sqrt{\frac{31 \cdot 18 + 35 \cdot 16}{32 + 36 - 2}} = 4.16 \quad t = \frac{94 - 92}{4.16 \sqrt{\frac{1}{32} + \frac{1}{36}}} = 1.98$$

Method of rejection regions: With $\alpha = 0.05$ and 2-sided test, we reject if $t < -1.96$ or $t > 1.96$. The test statistics is equal to 1.98, thus H_0 is rejected.

We conclude that type of Treatment (A or B) and Response are statistically associated (at 5% significance level).



Association of a continuous and a categorical variable: testing

- Observing k independent groups of statistical units, we want to verify the hypothesis that the distribution of Y in each group is the same, versus the hypothesis that at least 1 of the k groups is different. The main methods are:

	<i>Parametric</i>	<i>Non-parametric</i>
<i>$k=2$ groups</i>	T-test	Wilcoxon; Mann-Whitney
<i>$k>2$ groups</i>	ANOVA	Kruskal-Wallis

- T-test and ANOVA need that the **hypotheses of Normality and Omoschedasticity i.e. equal variances in the groups** are satisfied. This can be verified by applying specific tests or other methods:
 - Normality: graphical methods (eg. Normal Probability Plot); Test of Kolmogorov-Smirnov or Shapiro-Wilks
 - Omoschedasticity: Test F; Levene test
- Transforming the data before testing can make the assumptions valid (e.g. if Y has a right skewed distribution, $\log Y$ can have a symmetric distribution)



Parametric and non-parametric methods

- The tests called "**parametric**" are valid under certain hypotheses on the distribution of the population (e.g. that it is Normal) or if the sample size is large.
- The tests called "**non-parametric**" do not rely their validity on restrictive assumptions on the distribution. They are to be preferred when the hypotheses of the parametric methods do not hold or cannot be assessed, and with small samples.
- However, using a parametric test under correct hypothesis gives more power than using a non-parametric test in the same situation. The non-parametric tests are in other terms more "conservative" in the sense that only a strong evidence against H_0 leads to its rejection, even if H_0 is false.

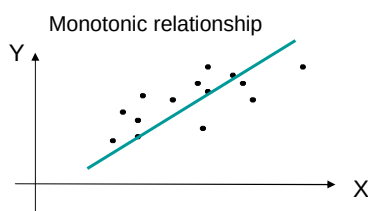


Comparing: paired or independent groups

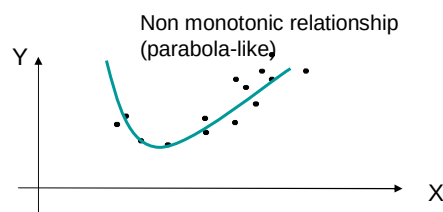
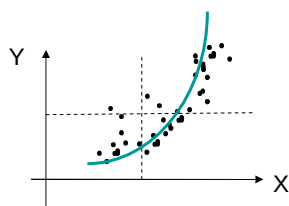
- Two groups are **independent** when the statistical units of the first group are different from the units of the second and neither related to them by sharing common features; the two groups can have different size
- Two groups are **dependent or paired** when they include the same units, or they include units that are dependent or share relevant factors, eg:
 - Measuring the weight (or WBC; or pressure; ...) before and after therapy
 - Comparing groups of siblings (they share genetic / environmental factors)
 - Comparing organs or parts of organs of the same patient (eg the eyes, two pieces of skin, etc)
 - Observing outcomes in experimental studies with cross-over, or in studies with individual matching (see Epidemiology)
- There exist specific tests for paired samples

Association between two continuous variables

- Two continuous variables show a degree of association if in the **scatter-plot it is possible to detect a trend for being distributed according to a “regular” relationship, that could be represented by a mathematical function: $Y=f(X)$**



The **line** is the simplest functional relationship to represent the association between X and Y



The line

A line is a curve characterized by having a constant slope: when the abscissa is increased by 1 unit, the ordinate is increased (or decreased, for descending line) by a fixed quantity, indicated by a number b (slope. b =tangent of the angle created with the horizontal axis)

The value of the ordinate when $x=0$ is indicated by a (intercept).

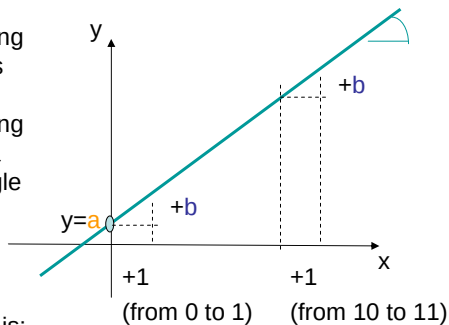
The equation that describes the curve is:

$$y = a + b \cdot x$$

$b > 0$: increasing line

$b < 0$: decreasing line

$b = 0$: horizontal line parallel to the abscissa



x	y	Δx	Δy
0	a		
1	a+b	1	b
10	a+10b		
11	a+11b	1	b

The line

A line is a curve characterized by having a constant slope: when the abscissa is increased by 1 unit, the ordinate is increased (or decreased, for descending line) by a fixed quantity, indicated by a number b (slope. b =tangent of the angle created with the horizontal axis)

The value of the ordinate when $x=0$ is indicated by a (intercept).

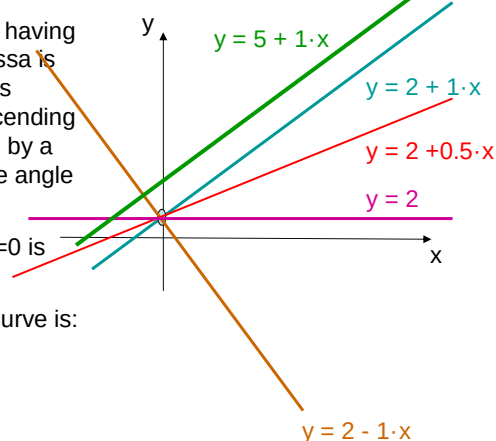
The equation that describes the curve is:

$$y = a + b \cdot x$$

$b > 0$: increasing line

$b < 0$: decreasing line

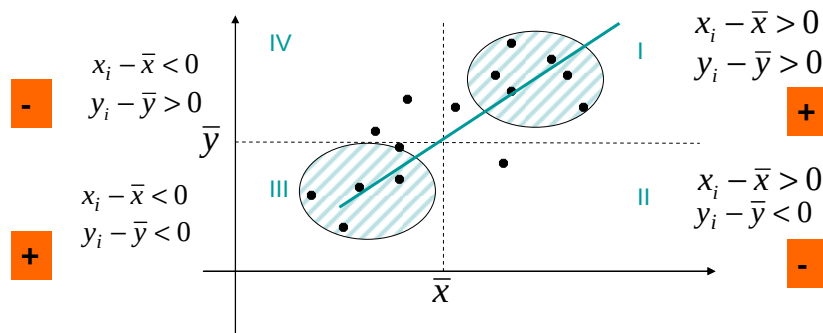
$b = 0$: horizontal line parallel to the abscissa



Measuring linear association (1)

An overall measure is the covariance: $\text{cov}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$

Divide the plane in 4 sectors by drawing lines in the two means of X and Y. When the couples (x,y) are distributed along a line, they tend to belong to sector I and III if the line is increasing, and to the sector II and IV if the line is decreasing. In the first case the products of the errors (distance from the average) are positive, in the second case they are negative.



Practicality

Computing covariance

General formula

$$\text{cov}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

1 – compute the errors for each unit

2 – compute the products

3 – sum

4 – divide by n

id	Pressure (X)	kcal/die (Y)	x-m(x)	y-m(y)	product
1	145	1850	-17.4	-66.7	1163
2	180	1960	17.6	43.3	760.74
3	150	1790	-12.4	-126.7	1576.3
4	155	1680	-7.4	-236.7	1761.9
5	148	2090	-14.4	173.3	-2504
6	160	2000	-2.4	83.3	-203.7
7	175	2100	12.6	183.3	2301.9
8	180	1900	17.6	-16.7	-292.6
9	169	1880	6.6	-36.7	-240.4
sum	1462.0	17250.0	0.0	0.0	4323.3
sum/n	162.4	1916.7	0.0	0.0	480.4

Mean of X and Y

cov_{xy}

Computing covariance faster

"Fast" formula
$$\text{COV}_{xy} = \frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \cdot \bar{y}$$

id	Pressure (X)	kcal/die (Y)	xy
1	145	1850	268250
2	180	1960	352800
3	150	1790	268500
4	155	1680	260400
5	148	2090	309320
6	160	2000	320000
7	175	2100	367500
8	180	1900	342000
9	169	1880	317720
sum	1462.0	17250.0	2806490.0
sum/n	162.4	1916.7	311832.2

Mean of X and Y

1 – compute the products for each unit

2 – sum

3 – divide by n

4 – subtract the product of the means

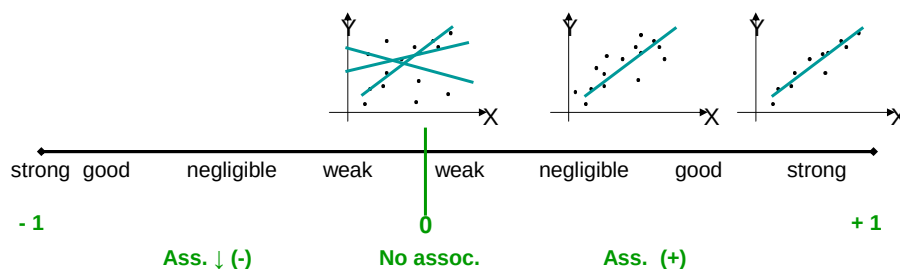
$150 \cdot 1790$

$$\text{COV}_{xy} = 311832.2 - 162.4 \cdot 1916.7 = 480.4$$

Measuring linear association (2)

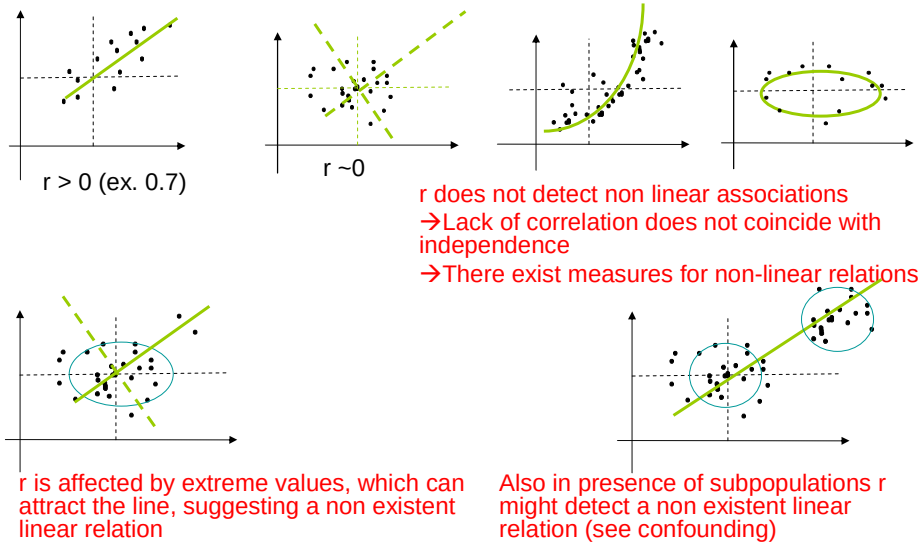
The covariance is not readily assessable as being "high" or "low". A better measure of the degree of linear relation is the **linear correlation coefficient**, which varies between -1 e +1, being 0 if there is no linear relation, and going to the extreme values in case of perfect linear relation, respectively positive (↑) or negative (↓)

$$r_{xy} = \frac{\text{COV}_{xy}}{\text{std}_x \cdot \text{std}_y}$$



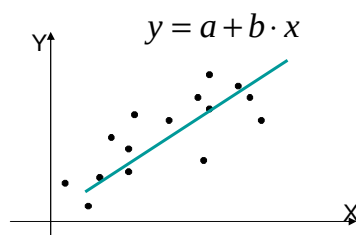
Also for r it is possible to compute a 95%CI and to perform a significance test to reject the null hypothesis of no linear association between X and Y ($H_0: r=0$)

Interpreting the correlation coefficient



Identifying the equation of the line

We can identify the equation of the line passing through the scatter-plot, thus returning a **measure of the impact of X on Y**



b represents the effect of a unit increase by X on Y

If $b=0$: line parallel to the abscissa: X does not have any effect on Y

→ The slope of the line is estimated by the **regression coefficient**, which has a direct relation to the correlation coefficient r

$$\hat{b} = r_{xy} \frac{std_y}{std_x}$$

→ The intercept of the line is estimated by:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Also for a and b it is possible to compute 95%CI and to perform significance tests. In particular, for b the null hypothesis is: no impact of X on Y ($H_0: b=0$)



Computing r and the regression line

We need the covariance and the two standard deviations. They are:

X: 13.9

Y: 137.2

$$r_{xy} = \frac{\text{cov}_{xy}}{\text{std}_x \cdot \text{std}_y} = \frac{480.4}{13.9 \cdot 137.2} = 0.283$$

id	Pressure (X)	kcal/die (Y)
1	145	1850
2	180	1960
3	150	1790
4	155	1680
5	148	2090
6	160	2000
7	175	2100
8	180	1900
9	169	1880

$$\hat{b} = r_{xy} \frac{\text{std}_y}{\text{std}_x} = 2.79$$

$$\text{Equivalent: } \hat{b} = \frac{\text{cov}_{xy}}{(\text{std}_x)^2} = \frac{\text{cov}_{xy}}{\text{var}_x}$$

finally:

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 1916.7 - 2.79 \cdot 162.4 = 1462.5$$

E.g. for pressure X=150 we expect a theoretical value of calories Y equal to $1462.5 + 2.79 \times 150 = 1881.9$

Interpretation of r, b, the regression line

Ex: We studied the relationship between X=Height(cm) and Y=FEV(l) in 20 students. Some results:

a= -9.19 b=0.074 [95%CI=(0.023,0.126); p=0.007] r=0.58 [95%CI=0.18,0.81]

- Q: Is there a relationship between Height and FEV?
- A: Yes, there is a (highly) significant linear relationship, with FEV increasing when Height increases.
- Q: What is the impact of an increase on 1 cm on FEV?
- A: FEV increases by 0.074 litres (minimum estimated increase: 0.023; maximum: 0.126)
- Q: What is the difference of FEV between one student 180cm tall and another student 181cm tall?
- A: See above: the difference is 0.074
- Q: What is the difference of FEV between one student 180cm tall and another student 195cm tall?
- A: the difference is $0.074 \cdot (195-180) = 1.11$ litres
- Q: What is the FEV of a student 160cm tall?
- A: $y = -9.19 + 0.074 \cdot 160 = 2.65$ litres
- Q: Is there a strong linear relationship?
- A: It is quite good, but not very strong ($r=0.58$). It means that on the scatterplot the points are quite scattered around the regression line (not very good "fit" of the model to the data)

Association of two continuous variables: overview

- The main methods are:

	Measuring the degree of association and related tests	
linear relation	(Covariance)	$\text{cov}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \cdot \bar{y}$
	Pearson Linear Correlation Coefficient (r)	$r_{xy} = \frac{\text{cov}_{xy}}{\text{std}_x \cdot \text{std}_y}$
	Regression Coefficient (=slope of the regression line)	$\hat{b} = r_{xy} \frac{\text{std}_y}{\text{std}_x} \quad \text{Intercept: } \hat{a} = \bar{y} - \hat{b}\bar{x}$
non-linear relation	Spearman's Rho or Kendall's Tau	

Test on the slope of a regression line

$H_0: \beta = 0$ i.e. no association between X and Y

Compute the slope b and its standard error SE(b):

$$s_{RES}^2 = \frac{\sum (y - \hat{y})^2}{n - 2} \quad SE(b) = \frac{s_{RES}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Test statistic:

the sample slope, standardized:

$$t = \frac{b}{SE(b)}$$

Compute the p-value on the table of N(0,1) [only n large]

Or

Fix α and compare t with the threshold(s) of the rejection region delimited by the quantile(s):

$$\begin{array}{ll} H_1 & z \\ \beta > 0 (\beta < 0) & z_\alpha (-z_\alpha) \\ \beta \neq 0 & \pm z_{\frac{\alpha}{2}} \end{array}$$

Notice: when the sample is small the quantiles are taken from the Student T distribution with (n-2) degrees of freedom

Some elements about the interpretation of an analysis of relationships

- We have seen some methods to describe an association between a variable X and a variable Y and assess the significance
- There are actually several issues important for a correct interpretation, and many more methods for an insight into the investigation
- We will now have a quick look to the following issues:
- CONFOUNDING (CAUSALITY)
- RELEVANCE vs. SIGNIFICANCE

Studying a relationship – an example

Suppose we have the following data on the outcome of a cardiac surgery:

	alive	dead	tot
hospital A	37	13	50
hospital B	80	32	112
	117	45	162

Is there a relationship between Hospital and Outcome?

I.e. is there an hospital with higher mortality rate?

Studying a relationship – an example

To answer, we can look at the row percentages

	alive	dead	tot
hospital A	37	13	50
hospital B	80	32	112
	117	45	162

Mortality rate in hospital A: $13/50 = 0.26 = 26\%$ RR=1.11

Mortality rate in hospital B: $32/112 = 0.29 = 29\%$

Two questions:

- Is this difference due to chance? Or “is it true” that mortality is higher in hospital B than A? → **Hypothesis test for significance (Chi-Squared)**
- Is this “causal”, i.e. hospital B “is worse” than hospital A? or is the difference attributable to some other phenomenon? → **Analysis of confounding**

Confounding in an example

	alive	dead	tot
hospital A	37	13	50
hospital B	80	32	112
	117	45	162

A: 26%

B: 29%

Young patients

	alive	dead	tot
hospital A	32	8	40
hospital B	10	2	12
	42	10	52

A: 20%

B: 17%

Overall, in
Young:

$10/52=19\%$

Elderly patients

	alive	dead	tot
hospital A	5	5	10
hospital B	70	30	100
	75	35	110

A: 50%

B: 30%

Overall, in
Elderly:

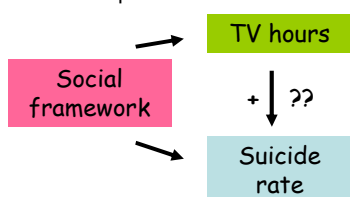
$35/110=32\%$

Confounding in an example

- The mortality seems higher in relation to the hospital ($B > A$)
- Actually, mortality depends on the Age of the patient: elderly have higher mortality (around 32%) than young patients ($\approx 19\%$)
- The point here is that hospital B treated many more elderly patients than hospital A. That's why overall the relative number of deaths was higher in B than in A
- But within the same Age group, B has lower mortality than A!
- We say that the relationship of Hospital and Outcome is confounded by Age
- We learn that **when we assess relationships between two variables X and Y, we need to control for other variables**
- There are several statistical methods to control for confounding
- Statistics can measure an association, adjusting for some confounding, and assess whether it is statistically significant or not. Still, the nature of the association (causal or not) needs to be discussed

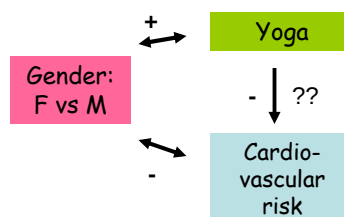
Non causal relationships between X and Y

Study on suicide incidence among children in Japan in different calendar period



➤ The third factor is a causal factor for both X and Y, but there is no causal relation between X and Y, although our statistical tool detects it

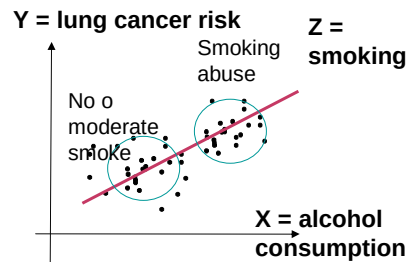
"Beneficial effect of yoga to reduce cardiovascular risk found in young adults"



➤ The third factor is (non-causally) associated to both X and Y. Again we can mis-interpret as causal the relationships between X and Y

Confounding (ex. 1)

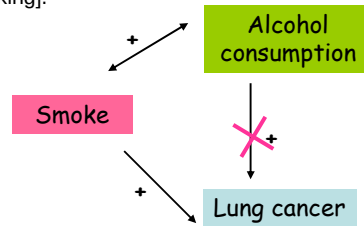
We observe an increasing relationship between alcohol consumption and lung cancer risk.



Due to social-cultural framework people who abuse alcohol tend to abuse smoking too

Smoking abuse is a proved causal risk factor of lung cancer

The relationship between Alcohol and Lung cancer is only apparent, disappearing when observed within the groups defined according to smoking [conditional on smoking].

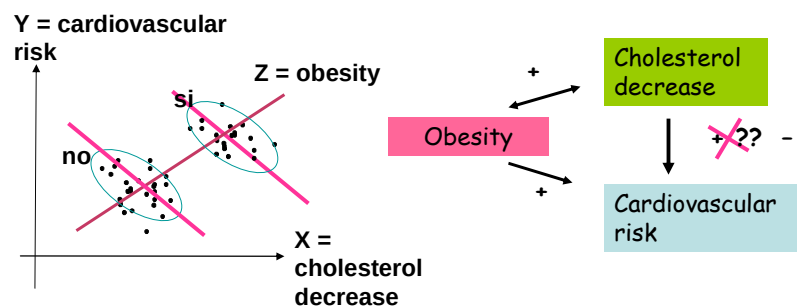


➤ The third factor Z is called “**confounder**” if it is **non-causally associated to X and causally associated to Y**

Confounding (ex. 2)

A trial aimed to demonstrate that cardiovascular risk could be decreased by decreasing cholesterol with better lifestyle. Unexpectedly, patients who had larger decrease of cholesterol had also higher risk measured at the end of the trial.

Explanation: Obese patients had higher cholesterol level at start and thus got larger decrease; but being obese they had anyway higher cardiovascular risk



The positive relationship between X and Y was only apparent, it is negative if observed within the groups [strata] of obese and not obese.

Significance, relevance and sample size

	alive	dead	tot
hospital A	37	13	50
hospital B	80	32	112
	117	45	162

A: 26%

B: 29%

RR=1.11
P=0.736

Young patients

	alive	dead	tot
hospital A	32	8	40
hospital B	10	2	12
	42	10	52

A: 20%

B: 17%

RR=0.85
P=0.797

Elderly patients

	alive	dead	tot
hospital A	5	5	10
hospital B	70	30	100
	75	35	110

A: 50%

B: 30%

RR=0.60
P=0.195

In this example none of the Risk Ratios is significant, which means that we cannot conclude that hospital A and B are different in terms of risk of death.

This is quite "disappointing" since the decrease of risk of death in hospital B within the Young (RR=0.85) and especially within the Elderly (RR=0.60) seem to be clinically relevant.

On the contrary, it seems natural that in the first table (comparison not controlled for age) the RR=1.11 is not-significant.

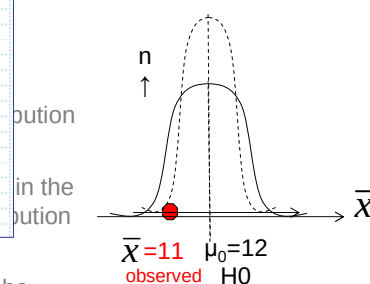
See in the next slide what happens when the sample size is increased, keeping all RR the same.

under H_0 , delimited by the observed value.

Keeping the observed value fixed, this area will be smaller with larger sample size. Thus relevance is the same, significance increases with n .

Significance depends on the sample size

RELOADED



Significance, relevance and sample size

	alive	dead	tot
hospital A	3700	1300	5000
hospital B	8000	3200	11200
	11700	4500	16200

A: 26%

B: 29%

RR=1.11
P=0.0007

Young patients

	alive	dead	tot
hospital A	3200	4000	7200
hospital B	1000	200	1200
	4200	4200	8400

A: 20%

B: 17%

RR=0.85
P<0.0001

Elderly patients

	alive	dead	tot
hospital A	500	500	1000
hospital B	7000	3000	10000
	7500	3500	11000

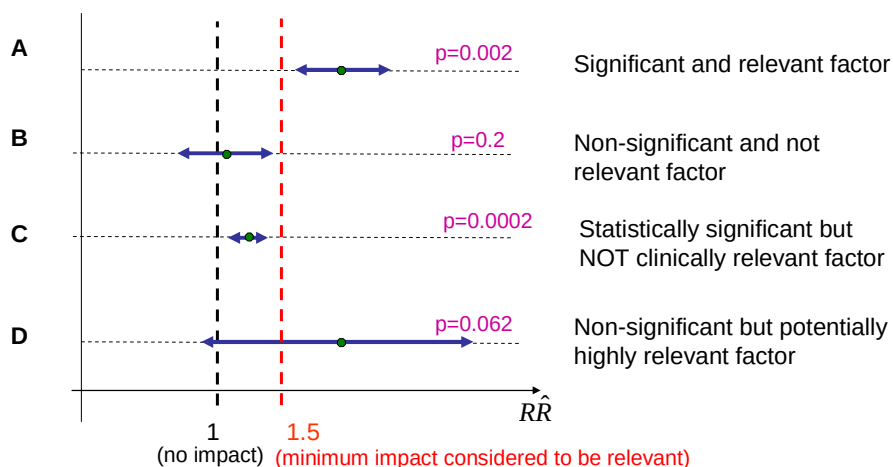
A: 50%

B: 30%

RR=0.60
P<0.0001

CI and p – Relevance vs. Significance

We have four different risk factors for a disease. For each factor we assess the impact on the risk of disease by computing the **Risk Ratio**. We compute a **95% CI** for the RR and we **test** for $H_0: RR=1$ vs. $H_1: RR \neq 1$. In this example, a risk factor is clinically relevant if it increases the risk by at least 50% (i.e. if $RR \geq 1.5$)



Final messages

- CONFOUNDING, CAUSALITY
 - Other factors might explain the (apparent) relationship between X and Y, or anyhow might help identifying the true effect of X on Y
 - Finding a statistical association between X and Y does not imply that there is a causal relationship between X and Y
- RELEVANCE vs. SIGNIFICANCE
 - Significance is important, however it strongly depends on the sample size
 - “statistically significant” does not mean “relevant, important on a clinical level”
 - The relevance of an association is better appraised by looking at a confidence interval for its measurement

Appendix

Mathematics Prerequisites



- (Very) basic notions:
 - Rounding
 - Basic computing
 - Sum Σ
 - The logarithm function

Rules for rounding

- Rounding a number means reducing the figures after the decimal separator.

Original value	→1 decimal	→2 decimals
12.422	12.4	12.42
11.237	11.2	11.24
10.251	10.2	10.25
10.257	10.3	10.26
14.0	14.0	14.00

If the decimal number after the position we want to round to is:

<5 → truncate to floor

>5 → round to the next number (floor + 1)

=5 → look one number further until a number ≠5 is found and apply the same rules

Basic notions of mathematics ...

$$\overbrace{a + a + \dots + a}^{k \text{ times}} = k \cdot a$$

$$\overbrace{a \cdot a \cdot \dots \cdot a}^{k \text{ times}} = a^k$$

$$k \cdot (a + b) = k \cdot a + k \cdot b$$

$$-(a + b) = -a - b$$

$$\frac{1}{k} \cdot (a + b) = \frac{1}{k} \cdot a + \frac{1}{k} \cdot b$$

$$a^2 = b \Leftrightarrow a = \sqrt{b}$$

$$\left(\frac{a}{b}\right)^2 = \frac{a^2}{b^2}$$

$$a = b + c \Leftrightarrow a - b = c \Leftrightarrow a - b - c = 0 \Leftrightarrow -a + b + c = 0$$

$$\frac{a}{b} = c \Leftrightarrow a = b \cdot c \Leftrightarrow b = \frac{a}{c}$$

$$\log_a b = x \Leftrightarrow a^x = b$$

... And a bit more

Factorial
numbers

$$k! = k \cdot (k-1) \cdot (k-2) \cdot \dots \cdot 2 \cdot 1$$

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

$$0! = 1$$

Binomial
coefficient
("N over k")

$$\binom{N}{k} = \frac{N!}{(N-k)! \cdot k!}$$

The sum Σ

$$\overbrace{a + a + \dots + a}^{k \text{ times}} = k \cdot a$$

$$a_1 + a_2 + a_3 + \dots + a_i + \dots + a_k = \sum_{i=1}^k a_i$$

Read: "sum of a with i for i from 1 to k"

$$= (a_1 + a_2 + a_3) + (a_4 + \dots + a_k) = \sum_{i=1}^3 a_i + \sum_{i=3}^k a_i$$

This is the sum of k
terms all equal

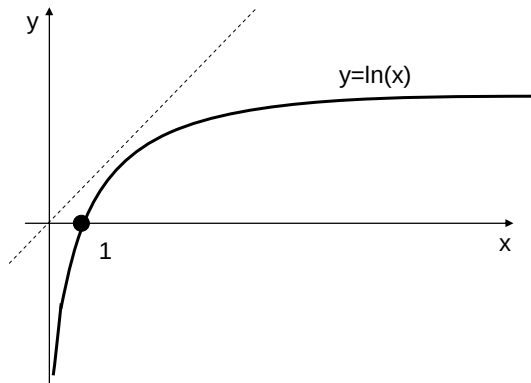
This is the sum of k
terms that can be
different from each
other: we use the
symbol Σ for the sum
and a with a subscript i
to indicate the general
addendum

The logarithm function

$$\log_a b = x \Leftrightarrow a^x = b$$

$$\ln b = \log_e b = x \Leftrightarrow e^x = b \quad (\text{where } e = 2.718\dots)$$

Natural logarithm
(the basis is Nepero's number, $e=2.718\dots$)



Transforming X to its logarithm, $Y = \ln X$, implies reducing high values

$$\ln(1) = 0$$

$$\ln(x) < 0 \text{ per } 0 < x < 1$$

$\ln(x)$ not defined for $x \leq 0$