Statistica medica

Metodologia medico scientifica e metodi quantitativi in biomedicina

Anno accademico 2020/2021

Statistica…che cosa è? Perché viene utilizzata?

È una disciplina che concerne la raccolta, l’organizzazione, la sintesi e l’analisi di dati

Nasce dalla necessità di trarre conclusioni e informazioni da un insieme di dati che costituisce solo una parte dell’intera popolazione

Statistica…che cosa è? Perché viene utilizzata?

Viene utilizzata in quelle situazioni nelle quali occorre prendere decisioni in condizioni di incertezza

Non abbiamo mai la sicurezza e la correttezza assoluta

Interviene per controllare e limitare gli errori

Fasi del progetto di ricerca

1. **Ideazione**

   Il ricercatore definisce l’obiettivo di ricerca. Deve essere chiaro, ben strutturato e preciso. Il progetto di ricerca deve essere fattibile in termini di tempo, in termini economici, di personale e di risorse.

2. **Progettazione**

   Deve essere scritto un *protocollo* dove vengono definiti tutti gli aspetti pratici dello studio

3. **Attuazione**

   Il ricercatore mette in pratica ciò che ha scritto nel protocollo di ricerca, raccogliendo informazioni e traendo poi conclusioni sul raggiungimento dell’obiettivo di ricerca

Statistica…quando interviene?

La statistica interviene in due fasi del progetto di ricerca:

- **Nella fase di progettazione dello studio** dove lo statistico collabora con il ricercatore e l'epidemiologo per identificare la popolazione di studio, il sottoinsieme da studiare e come estrarre questo sottoinsieme. Deve anche capire quanto questo sottoinsieme deve essere numeroso. Deve dare indicazioni su come raccogliere e organizzare informazioni

Statistica…quando interviene?

La statistica interviene in due fasi del progetto di ricerca:

- **Nella fase finale dell'attuazione dello studio** dove lo statistico deve sintetizzare le informazioni raccolte e trarne informazioni e conclusioni.

Statistica…perché è utile?

Sarà utile per comprendere, leggere e analizzare un articolo in modo da essere consapevoli se ciò che viene scritto è attendibile oppure no

Ci aiuterà a comprendere e studiare il fenomeno che ci interessa (ruolo attivo del ricercatore)

Ci fornirà strumenti utili per il supporto alle decisioni

Attenzione!!!

LA STATISTICA NON FA MIRACOLI!!

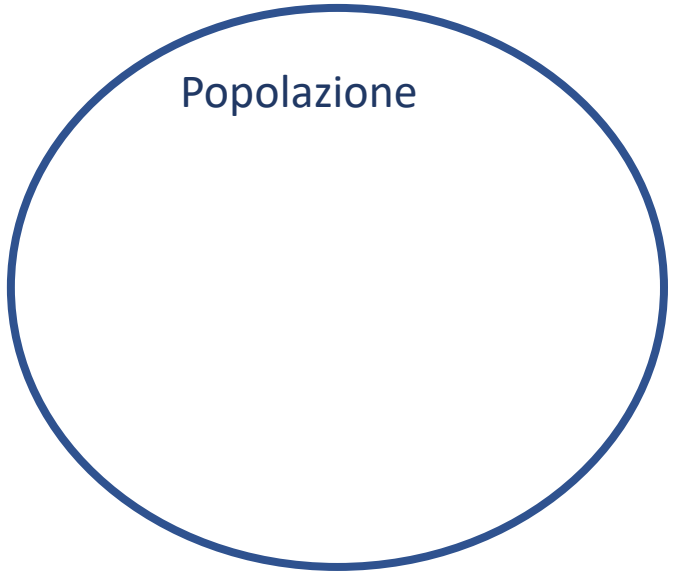
Se ogni fase della ricerca scientifica che va dalla fase di progettazione dello studio alla raccolta dati non viene eseguita con rigore e accuratamente, una bella statistica finale non darà valore a ciò che è stato fatto.

PERCIÒ VARRÁ SEMPRE IL PRINCIPIO GIGO!

GARBAGE IN → GARBAGE OUT

Popolazione e campione

Popolazione e campione



- La **popolazione** è un insieme di elementi oggetti di studio
- Può essere finita, se le unità sono un numero finito e po’ essere stilata una lista, o infinita, insieme di unità potenzialmente osservabili e non necessariamente già esistenti.

Popolazione e campione

Viene definita a partire dai **criteri di inclusione e esclusione**

Sono enunciati che ci permettono di dire chi può o non può far parte della popolazione

I principali criteri di esclusione sono per esempio: essere analfabeti, essere carcerati, .. → si potrebbero sentire obbligati a rispondere o fare certe cose.

Tutti i requisiti clinici per cercare di avere una popolazione il più omogenea possibile, se studiassimo la comparsa di aritmie dovute da un farmaco escluderei i pazienti con pregresse aritmie.

Popolazione e campione

Per quanto riguarda la popolazione è doveroso fare una distinzione tra:

- **Popolazione obiettivo/di riferimento/target:** è la popolazione che vorremmo studiare
- **Popolazione di studio:** è la popolazione che viene realmente studiata

Idealmente le due popolazioni coincidono ma a livello pratico non accade

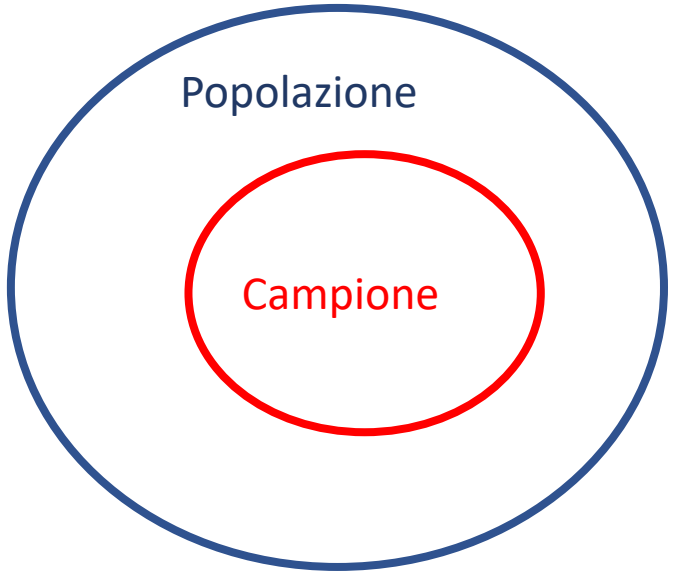
Popolazione e campione

Es. studio dell’asma:

→la popolazione obiettivo è composta da soggetti che soffrono di asma

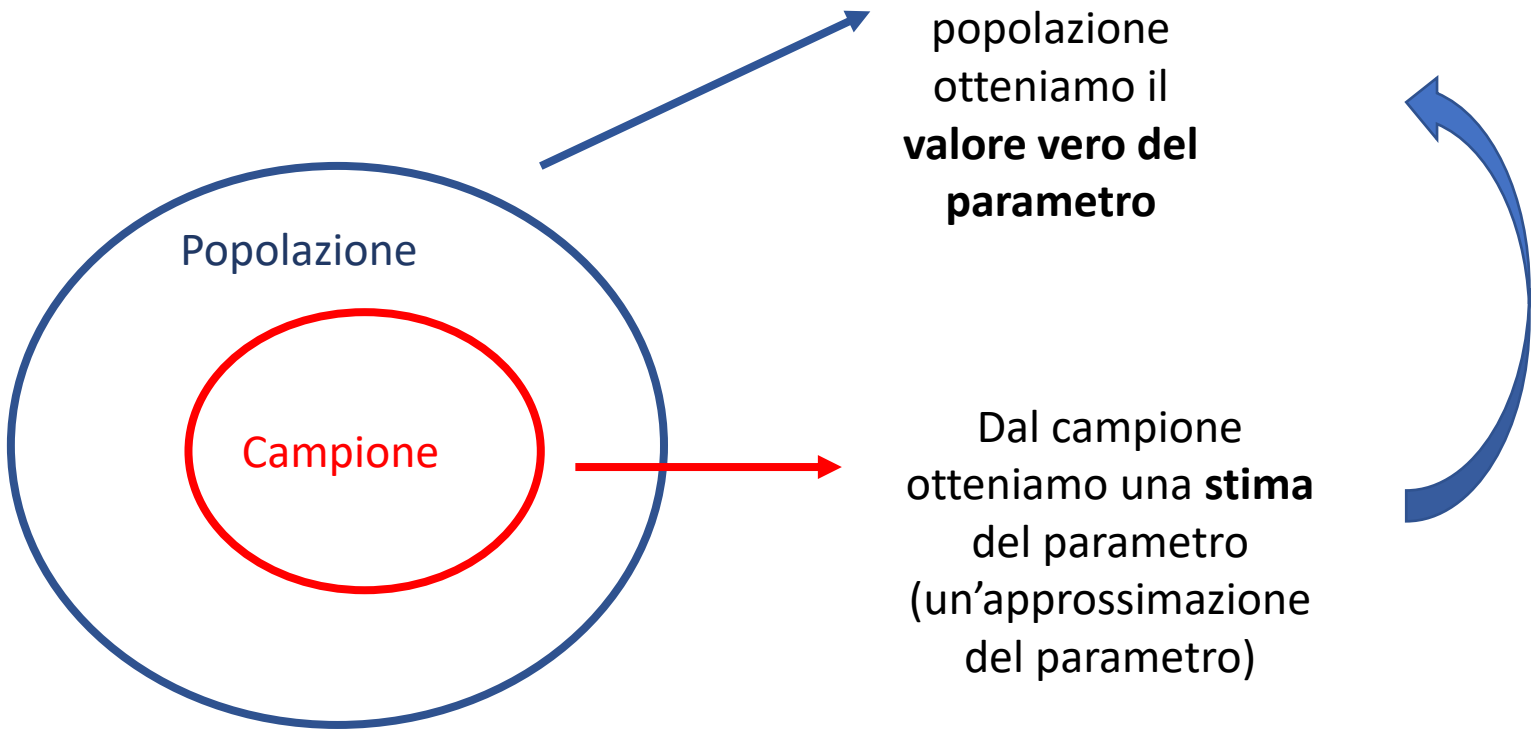
→alcuni pazienti potrebbero non sapere di soffrire di asma, perciò non vengono inclusi nella popolazione oppure non è possibile estrarre il campione da tutta la popolazione obiettivo perciò selezioneremo il campione da chi soffre di asma e fa delle visite all'ospedale di Siena

Popolazione e campione



- Il **campione** è un sottoinsieme della popolazione.
- Viene utilizzato perché non è sempre possibile fare valutazioni sulla popolazione (cosa che fa il censimento)
- La tecnica di estrazione di un campione viene detta **campionamento**

Popolazione e campione



Dalla popolazione otteniamo il **valore vero del parametro**

Dal campione otteniamo una **stima** del parametro (un'approssimazione del parametro)

La stima ottenuta sul campione può essere generalizzata alla popolazione. Questo processo si chiama **inferenza**.
Può essere fatto a patto che il campione sia **rappresentativo** della popolazione

Campionamento

Probabilistico

- Conosciamo a priori la probabilità che ogni individuo ha di entrare a far parte del campione, questa probabilità si chiama **probabilità di inclusione.**
- Se il campionamento è ben strutturato allora possiamo essere confidenti sul fatto che il campione è rappresentativo della popolazione
- Può essere fatta inferenza
- Non sono facilmente attuabili in termini economici e di tempistiche
- È necessaria la lista completa della popolazione

Ragionato/di convenienza

- Non conosciamo le probabilità di inclusione
- I campioni non sono rappresentativi della popolazione.
- Non può essere fatta inferenza. Le stime ottenute su questi campioni potrebbero essere affette da distorsioni dovute all'interessamento dell'unità statistica allo studio.

  Es. se volessimo condurre uno studio sull'efficacia di una crema antirughe utilizzando un campione di volontari, sicuramente le donne più adulte sarebbero più interessate, non parteciperebbero uomini o donne più giovani. Perciò l'efficacia del farmaco verrà testata solo in una sottopopolazione e non varrà per tutti.
- Sono facilmente attuabili in termini economici e di tempistiche
- Non è necessario conoscere la popolazione

Campionamento

Probabilistico

- Campionamento casuale semplice
- Campionamento stratificato
- Campionamento sistematico
- Campionamento a cluster

Ragionato/di convenienza

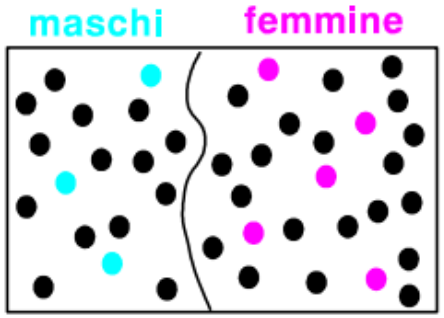
- Campionamento di volontari
- Campionamento a palla di neve
- Campionamento per quote

Campionamenti probabilistici

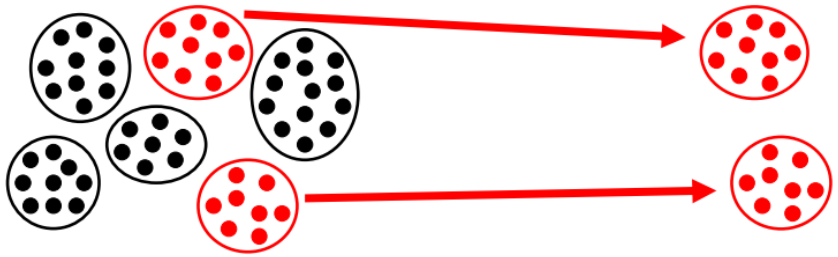
Casuale semplice

Sistematico

Stratificato proporzionale



A grappolo/a cluster



Simbologia

N = numerosità della popolazione

n = numerosità campionaria

π = probabilità di inclusione

k = passo di campionamento

h = strato

Campionamento casuale semplice

È equivalente all'estrazione random di palline da un'urna senza reinserimento

La probabilità di inclusione per ciascuna unità di analisi è $\pi = \frac{1}{N}$

Un campionamento in cui tutte le unità hanno la stessa probabilità di inclusione viene detto **autoponderante**

Dobbiamo fare attenzione ad usare la corretta terminologia!

RANDOM = selezione casuale di un’unità di analisi

RANDOMIZZAZZIONE = assegnazione casuale del soggetto ad un gruppo di trattamento o ad un gruppo di controllo

Campionamento casuale semplice

Per effettuare un campionamento casuale semplice possiamo numerare tutte le unità della popolazione, mettere in un'urna tante palline numerate, tutte uguali fra loro, quante sono le unità della popolazione e quindi sorteggiare da tale urna le palline per formare il campione.

Invece dell'urna si preferisce oggi ricorrere a una tavola di numeri casuali. Le tavole dei numeri casuali si costruivano, un tempo, con metodi empirici; attualmente si utilizzano fogli di calcolo come Excel

Campionamento casuale semplice

In medicina/pratica clinica è piuttosto difficile attuare un campionamento casuale semplice perché le popolazioni sono tendenzialmente infinite e non stabili dal punto di vista geografico e temporale.

Viene considerato assimilabile ad un campionamento casuale semplice la seguente tipologia di selezione:

«il campione è determinato da tutti i soggetti che si presentano ad una data Unità operativa di uno specifico ospedale da … a ..»

Campionamento casuale semplice

Esempio:

Il campione è costituito da tutti i pazienti che rispettano i criteri di inclusione e esclusioni e che hanno sostenuto una visita per sospetto carcinoma presso l’Unità di Endocrinologia dell’Azienda Ospedaliera Senese dal 01/01/2018 al 01/01/2020

Campionamento sistematico

Selezioniamo una unità ogni k elementi.

Il passo di campionamento K è calcolato come k=$\frac{N}{n}$

Campionamento sistematico

1. Estraiamo in modo casuale la prima unità tra i primi k elementi seguendo le procedure del campionamento casuale semplice

2. Inseriamo nel campione un’unità ogni k elementi

Campionamento sistematico

ATTENZIONE!

Esistono rischi di scarsa rappresentatività nella selezione del campione nel caso in cui la lista presenti un qualche ordinamento di tipo ciclico o stagionale che, a causa di un passo di campionamento fisso, porterebbe alla selezione di unità del tutto simili.

Campionamento sistematico

Esempio:

Ordino la lista dei soggetti inserendo un maschio e poi una femmina. La prima unità estratta è un maschio e ho un passo di campionamento pari. Il campione sarà composto da sole unità di sesso maschile.

Campionamento stratificato proporzionale

La popolazione viene suddivisa in sottopopolazioni dette **«strati»** da cui andiamo ad estrarre il campione

Da ciascuno strato si estraggono le unità attraverso un campionamento casuale semplice. Queste unità vengono riunite in un unico campione.

Questo tipo di campionamento risulta essere autoponderante

Campionamento stratificato proporzionale

- Gli strati devono essere il più omogenei possibili rispetto alla variabile di stratificazione

- Non devono avere una numerosità eccessivamente piccola e soprattutto devono contenere almeno una unità.

- Possono essere utilizzate più caratteristiche in contemporanea per stratificare. Più condizioni di stratificazione (e quindi più strati) ci sono e più al suo interno le unità saranno omogenee tuttavia diminuirà la numerosità interna dello strato e aumenterà la variabilità delle stime.

Campionamento stratificato proporzionale

Viene utilizzato per:

- Evidenziare insiemi di unità particolari
- Separare strati con caratteristiche diverse

Alcune variabili per cui di solito stratifichiamo sono il sesso, l’età, la condizione socioeconomica,..

Campionamento stratificato proporzionale

Strati h=1,…,H

Peso di ciascun strato $\boldsymbol{W_h} = \frac{\boldsymbol{N_h}}{\boldsymbol{N}}$ perciò $\sum_h W_h = 1$

Numerosità dello strato nel campione $\boldsymbol{n_h} = \boldsymbol{W_h} * \boldsymbol{n}$

Campionamento stratificato proporzionale

Da una lista di 1872 studenti iscritti ad un corso ne vengono selezionati 250 per stimare la proporzione di quelli che hanno svolto correttamente una determinata prova di esame ed il numero medio di ore giornaliere passate da ciascuno studente davanti alla televisione. Facciamo l'ipotesi che la lista sia divisa in quattro sottoliste corrispondenti ai quattro anni in cui si articola il corso; gli anni di corso rappresentano gli strati dai quali vengono selezionati quattro campioni indipendenti

Strato	(1) N_h	(2) W_h	(3) n_h
I anno	524	0,28	70
II anno	487	0,26	65
III anno	449	0,24	60
IV anno	412	0,22	55
Totali	1872	1,00	250

Campionamento a cluster

Diversamente dagli altri campionamenti non prevede l'estrazione di singole unità dalla popolazione di riferimento, ma di *grappoli/cluster*, cioè agglomerati di unità statistiche che per natura hanno una forma raggruppata.

Tutte le unità che costituiscono il grappolo estratto entrano a far parte del campione

Il maggior vantaggio che si può trarre dal campionamento a grappoli si ha quando ogni grappolo al suo interno non è omogeneo, cioè le unità che compongono il grappolo differiscono tra loro

Campionamento a cluster

Il campionamento a cluster viene spesso combinato agli altri tipi di campionamento dando forma ad un **campionamento a due o più stadi**

Un esempio di campionamento a due stadi si ha applicando un primo campionamento a cluster e successivamente il campionamento casuale per selezionare solamente alcune unità dei grappoli selezionati.

Campionamento a cluster

Degli esempio di grappoli sono i comuni, le famiglie, le scuole o le classi, le usl,..

I grappoli possono essere selezionati in modo casuale o in modo sistematico

Campionamenti a due stadi

Esempio

HBSC (Health Behaviour in School-aged Children) è un questionario per valutare lo stato di salute dei giovani (di 11, 13 e 15 anni) e il loro contesto sociale. È uno studio internazionale il cui promotore è l'OMS.

La somministrazione di questo questionario avviene attraverso un campionamento a più stadi. Attraverso un campionamento sistematico vengono estratte alcune delle scuole italiane, successivamente con un campionamento a cluster vengono estratte le classi che dovranno compilare tale questionario.

Campionamenti non probabilistici

- **Campionamento di volontari**

  Entrano a far parte del campione tutti coloro che ne sono interessati

- **Campionamento a palla di neve**

  Viene reclutato il primo soggetto e gli viene affidato il compito di arruolare altri soggetti che a loro volta possono fare la stessa cosa

- **Campionamento per quote**

  Il ricercatore inserisce nel campione soggetti fino al raggiungimento di proporzioni fissate a partire dalla popolazione (come viene fatto per il campionamento stratificato anche se ben lungi dall'assomigliargli)

Numerosità campionaria

- Per avere un «buon» campione è necessario avere un'indicazione sulla numerosità campionaria, perciò viene stimata.
- I parametri, stabiliti a priori, che intervengono per la stima della numerosità sono:
  - L'obiettivo da valutare/ipotesi di ricerca
  - La tipologia di studio (sperimentale, caso-controllo, prospettico, trasversale, ecc..)
  - Errore di primo tipo (i valori generalmente utilizzati sono 0.10, *0.05*, 0.01)
  - Potenza (i valori generalmente utilizzati sono *0.80*, 0.85, 0.90)
  - Effect size (sintesi di variabilità e differenza di valori che vogliamo trovare statisticamente significativa). Possiamo sintetizzare questa misura in piccola, media o grande.

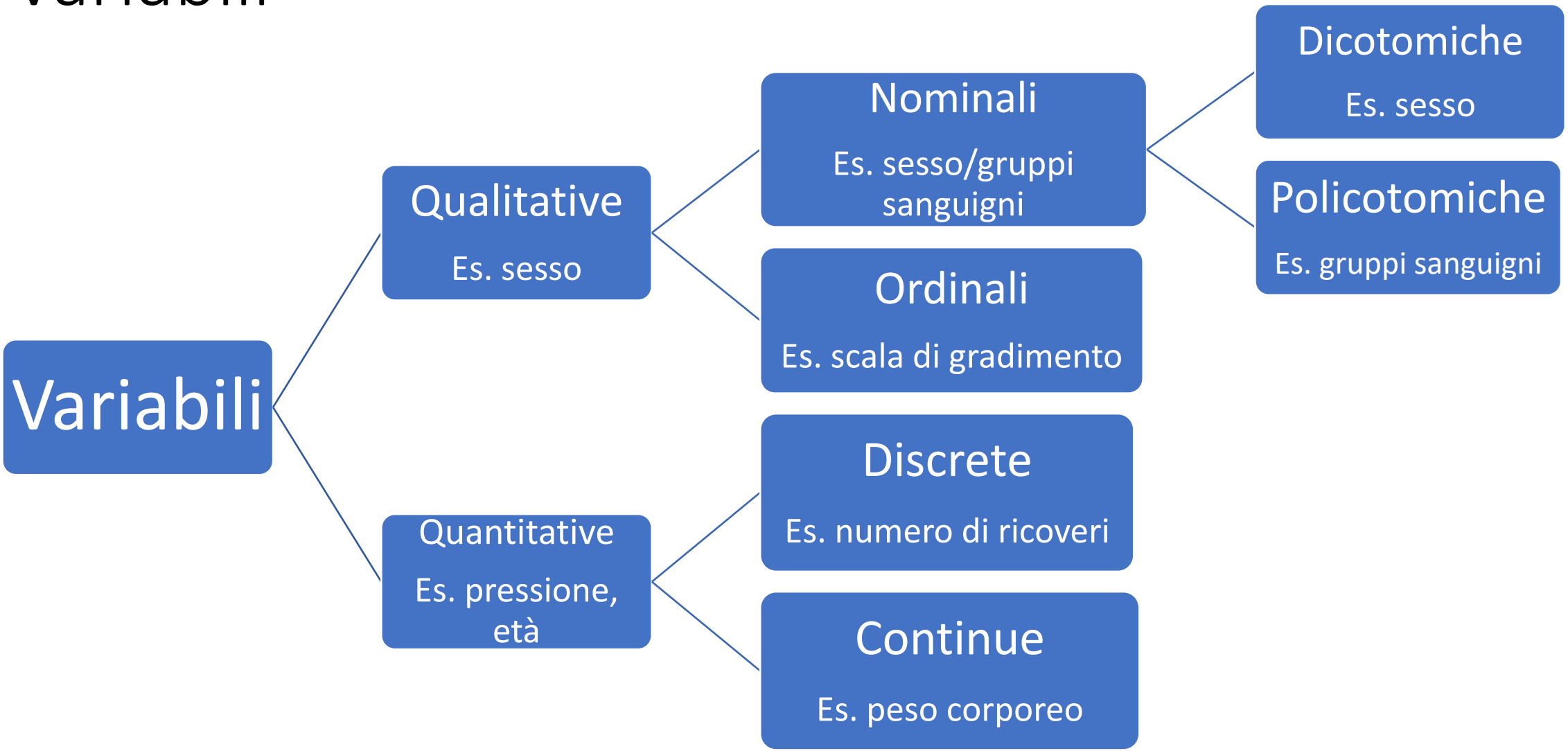
Statistica descrittiva

Definizioni

	Sesso	Età
Maria	F	34
Filippo	M	23
Giulia	F	19

- La tabella complessiva si chiama **database** (raccolta ben strutturata di informazioni)
- Maria, Filippo, Giulia sono **unità statistiche** (tutte le informazioni relative ad una specifica unità statistica sono riportate sulla stessa riga)
- Sesso e età sono **variabili**
- F/M,34,23,19 sono **dati o osservazioni** (realizzazione della variabile, ovvero il valore che assume la variabile su quella specifica unità statistica).
- Nel caso delle variabili che esprimono qualità come il sesso, i «valori» che può assumere la variabile sono detti **modalità** o **categorie**

Variabili


Variabili
Qualitative
Es. sesso
Quantitative
Es. pressione, età
Nominali
Es. sesso/gruppi sanguigni
Ordinali
Es. scala di gradimento
Discrete
Es. numero di ricoveri
Continue
Es. peso corporeo
Dicotomiche
Es. sesso
Policotomiche
Es. gruppi sanguigni

Variabili quantitative

Continue:

Peso: 65,7kg 80,2kg 56,1kg …

Diametro: 2,3cm 1,7cm 2,2cm..

Altezza: 170cm 165cm 183cm..

Discrete:

Accessi al PS: 45 72 121 3 …

Numero di pz visitati al mese con sospetto melanoma: 2 45 31 12..

Variabili qualitative

Nominali policotomiche:

Gruppi sanguigni: A, B, AB, 0
Fumo: Si, No, Ex fumatore
Regioni: Toscana, Lombardia, Sicilia,..

Nominali dicotomiche:

Sesso: Maschio, Femmina
Fumo: Si, No
Istologia: nodulo benigno, tumore

Ordinali:

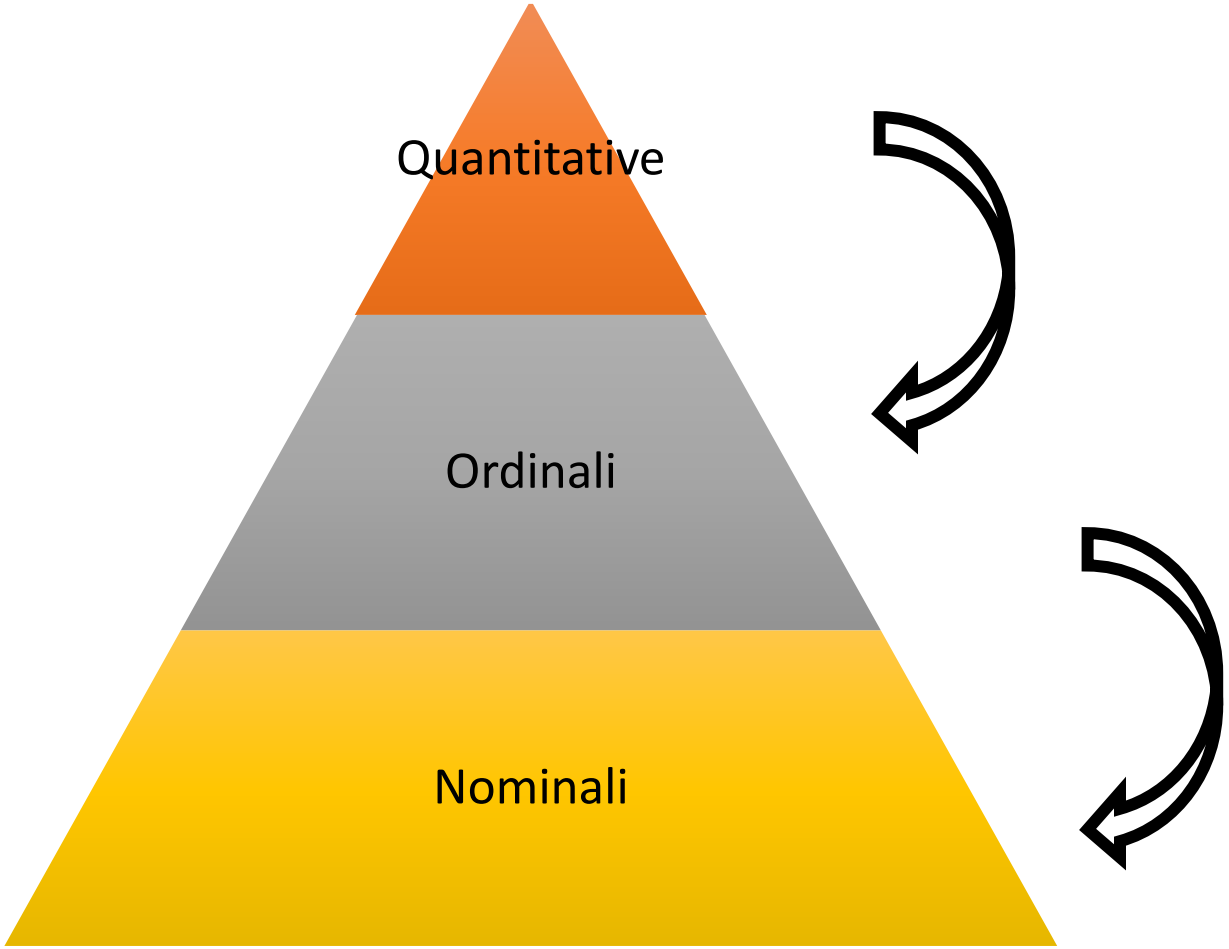
Punteggi e scale

Scala NRS: 0,1,2,3,4,5,6,7,8,9,10
Indice di Apgar: 0,1,2,3,4,5,6,7,8,9,10

Qualità ordinabili in modo crescente o descrescente

BMI: sottopeso, normopeso, sovrappeso, obeso
Esposizione al sole: scarsa, media, alta

Gerarchia delle variabili



La gerarchia è dovuta all'apporto di informazione che danno le variabili e alla possibilità di trasformare le variabili degli strati più alti in variabili di livello più basso.

Statistica descrittiva

- Frequenze
- Tabelle di contingenza
- Indici di sintesi (indici di centralità e dispersione)
- Grafici

Frequenze e tabelle di contingenza

Frequenze

- **Assolute**: numero di volte in cui si presenta una modalità della variabile

- **Relative**: vengono calcolate come rapporto tra le frequenze assolute e la numerosità campionaria.

- **Percentuali:** vengono calcolate come prodotto delle frequenze relative per 100

Frequenze assolute

Variabile qualitativa policotomica fumo (0=no, 1=si, 2=ex fumatore)

Osservazioni sulle unità statistiche: 0,1,2,0,0,0,1,2,1,2,1,0,1,2,0,1,2,0,1,1,1

	Frequenze assolute
0 (SI)	7
1 (NO)	9
2 (ex fumatore)	5
Totale	21

Conteggi delle categorie della variabile

Frequenze relative

Variabile qualitativa policotomica fumo (0=no, 1=si, 2=ex fumatore)

Osservazioni sulle unità statistiche: 0,1,2,0,0,0,1,2,1,2,1,0,1,2,0,1,2,0,1,1,1

	Frequenze assolute	Frequenze relative
0 (SI)	7	0.33
1 (NO)	9	0.43
2 (ex fumatore)	5	0.24
Totale	21	1

Rapporto delle frequenze assolute per il totale:
7/21=0.33
9/21=0.43
5/21=0.24

Frequenze percentuali

Variabile qualitativa policotomica fumo (0=no, 1=si, 2=ex fumatore)

Osservazioni sulle unità statistiche: 0,1,2,0,0,0,1,2,1,2,1,0,1,2,0,1,2,0,1,1,1

	Frequenze assolute	Frequenze relative	Frequenze percentuali
0 (SI)	7	0.33	33%
1 (NO)	9	0.43	43%
2 (ex fumatore)	5	0.24	24%
Totale	21	1	100%

Prodotto delle frequenze relative per 100%:
(0.33x100)%=33%
(0.43x100)%=43%
(0.24x100)%=24%

Tabelle di contingenza

Modalità o categorie della variabile

Outcome/variabile dipendente

	Mod1	Mod2	Totale di riga
Mod1			R_1
Mod2			R_2
Totale di colonna	C_1	C_2	Totale (n)

Exposure /variabile indipendente

- Frequenze assolute
- Frequenze relative
- Frequenze percentuali

Numerosità campionaria

Tabelle di contingenza

- **Assolute:** numero di volte in cui compare una coppia di modalità
- **Relative/Percentuali:** rapporto tra frequenze assolute e numerosità campionaria, in questo caso è possibile calcolarle rapportando le frequenze assolute al totale di riga o di colonna (es studi caso-controllo rapporto con totale di colonna, studi prospettici rapporto con totale di riga)
- **Attese**: vengono calcolate nelle tabelle di contingenza. Non vengono utilizzate per sintetizzare i dati. Sono necessarie però per stabilire quale test utilizzare. La frequenza attesa di una cella viene calcolata come:

$$\frac{C_i * R_i}{n}$$

Tabelle di contingenza

Voglio sintetizzare congiuntamente le variabili Fumo e Cancro al polmone per vedere se il cancro è più presente o meno nei fumatori rispetto ai non fumatori

id	Fumo	Cancro al polmone
1	No	Si
2	Si	No
3	No	No
4	Si	Si
5	No	Si
6	Si	No
7	No	Si
8	No	Si
9	Si	No
10	No	No
11	No	Si
12	Si	No
13	Si	No
14	No	No
15	No	No
16	Si	Si
17	Si	No
18	No	Si

Frequenze assolute

	Cancro al polmone		
Fumo	Si	No	Totale
Si	**52**	**27**	79
No	**34**	**77**	111
Totale	86	104	190

Numero di volte in cui compaiono le coppie
No-No
No-Si
Si-No
Si-Si

Frequenze relative

	Cancro al polmone		
Fumo	Si	No	Totale
Si	**52**	**27**	79
No	**34**	**77**	111
Totale	86	104	190

Rapporto con numerosità campionaria

	Cancro al polmone		
Fumo	Si	No	Totale
Si	**0.27**	**0.14**	0.41
No	**0.18**	**0.41**	0.59
Totale	0.45	0.55	1

Rapporto con totale di riga

	Cancro al polmone		
Fumo	Si	No	Totale
Si	**0.66**	**0.34**	1
No	**0.31**	**0.69**	1

Rapporto con totale di colonna

	Cancro al polmone	
Fumo	Si	No
Si	**0.60**	**0.26**
No	**0.40**	**0.74**
Totale	1	1

Frequenze percentuali

	Cancro al polmone		
Fumo	Si	No	Totale
Si	**52**	**27**	79
No	**34**	**77**	111
Totale	86	104	190

Rapporto con numerosità campionaria

	Cancro al polmone		
Fumo	Si	No	Totale
Si	**27%**	**14%**	41%
No	**18%**	**41%**	59%
Totale	45%	55%	100%

Rapporto con totale di riga

	Cancro al polmone		
Fumo	Si	No	Totale
Si	**66%**	**34%**	100%
No	**31%**	**69%**	100%

Rapporto con totale di colonna

	Cancro al polmone	
Fumo	Si	No
Si	**60%**	**26%**
No	**40%**	**74%**
Totale	100%	100%

Indici di sintesi

Indici di centralità e dispersione

Indici di centralità

Media: media aritmetica dei dati osservati.

$$\underbrace{2, 3, 4, 4, 6, 12, 5, 6}_{\text{8 valori}} \rightarrow \frac{2+3+4+4+6+12+5+6}{8} = 5.25$$

Indici di centralità

Mediana: numero centrale (per calcolarla la serie deve essere ordinata)

La mediana individua e separa il 50% dei valori.

Non è sensibile agli outlier diversamente dalla media

Viene calcolata in modo diverso in base al numero di elementi della serie

Indici di centralità

- **Numero pari di elementi:**

  calcolo della media tra gli elementi in posizione $\frac{n}{2}$ e $\frac{n+2}{2}$

  2, 3, 4, $\boxed{4, 5}$, 6, 6, 12 → media dei due valori centrali → mediana=4.5

- **Numero dispari di elementi:**

  selezioniamo l'elemento in posizione $\frac{n+1}{2}$

  2, 3, 4, 4, $\boxed{5}$, 6, 6, 12, 13 → valore centrale → mediana=5

Indici di centralità

Moda: è il valore che compare più volte.

Nel caso in cui più valori distinti compaiono lo stesso numero di volte allora la moda non è unica ma sono tante quante questi valori

2, 3, 4, 4, 5, 6, 8, 12 → moda: 4

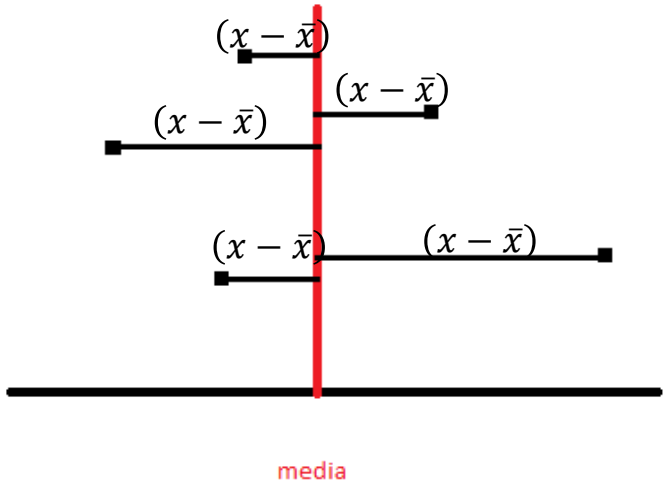
2, 3, 4, 4, 5, 6, 6, 12 → mode: 4,6

Indici di dispersione

- **Varianza**: scarto quadratico medio, indice di centralità associato alla media

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n}$$

- **Deviazione standard:** radice quadrata della varianza, è l'indice di dispersione usualmente associato alla media dato che hanno la stessa unità di misura



Indici di dispersione

Ce ne sono altri non frequentemente utilizzati come indici di sintesi me molto utili. Tra questi ci sono:

- Lo **standard error**

Quantifica il grado di certezza col quale il parametro stimato su un campione casuale stima la vera media della popolazione da cui il campione è tratto

$$\textbf{SE}=\frac{\sigma}{\sqrt{n}}$$

- Il **coefficiente di variazione**

Permette di valutare la dispersione dei valori attorno alla media indipendentemente dall'unità di misura (grandezza adimensionale). Più è basso più i dati sono concentrati intorno alla media.

$$\textbf{CV}= \frac{\sigma}{media} * 100$$

Indici di dispersione

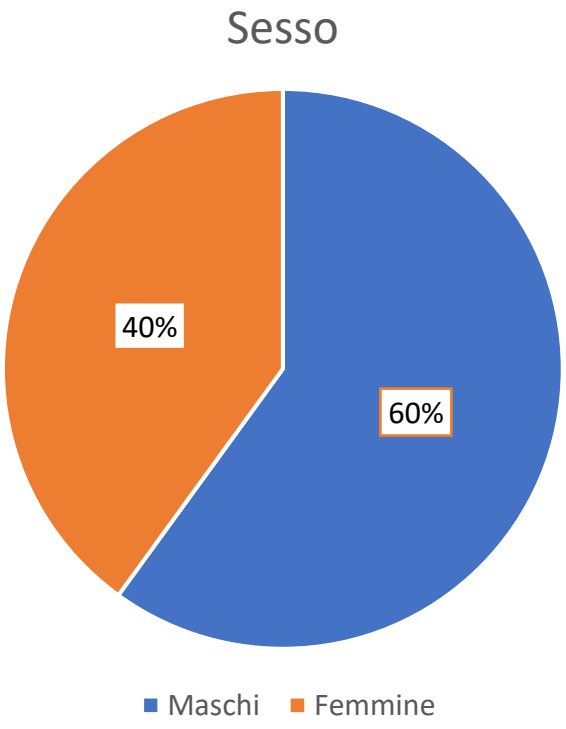
- **Range interquartile:** range compreso tra il primo e il terzo quartile (primo quartile ha a sinistra il 25% delle osservazioni, il terzo quartile il 75%); è l'indice di dispersione associato alla mediana

- **Range minimo-massimo:** range compreso tra il valore più piccolo osservato e quello più grande, anche questo viene associato alla mediana

Grafici

Grafici per variabili qualitative

Grafico a torta



Viene utilizzato per rappresentare un'unica variabile qualitativa

Gli spicchi sono tanti quanti le modalità della variabile qualitativa

È preferibile utilizzarlo quando le modalità della variabile non sono molte

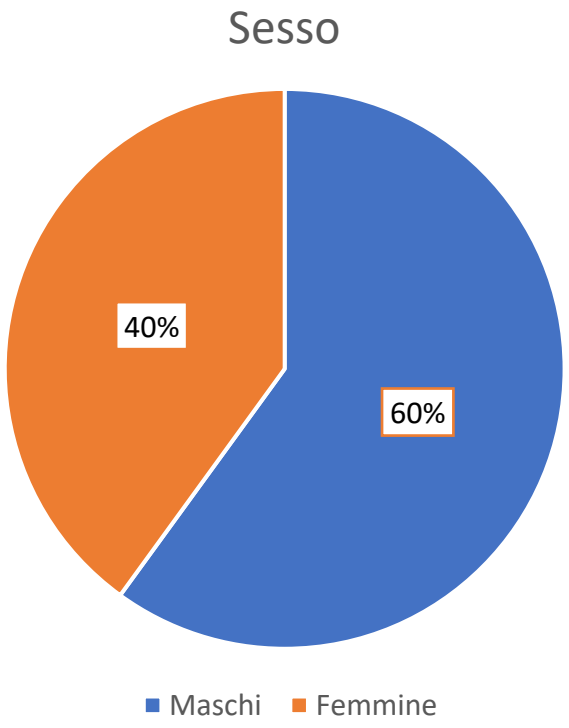
Possiamo decidere di inserire nel grafico le frequenze assolute o relative o percentuali

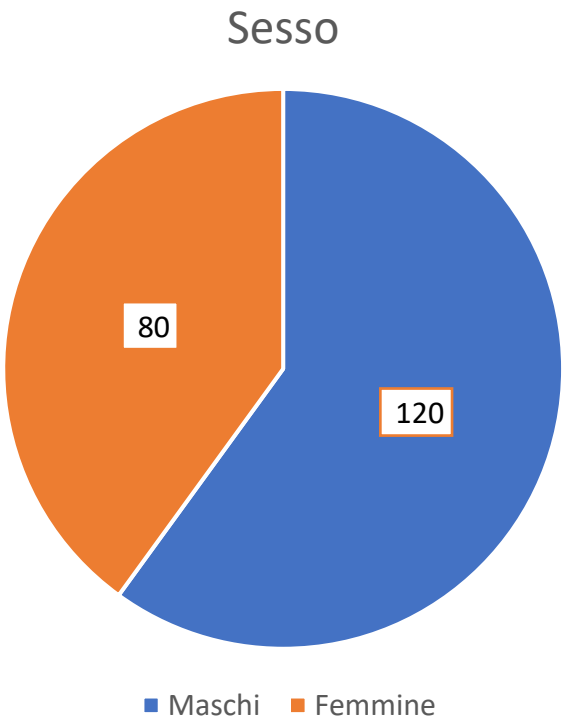
Gli spicchi ricoprono un'area del cerchio pari alla frequenza percentuale

Grafici per variabili qualitative

Grafico a torta

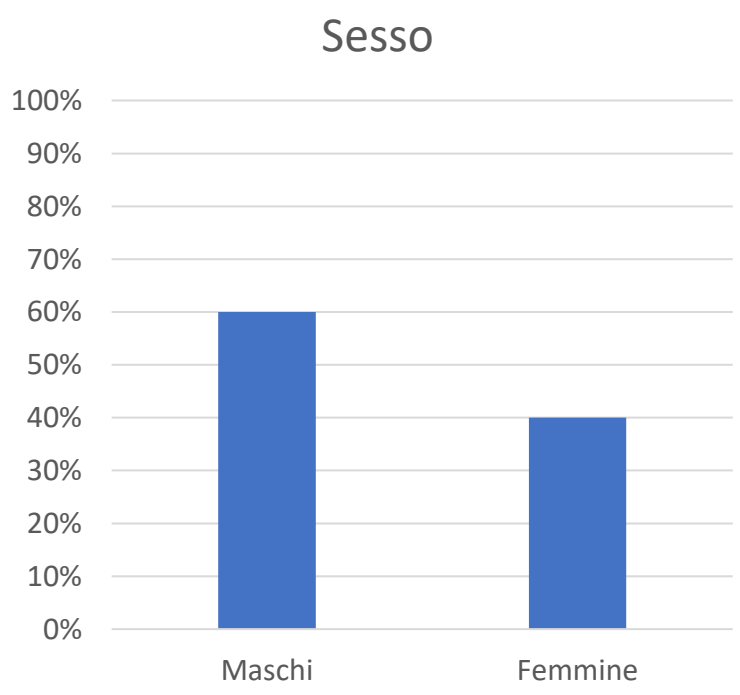
Sesso		80
M	120	60%
F	80	40%
Totale	200	100%





Grafici per variabili qualitative

Grafico a barre



Può sintetizzare una o due variabili qualitative

Se sintetizza una sola variabile qualitativa le barre sono tante quante le modalità della variabile

Nell'asse delle ordinate possiamo decidere di inserire le frequenze assolute, relative o percentuali

Le altezze dei rettangoli dipendono dalle frequenze assolute/relative/percentuali delle categorie

Questo grafico può essere utilizzato in presenza di molte modalità

Grafici per variabili qualitative

Grafico a barre

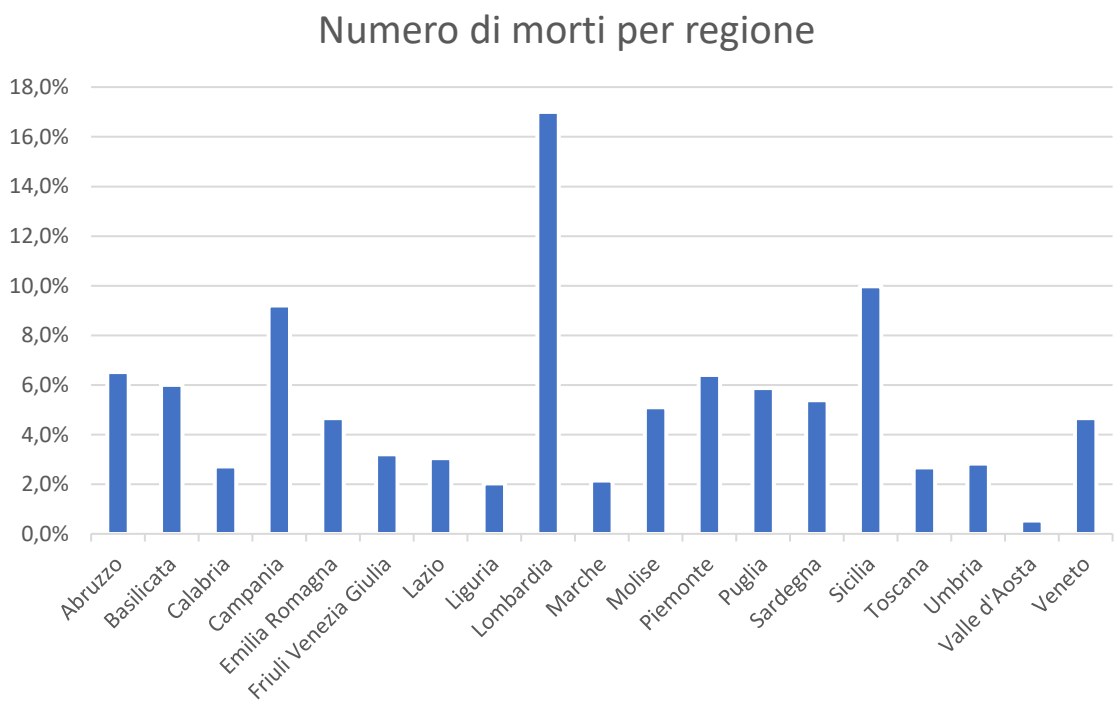
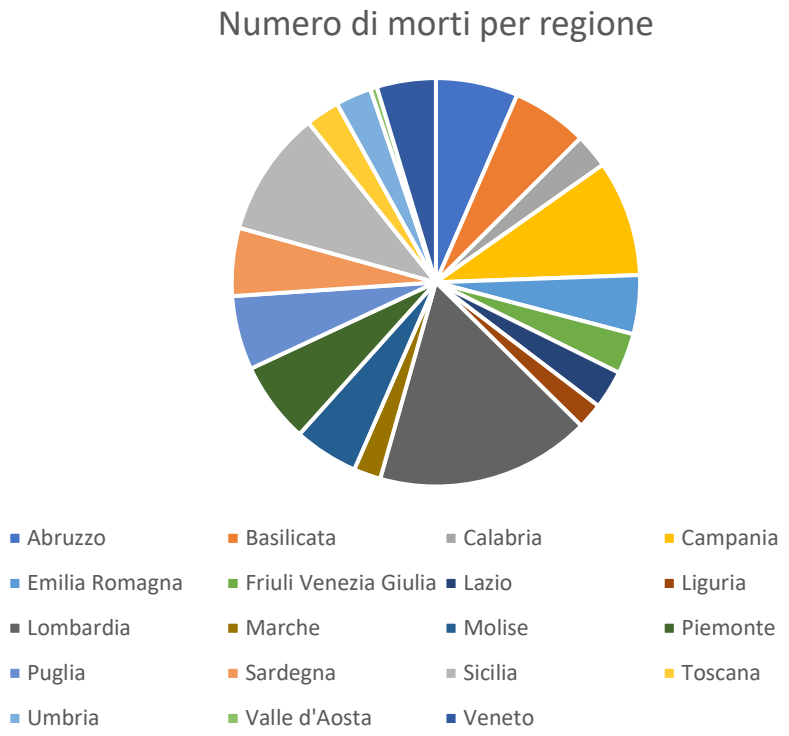


Grafico a torta



Grafici per variabili qualitative

Grafico a barre

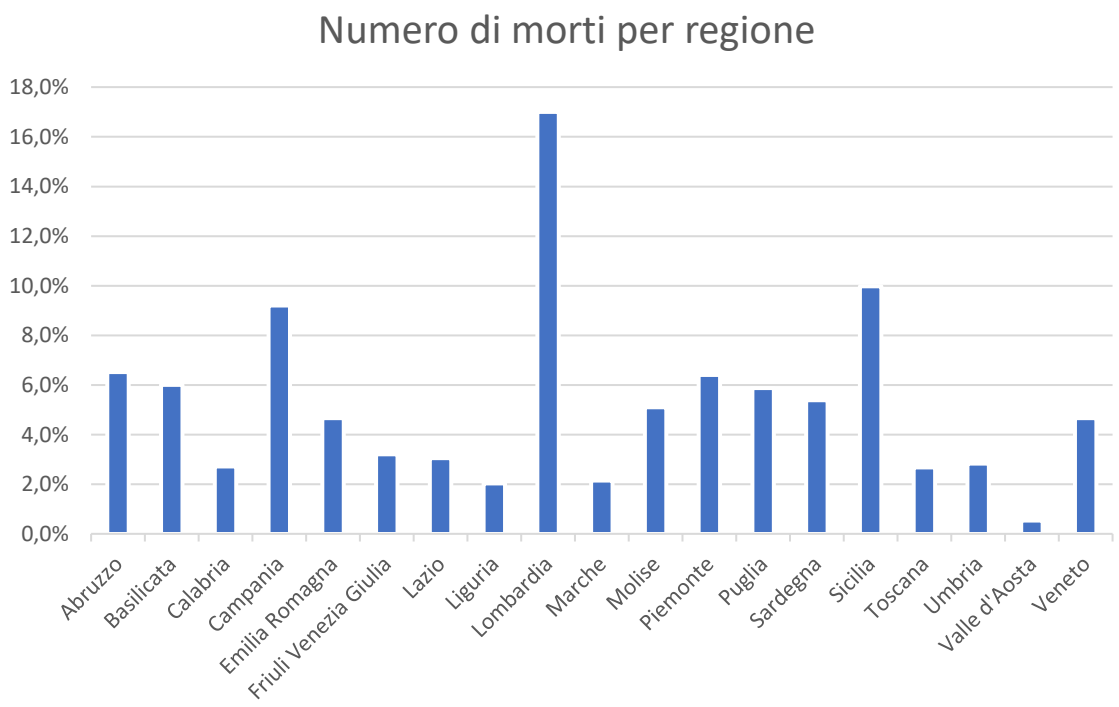
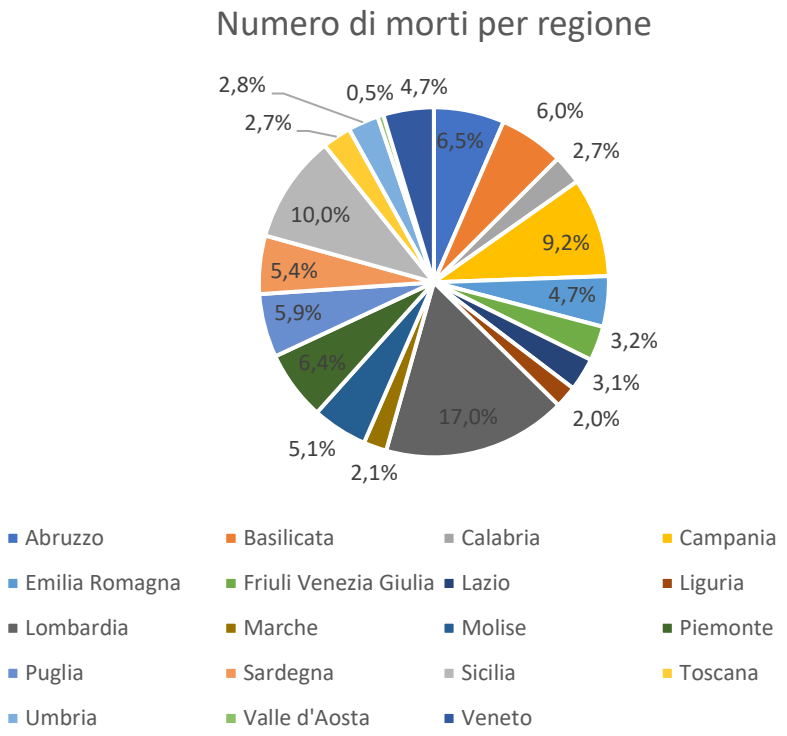
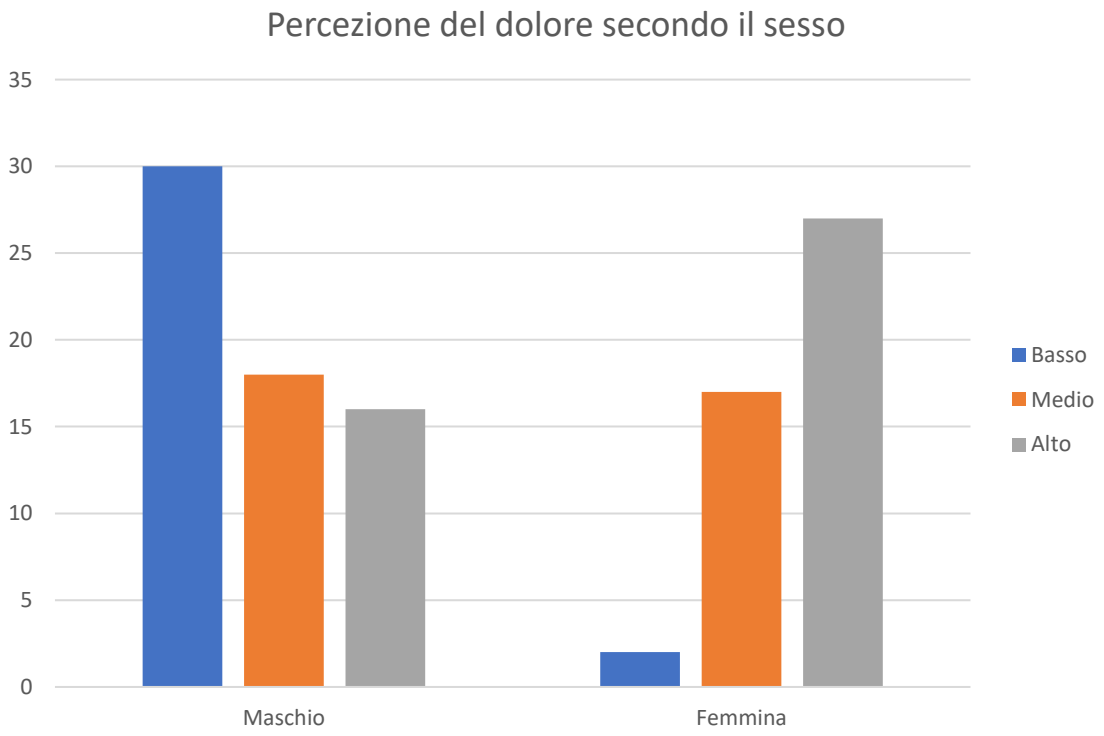


Grafico a torta



Grafici per variabili qualitative

Grafico a barre per due variabili qualitative



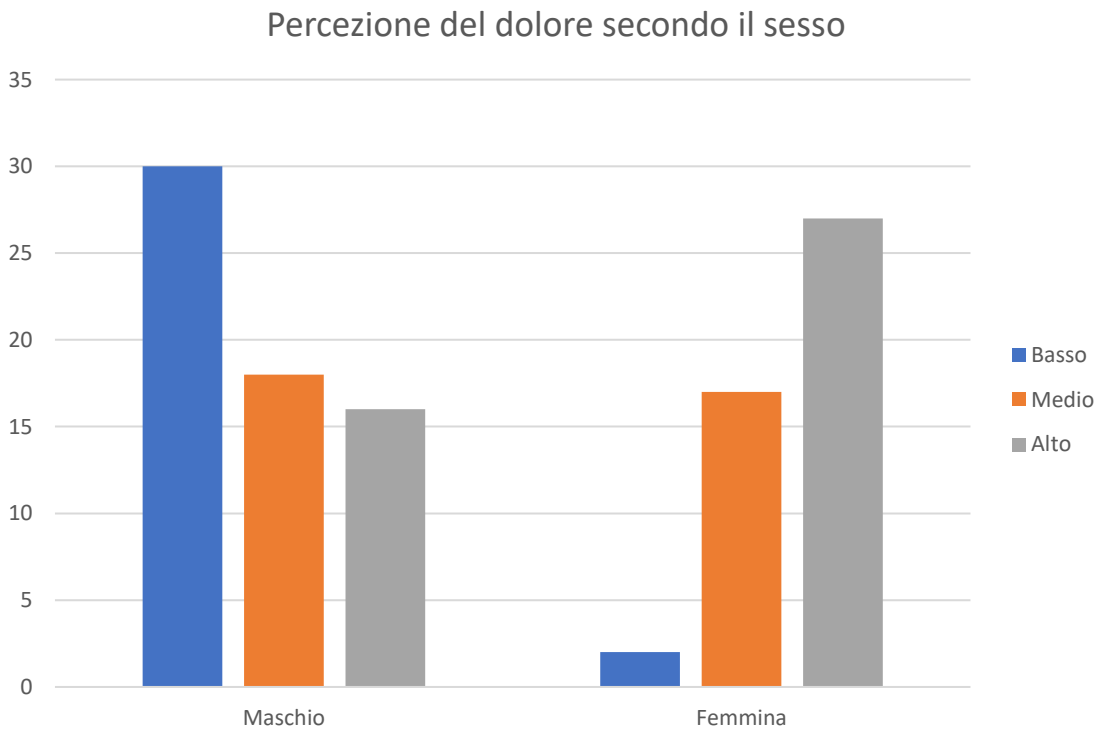
Questo grafico sintetizza una tabella di contingenza

La variabile di exposure viene messa sull'asse orizzontale

La variabile di outcome definisce i rettangoli per ciascun gruppo identificato dalla variabile di exposure

Grafici per variabili qualitative

Grafico a barre per due variabili qualitative

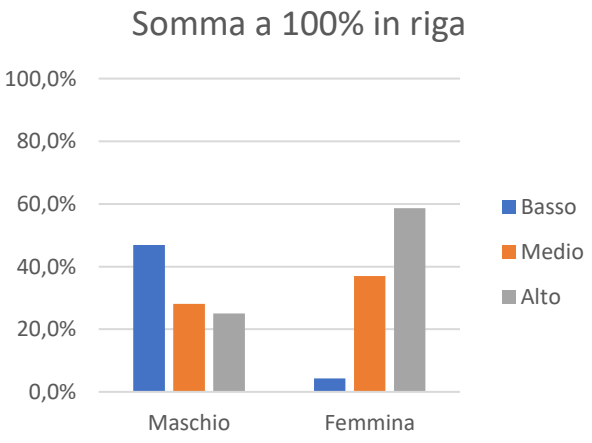


Se la tabella di contingenza ha dimensione 2x3 allora le barre saranno 2x3

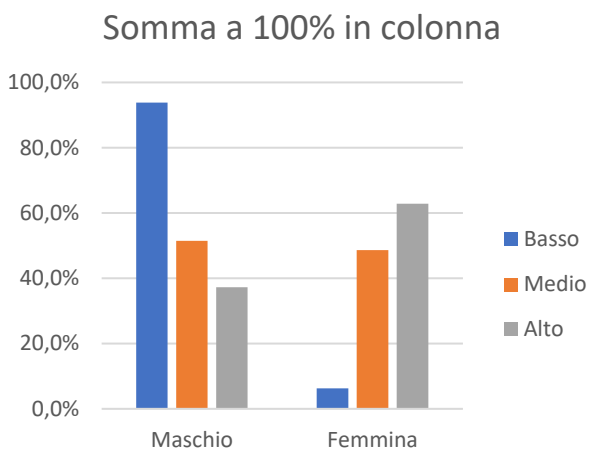
	Scala di dolore		
Sesso	Basso	Medio	Alto
Maschio	30	18	16
Femmina	2	17	27

Grafici per variabili qualitative

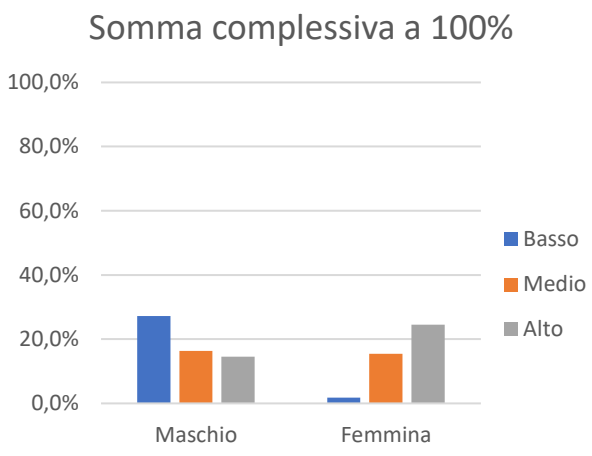
3 alternative in base a come vogliamo calcolare le frequenze percentuali



	Scala di dolore		
Sesso	Basso	Medio	Alto
Maschio	46,9%	28,1%	25,0%
Femmina	4,3%	37,0%	58,7%



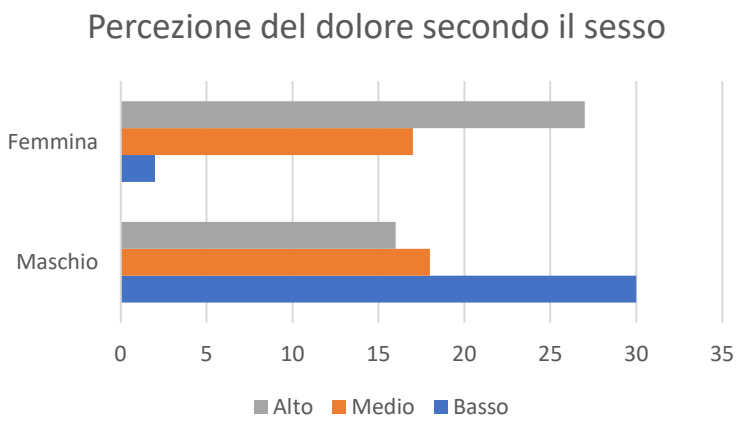
	Scala di dolore		
Sesso	Basso	Medio	Alto
Maschio	93,8%	51,4%	37,2%
Femmina	6,3%	48,6%	62,8%

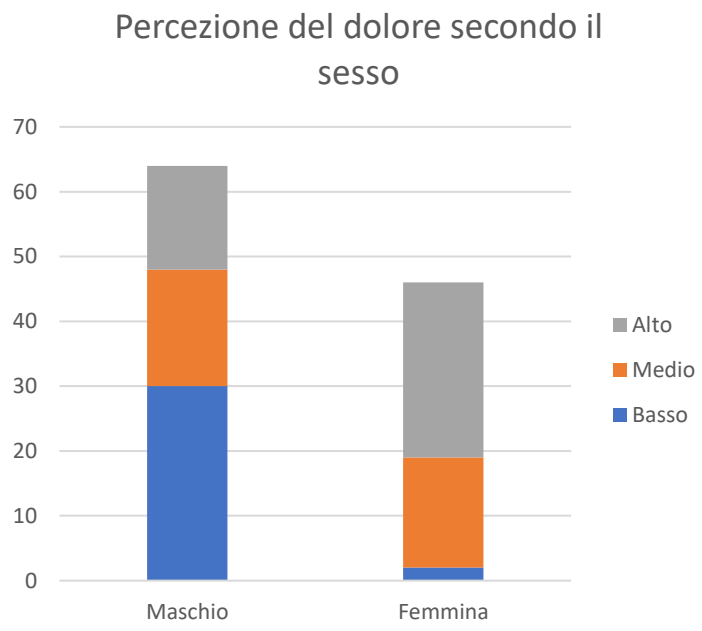


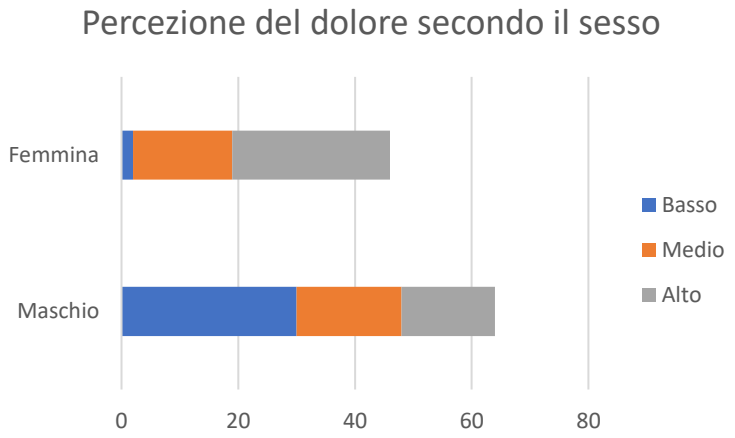
	Scala di dolore		
	Basso	Medio	Alto
Maschio	27,3%	16,4%	14,5%
Femmina	1,8%	15,5%	24,5%

Grafici per variabili qualitative

Altre rappresentazioni del grafico a barre

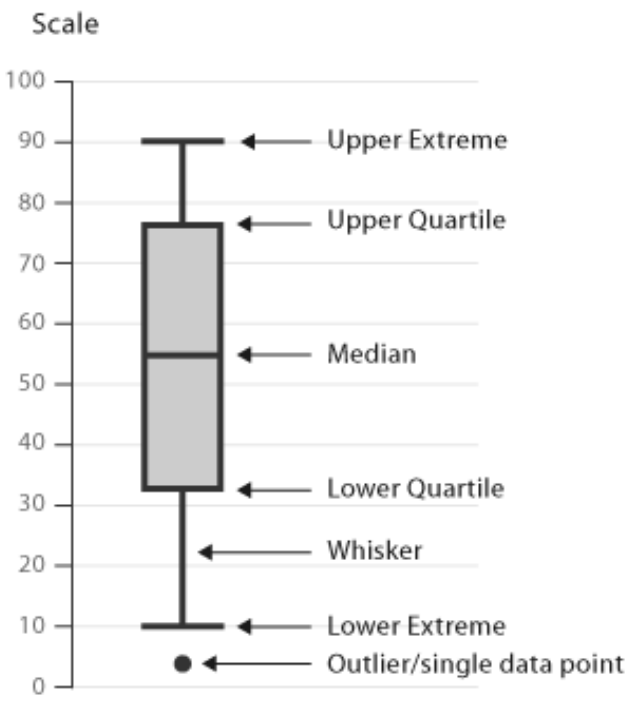






Grafici per variabili quantitative

Box-Whisker



Questo grafico viene utilizzato per sintetizzare una variabile di tipo quantitativo

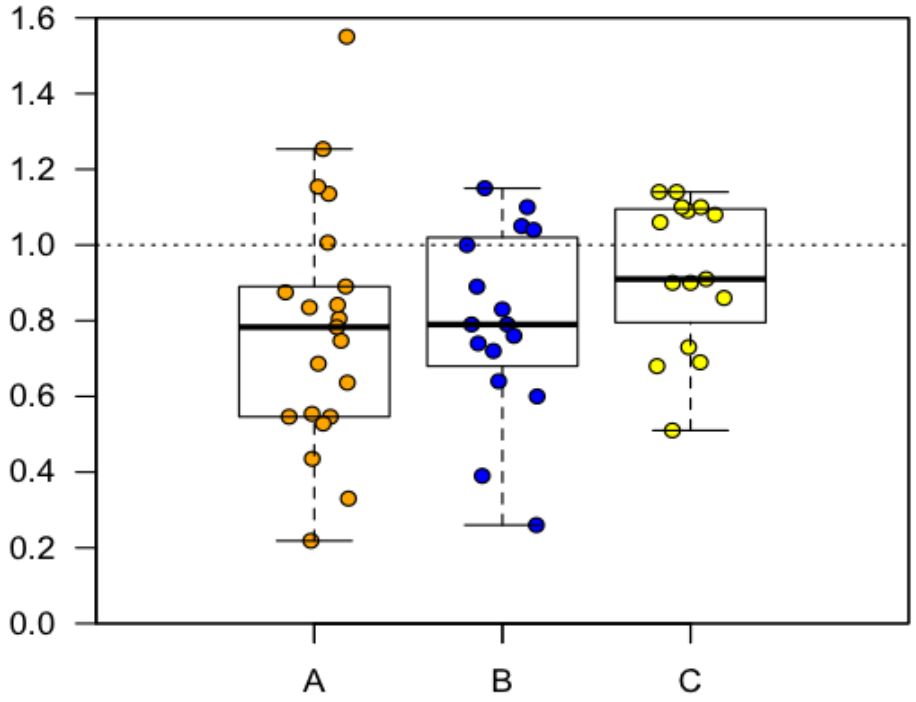
È necessario conoscere i valori della mediana e del range interquartile

È un grafico molto utile per capire se ci sono outlier/valori anomali.

Questi grafici ci aiutano a capire come e dove sono posizionate le osservazioni

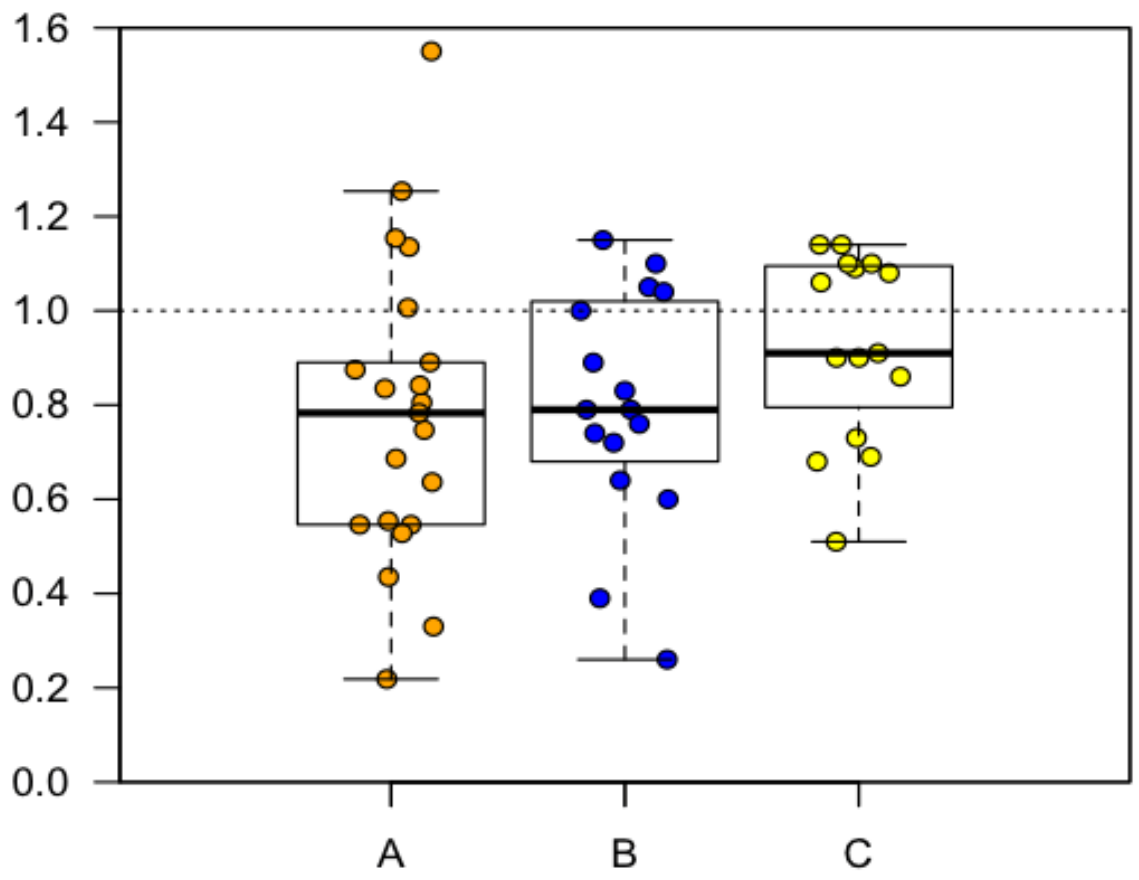
Grafici per variabili quantitative

Box-Whisker – variabile quantitativa e qualitativa



Nel caso in cui volessimo valutare la relazione di una variabile qualitativa e una qualitativa potremmo creare tanti grafici box-whisker quante sono le modalità della variabile qualitativa

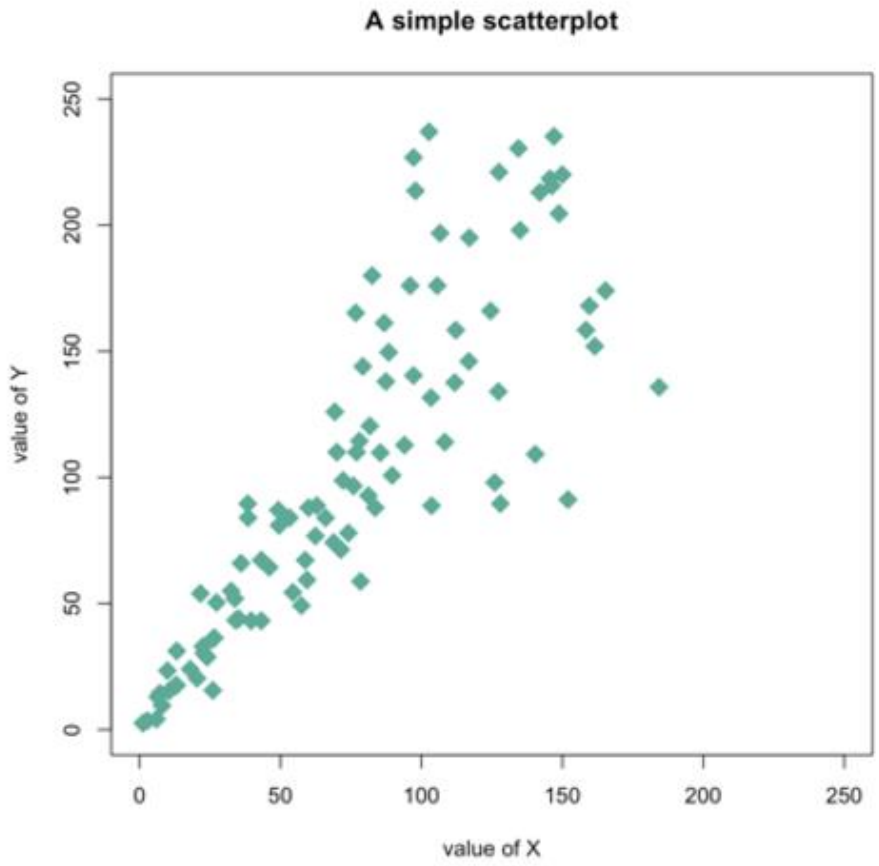
Grafici per variabili quantitative



id	terapia	magnesio
1	A	0,5
2	A	1,2
3	A	1,1
4	A	0,6
5	A	0,2
6	A	0,3
7	A	0,9
8	B	0,3
9	B	1,6
10	B	1,5
11	B	1,2
12	B	0,9
13	B	0,2
14	B	0,7
15	B	1,1
16	B	0,0
17	B	1,8
18	C	1,3
19	C	1,4
20	C	0,8
21	C	0,7
22	C	0,8
23	C	0,6
24	C	1,0
25	C	1,3
26	C	1,1
27	C	0,7

Grafici per variabili quantitative

Scatterplot



Sintetizza due variabili quantitative

La variabile di exposure viene messa in ascissa mentre quella di outcome in ordinata

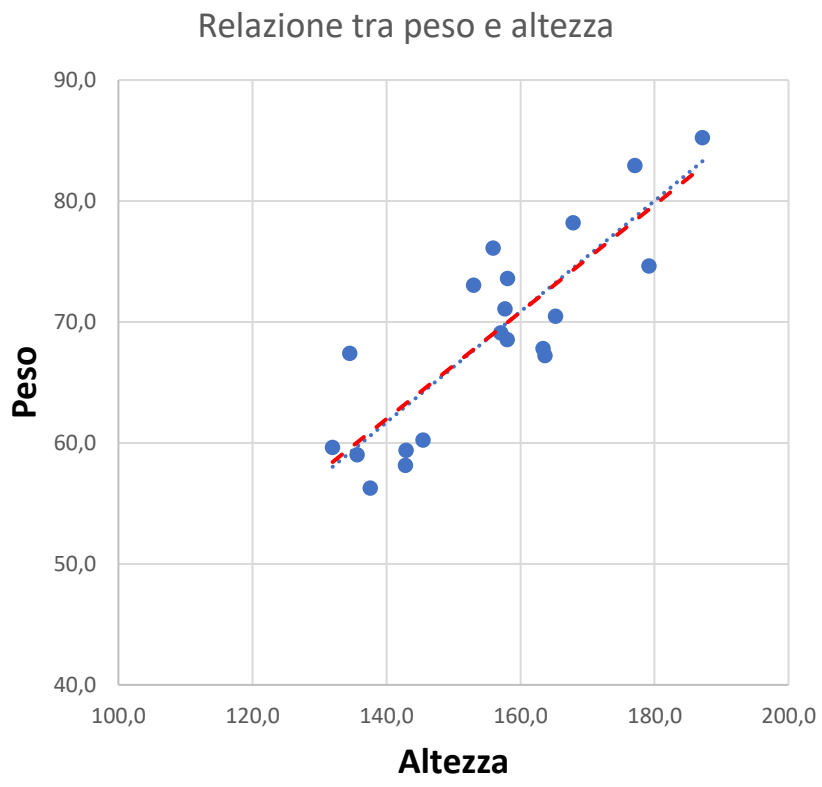
I punti rappresentano le unità statistiche

Possiamo vedere come varia l'outcome in base all'exposure

Grafici per variabili quantitative

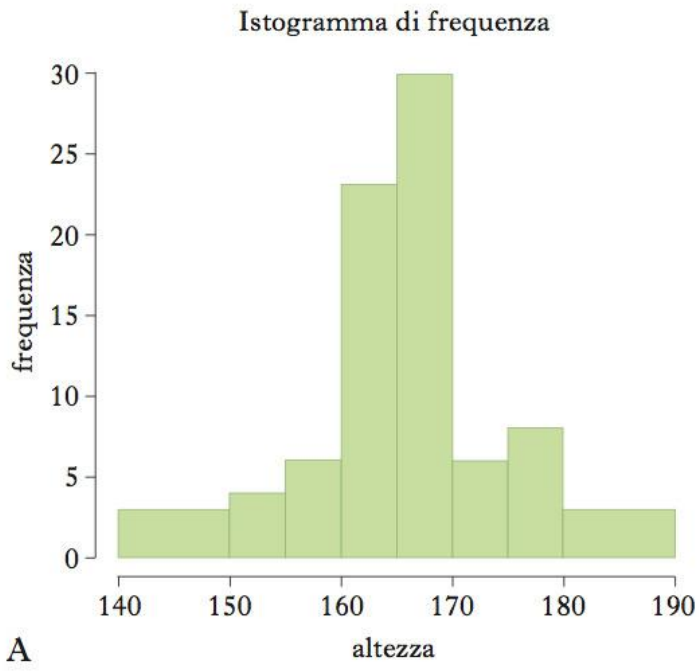
Come possiamo osservare graficamente come varia il peso in funzione dell'altezza?

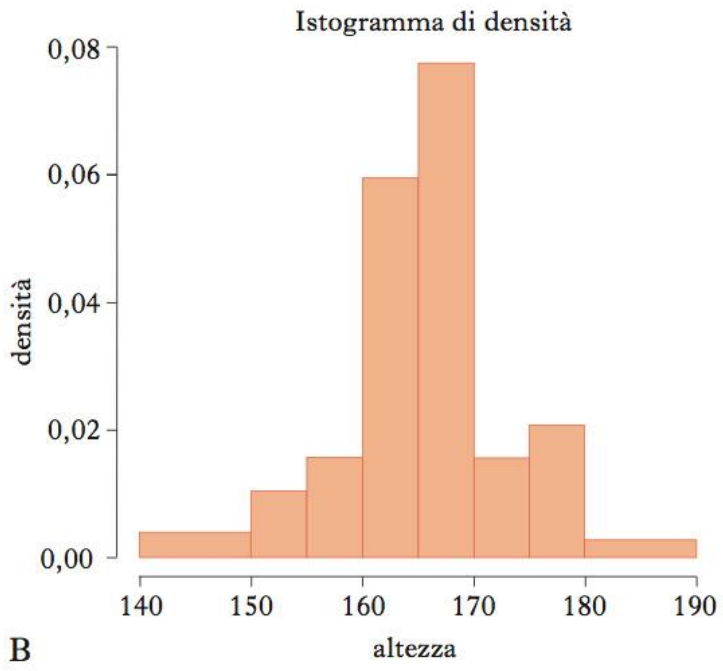
Altezza	Peso
142,9	59,4
163,7	67,2
153,0	73,0
167,9	78,2
155,9	76,1
177,1	82,9
163,4	67,8
158,0	73,6
157,7	71,1
187,2	85,2
135,6	59,0
179,2	74,6
165,2	70,5
132,0	59,6
145,5	60,2
157,0	69,1
142,9	58,2
134,5	67,4
137,6	56,3
158,0	68,5



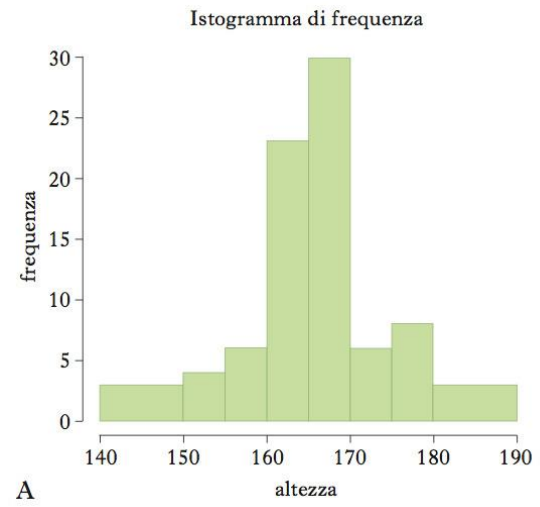
Grafici per variabili quantitative

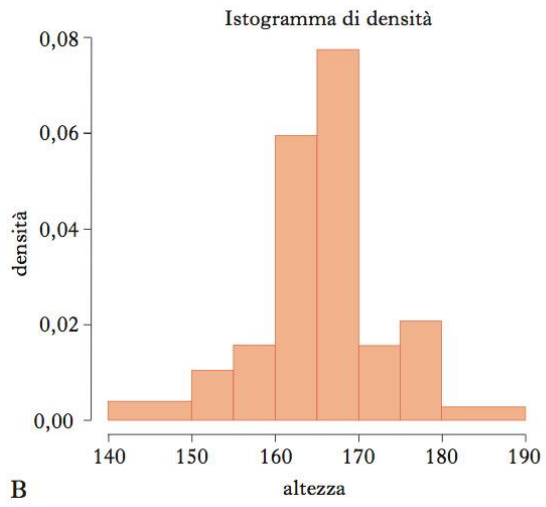
Istogramma





Istogramma

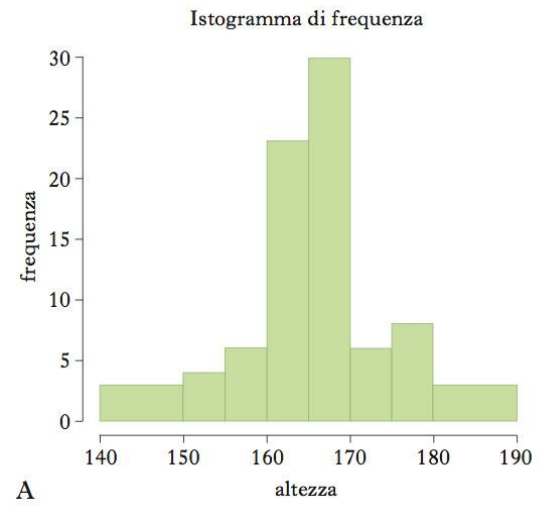


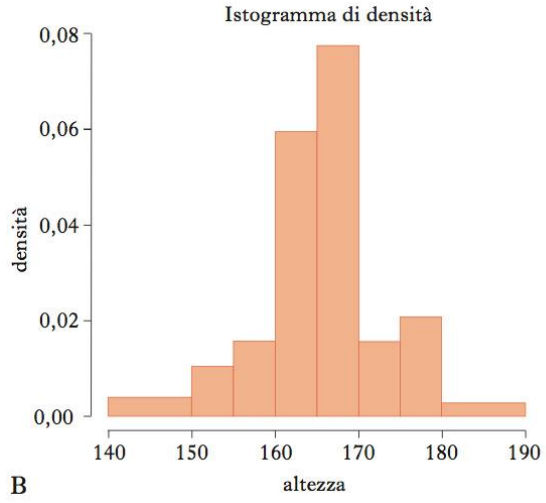


La variabile viene suddivisa in classi di ampiezza costante o variabile e viene posta sull'asse delle ascisse. L'ordinata (altezza del rettangolo), invece, è uguale :

- Alla frequenza (il numero delle osservazioni che ricadono in quell'intervallo) se le classi hanno ampiezza costante
- Al rapporto della frequenza fratto l'ampiezza della base se gli intervalli non hanno ampiezza costante

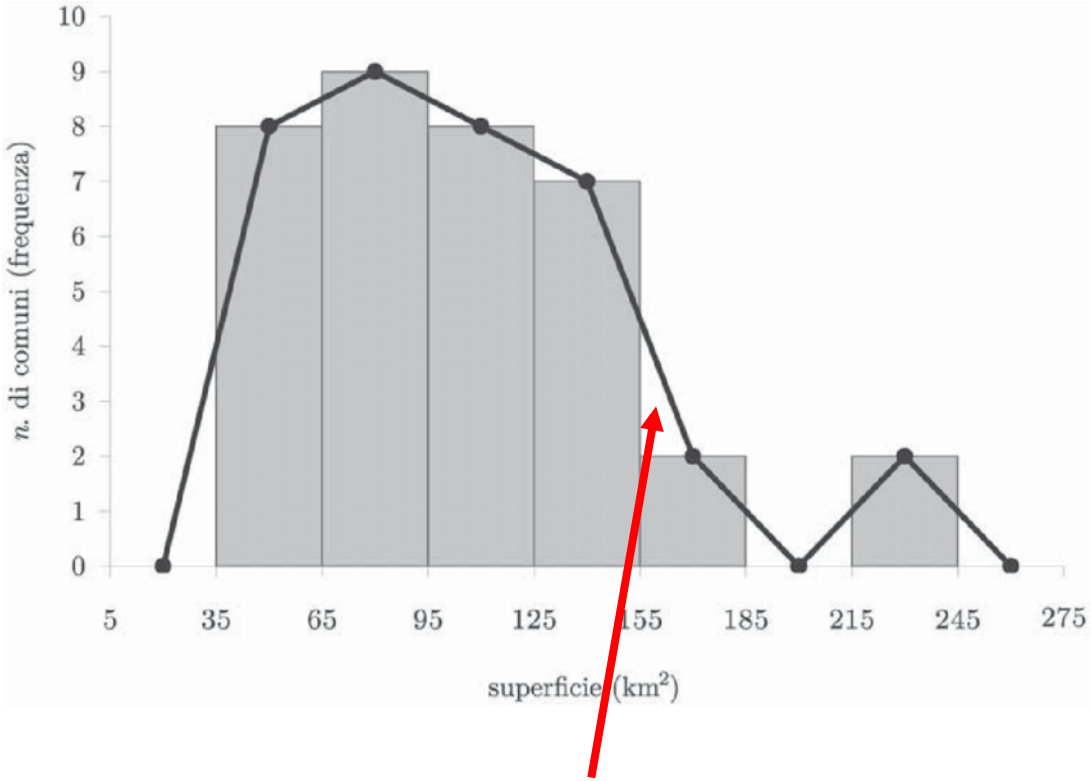
Istogramma





Il rapporto della frequenza per l'ampiezza della base viene chiamato **densità,** perciò in caso di classi di ampiezza non costante è necessario utilizzare le densità in ordinata, mentre per classi di egual misura è indifferente

Istogramma



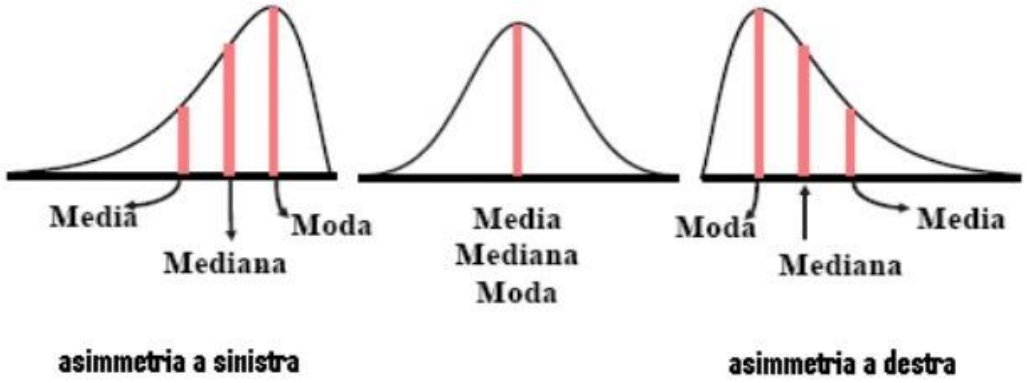
Poligono di frequenza

- Stringendo sempre più gli intervalli otterremo la distribuzione della variabile
- Il poligono diventerà una curva
- L'istogramma ci aiuta a vedere graficamente la «forma» della distribuzione

Proprietà delle distribuzioni

- f(x)≥0 per ogni valore di x
- $\int f(x)dx = 1$ → l'area sottostante alla curva è uguale a 1

Distribuzione della variabile

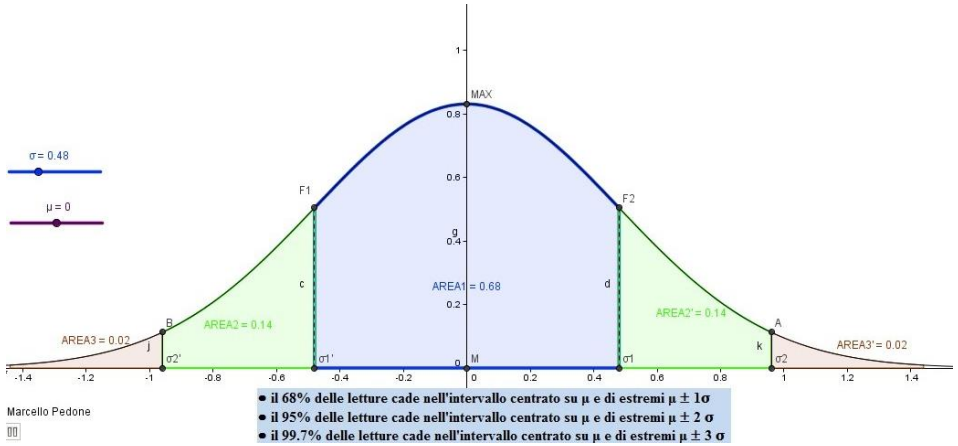


Particolari tipi di distribuzione simmetrica sono la distribuzione **normale** o **Gaussiana** e la distribuzione **T di Student**

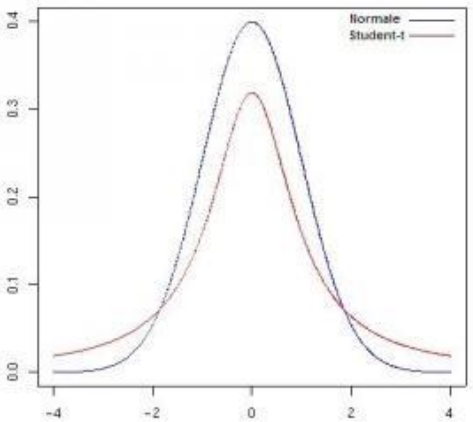
Particolari tipi di distribuzione asimmetrica a destra sono la distribuzione **chi-quadrato** e la distribuzione **F di Fisher-Snedecor**

Distribuzioni particolari

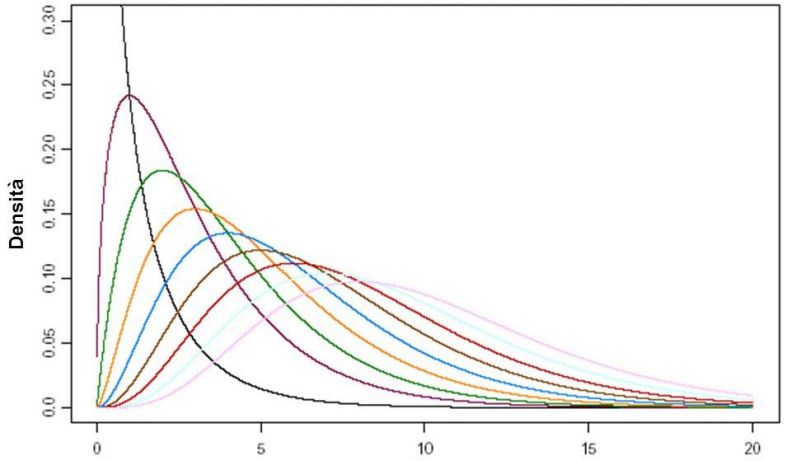
Distribuzione Gaussiana



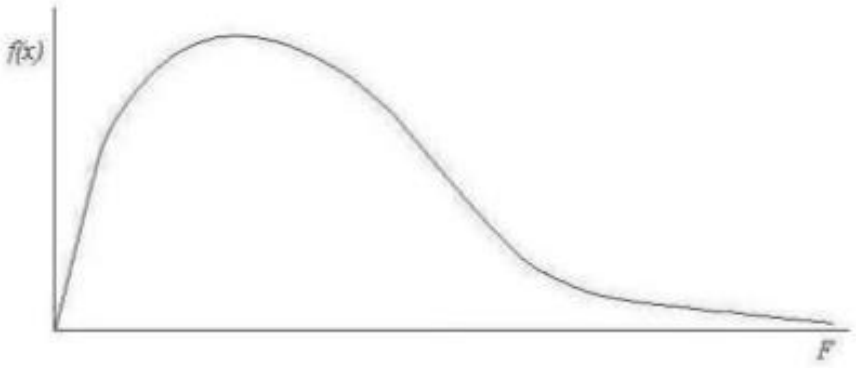
Distribuzione T di Student



Distribuzione Chi quadrato

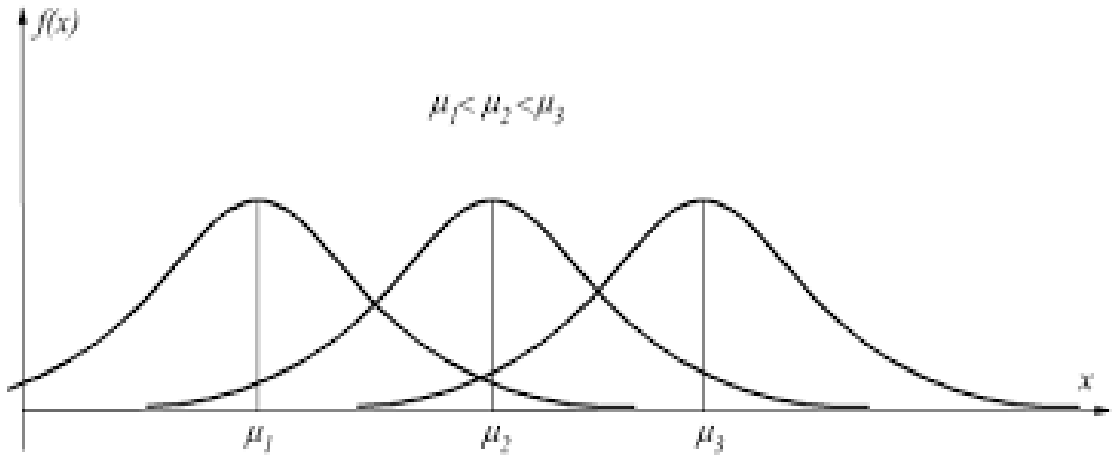


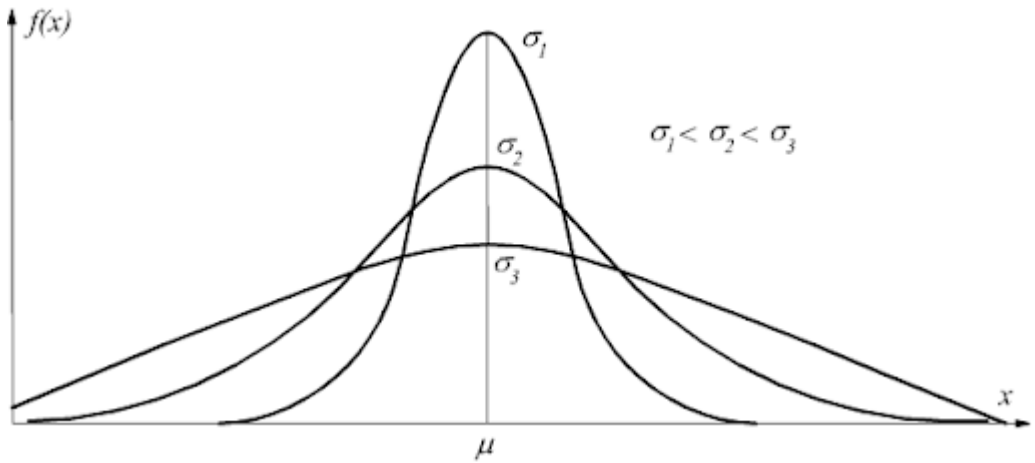
Distribuzione F di Fisher-Snedecor



Cos'è la Statistica - G. Garau, L. Schirru

Distribuzione normale

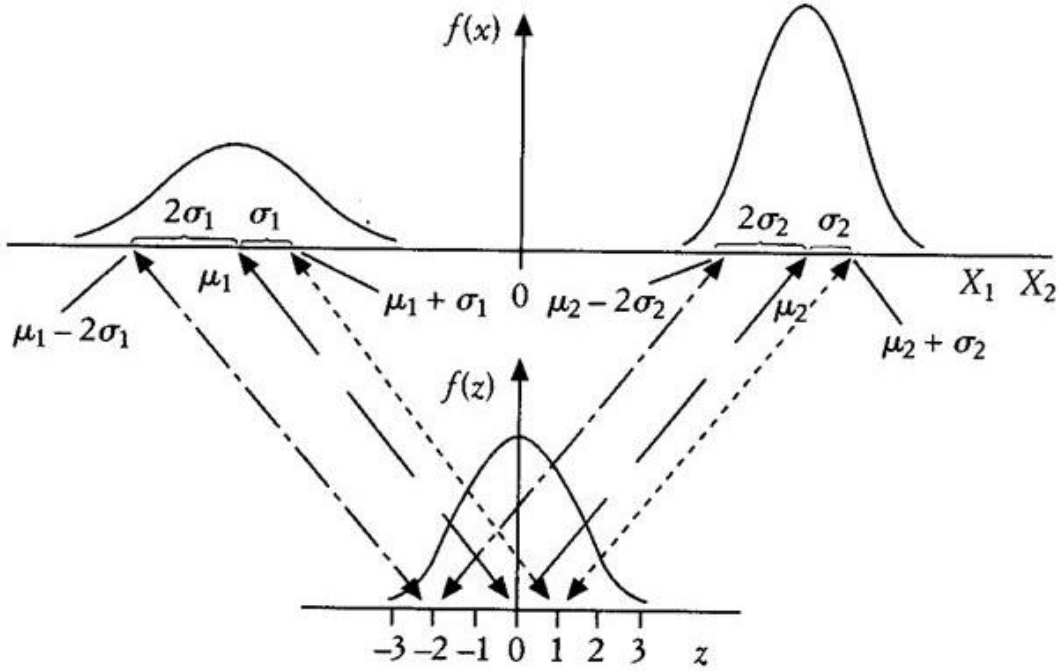




Questa distribuzione è completamente spiegata dai parametri μ (media) e σ (deviazione standard)

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

Distribuzione normale



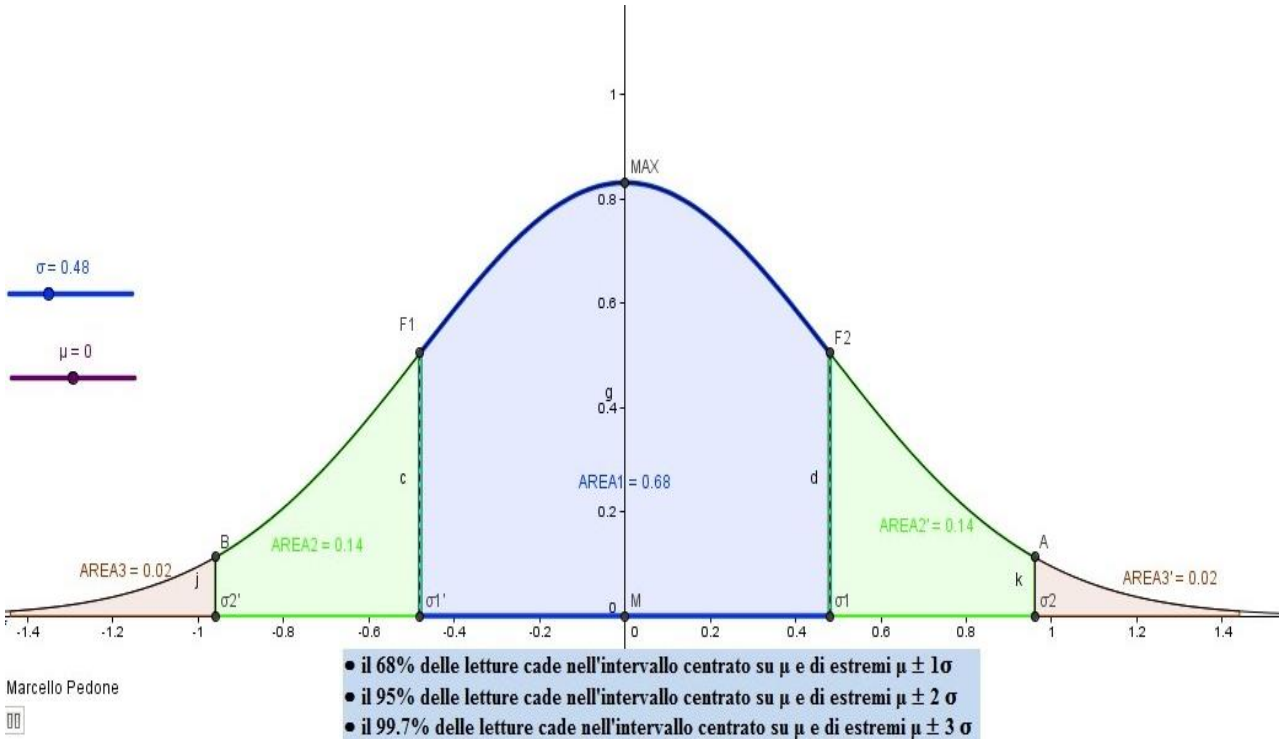
Ogni variabile normale può essere trasformata in una **normale standardizzata**: la media è pari a 0 e la deviazione standard a 1

$$z = \frac{X - \mu}{\sigma}$$

La funzione diventa

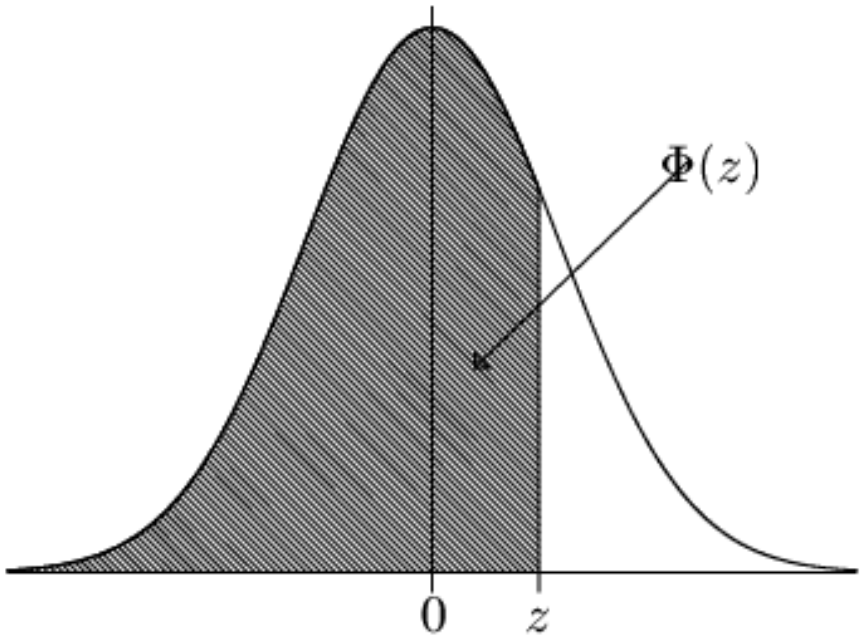
$$f(x) = \frac{1}{\sqrt{2\pi}} * e^{-\frac{1}{2}z^2}$$

Distribuzione normale



- Nell'intervallo [μ-1σ; μ+1σ] troviamo il 68% dei valori (equivalente a dire che l'area individuata dai due estremi dell'intervallo è il 68% dell'area sottesa alla curva)
- Nell'intervallo [μ-2σ; μ+2σ] troviamo il 95% dei valori
- Nell'intervallo [μ-3σ; μ+3σ] troviamo il 99.7% dei valori

Distribuzione normale



z = valore critico o percentile

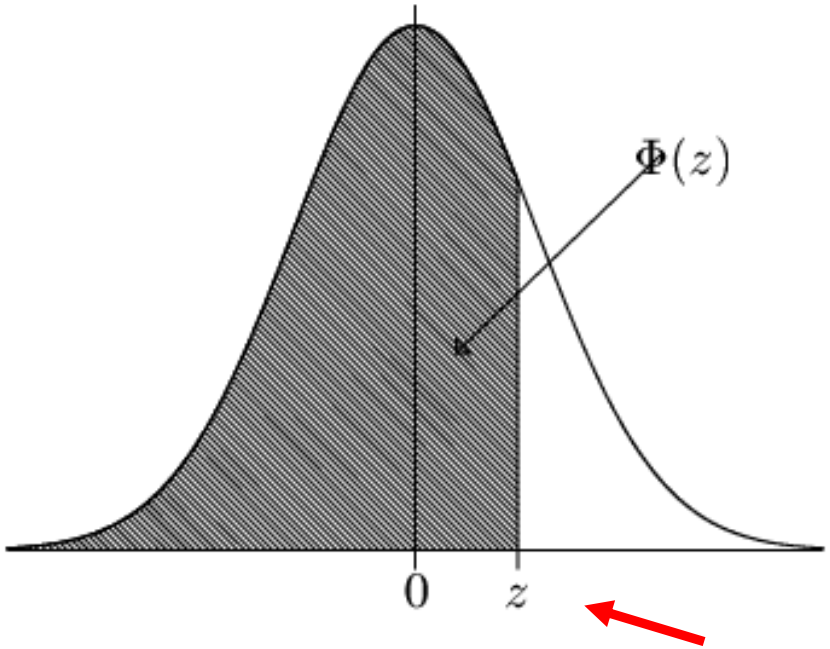
ϕ(z) = p(X<z) probabilità di osservare un valore più piccolo di z (area delimitata da curva, retta verticale passante per z e asse delle ascisse)

Tavole della distribuzione Gaussiana

z	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417
$\Phi(z)$	0.90	0.95	0.975	0.99	0.995	0.999	0.9995	0.99995	0.999995
$2[1-\Phi(z)]$	0.20	0.10	0.05	0.02	0.01	0.002	0.001	0.0001	0.00001

	Seconda cifra decimale di z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643

Valore critico/percentile



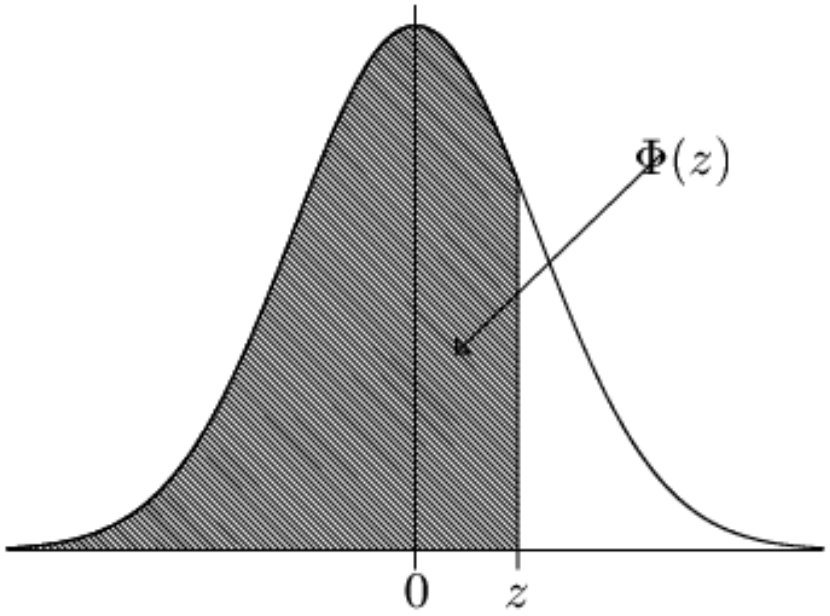
Tavole della distribuzione Gaussiana

z	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417
$\Phi(z)$	0.90	0.95	0.975	0.99	0.995	0.999	0.9995	0.99995	0.999995
$2[1-\Phi(z)]$	0.20	0.10	0.05	0.02	0.01	0.002	0.001	0.0001	0.00001

	Seconda cifra decimale di z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643

Valore critico/percentile

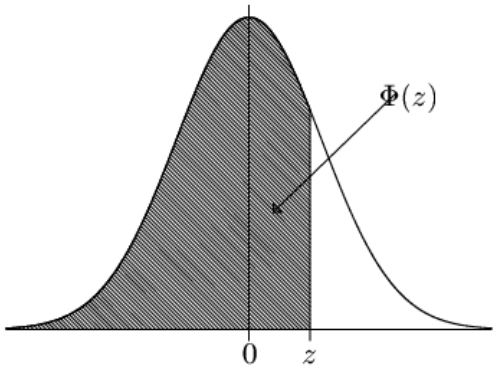
% di area sottostante la curva e a sinistra del valore critico



Esempio:

Calcolare p(X<1.96)

z	Seconda cifra decimale di z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643

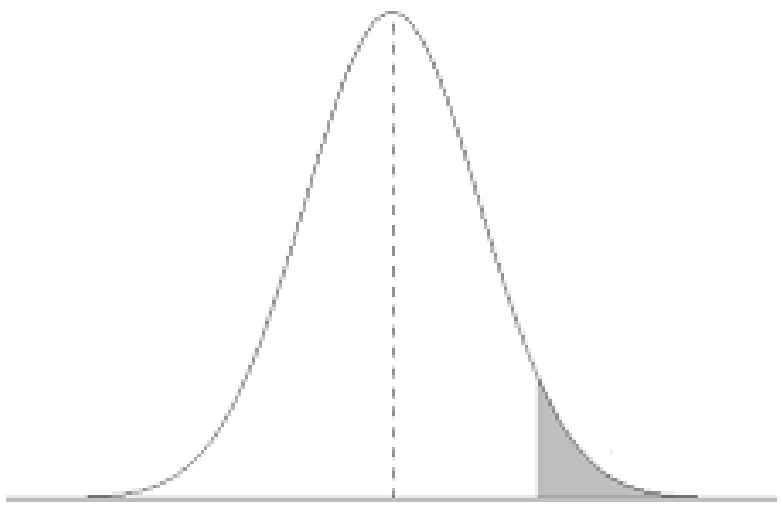


p(X<1.96)=0.975

Esempio:

Calcolare p(X>1.96)

z	Seconda cifra decimale di z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643

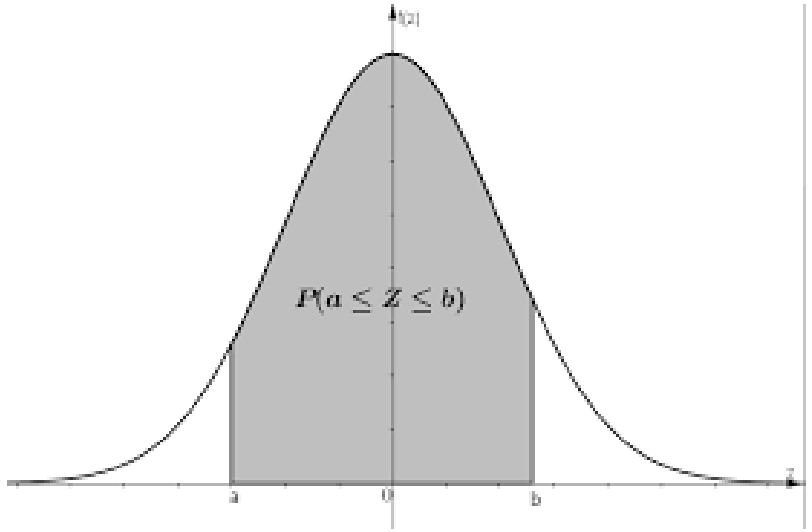


1-p(X<1.96)=0.025

Esempio:

Calcolare p(1.64<X<1.96)

	Seconda cifra decimale di z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643



p(X<1.96)-p(X<1.64)=
0.975-0.9495=0.0255

Test di ipotesi

Che cosa sono le ipotesi e i test di ipotesi?

- Un’ipotesi è una formulazione di una congettura/ un’assunzione sulla popolazione che deve essere verificata e può essere vera o falsa
- Le conclusioni che traiamo sulla popolazione vengono valutate esaminando un campione estratto da quella popolazione
- Le ipotesi fatte sono relative a parametri di quella popolazione
- Il **test di ipotesi** è la procedura che ci permette di stabilire se l’ipotesi è vera o falsa

Ipotesi

Nei test di ipotesi le ipotesi sono due: l’**ipotesi nulla** e l’**ipotesi alternativa**.

L’ipotesi nulla è quella di uguaglianza o mancanza dell’effetto, quella alternativa è quella di differenza/diversità

Ipotesi

L'ipotesi nulla in genere è la negazione dell'ipotesi di ricerca (che coincide perciò con quella alternativa)

Es. Se vogliamo valutare se una moneta è truccata (perciò la probabilità che esca testa non è uguale alla probabilità che esca croce) allora formuleremo l'ipotesi come:

Ipotesi nulla: $p_{testa} = p_{croce}$ Ipotesi alternativa: $p_{testa} \neq p_{croce}$

Oppure

Ipotesi nulla: $p_{testa} = 0.5$ Ipotesi alternativa: $p_{testa} \neq 0.5$

Ipotesi

bilaterale

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$$

unilaterale

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases} \qquad \begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$$

H_0 è l'ipotesi nulla

H_1 è l'ipotesi alternativa

→ Preferiamo rifiutare l'ipotesi nulla a favore di quella alternativa

Ipotesi

bilaterale

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$$

unilaterale

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases} \qquad \begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$$

H_0 è l'ipotesi nulla

H_1 è l'ipotesi alternativa

Non rifiutare H_0 non significa che la condizione descritta dall'ipotesi nulla è vera, ma significa che non abbiamo sufficiente evidenza per rifiutarla!

La statistica test

- Per stabilire se rifiutare o non rifiutare l'ipotesi nulla dobbiamo calcolare la **statistica test**

- In generale la statistica test è calcolata come:

$$T = \frac{stima - valore\ supposto}{errore}$$

- Ogni statistica test ha una distribuzione nota

La statistica test della media

Consideriamo la seguente ipotesi:

«la pressione arteriosa media negli anziani è maggiore di 120»

Come è strutturato il test di ipotesi?

$$\begin{cases} H_0: \mu = 120 \\ H_1: \mu > 120 \end{cases}$$

La statistica test della media

Consideriamo la seguente ipotesi:

«la pressione arteriosa media negli anziani è maggiore di 120»

Che cosa facciamo per condurre la verifica di questa ipotesi??

1. Estraggo un campione dalla popolazione
2. Su tutte le unità valuto la pressione arteriosa
3. Calcolo la statistica test della media che è data da:

$$T = \frac{stima - valore\ supposto}{standard\ error} \rightarrow T = \frac{\bar{X} - 120}{\frac{\sigma}{\sqrt{n}}}$$

4. La statistica test per la media ha una distribuzione Gaussiana di media 0 e deviazione standard 1

La regione di accettazione e rifiuto

- Per capire se rifiutare o non rifiutare l'ipotesi nulla devo vedere se il valore assunto dalla statistica test appartiene alla **regione di accettazione** o alla **regione di rifiuto**
- La regione di accettazione è l'insieme di valori per cui non rifiuto H_0
- La regione di rifiuto è l'insieme di valori per cui rifiuto H_0

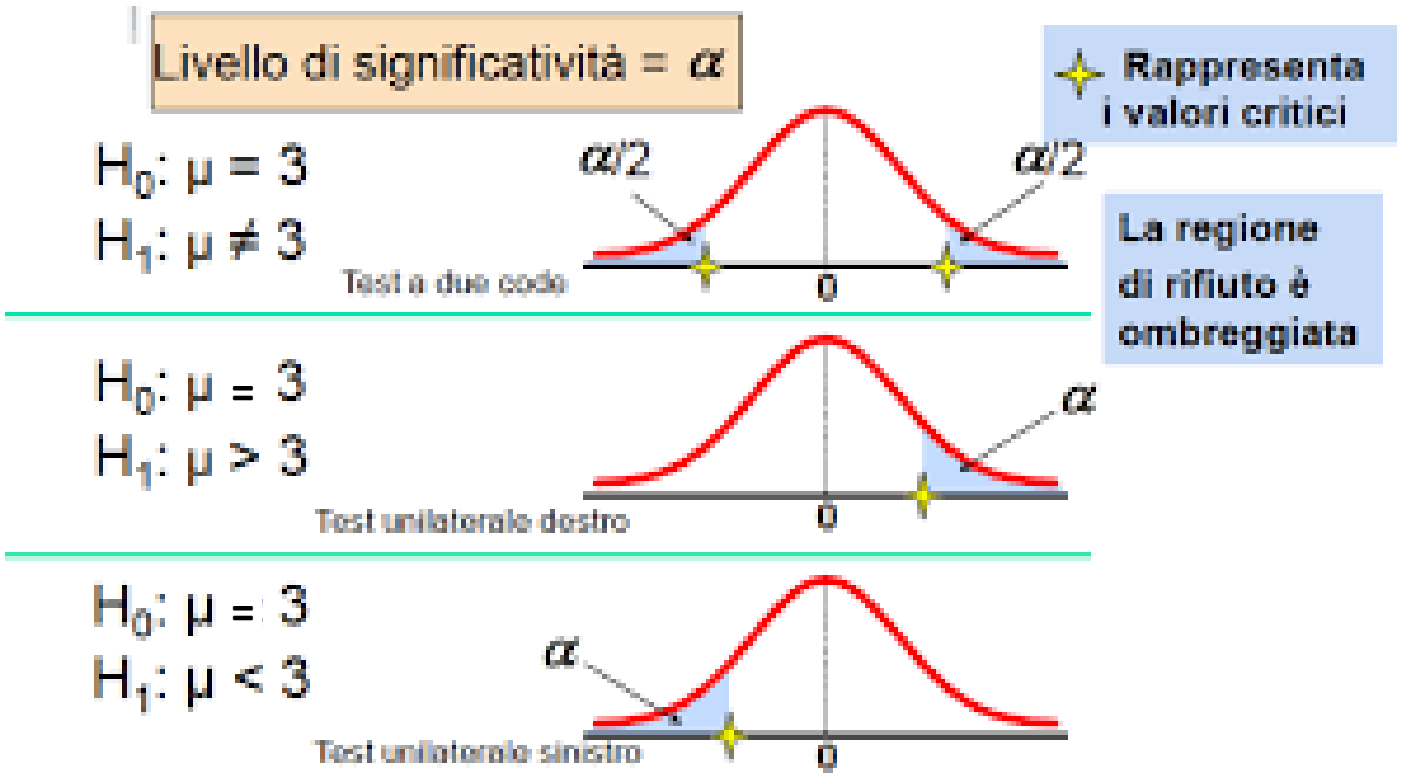
Livello di significatività α

- L’insieme dei valori che appartengono alla regione di rifiuto/di accettazione è definito dal **livello di significatività (α)**

- **Il livello di significatività è una probabilità, cioè la probabilità di rifiutare l’ipotesi nulla quando questa è vera**

- α viene detto anche **errore del 1°tipo**, e in genere assume valore 0.05 ma può assumere anche valore 0.10 o 0.01 in base all’entità dell’errore che siamo disposti a commettere

Livello di significatività α



La collocazione di α segue l'ipotesi alternativa, perciò se nell'ipotesi alternativa troviamo il ≠ posizioneremo metà errore nella coda sinistra e metà nella coda destra, se è presente il < allora α è tutto a sinistra, con il a destra come nell'immagine.

Se il valore della statistica test ricade nell'area bianca allora non rifiuto l'ipotesi nulla, se ricade nella zona celeste rifiuto l'ipotesi nulla

Errore del 2° tipo

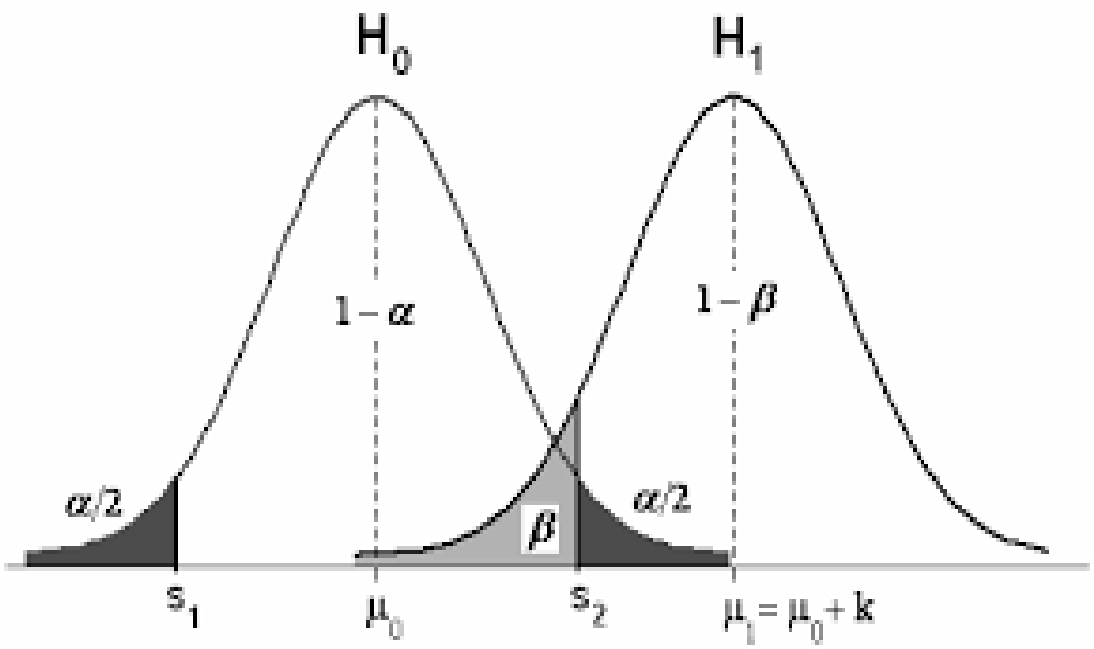
- **errore del secondo tipo** viene indicato con **β**
- indica la probabilità di accettare l'ipotesi nulla quando andrebbe rifiutata
- è un errore meno grave di quello del primo tipo
- È in genere più elevato rispetto a quello del primo tipo

Errore del 2° tipo

- in genere viene settato a 0.20, 0.15 o 0.10. Nella ricerca scientifica 0.20 è il valore più frequentemente utilizzato. Non dovremmo considerare valori più elevati di 0.20

- **β** non è noto quando conduciamo un test (infatti è un parametro che non interviene direttamente)

- 1-**β** viene detta **potenza** e questo è un parametro fondamentale da considerare nella stima della numerosità campionaria



Riassumendo..

Probabilità		Realtà	
		Ipotesi nulla vera	ipotesi alternativa vera
Decisione	non rifiuto l'ipotesi nulla	1-α	errore di II tipo β
	rifiuto l'ipotesi nulla	errore di I tipo α	potenza 1-β

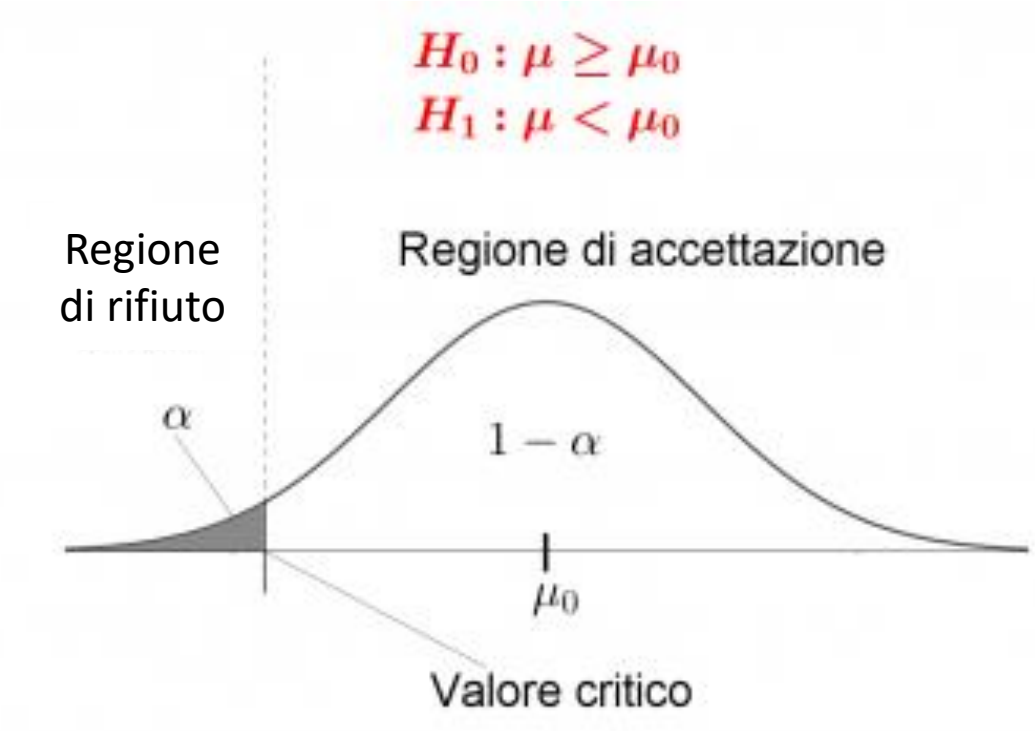
Regione di accettazione e rifiuto

Indichiamo con *z* il percentile che individua α, allora:

- Per ipotesi unilaterale a sinistra (<)

  Regione di accettazione: valori > z

  Regione di rifiuto: valori < z



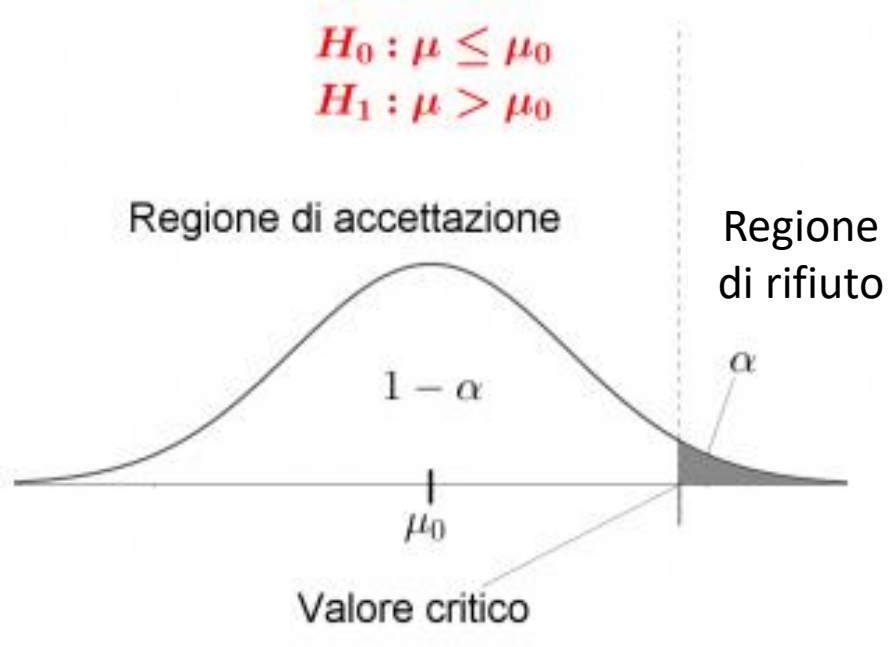
Regione di accettazione e rifiuto

Indichiamo con *z* il percentile che individua α, allora:

- Per ipotesi unilaterale a destra (>)

  Regione di accettazione: valori < z

  Regione di rifiuto: valori > z



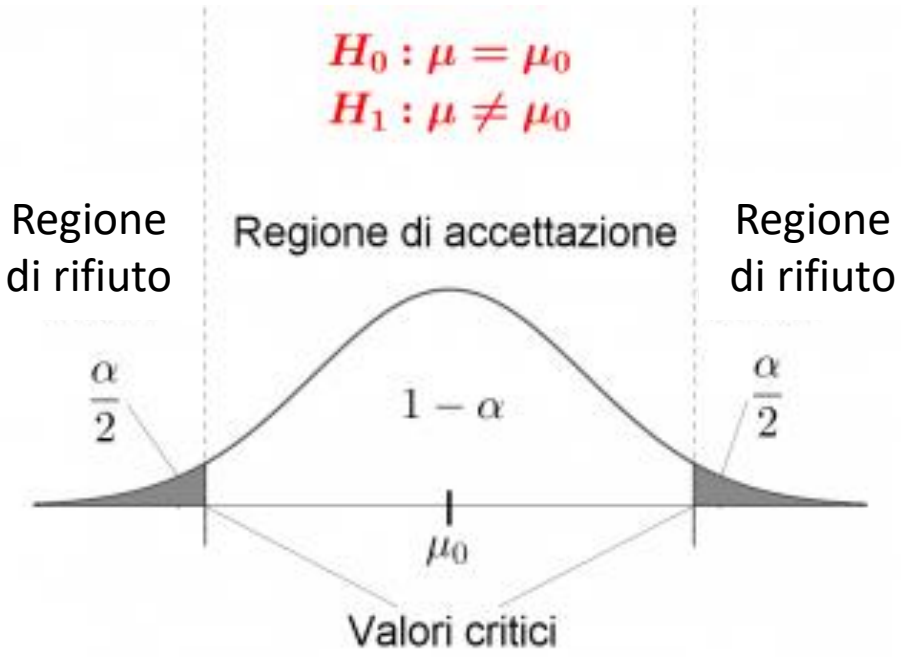
Regione di accettazione e rifiuto

Indichiamo con *z* il percentile che individua α, allora:

- Per ipotesi bilaterale (≠)

  Regione di accettazione: $z_1 <$ valori < z_2

  Regione di rifiuto: valori < z_1 e valori > z_2



La statistica test della media

Consideriamo la seguente ipotesi:

«la pressione arteriosa media negli anziani è maggiore di 120»

Che cosa facciamo per condurre la verifica di questa ipotesi??

1. Stabilire il livello di significatività ad esempio 0.05
2. Individuare il valore critico

z	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417
$\Phi(z)$	0.90	0.95	0.975	0.99	0.995	0.999	0.9995	0.99995	0.999995
$2[1-\Phi(z)]$	0.20	0.10	0.05	0.02	0.01	0.002	0.001	0.0001	0.00001

3. Costruire le due regioni di accettazione/rifiuto e vedo a quale delle due appartiene la statistica test:
   - Se $z<1.645$ non rifiuto l'ipotesi nulla
   - Se $z\geq1.645$ rifiuto l'ipotesi nulla

p-value e decisione del test

- Per stabilire se accettare o rifiutare l’ipotesi nulla possiamo anche calcolare il ***p-value***

- Il p-value è la probabilità di osservare quel risultato ottenuto o un valore più estremo

- Il p-value è l’area individuata dal valore calcolato della statistica test. Anche il p-value viene posizionato in base all’ipotesi alternativa.

p-value e decisione del test

- Il p-value deve essere confrontato con l'errore del primo tipo/livello di significatività e se:

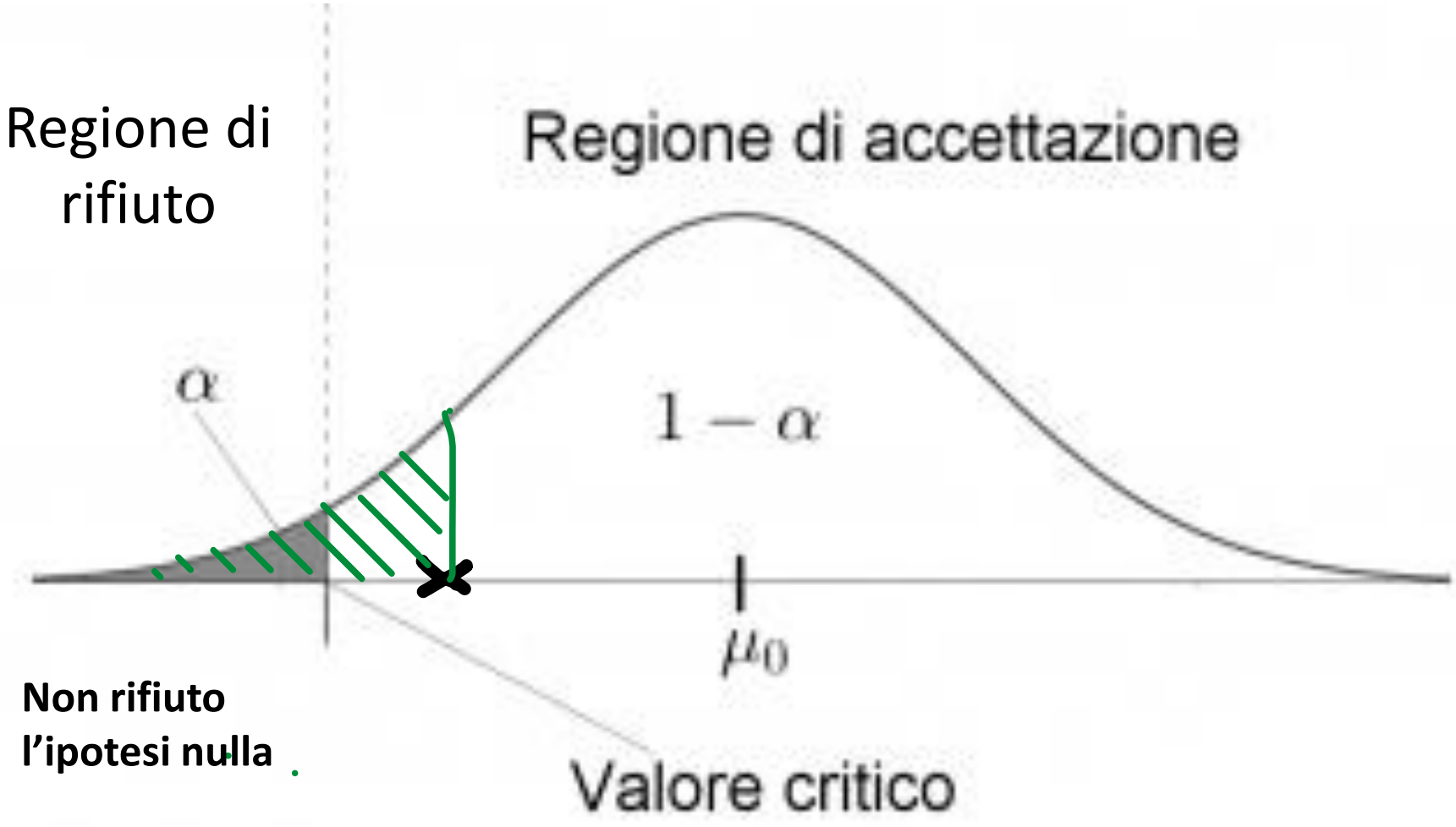
$p<\alpha$ → rifiuto l'ipotesi nulla

$p\geq \alpha$ → non rifiuto l'ipotesi nulla

p-value e decisione del test

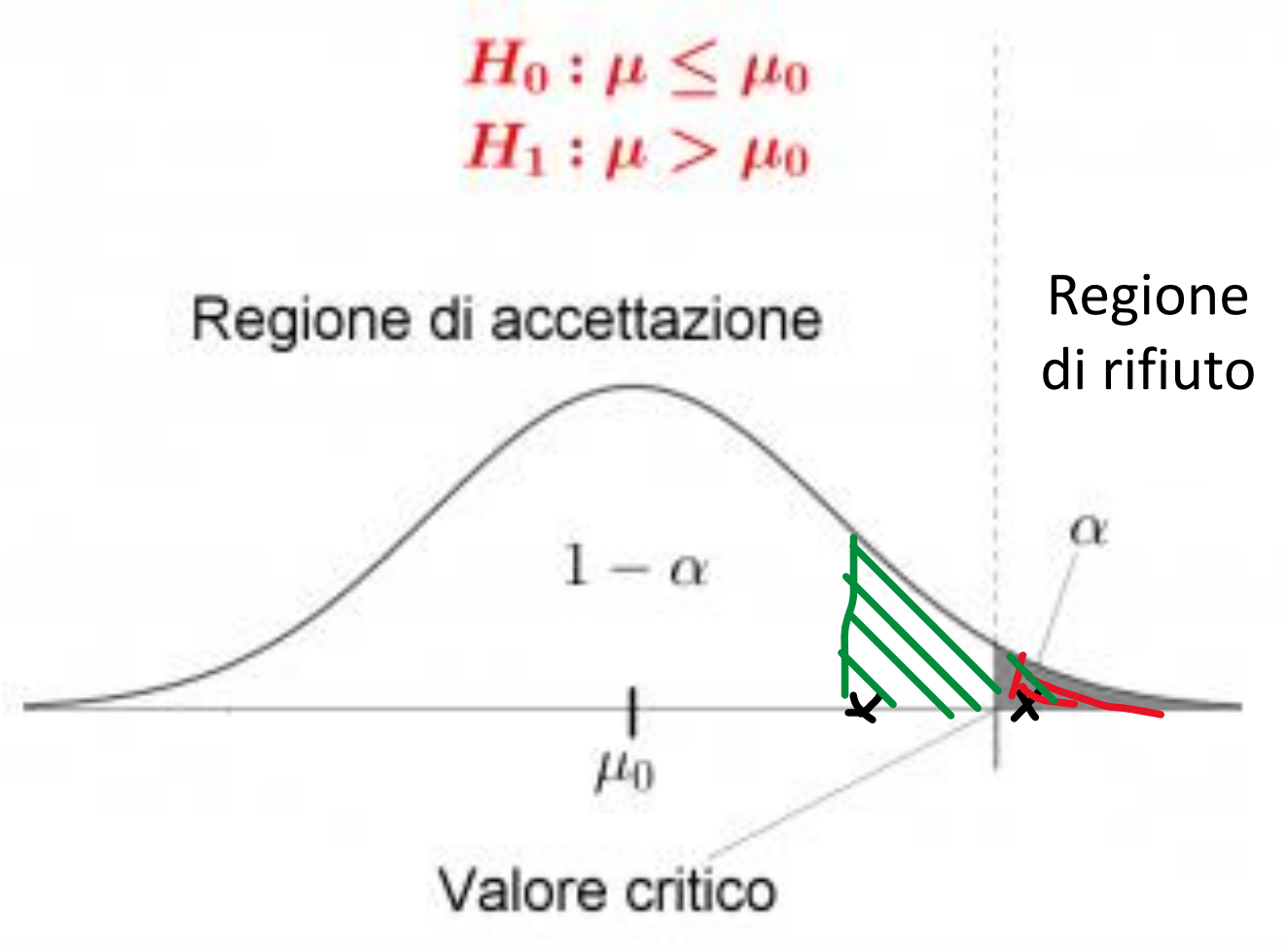
La crocetta indica il valore della statistica test calcolato e il p-value è l'area a sinistra
Rifiuto l'ipotesi nulla

p-value e decisione del test



Non rifiuto l'ipotesi nulla

p-value e decisione del test



Principali ipotesi:

- **Ipotesi sulla distribuzione della variabile**
- **Ipotesi sulle varianze**
- **Ipotesi riguardanti le «frequenze»:**
  - ❑ Ipotesi sulle proporzioni
    - → Vogliamo valutare se il verificarsi di un evento segue una certa proporzione
  - ❑ Ipotesi di associazione
    - → Vogliamo valutare la relazione tra due variabili qualitative
- **Ipotesi riguardanti le medie**
  - ❑ Ipotesi sulla media
    - → Vogliamo vedere se la media di una variabile è uguale o meno a un valore prefissato
  - ❑ Ipotesi sulla differenza tra medie
    - → Vogliamo vedere se le medie di due o più gruppi sono uguali oppure no
- **Ipotesi di correlazione**

Test di ipotesi

I test di ipotesi possono essere suddivisi in due gruppi:

Test parametrici

Test non parametrici o distribution free

I test parametrici sono particolari test per i quali, prima di applicarli, devono essere verificate delle ipotesi stabilite a priori generalmente sulla distribuzione di una variabile.

I test non parametrici invece non hanno ipotesi a priori da verificare e possono essere sempre applicati.

Test di ipotesi

Tuttavia nel caso in cui le ipotesi dei test parametrici siano soddisfatte è preferibile utilizzare questi test data la maggior potenza, cioè a parità di numerosità campionaria sono in grado di evidenziare differenze più piccole come significative.

Ipotesi sulla distribuzione

- Per capire preliminarmente quale potrebbe essere la distribuzione di una variabile potremmo utilizzare l'istogramma

- La distribuzione di una variabile viene verificata attraverso il test di **Kolmogorov-Smirnov.** In genere questo test viene utilizzato per valutare se la distribuzione ha una forma Gaussiana oppure no (ipotesi preliminare di tutti i test parametrici che studieremo).

Ipotesi sulla distribuzione

- L’ipotesi nulla è: la distribuzione empirica/osservata segue la distribuzione teorica presupposta

- L’ipotesi alterntiva è: la distribuzione empirica/osservata non segue la distribuzione teorica presupposta

Ipotesi sulle varianze

L'omogeneità/omoschedasticità delle varianze viene valutata tramite il test di Bartlett o di Levene oppure attraverso una regola pratica

$$\frac{\sigma_{max}^2}{\sigma_{min}^2} < 2$$

Anche questa ipotesi viene utilizzata come ipotesi preliminare di alcuni test parametrici

Ipotesi sulle varianze

- L’ipotesi nulla è: le varianze sono omogenee

- L’ipotesi alternativa è: le varianze non sono omogenee

Ipotesi sulle frequenze

1. Ipotesi sulle proporzioni
2. Ipotesi di associazione

Ipotesi sulle frequenze

Proporzioni

- Uguaglianza a una data proporzione
- Differenza di proporzioni

Associazione tra due variabili

- Test di McNemar
  (tabelle di contingenza 2x2)
- Chi-quadrato
  (tabelle di contingenza 2x2 più grandi di 2x2)
- Chi-quadrato con correzione di Yates
  (tabelle di contingenza 2x2)
- Test esatto di Fisher
  (tabelle di contingenza 2x2)
- Chi-quadrato per il trend
  (tabelle di contingenza 2xK)

Ipotesi sulla proporzione

Vogliamo valutare se il verificarsi di un evento segue una certa proporzione prefissata

L'evento è dicotomico perciò può essere sintetizzato come successo/insuccesso, si/no, avvenuto/non avvenuto

Questa proporzione viene calcolata come:

$$\frac{n^\circ\ di\ volte\ in\ cui\ si\ verifica\ l'evento}{n}$$

Ipotesi sulla proporzione

Per verificare se l'evento rispetta una proporzione prefissata abbiamo due possibilità in base al fatto che il campione è piccolo o grande:

- Se il campione è piccolo possiamo utilizzare il **test binomiale** (confronto diretto del numero di eventi osservato con la distribuzione binomiale)
- Se il campione è sufficientemente grande allora possiamo utilizzare la seguente formula:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Ipotesi sulla proporzione

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Confrontiamo questa statistica test con i valori critici della normale standardizzata!

Dove:

$\hat{p}$= probabilità/proporzione stimata

p_0= proporzione supposta

$\hat{p}(1-\hat{p})$ = varianza della proporzione

Ipotesi sulla proporzione

Esempio: voglio valutare se la moneta è truccata e la probabilità che esce testa è pari a 0.25; considerando che ho lanciato la moneta 22 volte ed è uscita testa 5 volte.

- Con il test binomiale inserisco questi valori nella funzione della distribuzione binomiale e calcolo i p-value
- Con la formula

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \rightarrow \hat{p} = \frac{5}{22} = 0.23 \quad \rightarrow \frac{0.23 - 0.25}{\sqrt{\frac{0.23(1-0.23)}{22}}}$$

Ipotesi sulla proporzione

Nel caso in cui volessimo comparare due proporzioni e capire se c'è differenza allora potremmo utilizzare la formula:

$$z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}}$$

Esempio: vogliamo verificare se la probabilità che esca testa dipende dal fatto che i lanci della moneta vengono fatti con una moneta d'oro o di argento

Ipotesi di associazione

- Coinvolgono due variabili di tipo qualitativo
- La tecnica descrittiva adeguata per sintetizzare i dati è la tabella di contingenza
- La tecnica grafica più adatta è il diagramma a barre

Ipotesi di associazione

Il fumo è associato con al cancro al polmone?

	Cancro al polmone		
Fumo	Si	No	Totale
Si	**52**	**27**	79
No	**34**	**77**	111
Totale	86	104	190

Ipotesi di associazione

Ipotesi nulla: non c’è associazione

La variabile che ha funzione di exposure non produce alcun cambiamento nell’outcome.

Ipotesi alternativa: c’è associazione

La variabile che ha funzione di exposure comporta cambiamenti nelle frequenze dell’outcome.

Ipotesi di associazione

Per capire quale test utilizzare per verificare un'ipotesi di associazione dobbiamo tenere in considerazione i seguenti aspetti:

- Tipologia di variabili qualitative coinvolte
- Misura della tabella di contingenza (numero di categorie per variabile)
- Frequenze attese
- Appaiamento dei dati (i dati vengono definiti appaiati quando sulle stesse unità di analisi viene ripetuta la misurazione in tempi diversi)

Test di Mc Nemar

Test che si applica solamente a tabelle di dimensione 2x2

Test non parametrico

Viene applicato quando i dati sono appaiati, perciò la valutazione della variabile qualitativa viene ripetuta prima e dopo un certo evento/trattamento

Nel caso di frequenze attese basse è preferibile utilizzare il test binomiale

Test di Mc Nemar

Esempio: consideriamo 200 persone che vengono classificate come fumatori/non fumatori prima dell'uso della sigaretta elettronica. Siamo interessati a vedere se dopo un periodo di uso della sola sigaretta elettronica i pazienti tornano o meno a fumare sigarette

	Fumo post	
Fumo pre	SI	NO
SI	90	30
NO	20	60

La somma delle numerosità di queste celle non deve essere piccola

Test del chi-quadrato generico

Le due variabili possono essere di qualunque tipo (policotomiche o ordinali), è sufficiente che la tabella di contingenza associata sia di dimensione maggiore di 2x2

I dati non devono essere appaiati

È piuttosto sensibile alle frequenze attese basse. Le regole pratiche sono:

- L’80% delle frequenze attese deve essere più grande di 5
- Tutte le frequenze attese devono essere maggiori di 1

Test del chi-quadrato generico

Esempio: il tipo di trattamento (combinazione di morfina con antidolorifici o solo morfina o solo antidolorifici) è associato con il dolore (basso, medio o alto)?

Frequenze osservate	**Dolore**			
Terapia	Basso	Medio	Alto	Totale
Morfina	12	34	27	73
Antidolorifici	16	10	3	29
Morfina+ antidolorifici	14	27	31	72
Totale	42	71	61	174

Frequenze attese	**Dolore**			
Terapia	Basso	Medio	Alto	Totale
Morfina	18	30	26	73
Antidolorifici	7	12	10	29
Morfina+ antidolorifici	17	29	25	72
Totale	42	71	61	174

Test del chi-quadrato generico

Esempio: il tipo di trattamento (combinazione di morfina con antidolorifici o solo morfina o solo antidolorifici) è associato con il dolore (basso, medio o alto)?

Frequenze osservate	**Dolore**			
Terapia	Basso	Medio	Alto	Totale
Morfina	2	2	8	12
Antidolorifici	10	10	3	23
Morfina+ antidolorifici	5	7	1	13
Totale	17	19	12	48

Frequenze attese	**Dolore**			
Terapia	Basso	Medio	Alto	Totale
Morfina	4	5	3	12
Antidolorifici	8	9	6	23
Morfina+ antidolorifici	5	5	3	13
Totale	17	19	12	48

Test del chi-quadrato generico

Esempio: il tipo di trattamento (combinazione di morfina con antidolorifici o solo morfina o solo antidolorifici) è associato con il dolore (basso, medio o alto)?

L'80% di 9 celle è 7 ma in questo caso sono solamente 6 le celle con frequenze più elevate di 5

→ Non possiamo utilizzare il chi-quadrato generico

→ L'unica alternativa che abbiamo è accorpare le categorie di significato simile

Frequenze attese	**Dolore**			
Terapia	Basso	Medio	Alto	Totale
Morfina	4	5	3	12
Antidolorifici	8	9	6	23
Morfina+ antidolorifici	5	5	3	13
Totale	17	19	12	48

Test del chi-quadrato generico

In questo esempio non sarebbe stato sensato accorpare le terapie perché molto diverse.

Abbiamo accorpato le categorie del dolore e abbiamo risolto il problema delle frequenze attese.

Frequenze attese	**Dolore**		
Terapia	Basso	Medio-Alto	Totale
Morfina	4	8	12
Antidolorifici	8	15	23
Morfina+ antidolorifici	5	8	13
Totale	17	31	48

Test del chi-quadrato con correzione di Yates

Questo test viene applicato solamente per tabelle di contingenza di dimensione 2x2, perciò le variabili sono entrambe dicotomiche.

I dati non sono appaiati

Questo test in genere viene utilizzato quando le osservazioni non sono inferiori a 40

Test esatto di Fisher

Questo test viene applicato per tabelle di contingenza di dimensione 2x2 , perciò le variabili sono entrambe dicotomiche.

I dati non sono appaiati

Questo test può essere sempre applicato perché non è influenzato dalle frequenze attese basse.

Test esatto di Fisher

Questo test si definisce «esatto» perché calcola la probabilità esatta di ottenere la tabella osservata fra tutte quelle che hanno gli stessi valori di totali di riga e colonna (è per questo che viene utilizzato per frequenze osservate basse)

Es

13	23	**36**
7	12	**19**
20	**35**	

11	25	**36**
9	10	**19**
20	**35**	

Queste due tabelle hanno frequenze assolute diverse ma i totali di riga e colonna uguali

Ne esiste un’approssimazione per tabelle di contingenza più grandi di 2x2

Test del chi quadrato per il trend

È detto anche test di Armitage

Le variabili sono una di tipo dicotomico e l’altra di tipo ordinale

La tabella di contingenza associata ha dimensione 2xk

Vogliamo valutare se esiste un trend lineare nelle proporzioni/frequenze dell’evento all’aumentare/diminuire della variabile ordinale

Test del chi quadrato per il trend

Esempio: verificare se il numero di cavie morte aumenta dopo somministrazione di dosi crescenti di un certo farmaco sperimentale

	0.1g/l	0.15g/l	0.20g/l	0.25g/l
Morto	19	29	24	20
Vivo	497	560	269	147
	0.037	0.049	0.082	0.120

Ci sono altri tipi di analisi diverse dall'associazione che vengono svolte sulle tabelle di contingenza (di dimensione 2x2 o più grandi ma sempre quadrate.)

Queste analisi dipendono dall'obiettivo di ricerca:

- Se vogliamo valutare la performance di uno strumento o di una classificazione, la capacità di individuare delle condizioni, dobbiamo fare analisi di sensibilità e specificità (gradi di errore)
  Es. confronto tra diagnosi clinica del melanoma e l'istologia

- Se invece vogliamo capire l'affidabilità di una classificazione o di uno strumento utilizziamo analisi di concordanza.
  Es. vogliamo capire se due ricercatori hanno lo stesso giudizio sulla benignità/malignità di un nodulo utilizzando un'ecografia

Concordanza

Valuta l’affidabilità di una classificazione.

Valuta il **grado di accordo** tra due valutatori o tra un valutare/medico e uno strumento o tra due strumenti o valutazioni dello stesso soggetto fatte in tempi differenti.

Gli strumenti/valutatori devono dare una valutazione sullo stesso fenomeno.

Concordanza

Il coefficiente che misura il grado di accordo è la **kappa di Cohen**

In base al suo valore viene misurato il grado di accordo

Se:

- $k<0$ non c'è concordanza
- $0<k<0.4$ concordanza scarsa
- $0.4<k<0.6$ concordanza discreta
- $0.6<k<0.8$ concordanza buona
- $0.8<k<1$ concordanza ottima

Concordanza

$$k = \frac{accordo(osservato) - accordo(atteso)}{1 - accordo(atteso)}$$

Frequenze osservate

		Osservatore 2		Totale
		SÌ	NO	
Osservatore 1	SÌ	*a*	*b*	**a + b**
	NO	*c*	*d*	**c + d**
Totale		**a + c**	**b + d**	**N**

Accordo osservato $A_o = \frac{a+d}{N}$

Frequenze attese

		Osservatore 2		Totale
		SÌ	NO	
Osservatore 1	SÌ	$\frac{(a+c)*(a+b)}{N}$	$\frac{(b+d)*(a+b)}{N}$	**a + b**
	NO	$\frac{(a+c)*(c+d)}{N}$	$\frac{(b+d)*(c+d)}{N}$	**c + d**
Totale		**a + c**	**b + d**	**N**

Accordo atteso $A_a = \frac{[(a+c)*(a+b)]+[(b+d)*(c+d)]}{N^2}$

Concordanza

Esempio. Valutazione della presenza/assenza di lesioni addominali valutate da due medici sulla base di tomografie computerizzate effettuate su 300 pazienti

	Frequenze osservate			
	Presenza	Assenza	Totale	
Presenza	14	20	34	
Assenza	24	242	266	
Totale	38	262	300	
	Frequenze attese			
	Presenza	Assenza	Totale	
Presenza	4	30	34	
Assenza	34	232	266	
Totale	38	262	300	

accordo osservato=(14+242)/300=0.85

accordo atteso=(4+232)/300=0.79

$$k = \frac{0.85-0.79}{1-0.79} = 0.31$$

Ipotesi sulle differenze

Ipotesi sulle medie

Singola media

- Test t di Student per uguaglianza a una data media
- Test di Mann-Whitney per uguaglianza a una data media

Differenza tra medie

- Test t di Student (per dati appaiati o indipendenti)
- Test di Mann-Whitney
- Test di Wilcoxon
- Test ANOVA (per dati appaiati o indipendenti)
- Test Kruskall-Wallis
- Test di Friedmann

Nota: In blu i test parametrici, in verde i test non parametrici

Ipotesi su una singola media

Ipotesi nulla: la media è uguale a un valore prefissato

Ipotesi alternativa: la media è diversa/maggiore/minore di un valore prefissato

Ipotesi su una singola media

- Possiamo utilizzare il test t di Student o il test di Mann-Whitney
- Il primo test è di tipo parametrico e l’unica ipotesi da soddisfare è la normalità della distribuzione
- Il secondo test è un equivalente non parametrico

Test t di Student

La formula di questo test è:

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

Il risultato deve essere confrontato con i valori critici della distribuzione normale standard

$\bar{X}$ = media aritmetica

μ_0 = valore supposto

$\frac{\sigma}{\sqrt{n}}$ = standard error della media

Ipotesi su una singola media

Esempio: Da letteratura so che il valore medio del magnesio nei soggetti adulti è pari a 2.4. Dato che sto testando un nuovo farmaco, voglio verificare se questo tende ad aumentare o diminuire i livelli di magnesio.

Test di ipotesi: $\begin{cases} H_0: \mu = 2.4 \\ H_1: \mu \neq 2.4 \end{cases}$

Ipotesi su una singola media

id	magnesio
1	1,54
2	1,76
3	1,33
4	1,35
5	1,49
6	1,90
7	1,95
8	1,27
9	1,97
10	1,47
11	1,52
12	1,58
13	1,58
14	1,97
15	2,00
16	1,92
17	1,70
18	1,96
19	1,17

$\bar{X}$ = 1.65

μ_0 = 2.4

$\frac{\sigma}{\sqrt{n}} = \frac{0.27}{\sqrt{19}} = 0.06$

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{1.65 - 2.4}{0.06} = -12.09$$

Ipotesi su una singola media

Test di ipotesi: $\begin{cases} H_0: \mu = 2.4 \\ H_1: \mu \neq 2.4 \end{cases}$

Stiamo svolgendo un test bilaterale e decido che α=0.05

→i valori critici sono -1.96 e +1.96

→-12.09<-1.96 perciò rifiuto l'ipotesi nulla perché la statistica test ricade nella regione di rifiuto.

Ipotesi su una singola media

Dai valori ottenuti ottengo un valore di p-value<0.001

→Il p-value ottenuto è più piccolo di 0.05 (l'alpha stabilito)

→Rifiuto l'ipotesi nulla a favore di quella alternativa

Ipotesi su una singola media

Il farmaco riduce in modo **significativo** la concentrazione media del magnesio!!

(la differenza osservata non è frutto del caso)

Test per la differenza tra medie

Ipotesi nulla: non c’è differenza tra le medie

L’ipotesi alternativa cambia in base al numero di medie da confrontare

Se sono 2:
Ipotesi alternativa: le due medie sono differenti

Se sono 3 o più di 3:
Ipotesi alternativa: esiste **almeno** una media differente dalle altre

Test per la differenza tra medie

Esempi:

in media gli uomini sono più alti delle donne?

Un farmaco produce una diminuzione media dei livelli di magnesio?
(dovremo confrontare chi non utilizza il farmaco e chi lo utilizza)

Test per la differenza tra medie

Necessitano 2 variabili: una quantitativa (outcome) e una qualitativa (exposure)

Test t di Student, test di Mann-Whitney, test di Wilcoxon vengono applicati con variabili qualitative dicotomiche

ANOVA, Friedmann e Kruskal-Wallis vengono applicati con variabili qualitative policotomiche

Test per la differenza tra medie

Test parametrici		**Test non parametrici**
Test t di Student per dati non appaiati	→	Test di Mann-Whitney
Test t di Student per dati appaiati	→	Test di Wilcoxon

Test per la differenza tra medie

Test parametrici		**Test non parametrici**
Test ANOVA per dati non appaiati	→	Test di Kruskal-Wallis
Test ANOVA per dati appaiati	→	Test di Friedmann

Test per la differenza tra medie

Per il test t di Student per la differenza tra medie e il test ANOVA i requisiti da soddisfare sono:

1. Distribuzione normale della variabile in ogni gruppo

2. Omoschedasticità delle varianze tra i gruppi

Test per la differenza tra medie

Prima di applicare il test t di Student o il test ANOVA è necessario valutare:

Il test di Kolmogorov-Smirnov per verificare che la distribuzione sia normale.

→ Viene applicato tante volte quanti sono i gruppi (le categorie della variabile quaitativa)

Il test di Bartlett/Levene per l'omoschedasticità delle varianze

Test post hoc

Per i test per la differenza di 3 o più medie l’ipotesi alternativa è data da: esiste almeno una media differente dalle altre

Per capire qual è/quali sono le medie differenti dobbiamo effettuare dei test detti ***post hoc.***

Li effettuiamo **solamente** quando rifiutiamo l’ipotesi nulla del test ANOVA/Kruskal-Wallis/Friedmann

Test post hoc

Primo test		**Test post hoc**
ANOVA	→	Procedura di Tukey
Kruskal-Wallis	→	Test di Dunn

Test post hoc

Quando valutiamo la differenza tra tre medie e vogliamo capire quali sono quelle differenti dovremmo fare un certo numero di confronti a coppie

Per esempio per il confronto di 3 medie dovremmo fare 3 confronti:

1-2 1-3 2-3

Per ciascun confronto commettiamo un errore α, perciò alla fine l'errore commesso sarebbe il triplo di quello stabilito!

I test post hoc ci permettono di gestire questo problema lasciando come livello generale di errore pari a α

Test per la differenza tra medie

Dal punto di vista descrittivo possiamo sintetizzare i dati attraverso:

- Media e deviazione standard calcolati su ciascun gruppo definito dalla variabile qualitativa

- Grafico box-whisker

Test per la differenza tra medie

Esempio:

Un farmaco produce una diminuzione media dei livelli di magnesio?

ID	farmaco	magnesio
1	si	1,93
2	si	1,10
3	si	1,69
4	si	1,06
5	si	1,95
6	si	0,96
7	si	0,92
8	si	0,62
9	si	1,17
10	si	1,18
11	si	1,11
12	no	2,84
13	no	2,56
14	no	2,30
15	no	3,59
16	no	2,43
17	no	3,28
18	no	3,12
19	no	2,40

Test per la differenza tra medie

Per sintetizzare i dati possiamo calcolare i valori di media e deviazione standard nei due gruppi

Farmaco si: 1.24±0.43 [media ± deviazione standard]

Farmaco no: 2.82±0.47 [media ± deviazione standard]

Test per la differenza tra medie

ID	farmaco	magnesio
1	si	1,93
2	si	1,10
3	si	1,69
4	si	1,06
5	si	1,95
6	si	0,96
7	si	0,92
8	si	0,62
9	si	1,17
10	si	1,18
11	si	1,11
12	no	2,84
13	no	2,56
14	no	2,30
15	no	3,59
16	no	2,43
17	no	3,28
18	no	3,12
19	no	2,40

Abbiamo una variabile qualitativa dicotomica e una variabile quantitativa.

I dati non sono appaiati, perciò possiamo applicare il test t di Student o il test di Mann Whitney

→dobbiamo verificare i requisiti per capire quale dei due poter applicare

Test per la differenza tra medie

Dobbiamo dividere la variabile quantitativa in due gruppi identificati dalla variabile qualitativa

ID	farmaco	magnesio
1	si	1,93
2	si	1,10
3	si	1,69
4	si	1,06
5	si	1,95
6	si	0,96
7	si	0,92
8	si	0,62
9	si	1,17
10	si	1,18
11	si	1,11

Ha una distribuzione normale?

ID	farmaco	magnesio
12	no	2,84
13	no	2,56
14	no	2,30
15	no	3,59
16	no	2,43
17	no	3,28
18	no	3,12
19	no	2,40

Ha una distribuzione normale?

Test per la differenza tra medie

Dobbiamo dividere la variabile quantitativa in due gruppi identificati dalla variabile qualitativa

ID	farmaco	magnesio
1	si	1,93
2	si	1,10
3	si	1,69
4	si	1,06
5	si	1,95
6	si	0,96
7	si	0,92
8	si	0,62
9	si	1,17
10	si	1,18
11	si	1,11

ID	farmaco	magnesio
12	no	2,84
13	no	2,56
14	no	2,30
15	no	3,59
16	no	2,43
17	no	3,28
18	no	3,12
19	no	2,40

Calcolo la varianza

Calcolo la varianza

Sono omogenee?

Quick Statistics Calculators

socscistatistics.com/tests/

App idscoreproject SimPLE Project Dart Programming... Scopus - Author de... ORCID labitem Sum

Statistics Calculators

Here you'll find a set of statistics calculators that are intuitive and easy to use. Included are a variety of tests of significance, plus correlation, effect size and confidence interval calculators.

If you're not sure what statistics calculator you require, check out our Which Statistics Test? wizard.

Significance Tests

- One-Way ANOVA Calculator for Independent Measures
- One-Way ANOVA Calculator for Repeated Measures
- Binomial Test Calculator
- Chi-Square Calculator for 2 x 2 Contingency Table
- Chi-Square Calculator for 5 x 5 (or less) Contingency Table
- Chi-Square Calculator for Goodness of Fit
- Fisher Exact Test Calculator for 2 x 2 Contingency Table
- The Friedman Test for Repeated Measures
- The Kolmogorov-Smirnov Test of Normality
- Kruskal-Wallis Test Calculator for Independent Measures
- Levene's Test of Homogeneity of Variance Calculator
- Mann-Whitney U Test Calculator
- Sign Test Calculator
- Standard Error Calculator
- T-Test Calculator for 2 Independent Means
- T-Test Calculator for 2 Dependent Means

Per effettuare questi test potete utilizzare:

← Calcolatori online

- Excel (chi-quadrato, test t, ANOVA, correlazione)
- Software (SPSS, STATA, R, SAS, Stat View, GraphPad)

The Kolmogorov-Smirnov Test of Normality

Success!

Interpreting the Result

The test statistic (D), which you'll see below, provides a measurement of the divergence of your sample distribution from the normal distribution. The higher the value of D, the less probable it is that your data is normally distributed. The *p*-value quantifies this probability, with a low probability indicating that your sample diverges from a normal distribution to an extent unlikely to arise merely by chance. Put simply, high D, low *p*, is evidence that your data *is not* normally distributed.

It's also worth taking a look at the figures provided for skewness and kurtosis. The nearer both these are to zero, the more likely it is that your distribution is normal.

Your Data

```
1.93
1.1
1.69
1.06
1.95
0.96
0.92
0.62
1.17
1.18
1.11
```

Distribution Summary

Count : 11

Mean: 1.24455

Median: 1.11

Standard Deviation: 0.427209

Skewness: 0.70271

Kurtosis: -0.426292

Result: The value of the K-S test statistic (D) is .29023.

The *p*-value is .25807. Your data does *not* differ significantly from that which is normally distributed.

Your Data

2.84
2.56
2.3
3.59
2.43
3.28
3.12
2.4

Distribution Summary

Count : 8

Mean: 2.815

Median: 2.7

Standard Deviation: 0.472259

Skewness: 0.565503

Kurtosis: -1.161635

Test di Kolmogorov-Smirnov sul secondo gruppo

Result: The value of the K-S test statistic (D) is .21811.

The *p*-value is .76824. Your data does *not* differ significantly from that which is normally distributed.

POSSO APPLICARE IL TEST T DI STUDENT PER DATI NON APPAIATI!!

Test di Levene

The *f*-ratio value is 0.29165. The *p*-value is .596168. The result is *not* significant at $p < .05$.

The requirement of homogeneity is met.

T-Test Calculator for 2 Independent Means

Enter the values for your two treatment conditions into the text boxes below, either one score per line or as a comma delimited list. Select your significance level and whether your hypothesis is one or two-tailed. Then give your data a final check, and press the "Calculate T and P Values" button.

Treatment 1 (*X*)

```
1.93
1.1
1.69
1.06
1.95
0.96
0.92
0.62
1.17
1.18
1.11
```

Treatment 2 (*X*)

```
2.84
2.56
2.3
3.59
2.43
3.28
3.12
2.4
```

Significance Level:

- ○ .01
- ◉ .05
- ○ .10

One-tailed or two-tailed hypothesis?:

- ○ One-tailed
- ◉ Two-tailed

No calculation has yet been performed.

Calculate T and P Values | Reset

Valore della statistica test

p-value

Decisione (in questo caso rifiutiamo l'ipotesi nulla→ il farmaco riduce i livelli medi di magnesio)

The *t*-value is -7.57275. The *p*-value is < .00001. The result is significant at *p* < .05.

Note: If you wish to calculate the effect size, this calculator will do the job.

Want to know how to report this *t*-test result in your work? (Opens in a new tab so you don't lose your calculation.)

How to report a *t*-test result (APA)

C'è una differenza *statisticamente significativa.* Significa che la differenza osservata tra la concentrazione del magnesio con e senza farmaco non è frutto del caso (ovvero non dipende dai soggetti selezionati) ma siamo confidenti che questa differenza esista realmente.

Ipotesi di correlazione

Ipotesi di correlazione

- Coinvolge due variabili di tipo quantitativo
- Una ha funzione di exposure e una di outcome
- La tecnica descrittiva che viene utilizzata è lo scatterplot

Ipotesi di correlazione

Ipotesi nulla: non c’è correlazione

Ipotesi alternativa: c’è correlazione

Ma che cosa è la correlazione?

Ipotesi di correlazione

La correlazione misura la forza della relazione tra le variabili

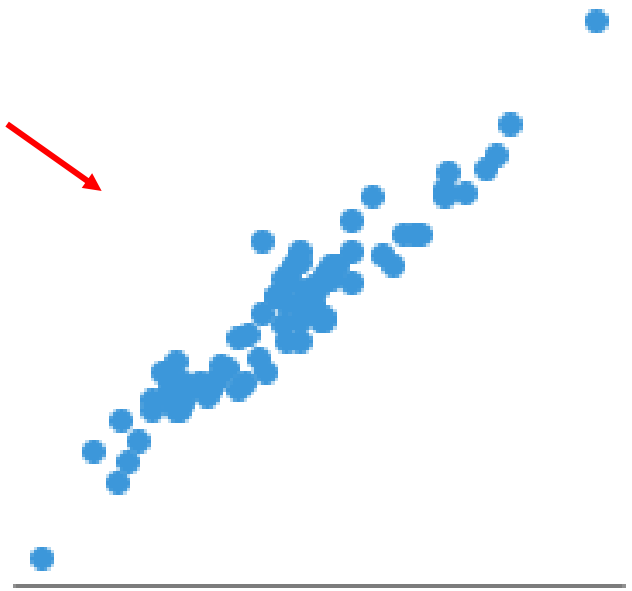
In modo visivo e pratico possiamo dire che la correlazione misura quanto le osservazioni sono vicine a una linea di tendenza

La linea di tendenza può essere una retta (**correlazione lineare**) o una curva (**correlazione non lineare**)

Ipotesi di correlazione

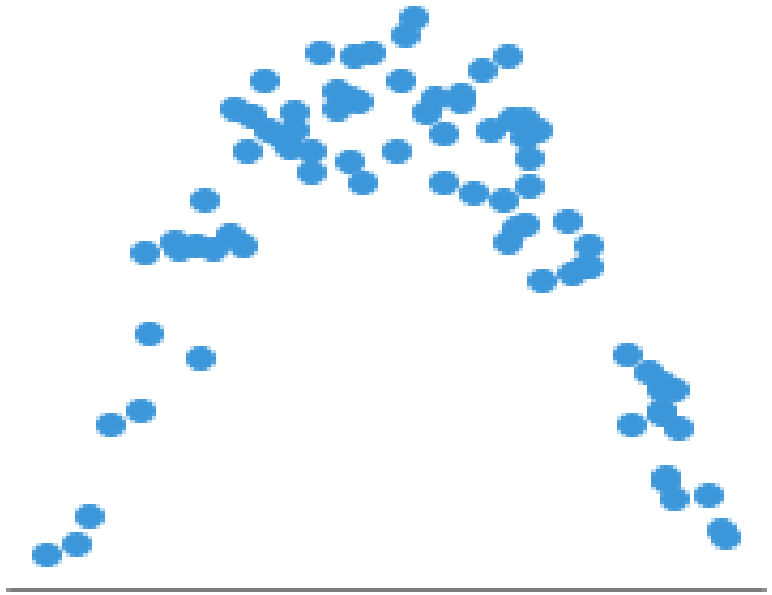
Correlazione lineare

Ci concentreremo su questi tipo di correlazione



strong, positive, linear

Correlazione non lineare

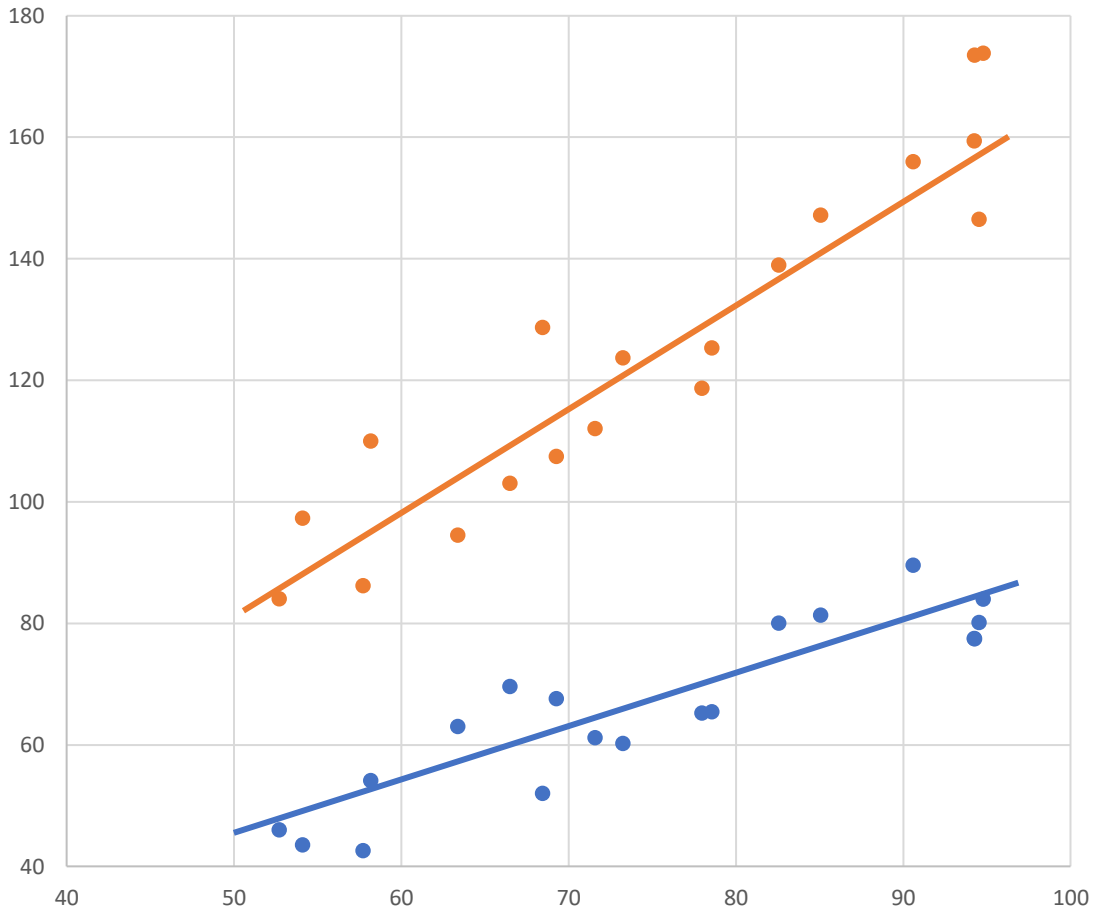


strong, non-linear

Ipotesi di correlazione

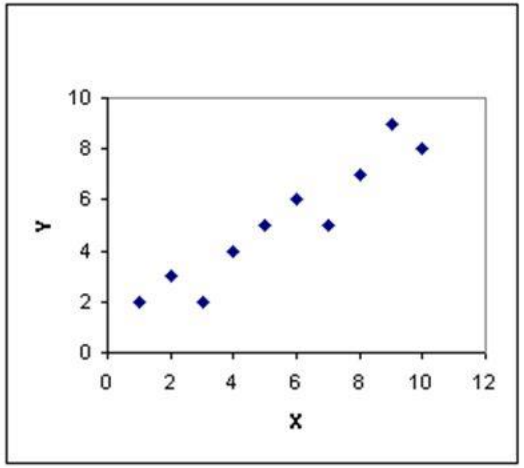
- Tuttavia non è importante l'equazione della linea di tendenza centrale ma solo quanto le osservazioni sono vicine ad essa.

- Due linee di tendenza centrale potrebbero avere la stessa correlazione

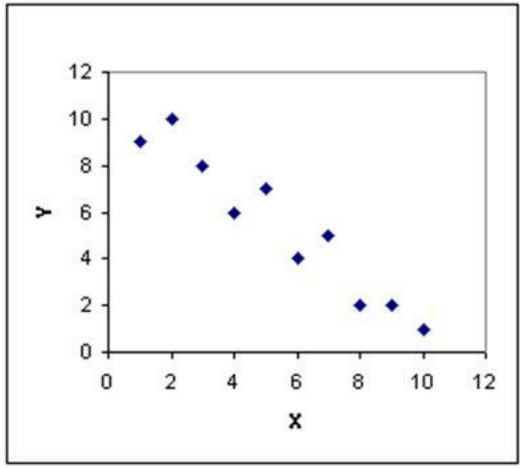


Ipotesi di correlazione

Correlazione positiva



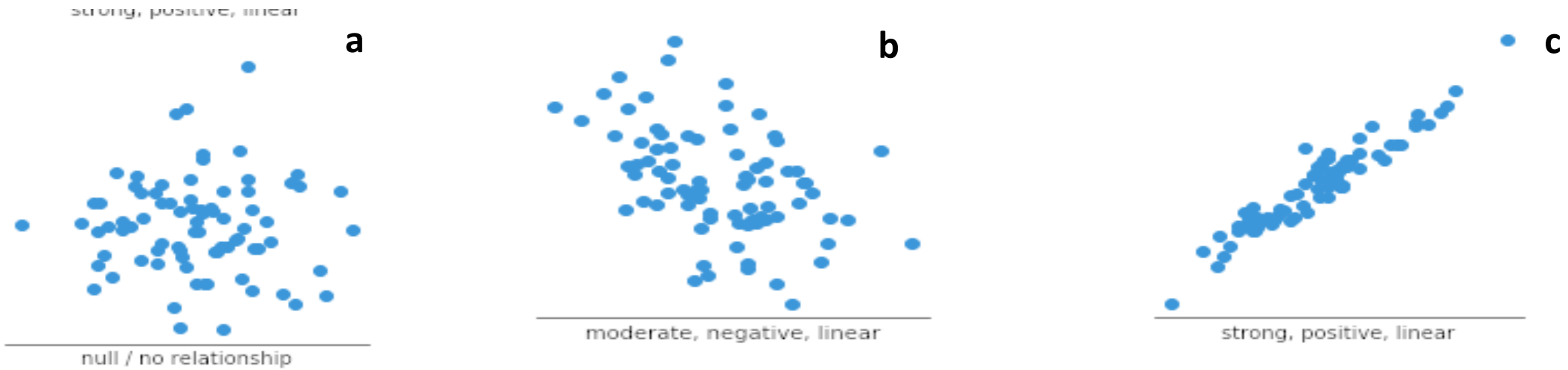
Correlazione negativa



- La correlazione può essere **positiva** o **negativa**

- **POSITIVA (o diretta):** all'aumentare dell'exposure aumenta l'outcome
- **NEGATIVA (o inversa):** all'aumentare della variabile di exposure diminuisce l'outcome

Ipotesi di correlazione



a) Non c'è correlazione, i punti sono una «nuvola casuale»
b) C'è una correlazione media, intravediamo un andamento e i punti iniziano ad avvicinarsi
c) C'è una correlazione elevata (quasi perfetta), i punti sono molto vicini ad una linea di tendenza centrale

Ipotesi di correlazione

Coefficienti di correlazione lineare *r*

Di Pearson

Di Spearman

Di Kendall

- I 3 coefficienti assumono valori tra -1 e 1
- Più il coefficiente si avvicina agli estremi più la correlazione è forte
- A 0 non c'è correlazione
- Per valori compresi tra -1 e 0 la correlazione è negativa
- Per valori compresi tra 0 e 1 la correlazione è positiva

Ipotesi di correlazione

Coefficienti di correlazione lineare *r*

Di Pearson

Di Spearman

Di Kendall

- Se $|r|<0.3$ la correlazione è **debole**
- Se $|r|<0.7$ la correlazione è **moderata**
- Se $|r|>0.7$ la correlazione è **forte**

Tuttavia nella clinica questa classificazione non è sempre rispettata infatti in alcuni campi correlazioni di 0.4 sono considerate forti

Ipotesi di correlazione

Coefficienti di correlazione lineare *r*

Di Pearson

Di Spearman

Di Kendall

- Il coefficiente di correlazione di Pearson fa parte delle tecniche parametriche
- Il requisito da soddisfare è che la distribuzione «congiunta» delle variabili sia normale (normale bivariata)
- Dato che questo requisito è più complesso da verificare, accettiamo che entrambe le variabili (separatamente) siano distribuite in modo normale

Ipotesi di correlazione

Coefficienti di correlazione lineare *r*

Di Pearson

Di Spearman

Di Kendall

- I coefficienti di correlazione di Spearman e di Kendall sono tecniche non parametriche

- Il coefficiente di correlazione di Spearman è sensibile ai valori uguali (**ties**) perché si basa sull'ordinamento. Potremmo ottenere un valore di correlazione un po' più elevato di quello che è realmente.

  ES. 1,1,1,2,2,2,2,2,3,3,3,3,4,4

Ipotesi di correlazione

Coefficienti di correlazione lineare *r*

Di Pearson

Di Spearman

Di Kendall

Questi coefficienti non sono test! Perciò non parliamo di test di Pearson o di Spearman

Tuttavia ci sono dei test che ci permettono di dire se la correlazione osservata e calcolata tramite i coefficienti di correlazione è frutto del caso oppure no.

Quando calcolerete questi coefficienti con i software otterrete anche il p-value del relativo test

Ipotesi di correlazione

Esempio:

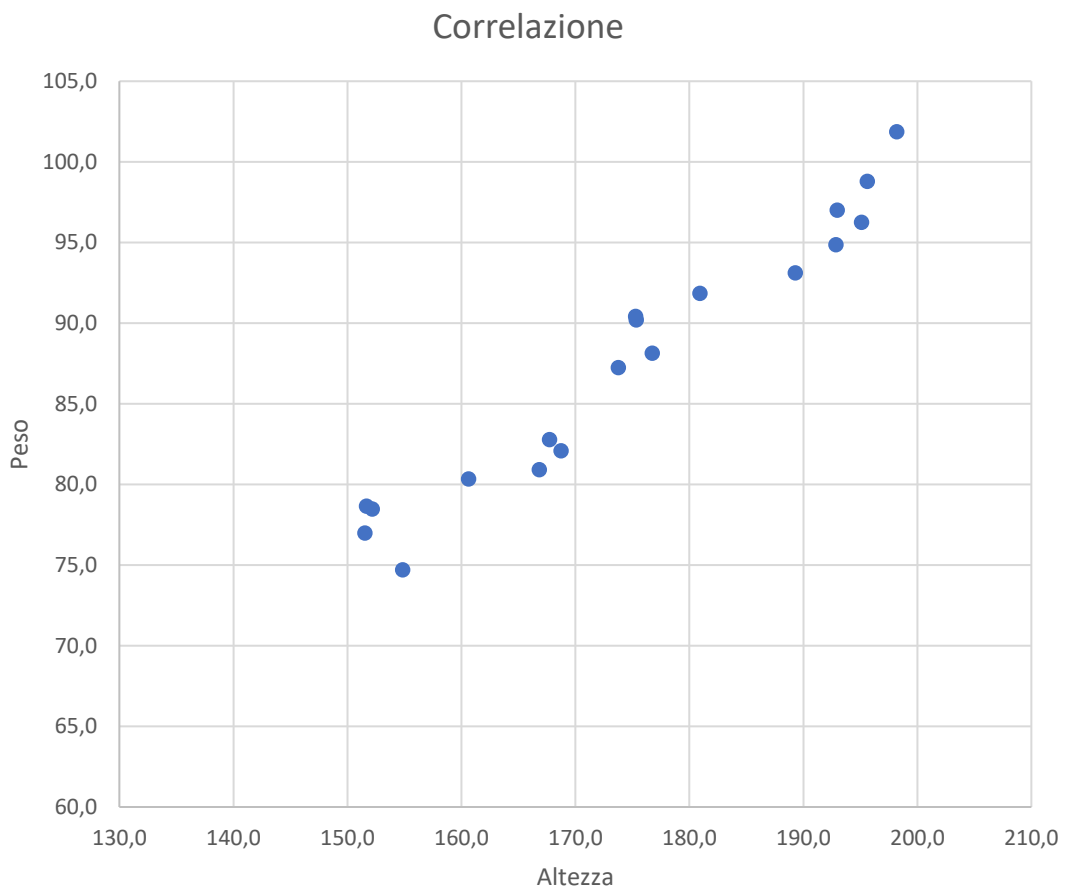
C’è correlazione tra peso e altezza?

Altezza → variabile di exposure

Peso → variabile di outcome

id	Altezza	Peso
1	175,3	90,4
2	173,8	87,2
3	198,2	101,9
4	175,3	90,2
5	166,8	80,9
6	168,7	82,1
7	176,7	88,1
8	195,1	96,3
9	195,6	98,8
10	167,7	82,8
11	192,8	94,9
12	154,8	74,7
13	180,9	91,8
14	192,9	97,0
15	152,2	78,5
16	189,3	93,1
17	160,6	80,3
18	151,7	78,7
19	151,5	77,0

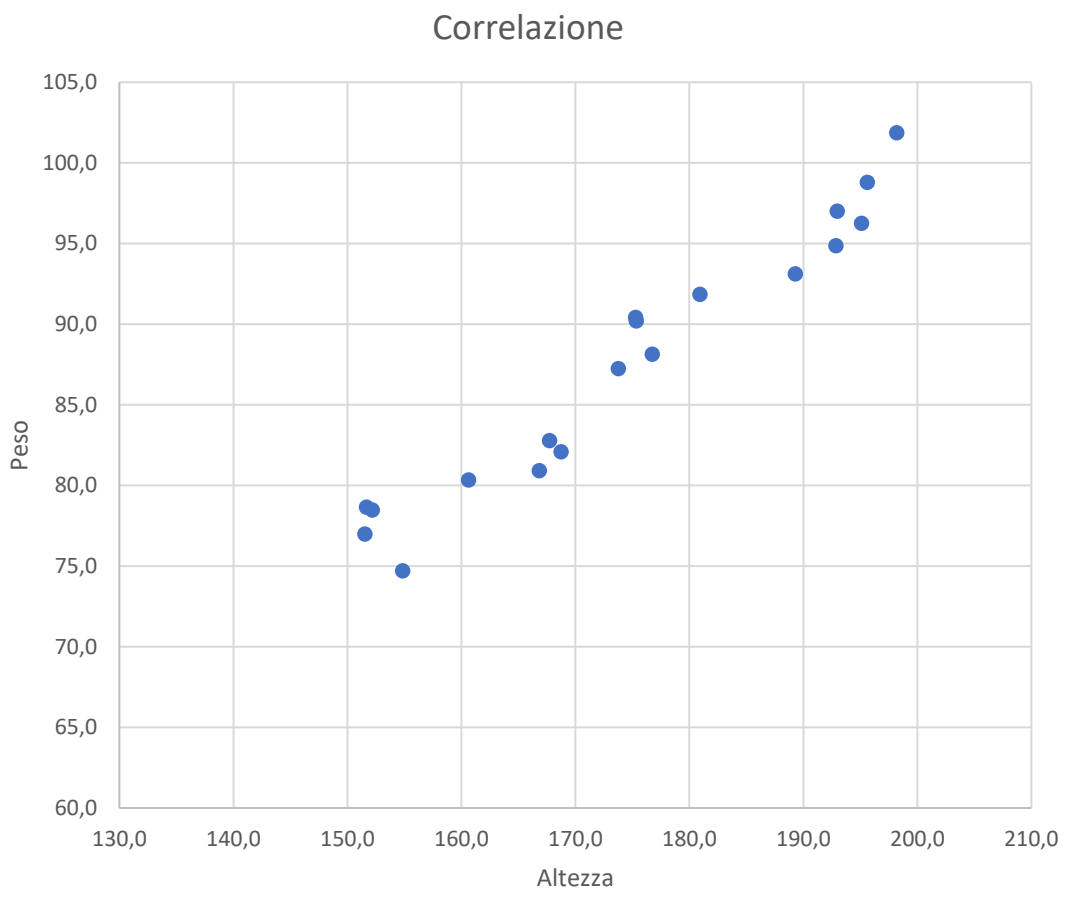
Ipotesi di correlazione



Calcolo per esempio il coefficiente di correlazione lineare di Pearson

r=0.97

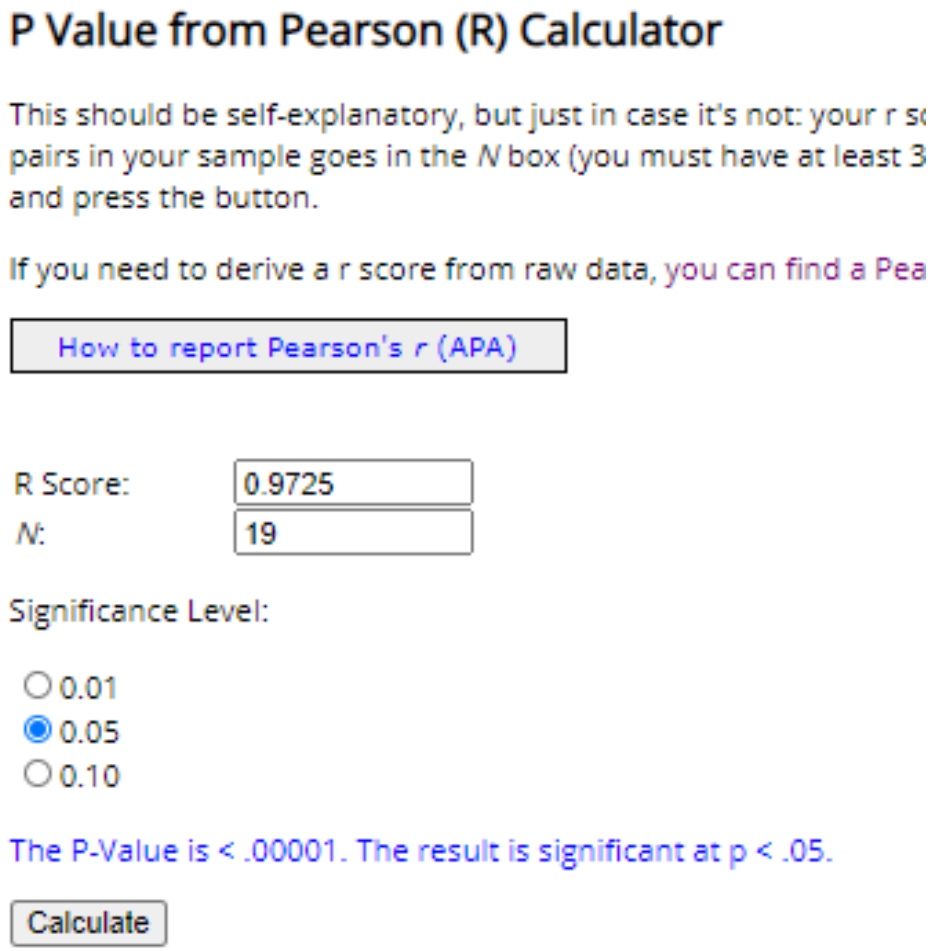
Ipotesi di correlazione



Questo coefficiente è *statisticamente significativo*? C'è realmente questa correlazione o è frutto del caso?

Dobbiamo fare il test per la verifica di questa ipotesi

Ipotesi di correlazione



Dal test otteniamo $p<0.05$

→il coefficiente di correlazione è statisticamente significativo

Perciò siamo confidenti che questa correlazione esista realmente

Stime puntuali e intervalli di confidenza

Stime

Puntuali

- Singolo valore che viene calcolato a partire dai dati campionari
- Non possono essere tratte conclusioni, non abbiamo informazioni aggiuntive ad eccezione del dato puntuale
- Avere più o meno informazione potrebbe non far cambiare il risultato della stima

Intervallari

- Intervallo di valori costruito attorno alla stima puntuale
- **Possiamo trarre conclusioni su precisione e significatività statistica (possiamo generalizzare la stima alla popolazione)**
- In questo caso più informazione abbiamo più la stima sarà precisa

Intervallo di confidenza

Un intervallo di confidenza è un intervallo che contiene il vero valore del parametro con probabilità prefissata 1-α

alias

Se estraiamo successivamente più campioni indipendenti dalla stessa popolazione e determiniamo i relativi intervalli di confidenza, il (1-α)% di intervalli conterrà il vero valore del parametro incognito

Intervallo di confidenza

N.B. Non sappiamo se uno specifico intervallo contiene oppure no il vero valore del parametro

Abbiamo un livello di fiducia/confidenza di (1- α)% di aver estratto un campione cui corrisponde un intervallo comprendente il parametro

Intervallo di confidenza

Una formula generica per il calcolo dell’intervallo di confidenza è:

Stima ± valore critico x errore della stima

Intervallo di confidenza

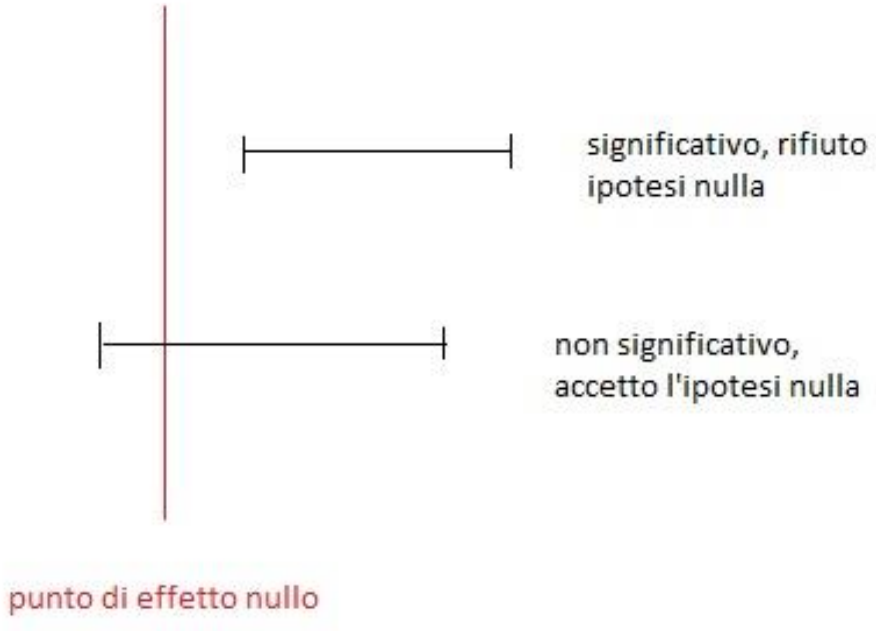
Ciò che possiamo dire per ogni intervallo di confidenza indipendentemente dal parametro da stimare è che:

- In base al tipo di parametro esistono dei punti detti di **effetto nullo** e se osserviamo la posizione di questi punti con l'intervallo di confidenza è come se conducessimo un test di ipotesi.

- I punti di effetto nullo sono 0 per ipotesi di correlazione e differenza o 1 per ipotesi su rapporti e associazione (come per l'OR e il RR).

Intervallo di confidenza

Ciò che possiamo dire per ogni intervallo di confidenza indipendentemente dal parametro da stimare è che:

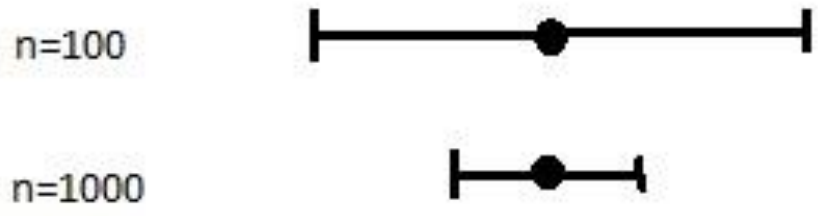


Se il punto di effetto nullo è contenuto nell'intervallo di confidenza è equivalente ad accettare l'ipotesi nulla del relativo test, se non è contenuto allora rifiutiamo l'ipotesi nulla

Intervallo di confidenza

Ciò che possiamo dire per ogni intervallo di confidenza indipendentemente dal parametro da stimare è che:

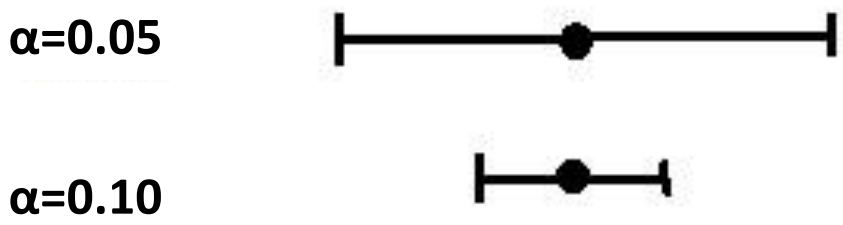
- Più la numerosità campionaria è elevata più l'intervallo di confidenza sarà preciso (=stretto)



Intervallo di confidenza

Ciò che possiamo dire per ogni intervallo di confidenza indipendentemente dal parametro da stimare è che:

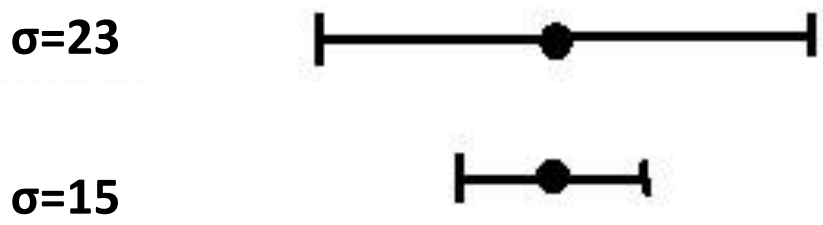
- Per valori di α maggiori l'intervallo di confidenza è più piccolo



Intervallo di confidenza

Ciò che possiamo dire per ogni intervallo di confidenza indipendentemente dal parametro da stimare è che:

- L'intervallo di confidenza aumenta all'aumentare della variabilità



Intervallo di confidenza per la media

L'intervallo di confidenza viene calcolato come:

$$\bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$$

$\bar{x}$ = stima puntuale

Z= valore critico della normale standardizzata che individua α (se consideriamo un intervallo di confidenza per la media z è uguale a 1.96 se α=0.05, 1.64 se α=0.10, 2.58 se α=0.01)

$\frac{\sigma}{\sqrt{n}}$ = errore standard

Intervallo di confidenza per la differenza tra medie

L'intervallo di confidenza viene calcolato come:

$$(\overline{x_1} - \overline{x_2}) \pm z * \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$\overline{x_1}$, $\overline{x_2}$ = stima puntuale in ciascuno dei due gruppi

Z= valore critico della normale standardizzata che individua α

σ_1^2, σ_2^2 = varianza dei due gruppi

n_1, n_1 =numerosità dei due gruppi

Intervallo di confidenza per la proporzione

L'intervallo di confidenza viene calcolato come:

$$\hat{p} \pm z * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$$

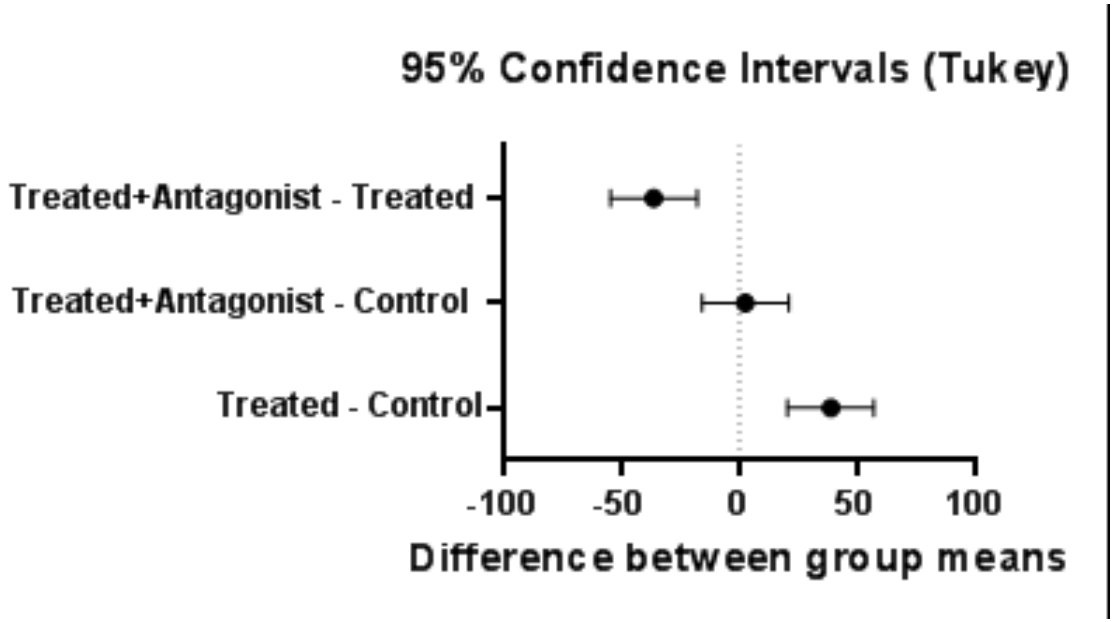
$\hat{p}$ = stima puntuale della proporzione

Z= valore critico della normale standardizzata che individua α

$\hat{p} * (1 - \hat{p})$ = varianza della proporzione

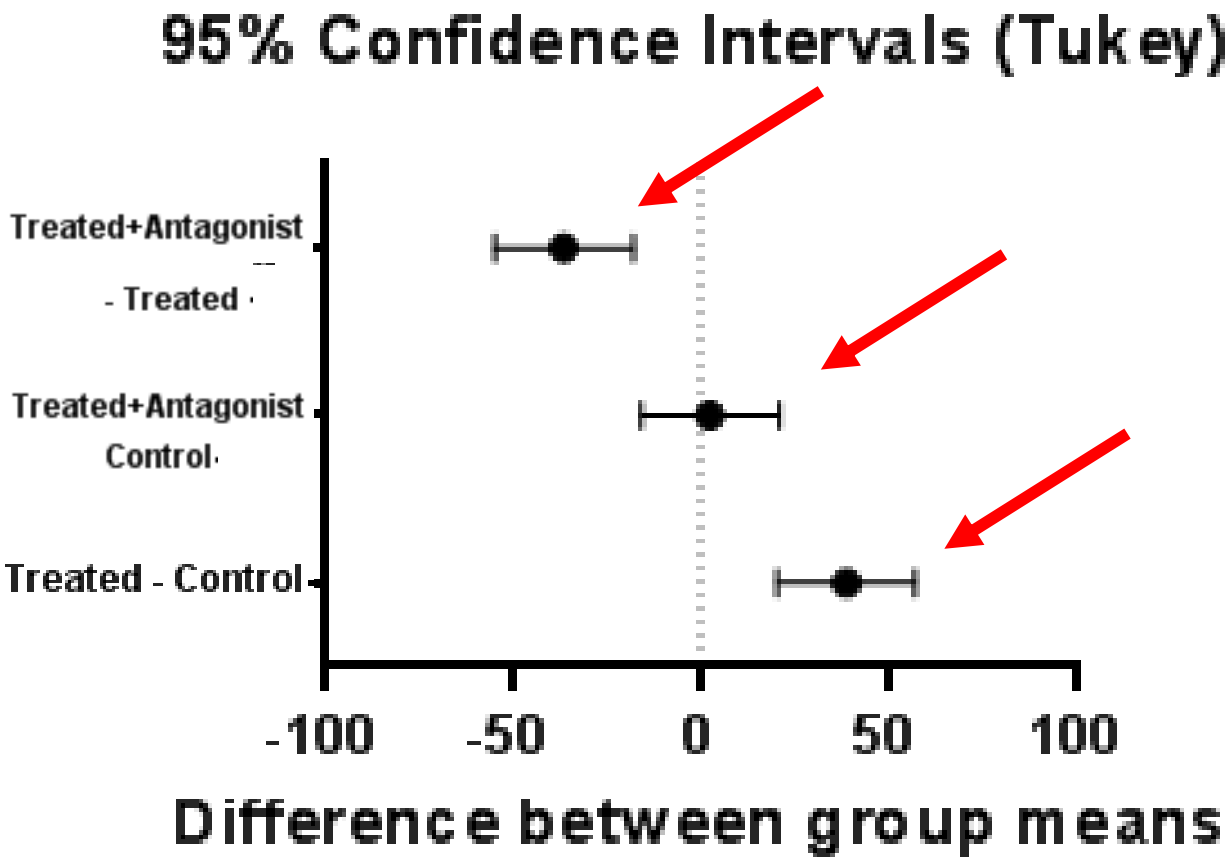
Intervallo di confidenza

Esempio: Vogliamo valutare se la concentrazione media di calcio è diversa tra un gruppo di controllo, un gruppo di trattamento e un gruppo che fa lo stesso trattamento più un farmaco



Dopo aver fatto un test ANOVA effettuiamo la procedura di Tukey (abbiamo rifiutato l'ipotesi nulla)

Intervallo di confidenza



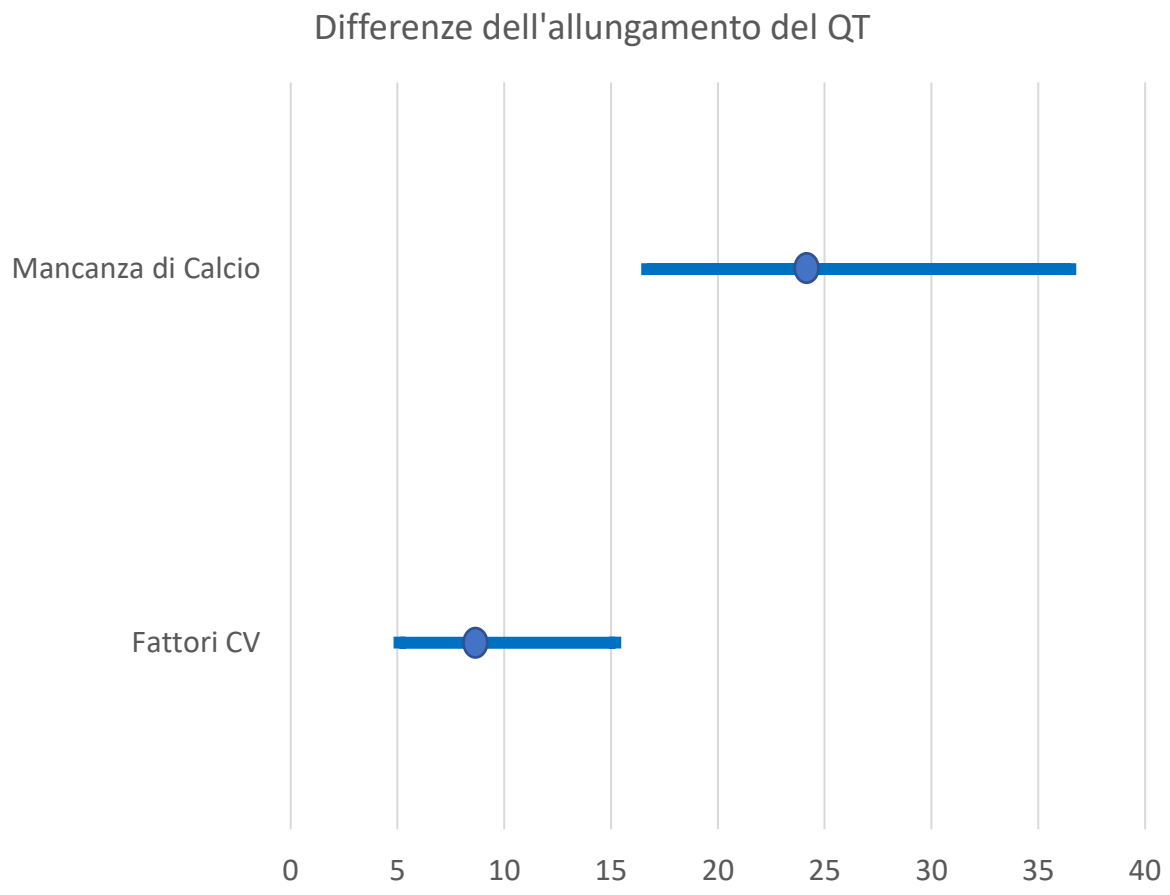
- Lo 0 è il punto di effetto nullo per le ipotesi di differenza
- Il secondo intervallo di confidenza ci dice che la differenza tra i due gruppi non è significativa
- Dal primo e il terzo intervallo invece deduciamo che le differenze sono statisticamente significative perché non contengono il punto di effetto nullo

Intervalli di confidenza

Vengono utilizzati molto spesso associati alla concordanza, agli Odds Ratio, ai Risk Ratio, agli Hazard Ratio

Gli intervalli di confidenza ci danno molte informazioni non solo sulla significatività statistica ma anche sulla rilevanza clinica

→ Tanto più l'intervallo di confidenza è distante dal punto di effetto nullo tanto più è forte la relazione/differenza/correlazione/associazione



Il primo fattore ha una rilevanza maggiore perché il fatto di avere mancanza di calcio produce una differenza media di QT di circa 25ms mentre avere patologie cardiovascolari comporta una differenza di soli 5ms

Entrambi sono comunque significativi e modificano il QT

Regressione e Statistica multivariata

Statistica

- Statistica descrittiva: coinvolge una due variabili, ci dà delle informazioni preliminari su quantità e distribuzioni.

- Statistica inferenziale: test di ipotesi e intervalli di confidenza. Coinvolge una/due variabili, ci dice se quelle relazioni che abbiamo osservato con la statistica descrittiva sono reali oppure frutto del caso. Viene anche detta analisi/statistica univariata.

Statistica

- Statistica Multivariata: è un insieme di tecniche statistiche che permette di valutare insieme un ampio set di variabili.

→ tutte queste variabili hanno un ruolo di exposure

→ in tutte le tecniche che vedremo avremo un unico outcome, ma esistono tecniche più complesse di statistica multivariata in cui gli outcome sono molteplici

Regressione

Che cosa è?

- È una tecnica utilizzata per capire la ***forma*** della relazione che intercorre

- Può avere un'unica variabile di esposizione (regressione semplice) o può avere più variabili di esposizione (regressione multipla)

  → questa tecnica è la tecnica principale della statistica multivariata

Regressione

Perché viene utilizzata?

- Per stimare e *prevedere* il valore che assume la variabile di outcome in corrispondenza di una combinazione di valori assunti dalla/dalle variabili exposure
- Capire se e quanto ciascuna variabile influisce sull’outcome
- Controllare per i fattori di confondimento e valutare l’interazione

Regressione

- Le variabili di exposure vengono dette «variabili esplicative» o «variabili predittive» o «regressori» (sono l’insieme delle variabili indipendenti)

- La variabile di outcome è la variabile dipendente e spesso viene chiamata anche «variabile risposta»

Regressione

In base alla tipologia della variabile dipendente esistono vari tipi di regressione:

- **Regressione lineare** (la variabile è quantitativa continua)
- **Regressione logistica** (la variabile è dicotomica)
- Regressione logistica multinomiale (la variabile di outcome è policotomica)
- Regressione logistica ordinale (la variabile è ordinale)
- Regressione di Poisson (la variabile è quantitativa discreta)
- Regressione di Cox (time-to-event, variabile dicotomica+tempo)

Regressione lineare

La formula della retta di regressione lineare è data da:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Y→VARIABILE DIPENDENTE

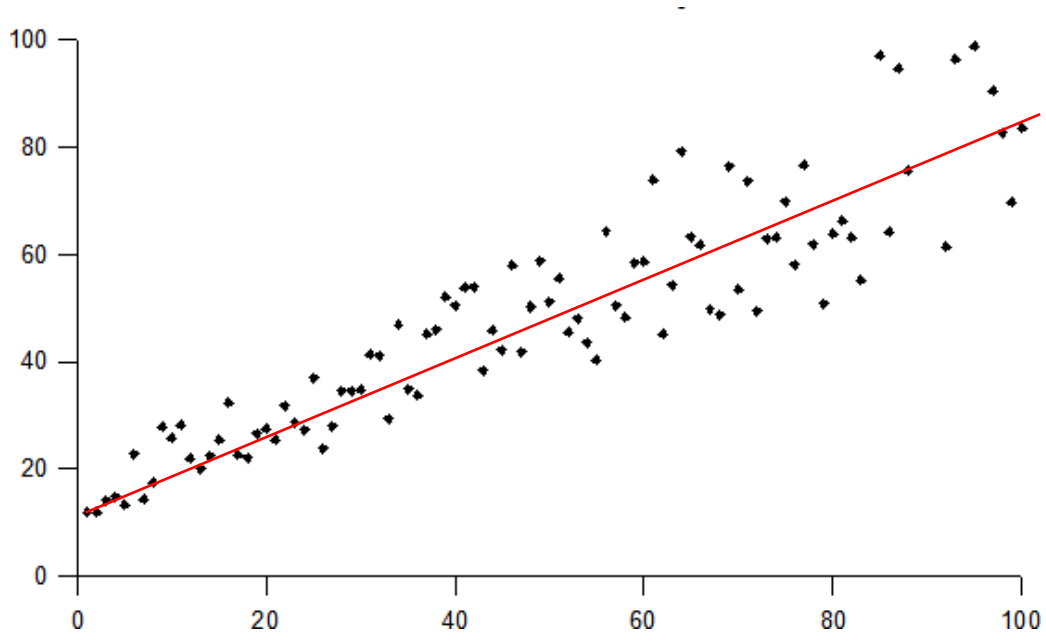
$X_1 \ldots X_n$→VARIABILI INDIPENDENTI

β_0→INTERCETTA

$\beta_1 \ldots \beta_n$→COEFFICIENTE DI REGRESSIONE

ε →ERRORE

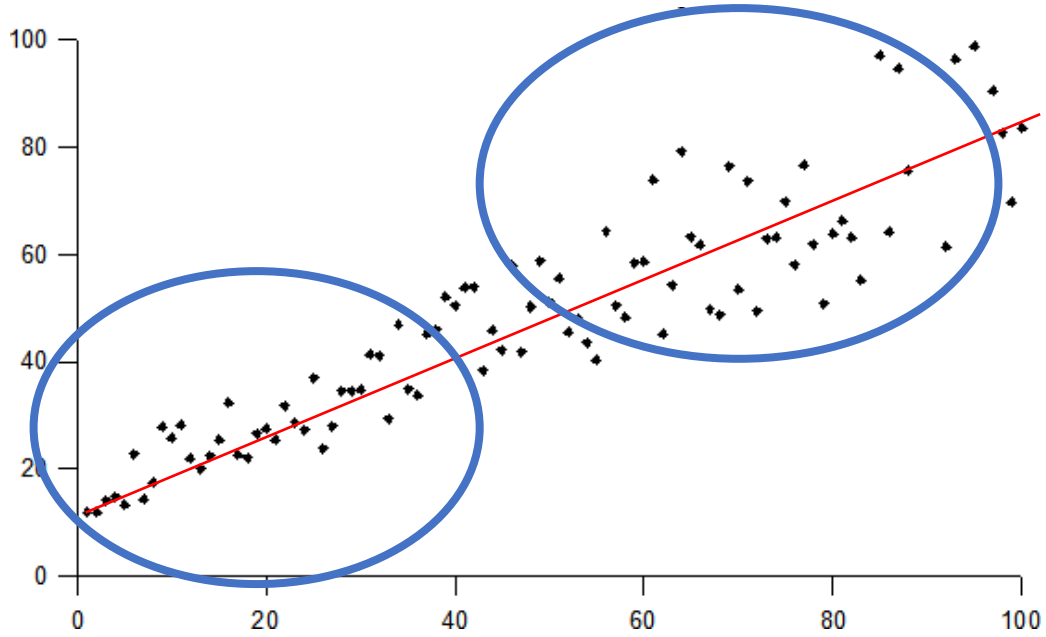
Regressione lineare



Ci sono dei casi in cui non è opportuno stimare un modello di regressione lineare:

1. Quando attraverso lo scatterplot costruito con la variabile dipendente e quella indipendente osserviamo questa distribuzione a imbuto

Regressione lineare

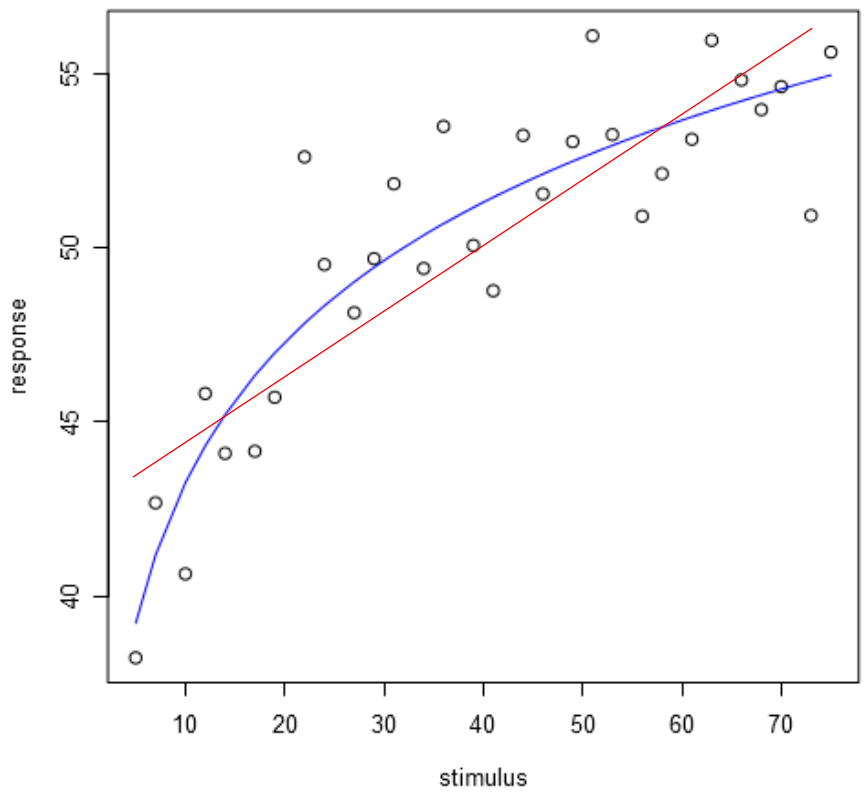


Ci sono dei casi in cui non è opportuno stimare un modello di regressione lineare:

I punti nella parte sinistra del grafico sono piuttosto vicini alla retta

Più ci spostiamo verso destra più si scostano

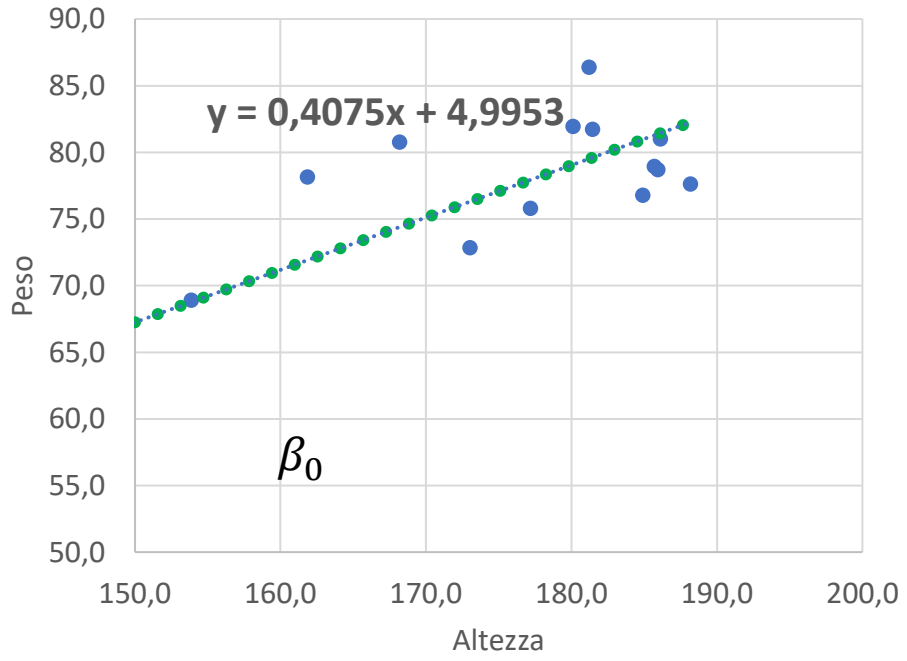
Regressione lineare



Ci sono dei casi in cui non è opportuno stimare un modello di regressione lineare:

2. Quando la distribuzione dei dati non ha un andamento lineare

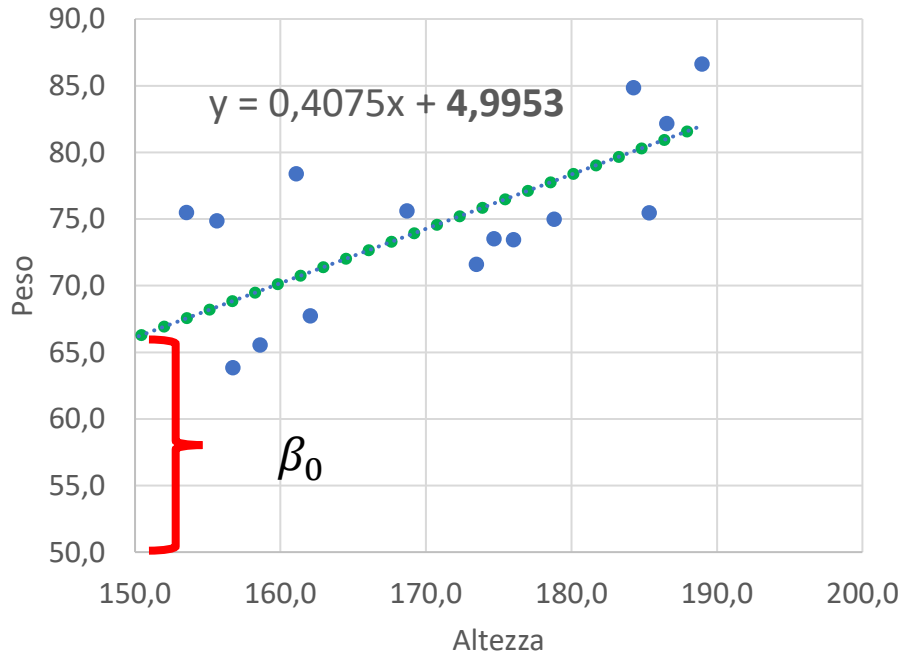
Regressione lineare



La retta verde è detta retta di regressione

(in particolare questa è una retta di regressione lineare semplice)

Regressione lineare



- β_0 **intercetta (4,9953)**
- è il valore in ordinata del punto di intersezione tra la retta e l'asse delle ordinate

Regressione lineare

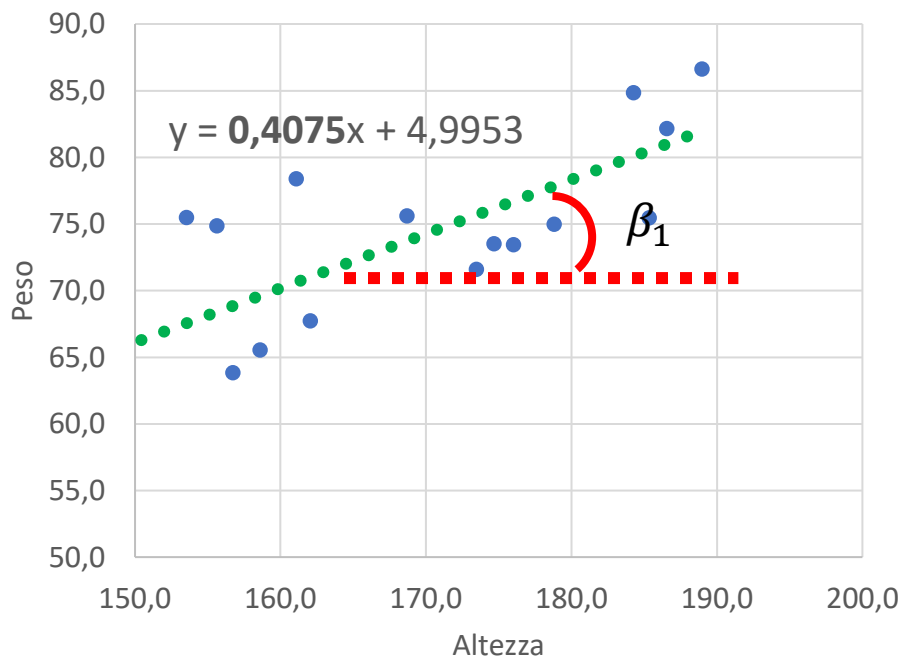
- L'intercetta è un valore che tendenzialmente sarà sempre positivo

- L'intercetta però può essere tolta e possiamo decidere quindi di stimare un modello che ha inizio dal punto di origine (0,0)

→ $Y = \cancel{\beta_0} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$

→ $Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$

Regressione lineare



- β_1 **coefficiente di regressione (0,4075)**
- misura il cambiamento medio nella variabile dipendente per una unità di incremento nella variabile dipendente
- All'aumentare dell'altezza di 1 centimetro il peso aumenta in media di 0,4075kg

Regressione lineare

I coefficienti di regressione possono assumere valori:

- **Positivi:** all’aumento unitario del regressore corrisponde un **aumento** della variabile dipendente

- **Negativi:** all’aumento unitario del regressore corrisponde una **diminuzione** della variabile dipendente

Regressione lineare

Interpretazione del coefficiente di regressione quando il regressore è una variabile dicotomica:

Il coefficiente di regressione misura la

differenza media

della variabile dipendente tra i due gruppi

Regressione lineare

Esempio: Peso=75.0+4.7*Sesso

- Prima di tutto dobbiamo aver stabilito un *livello di riferimento* per la variabile qualitativa (per esempio il sesso Femminile) prima di stimare il modello

- 4.7 significa che in media gli uomini pesano 4.7kg in più delle donne

Regressione lineare

Quando il regressore è di tipo qualitativo policotomico o ordinale dobbiamo trasformarlo in tante variabili dicotomiche quante sono il numero-1 delle categorie della variabile qualitativa

N-1 perché una categoria è utilizzata come riferimento

Un’unica variabile sarà trasformata in N-1 variabili e avremo N-1 coefficienti

Regressione lineare

Esempio: Peso – mangiare frutta (mai, saltuariamente, tutti i giorni)

Mangiare frutta:

Mai→ livello di riferimento (non ha coefficiente)

Saltuariamente→ se si 1 altrimenti 0

Tutti i giorni→ se si 1 altrimenti 0

Peso=75.0-1.5*Saltuariamente-2.7*TuttiIGiorni

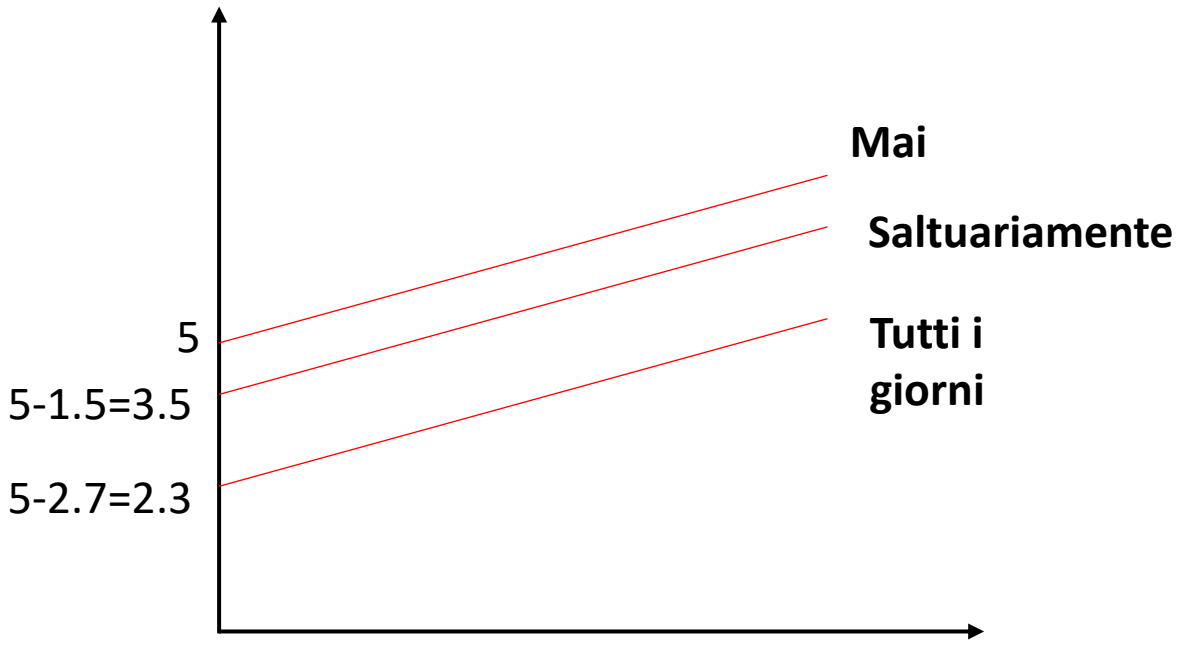
Regressione lineare

Peso=75.0-1.5*Saltuariamente-2.7*TuttiIGiorni

- Se un soggetto mangia frutta saltuariamente peserà in media 1.5 kg in meno rispetto a chi non la mangia mai

- Se un soggetto mangia frutta tutti i giorni peserà in media 2.7 kg in meno rispetto a chi non la mangia mai

Regressione lineare



Peso=5.0-1.5*Saltuariamente-2.7*TuttiIGiorni+0.45*Altezza

Le variabili qualitative traslano solamente la retta verso l'alto o verso il basso

I coefficienti vanno a sommarsi/sottrarsi all'intercetta

Regressione lineare

Esempio: Modello stimato utilizzato per scopo predittivo

Peso=5.0 +0.45*Altezza -2.3*MangiareDolci-0.3*MangiareFruttaSaltuariamente -1.0*MangiareFruttaSempre

Per ogni nuovo soggetto posso dire quanto dovrebbe pesare, con un certo grado di errore, senza farlo realmente basta che io sappia la sua altezza e le abitudini alimentari

Regressione lineare

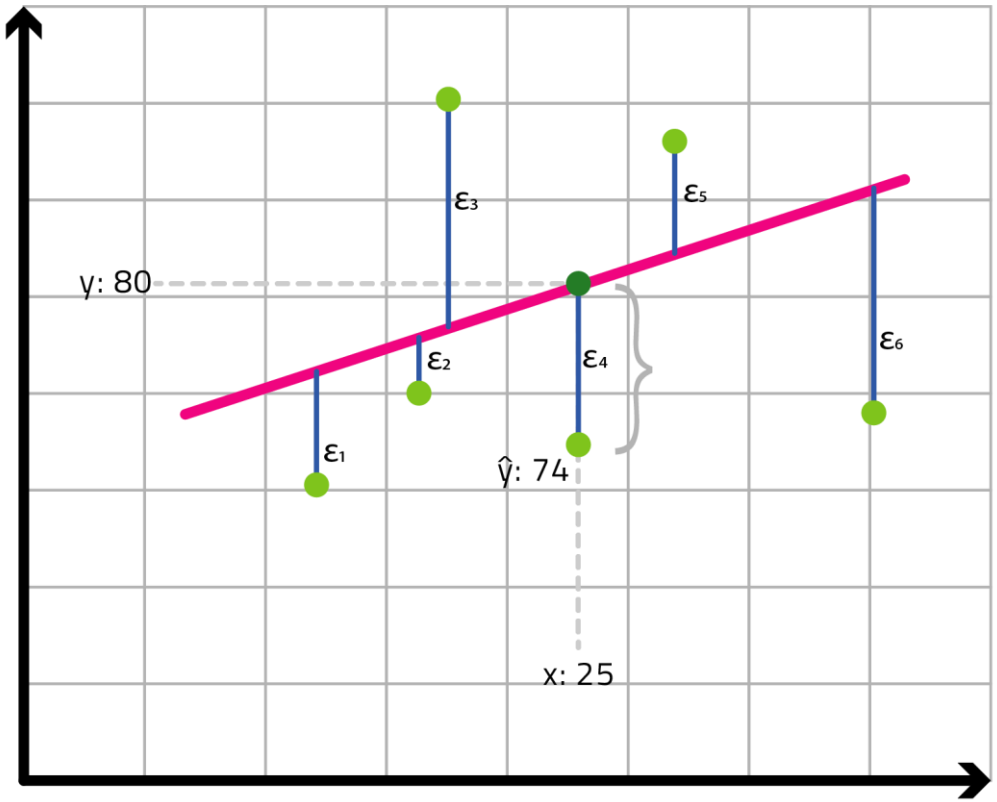
Esempio: Modello stimato utilizzato per scopo predittivo

Peso=**5.0** +**0.45***Altezza **-2.3***MangiareDolci-**0.3***MangiareFruttaSaltuariamente **-1.0***MangiareFruttaSempre

Il nuovo pz è alto 167.3cm non mangia mai dolci ma mangia frutta tutti i giorni

→il suo peso è P=**5.0**+**0.45***167.3**-2.3***0**-0.3***0**-1.0***1=79,3kg

Regressione lineare

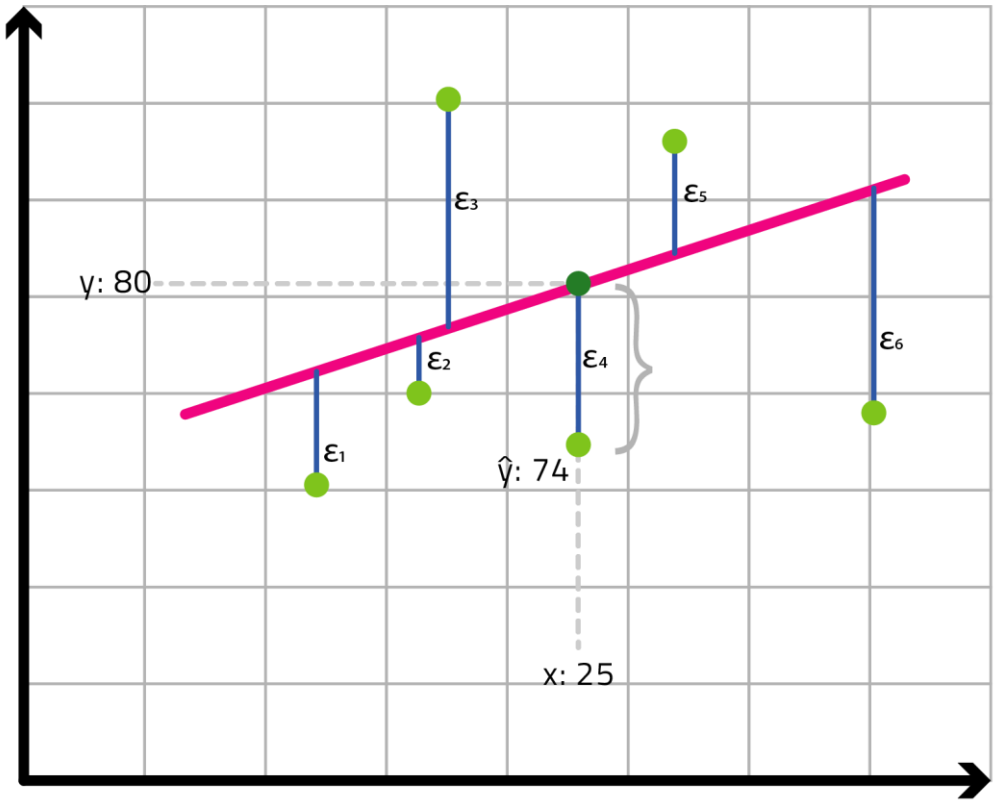


ε **errore**

- È la distanza tra i punti osservati e la retta (questa distanza è detta anche **scarto o residuo**)

- **I coefficienti di regressione vengono stimati in modo che sia minimizzata la somma di questi errori**

Regressione lineare



- Una volta stimato il modello questo parametro non è più presente (non è un valore stimato)
- Ma è dato dalla differenza tra il valore di Y osservato e Y predetto ($\hat{Y}$)

Regressione lineare

- Quando stimiamo l'intercetta e i coefficienti di regressione dobbiamo fare un **test di ipotesi** per capire se il coefficiente è significativo oppure no

- Ipotesi nulla: $\beta = 0$
- Ipotesi alternativa: $\beta \neq 0$

Regressione lineare

- Se non rifiutiamo l’ipotesi nulla dovremmo scartare quella data variabile dal set dei regressori

- Con i dati da noi ottenuti non siamo in grado di ottenere come significativa la relazione tra quella variabile e l’outcome

Regressione lineare

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

```
> modello <- lm(y ~ x2 + x3 + x4 + x5)
> summary(modello)
Call:
lm(formula = y ~ x2 + x3 + x4 + x5)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5804 -1.5223 -0.1098  1.4237  4.6214

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.820774  10.123152  -0.674   0.5062
x2           0.227246   0.099937   2.274   0.0311 *
x3           0.553726   0.369752   1.498   0.1458
x4          -0.149536   0.029229  -5.116 2.23e-05 ***
x5           0.154650   0.006446  23.992  < 2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

- intercetta e regressori
- Stime dei coefficienti
- p-value del test
- Non significativi

Regressione lineare

Come selezioniamo le variabili da inserire nel modello?

Nella pratica clinica selezioniamo le variabili da inserire nel modello secondo i seguenti criteri:

- Variabili che risultano associate/differenti/correlate all’analisi univariata

Regressione lineare

Come selezioniamo le variabili da inserire nel modello?

Nella pratica clinica selezioniamo le variabili da inserire nel modello secondo i seguenti criteri:

- Variabili che non risultano significative rispetto all'α stabilito ma che comunque non lo superano di molto

  Es. $\alpha=0.05$ → prendo tutte quelle con $p<0.10$ o $p<0.20$

Regressione lineare

Come selezioniamo le variabili da inserire nel modello?

Nella pratica clinica selezioniamo le variabili da inserire nel modello secondo i seguenti criteri:

- Quelle variabili che non sono risultate significative (con p-value molto elevati) all'analisi univariata ma che hanno un'elevata importanza clinica o sono noti come fattori confondenti

  →Potrebbe accadere che queste variabili non sono significative all'univariata ma nell'analisi multivariata si

Regressione lineare

Come selezioniamo le variabili da inserire nel modello?

Il numero delle variabili inserite nel modello tuttavia non dovrà essere eccessivamente elevato

Anche se il calcolo è approssimativo e semplicistico, dobbiamo considerare circa 15-20 unità statistiche per ogni regressore che vogliamo inserire nel modello

Regressione lineare

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

```
> modello <- lm(y ~ x2 + x3 + x4 + x5)
> summary(modello)
Call:
lm(formula = y ~ x2 + x3 + x4 + x5)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5804 -1.5223 -0.1098  1.4237  4.6214

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.820774  10.123152  -0.674   0.5062
x2           0.227246   0.099937   2.274   0.0311 *
x3           0.553726   0.369752   1.498   0.1458
x4          -0.149536   0.029229  -5.116 2.23e-05 ***
x5           0.154650   0.006446  23.992  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

- Come facciamo a togliere questa variabile e l'intercetta?
- Potremmo provare a ristimare il modello senza la terza variabile

Regressione lineare

- Ovviamente togliendo una variabile e ristimando il modello i coefficienti cambieranno

- Nonostante ogni coefficiente stimi il cambiamento della Y all'aumento unitario della X considerando costanti tutti gli altri (rapporto tra X e Y), i coefficienti vengono comunque stimati tenendo in considerazione tutte le variabili del modello.

- Perciò togliendo una variabile non è detto che le altre che prima erano significative non diventino non significative

Regressione lineare

Ci sono delle procedure ad hoc che permettono di selezionare un set minimo di variabili che danno la miglior accuratezza e precisione nel predire la Y

Queste procedure sono dette **stepwise** e si basano sul concetto di «selezionare la variabile che apporta informazione rispetto a tutte quelle già inserite»

Ad ogni step deve diminuire la somma dei residui

Regressione lineare

Esistono 3 tipi di procedure stepwise:

1. Forward: al modello composto da solamente l’intercetta aggiungiamo una variabile alla volta
2. Backward: al modello composto da tutti i regressori ne togliamo uno alla volta
3. Forward-Backword: composizione delle precedenti, ad ogni step può essere aggiunta o tolta una variabile

Regressione logistica

La formula della retta di regressione logistica è data da:

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

p=probabilità che si verifichi l'evento (probabilità che la variabile dicotomica predetta assuma valore 1)

1-p=probabilità che si verifichi l'evento (probabilità che la variabile dicotomica predetta assuma valore 1)

Regressione logistica

La formula della retta di regressione logistica è data da:

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

In questa forma la regressione logistica è molto simile alla regressione lineare

In particolare Y= $\ln\left[\frac{p}{1-p}\right]$

Regressione logistica

La formula della retta di regressione logistica è data da:

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

La reale funzione della regressione logistica è:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}, \quad i = 1, 2, \ldots, n.$$

Regressione logistica

- Tutte le considerazioni fatte sulla regressione lineare valgono anche per questo tipo di regressione

- I coefficienti devono essere valutati con test di ipotesi

- Possono essere applicate le procedure stepwise per selezionare il sottoinsieme di variabili che predicono meglio l’outcome

Regressione logistica

- I coefficienti β non sono facilmente interpretabili in questa forma

- Dobbiamo valutare e interpretare non β ma e^{β}

- **e^{β} è l'ODDS RATIO DELLA VARIABILE**

Regressione logistica

Interpretazione degli OR:

→se la variabile è **qualitativa:**

Ci dice quanto è più probabile che si verifichi l’outcome nei soggetti esposti rispetto a quelli non esposti

Regressione logistica

Interpretazione degli OR:

→se la variabile è **quantitativa:**

Ci dice quanto è più probabile che si verifichi l’outcome all’aumento unitario del regressore

Regressione logistica

La regressione logistica è molto utile quando dobbiamo calcolare gli OR aggiustati perché è sufficiente inserire i fattori confondenti nel modello e possiamo capire se e quanto le variabili che supponiamo fattori confondenti influenzino l'OR della variabile esplicativa

I software con cui svolete i calcoli restituiscono anche lo standard error dei coefficienti, perciò potete anche calcolare gli intervalli di confidenza

$$e^{\hat{\beta}_1 \pm z_{1-\alpha/2}\ \hat{SE}(\hat{\beta}_1)}$$

Regressione logistica

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.875732   0.720476   6.767 1.31e-11 ***
Farmaco      0.286738   0.065051   4.408 1.04e-05 ***
Age          0.029465   0.002438  12.086  < 2e-16 ***
Female       0.390083   0.068374   5.705 1.16e-08 ***
ca_mean     -0.551061   0.068880  -8.000 1.24e-15 ***
K_mean      -0.582286   0.096089  -6.060 1.36e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Coefficienti
- Standard error
- p-value del test

Regressione logistica

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.875732   0.720476   6.767 1.31e-11 ***
Farmaco             0.286738   0.065051   4.408 1.04e-05 ***
Age                 0.029465   0.002438  12.086  < 2e-16 ***
Female              0.390083   0.068374   5.705 1.16e-08 ***
ca_mean            -0.551061   0.068880  -8.000 1.24e-15 ***
K_mean             -0.582286   0.096089  -6.060 1.36e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Coefficienti

Farmaco

$e^{0.29}$=1.34

IC: $e^{0.29 \pm 1.96 * 0.07}$=1.17-1.53

Age

$e^{0.03}$=1.03

IC: $e^{0.03 \pm 1.96 * 0.002}$ =1.02-1.04