STATISTICA MEDICA

STATISTICA PER DISCIPLINE BIO-MEDICHE

Fabrizio QUARTA

Studio della statistica in medicina: motivi

L'approvazione di nuove terapie farmacologiche e/o pratiche cliniche, la validazione e l'utilizzo di test diagnostici, la prevenzione nei confronti delle malattie epidemiche, comportano l'assunzione di decisioni che necessitano del supporto metodologico ed analitico della statistica.

La statistica pervade la letteratura medica. Se il lettore di una rivista scientifica deve leggere questi articoli in modo intelligente per valutare i risultati ottenuti, egli deve avere una certa conoscenza della statistica.

Perché la statistica può essere utile ad un infermiere?

Rapporto sulla formazione degli infermieri (Comitato consultivo, Bruxelles 29/04/1981) - Principali obiettivi

capacità di pianificare, organizzare, dispensare e valutare i servizi di assistenza infermieristica (preventivi, curativi, educativi, ecc.);

capacità di partecipare alla ricerca;

capacità di contribuire alla promozione di una politica sanitaria efficiente.

L'importanza della statistica

La statistica è la scienza che analizza i fenomeni quantitativi, cercando di evidenziarne le caratteristiche salienti, le regolarità, le eccezioni.

Esempi:

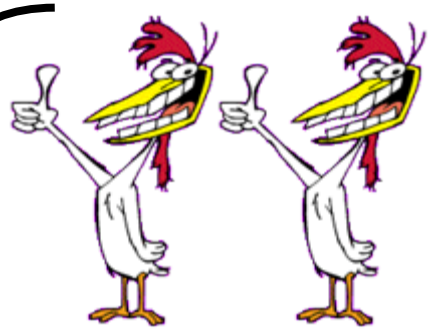
- **Rilevazione del numero di fratelli/sorelle per ogni studente presente in aula.**
- **Mese di nascita di ogni studente presente in aula.**
- **Rilevazione della temperatura a intervalli orari nella stazione meteorologica di Pratica di Mare.**
- **Rilevazione del primo numero estratto sulla ruota di Roma nelle ultime 1000 estrazioni del lotto.**
- **Mezzo di trasporto utilizzato da ciascuno studente per raggiungere l'Università.**

La disciplina statistica

Oggetto della Statistica sono dunque quei fenomeni che presentano caratteri di variabilità all'interno di un collettivo di riferimento **(popolazione statistica**), costituito da **unità statistiche o elementari.**

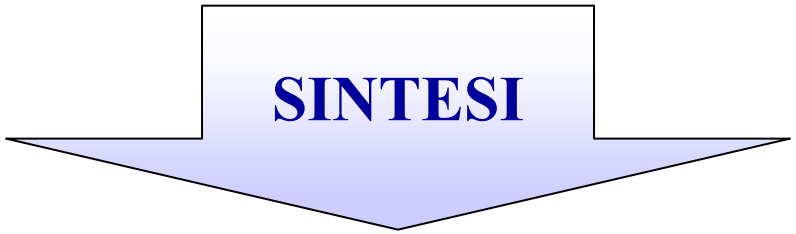
Il “dilemma” di TRILUSSA

“Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perché c'è un antro che ne magna due”

$$\left[\text{  } + 0 \right] / 2 = \text{  } \; (?)$$

La disciplina statistica

La Statistica, attraverso misure di sintesi (indici o parametri), non ci dice solo quanti "polli mangia" in media una popolazione, ma anche se esistono differenze "alimentari" tra gli individui



INDICI di POSIZIONE

INDICI di DISPERSIONE

Misure della Variabilità del fenomeno oggetto di studio nel collettivo di riferimento

Cose importanti da imparare

Affinché le informazioni deducibili da tali rilevazioni siano proficuamente utilizzate è necessario imparare a conoscere:

- **Gli obiettivi primari della statistica**
- **Il linguaggio statistico**
- **Come si organizza una indagine statistica**
- **Le principali tecniche di analisi e sintesi**
- **Il metodo statistico**

Obiettivi della statistica

- **Separare il “segnale” dal “rumore”.**
- **Valutare la complessità dei fenomeni.**
- **Saper “prevedere”.**

In base al tipo di obiettivi è necessario:

A. Saper predisporre, logicamente e praticamente, il tipo di indagine più adatta al conseguimento dei nostri obiettivi.

B. Definire con precisione qual è la popolazione di riferimento della nostra indagine.

C. Stabilire quali sono le caratteristiche della popolazione che “interessano”.

Le fasi di una indagine statistica

A. Definizione degli obiettivi
B. Pianificazione della raccolta dei dati
C. Rilevazione dei dati
D. Elaborazione metodologica
E. Presentazione dei risultati
F. Utilizzazione dei risultati della ricerca.

Un corso istituzionale di statistica deve soffermarsi (senza indugiare nel dettaglio) sui punti A-C, approfondire il punto D, sorvolare sul punto E (serve a vendere meglio la ricerca) lasciare all'esperto del settore il punto F.

Osservazione

Pur avendo tracciato un percorso logico lineare, nella realtà le cose sono più contorte. Spesso si parte con un obiettivo che sottintende una teoria, ma i dati a disposizione sembrano confutare la teoria, suggerendo al tempo stesso una interpretazione differente.

Esempio
TEST STATISTICO
Spesso si sottopone a verifica una ipotesi con l'intento di confutarla.
Sperimentazione di un nuovo farmaco A per curare gli stiramenti muscolari Confronto col trattamento standard B.
Se i risultati del campione sono a favore di A, tutto OK. Ma se A, nella sperimentazione manifesta effetti collaterali, essi possono esse-re riutilizzati per un ripensamento della composizione del farmaco e della sua utilizzazione.

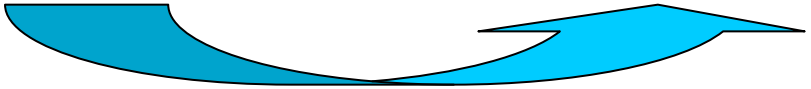
I due grandi rami della Statistica

Statistica Descrittiva

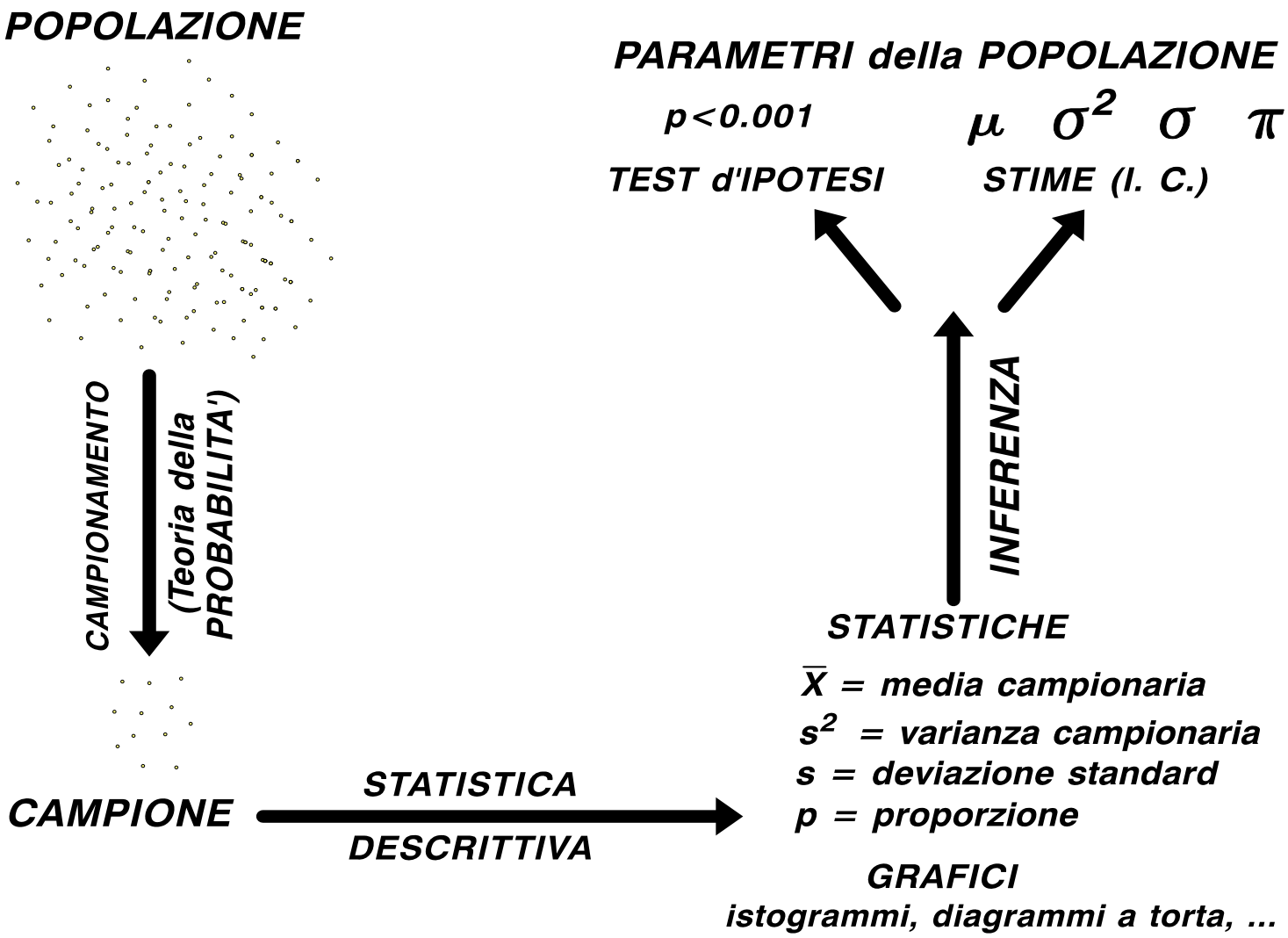
- **Metodo deduttivo** *(dal generale al particolare)*
- **Raccolta dei dati**
- **Sintesi dei dati di popolazione o del campione**
- **Presentazione dei risultati** *(Analisi esplorativa)*

Statistica Inferenziale

- **Metodo induttivo** *(dal particolare al generale)*
- **Rilevazioni parziali**
- **Stima dei parametri di popolazione**
- **Verifica delle ipotesi**
- **Previsione**



Schema logico



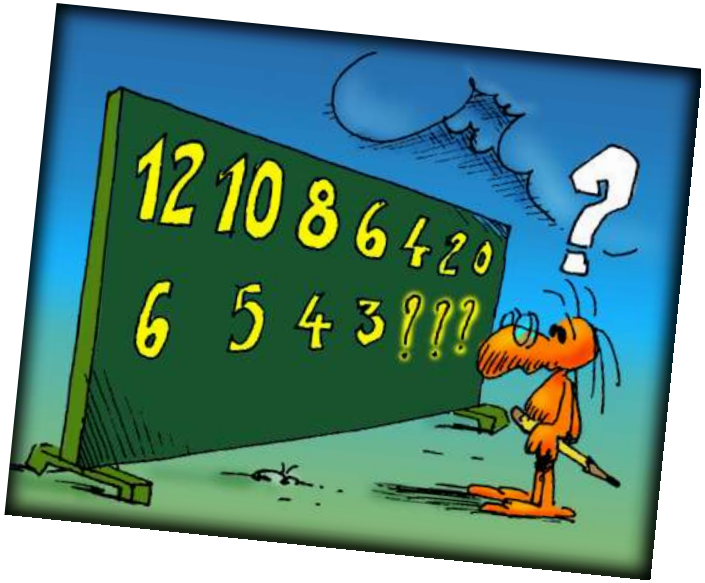
Fasi dell'indagine statistica - Rilevazione dati

CONSISTE IN QUELL'INSIEME DI OPERAZIONI PER CUI SI VIENE A CONOSCENZA DELLE UNITÀ CHE COMPONGONO UN DETERMINATO COLLETTIVO CONCRETO E DELLE NOTIZIE CHE RIGUARDANO TALI UNITÀ

Fasi dell'indagine statistica: rilevazione dei dati

DATI
OSSERVAZIONI FINALIZZATE

LA STATISTICA TRAE I SUOI RISULTATI DALLA ELABORAZIONE DEI DATI FORNITI DA UN INSIEME DI CASI OSSERVATI O DI ESPERIMENTI.



Fasi dell'indagine statistica: rilevazione dei dati

LE FONTI STATISTICHE

- ***IN MOLTE INDAGINI SI PARTE DA UN MATERIALE GIÀ RILEVATO:*** ISTAT - EUROSTAT - ONU - IMS

- ***IN ALTRE OCCORRE PROCEDERE AD UNA RILEVAZIONE***

Fasi dell'indagine statistica: rilevazione dei dati

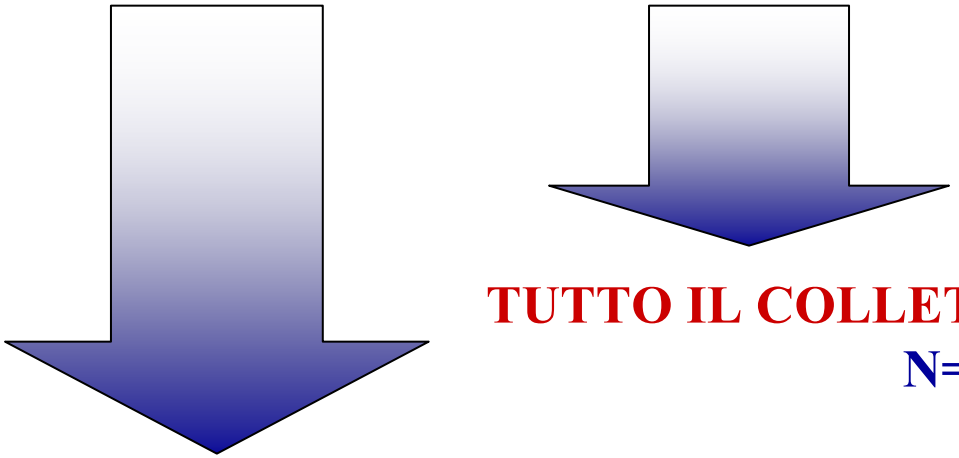
PIANO DELLA RILEVAZIONE
DEFINIZIONE DELLO SCOPO

- Definire l'unità statistica e l'unità di rilevazione
- Stabilire i caratteri quantitativi e qualitativi che interessa rilevare per ciascuna unità
- Indicare i mezzi tecnici per raccogliere le informazioni su detti caratteri (questionari)
- Fissare l'estensione della rilevazione in ordine al territorio, all'epoca, alle disponibilità finanziarie

Fasi dell'indagine statistica: rilevazione dei dati

POPOLAZIONE E CAMPIONE

LA RILEVAZIONE PUÒ RIGUARDARE



TUTTO IL COLLETIVO = POPOLAZIONE

N= numerosità del collettivo

UNA PARTE DI TUTTE LE UNITÀ STATISTICHE COSTITUENTI IL COLLETTIVO = CAMPIONE

n = numerosità del campione

Un po' di terminologia statistica

POPOLAZIONE

Qualsiasi insieme di elementi, reale o virtuale, che forma oggetto di studio.

Esempi

- **Tutti i residenti nel comune di Lecce il 30/09/2000**
- **Tutte le possibili sestine giocabili al Superenalotto**
- **Tutti i malati di *Sindrome della Statistica***

E' di fondamentale importanza (nonché indicatore di serietà della ricerca) definire esattamente la popolazione di riferimento della nostra indagine.

Un po' di terminologia statistica

UNITÀ STATISTICHE

Elemento di base della popolazione sulla quale viene effettuata l'indagine. E' indivisibile nell'ambito della ricerca ma non in senso assoluto (es.: famiglie).

CARATTERE (o VARIABILE)

Fenomeno oggetto di studio che è rilevato sulle unità statistiche. Esso si manifesta attraverso diverse modalità

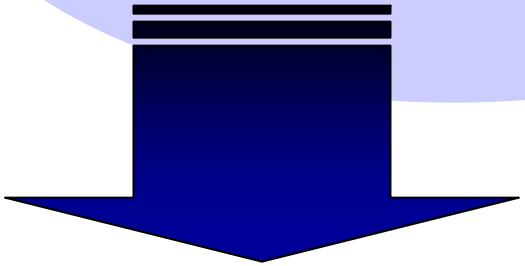
Esempi

- Il carattere **Sesso** si manifesta attraverso le modalità **M** e **F**.
- Il carattere **Tempo di spostamento** si manifesta attraverso infinite modalità
- Il carattere **Numero di fratelli** ha come modalità i numeri interi positivi e lo zero.

Nota: di una popolazione interessano soltanto le modalità del carattere che è oggetto di studio.

Collettivo statistico

COLLETTIVO STATISTICO **INSIEME DELLE UNITÀ STATISTICHE OGGETTO DI STUDIO**



IL FENOMENO COLLETTIVO È L'ASPETTO PARTICOLARE CHE INTERESSA STUDIARE DEL COLLETTIVO

Collettivo statistico

Definizione del fenomeno
Interventi chirurgici di cataratta

Collettivo statistico
Popolazione ospedaliera (Italia, 2002)

Unità statistica

Composta
Ricoverati nei reparti

Semplice
Singolo ricoverato

Carattere

Qualitativo

Quantitativo

Ordinabile
- Livello di istruz.
- Rischio operat.

Sconnesso
- Sesso
- Gruppo sang.

Continuo
- Altezza
- Pressione arter.

Discreto
- PL per reparto
- Giorni di degenza

Fasi dell'indagine statistica: Elaborazione

1 Acquisizione dati

procedimento attraverso il quale i dati rilevati vengono "Immessi" in un elaboratore

2 Controllo qualità

imperfezione degli strumenti di rilevazione, errori di caricamento

3 Organizzazione dell'informazione

creazione di un database

Raccomandazioni generali per le rappresentazioni (grafici o tabelle)

Ogni rappresentazione grafica deve contenere in sé tutte le indicazioni necessarie per la sua esatta interpretazione, indipendentemente dal testo

- **TITOLO** chiaro oggetto della rappresentazione
- **DATA** a cui si riferiscono i dati
- **AMBITO TERRITORIALE**
- **FONTE**
- **UNITA' di MISURA**

Fasi dell'indagine statistica: Presentazione

La sistemazione dei dati in tabelle

La sistemazione dei dati in tabelle ci permette di avere una visione della importanza relativa delle singole modalità dei caratteri investigati.

La frequenza

Il numero di volte che una data modalità si manifesta nel collettivo di riferimento

La distribuzione delle frequenze

La distribuzione delle frequenze descrive come il fenomeno in esame si manifesta nella popolazione (o campione). Distingueremo tra:

- ***Frequenze assolute***
- ***Frequenze relative o percentuali***
- ***Frequenze cumulate***
- ***Frequenze retrocumulate***

Frequenze assolute

La frequenza assoluta di una generica modalità x_i è il numero di volte n_i con cui si presenta quella modalità. L'insieme dei numeri rappresentanti le frequenze assolute n_i associate alle modalità x_i viene chiamato distribuzione di frequenza.

Modalità del carattere							
x_i	x_1	x_2	…	x_i	…	x_k	Totale
Frequenze assolute							
n_i	n_1	n_2	…	n_i	…	n_k	N

Disribuzione di frequenza teorica

Fasi dell'indagine statistica: Presentazione

Distribuzione per unità

Paziente	Dose (mg/die)
1	2,2
2	2
3	2,3
4	3,2
5	3,3
6	2
7	1,7
8	2,5
9	3
10	2,5
11	3
12	2
13	2,2
14	2
15	3
16	2,5
17	2,5
18	2
19	2
20	3
21	2,5
22	2,5
23	2,5
24	2,5

Distribuzione di frequenza per modalità

Dose (mg/die)	Pazienti da trattare
1,7	1
2	6
2,2	2
2,3	1
2,5	8
3	4
3,2	1
3,3	1
Totale	24

Dosi giornaliere di metadone in pazienti neoplastici con sintomatologia dolorosa (Modificata da: S. Mercadante et al., Annals of Oncology, 7: 613-617, 1996.)

Fasi dell'indagine statistica: Presentazione

Quando si ha un carattere quantitativo che si presenta con molte modalità è utile raggruppare i valori costruendo delle classi

Esempio

Su un collettivo di 40 studenti si registra il peso e le modalità vengono raggruppate in classi di ampiezza 5kg.

Classe	Frequenza
<60	7
[60,65)	9
[65,70)	10
[70, 75)	8
>= 75	6

Fasi dell'indagine statistica: Presentazione

Frequenze relative

$$f_i = \frac{n_i}{n}$$

La frequenza relativa esprime la frazione dei casi osservati che presentano una data modalità del carattere. Essa viene talora moltiplicata per 100; in tal caso viene denominata frequenza %

Tipo di esame	Numero analisi	Frequenze relative	Frequenze percentuali
Esami cardiologici	1.726.000	0,232	23,2
Analisi del sangue	2.590.000	0,348	34,8
Analisi delle urine	2.073.000	0,279	27,9
Altre analisi	1.051.000	0,141	14,1
Totale	7.440.000	1,000	100,0

Frequenze cumulate

In alcuni problemi concernenti caratteri con modalità ordinabili può interessare conoscere la frequenza dei casi che presentano un valore del carattere $< o = a\ x_i$.
Le frequenze cumulate possono essere di tre tipi:

- Assolute
- Relative
- Percentuali

Fasi dell'indagine statistica: Presentazione

Es.: nel collettivo degli studenti frequentanti il corso di statistica I il carattere *Numero di fratelli/sorelle* è così distribuito

Modalità	Freq. Assol.	Freq. Relat.	Freq. Cum.	Freq. Retr.
0	28	0.140	0.140	1.000
1	59	0.295	0.435	0.860
2	51	0.255	0.690	0.565
3	30	0.150	0.840	0.310
4	13	0.065	0.905	0.160
5	7	0.035	0.940	0.095
6	5	0.025	0.965	0.060
7	3	0.015	0.980	0.035
>7	4	0.020	1.000	0.020
Totale	200	1.000		

Fasi dell'indagine statistica: Presentazione

Serie storiche e territoriali

Esprimono la dinamica temporale o spaziale di un certo fenomeno (registrato istantaneamente (var. di stato) o in relazione a un certo periodo (var. di flusso)

Esempi:

- *Numero di nati vivi a Roma mese per mese*
- *Cambio lira - dollaro registrato giornalmente*
- *Il numero di assist di un giocatore di basket ad ogni partita durante una stagione*

Matrice di dati

Un modo generale di rappresentare i risultati di una rilevazione statistica, soprattutto quando i caratteri rilevati sono più di uno è la cosiddetta *matrice unità - variabili* composta da tante righe quante sono le unità osservate e su ogni riga vengono riportate le modalità specifiche per i diversi caratteri.

Esempio: Su un collettivo di 10 città si rilevano:

1. *Popolazione residente*
2. *Numero di ospedali pubblici*
3. *Numero di cinema/multisale*
4. *Numero di centri commerciali*

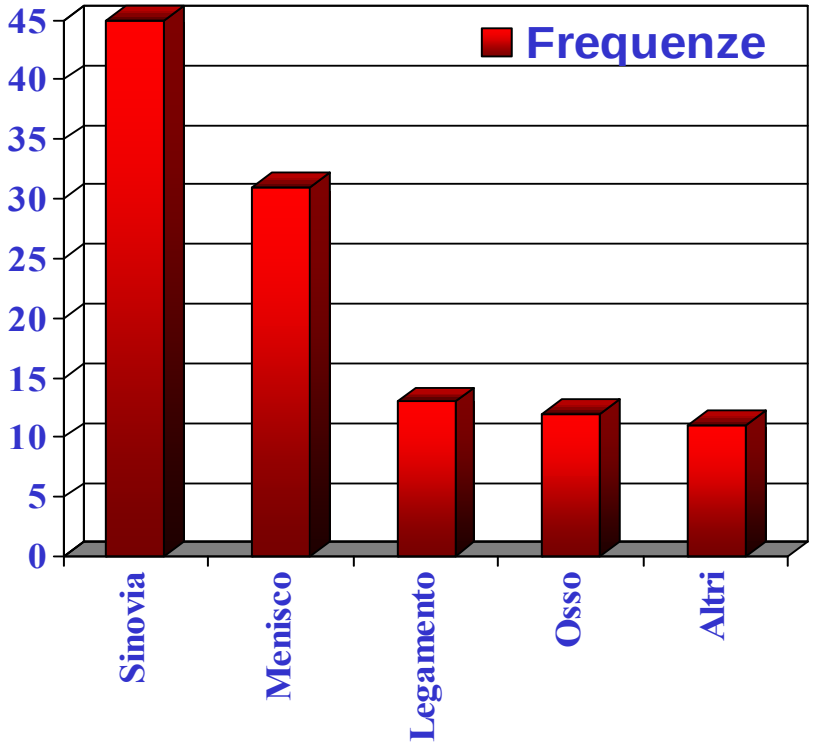
Città	Popolaz.	Ospedali	Cinema	C.Comm.
Roma	**3.824.000**	**18**	**72**	**14**
Milano	**2.726.000**	**16**	**53**	**18**
Napoli	**2.121.000**	**15**	**50**	**11**
Torino	**895.000**	**12**	**27**	**8**
Genova	**598.000**	**13**	**24**	**7**
Palermo	**680.000**	**10**	**21**	**6**

Rappresentazione grafica dei dati

Le rappresentazioni grafiche dei dati possono svolgere una funzione di sintesi illustrando il complesso dei valori in un "colpo d'occhio".

Grafici a colonne

Le frequenze o le quantità sono rappresentate da rettangoli con base simile e altezze proporzionali alle frequenze o alle quantità

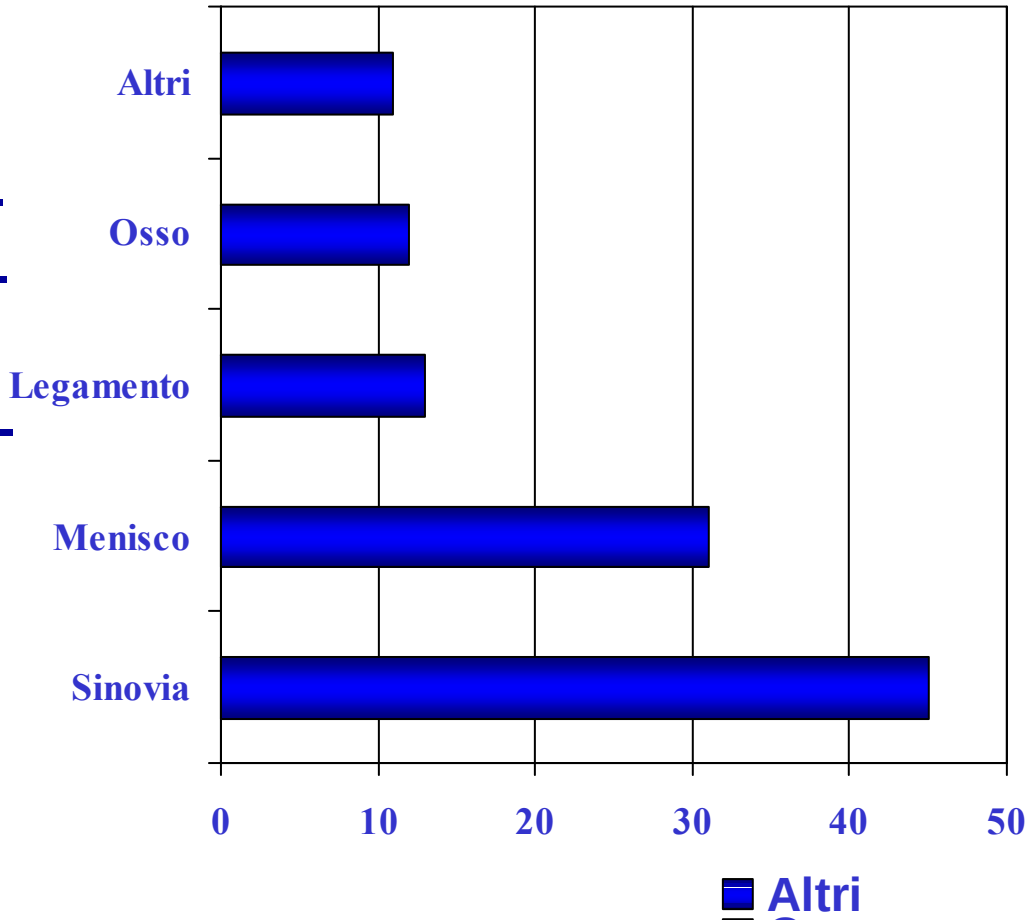


Reperti patologici riscontrati con la RMN in pazienti ricoverati per trauma recente del ginocchio.

Fasi dell'indagine statistica: Presentazione

Grafici a nastri

La frequenza o la quantità di ogni modalità o della situazione di riferimento è rappresentata da rettangoli aventi tutti la stessa altezza e basi proporzionali alle frequenze o alle quantità.



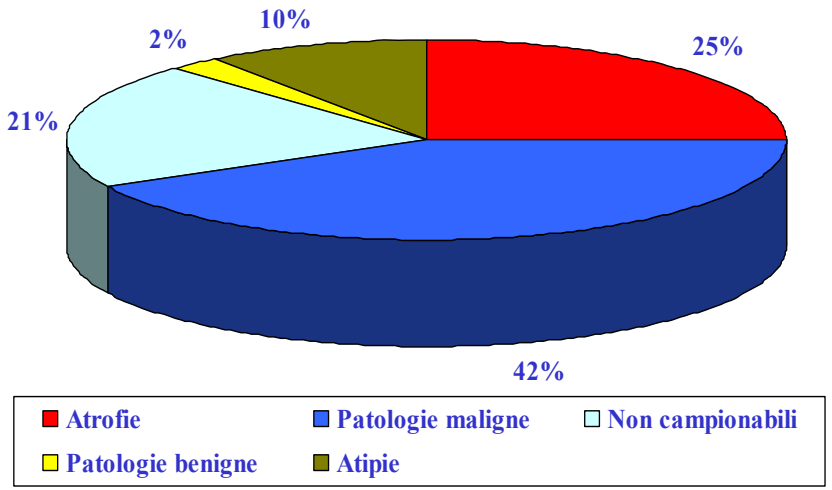
Reperti patologici riscontrati con la RMN in pazienti ricoverati per trauma recente del ginocchio.

Fasi dell'indagine statistica: Presentazione

Aerogrammi

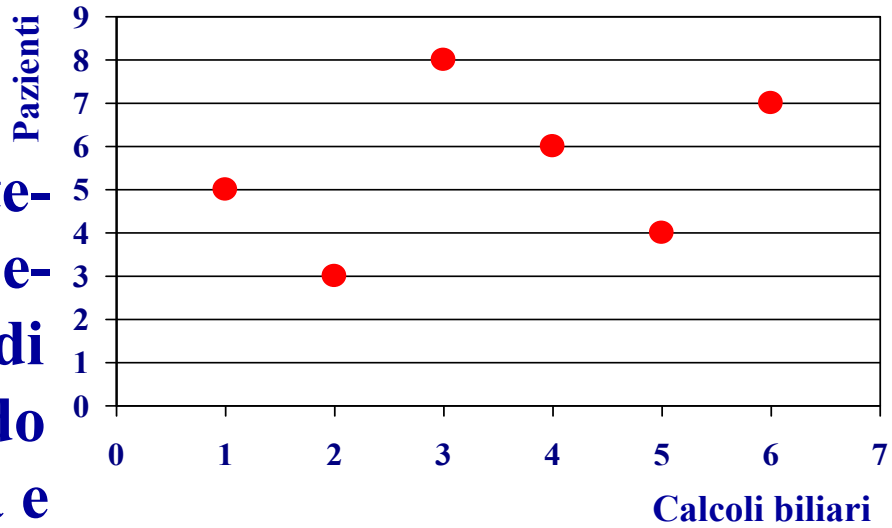
Sono grafici in cui le frequenze o le quantità di una distribuzione sono rappresentate da superfici di figure piane la cui superficie viene divisa proporzionalmente alle frequenze o alle quantità delle modalità.

Patologie riscontrate su T. endometriale in 966 donne



Diagrammi cartesiani

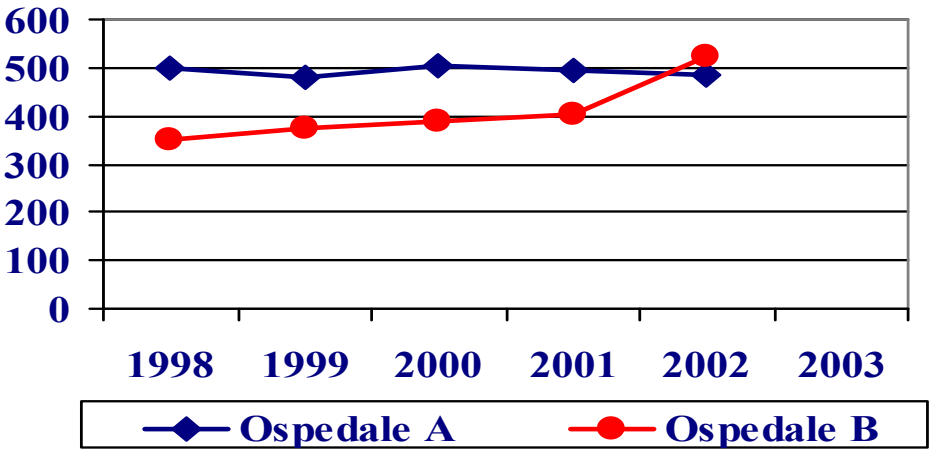
Se la distribuzione è di un carattere quantitativo discreto, può essere rappresentata con un sistema di assi cartesiani ortogonali ponendo sull'asse delle ascisse le modalità e sull'asse delle ordinate le frequenze.



Pazienti con calcoli biliari individuati con ecografia epatica.

Fasi dell'indagine statistica: Presentazione

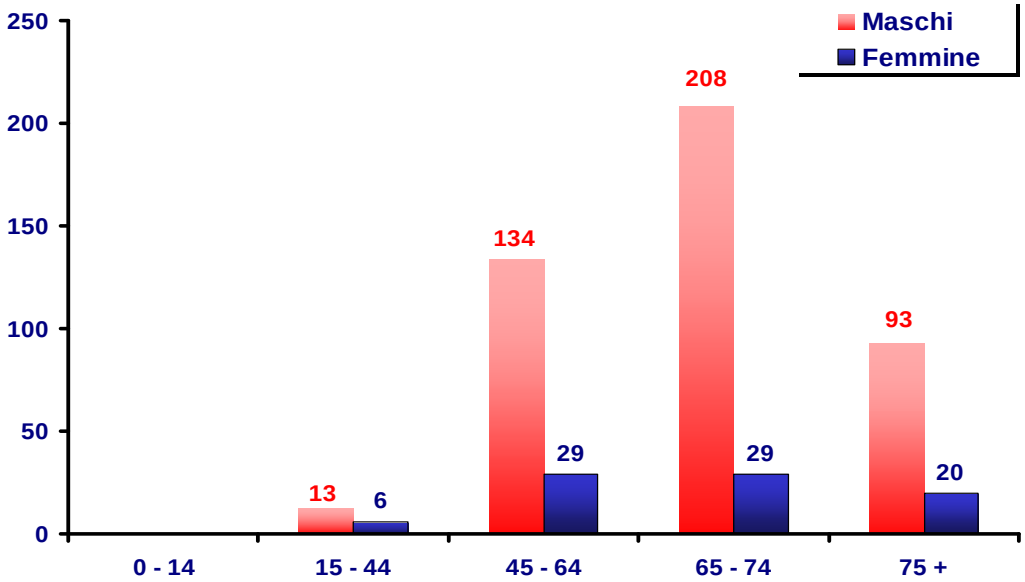
Se la distribuzione è secondo un carattere quantitativo continuo possiamo congiungere i punti per formare una linea continua.



N° di nuovi nati nell'ospedale A e nell'ospedale B

Istogrammi

Gli istogrammi sono grafici che si utilizzano per rappresentare distribuzioni di frequenza di variabili statistiche le cui modalità sono costituite da classi di valori.



Dimessi con diagnosi principale carcinoma polmonare.
Rappresentazione: età e sesso - SDO anno 2000

Fasi dell'indagine statistica: indici di posizione

Sintesi di dati quantitativi

- E' spesso necessario sintetizzare le informazioni fornite da una distribuzione attraverso un semplice indicatore.
- Diversi indicatori forniscono diversi tipi di sintesi
- Gli indicatori di posizione (location index) rappresentano un valore "rappresentativo" di tutti i valori della distribuzione.
- Per forza di cose, essi sacrificano delle informazioni

Misure di posizione

- media aritmetica
- media geometrica
- mediana
- moda

Media

Una media è un numero che esprime la sintesi di una distribuzione statistica.

Media aritmetica

La media aritmetica rappresenta il valore che ogni dato avrebbe se tutti i dati avessero lo stesso valore e se la somma dei valori dei dati rimanesse la stessa.

Il valore medio si rappresenta con $\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$

ed è pari alla somma dei valori di tutti i dati diviso per il numero dei dati:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

La media aritmetica, come tutte le altre medie, è espressa nella stessa unità di misura nella quale è espresso il carattere in studio.

Media aritmetica (segue)

Esempio: *i numeri di scarpa di 20 studenti maschi sono riportati di seguito*

38,40,45,42,42,39,41,43,42,41,46,40,42,41,42,42,40, 41,42,42

La media aritmetica è $\mu = \Sigma_i x_i /20 = 831/20 = 41.550$

Nel caso delle distribuzioni di frequenza: gli stessi dati potrebbero essere organizzati per frequenze Si può allora calcolare la media come (38*1 + 39*1+ 40*3+ 41*4 + 42*8+43*1+45*1+46*1)/20 = 41.550

Numero di scarpa	Frequenza
38	1
39	1
40	3
41	4
42	8
43	1
45	1
46	1
Totale	20

La media così ricavata prende il nome di media aritmetica ponderata con le frequenze

$$\sum_{i=1}^{k} x_i n_i$$

e la sua espressione algebrica è $\mu = \frac{\sum_{i=1}^{k} x_i n_i}{N}$

Media aritmetica (segue)

In presenza di una distribuzione di frequenza con dati raggruppati in classi per il calcolo della media aritmetica bisogna fare delle ipotesi su come si distribuiscono i valori all'interno delle classi. L'ipotesi più comunemente seguita è quella di immaginare che tutti i valori di una classe coincidano con il valore centrale della stessa.

Esempio: *Statura media in 50 studenti*

Statura	Numero studenti	Valore centrale di classe (statura)
160-165	4	162,5
166-175	16	170,5
176-185	19	180,5
186-190	9	188,0
191-195	2	193,0
Totale	50	

$$\mu = \frac{162{,}5*4 + 170{,}5*16 + 180{,}5*9 + 188*9 + 193*2}{50} = \frac{8885{,}5}{50} = 177{,}71$$

Media aritmetica (segue)

Principali proprietà della media aritmetica

1. Internalità

La media aritmetica è interna all'intervallo costituito dal più piccolo e dal più grande valore assunto dal carattere in studio.

$$x1 \leq \mu \leq xk$$

2. Somma degli scarti

La somma degli scarti di tutti i termini dalla media è nulla.

$$\Sigma_i (x_i - \mu) = 0$$

Moda o Norma

E' quella modalità della distribuzione che si presenta con la massima frequenza.

Esempio: *Distribuzione di 150 famiglie secondo il numero dei figli:*

		Moda			
Modalità	**0**	**1**	**2**	**3**	**> di 3**
Frequenze	**20**	**60**	**40**	**18**	**12**
		Max freq.			

- **La moda non è necessariamente unica se nell'esempio precedente anche la modalità 2 avesse avuto una frequenza pari a 60 avremmo avuto due mode.**
- **La moda è un indice molto rozzo perché non tiene conto di quello che avviene "dietro": distribuzioni molto diverse potrebbero avere la stessa moda pur essendo sostanzialmente diverse.**

Moda o Norma (segue)

- **Può avere un comportamento contro intuitivo come dimostra l'esempio seguente.**

Esempio:

X	1	2	3	4	5
Freq.	13	30	35	17	5

La moda è la modalità 3. Se ora spostiamo 20 unità dalla modalità 3 e le mettiamo alla modalità 5, otteniamo:

X	1	2	3	4	5
Freq.	13	30	15	17	25

E' innegabile che la distribuzione si è spostata verso valori più grandi ma la moda ora è 2 !!

Moda o Norma (segue)

Il caso di distribuzioni continue

In questo caso sappiamo che le modalità vengono raggruppate per classi. Occorre allora determinare la classe modale che non necessariamente corrisponde a quella di maggiore frequenza: infatti, bisognerà tenere conto dell'ampiezza delle classi.

Esempio:

X	[0,1)	[1,3)	[3,6)	[6,10)	[10,15]
Freq.	10	22	24	36	25
Fr./Am.	10	11	8	9	5

Qui la classe modale *non è* [6,10) bensì *è* [1,3).
In generale la classe modale è quella con maggiore *densità di frequenza*.

Mediana

La mediana (ME) di una distribuzione è il termine che bipartisce la graduatoria in modo da lasciare alla sua sinistra lo stesso numero di termini che lascia alla sua destra.

Essa può essere calcolata quando:

- **Il carattere è quantitativo**
- **Il carattere è qualitativo ordinabile**

Occorre prima ordinare le osservazioni in modo che le modalità osservate risultino in ordine crescente e poi definire la mediana come la modalità assunta dalla/e unità che occupano la posizione centrale.

Mediana (segue)

Esempio: Si osserva il carattere *numero di fratelli* su un collettivo di 9 studenti e, dopo aver ordinato i valori si ha

1,2, 3, 4, 4, 5, 5, 6, 7

In questo caso *n* è dispari e la mediana è la modalità relativa alla unità che occupa la posizione *(n+1)/2*, ovvero, in questo caso la quinta: la mediana è quindi la modalità Me=4.

Quando *n* è pari non esiste una sola unità mediana, bensì 2. Infatti, se nell'esempio precedente aggiungiamo un'osservazione pari, ad esempio a *5*, la distribuzione diventa

1,2, 3, 4, 4, 5, 5, 5, 6, 7

e le mediane sono ora le modalità osservate sulle unità che occupano le posizioni *n/2* e *n/2+1*.

Questo problema è importante quando *n* è piccolo ma perde importanza per grandi valori di *n*.

Mediana (segue)

- La mediana è di grande importanza in quanto non è influenzata dai valori estremi.
- Si usa dire che è più *robusta* della media aritmetica in quanto è meno influenzata dai dati anomali della media aritmetica.

Esempio: in un laboratorio è stata iniettata a dei topolini una sostanza chimica che aveva come effetto quello di addormentare i topolini.
La media aritmetica dei minuti di sonno è pari a $M = 65{,}2$ mentre la media delle prime 10 unità è solo 35,3.
Il topolino mediano è il $6° \; \frac{11+1}{2}$ a cui corrisponde la modalità 38
In tal caso si ritiene che la sintesi della distribuzione sia meglio espressa dalla mediana.

Unità	N. minuti di sonno
1	12
2	15
3	21
4	25
5	34
6	38
7	40
8	53
9	57
10	58
11	364

Mediana

Dato anomalo

Mediana (segue)

Il caso delle distribuzioni di frequenze

Quando la distribuzione è organizzata per frequenze, la mediana si calcola utilizzando la Funzione di ripartizione (o le Frequenze cumulate). La mediana è quella modalità xi per la quale risulta

- $F(x_{i-1}) < 0.5$
- $F(x_i) \geq 0.5$

Esempio:

X= Tit. di studio	n_i	F_i
Lic. Elementare	6	0.3
Lic. Media	2	0.4
Maturità	6	0.7
Laurea	6	1.0
Totale	20	

La mediana è dunque Me = Maturità perché la decima e la undicesima unità assumono entrambe questa modalità.

Mediana (segue)

Il caso delle variabili continue

Quando le modalità sono raggruppate per classi occorre:

- **Individuare la classe mediana**
- **Assumere una uniforme distribuzione all'interno delle classi**
- **Individuare esattamente la mediana all'interno della classe**

Es.: Classi di statura

Classi	*Val. centrale*	f_i	F_i
[155-164)	159.5	0.093	0.093
[164-169)	166.5	0.194	0.287
[169-174)	171.5	0.290	0.577
[174-179)	176.5	0.248	0.825
[179-184)	181.5	0.126	0.951
[184-194)	189	0.049	1.000
Totali		1.000	

La classe mediana è la classe [169-174). Assumendo distribuzione uniforme nella classe occorre allora individuare il valore che lascia alla sua Sx. esattamente il 50% delle unità. Il problema si risolve partendo

dalla classe mediana $(x_{i+1}, x_i]$ attraverso la formula

$$Me = xi\text{-}1 + (xi - xi\text{-}1)*(0.5 - F(xi\text{-}1)) / (F(xi) - F(xi\text{-}1))$$

difficile da ricordare ma semplice da ricavare. Nell'esempio si ha

$$Me = 169 + 5*(0.5\text{-}0.287)/(0.577\text{-}0.287) = 172.67$$

I percentili o quantili

Si può reinterpretare la mediana come *la più piccola modalità* che lascia alla sua sinistra il 50% delle unità statistiche. Si può effettuare lo stesso ragionamento cercando di individuare la modalità che lascia alla sua sinistra una percentuale di unità statistiche pari ad una frequenza relativa p. In questo senso la mediana diventa il quantile di ordine p=1/2.

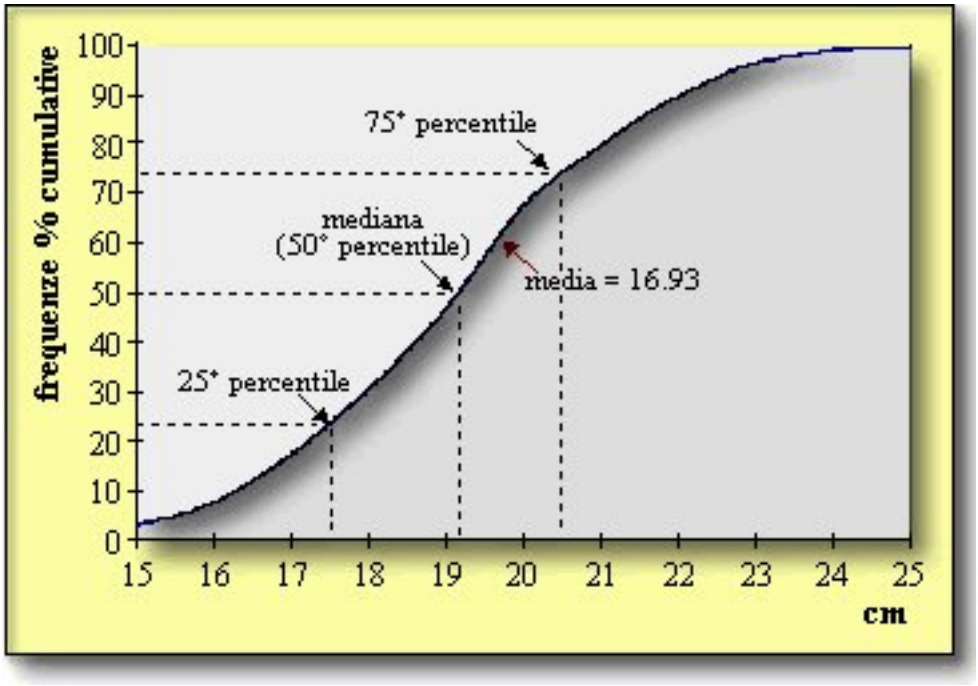
Più in generale si definisce quantile di ordine p la modalità xi tale che: - $F(x_{i-1}) < p$

- $F(xi) \geq p$

I quantili più utilizzati sono i percentili, soprattutto il 25-esimo, 50-esimo (mediana) e il 75-esimo.

Si definisce α-percentile quel valore a Sx del quale vi sono α% dei casi e quindi a Dx (1- α)% dei casi. Il primo quartile è quel valore che nella graduatoria crescente ha, a Sx, 25% dei casi. Il secondo quartile è la mediana; il terzo quartile ha, a Sx, il 75% dei casi.

Fasi dell'indagine statistica: indici di posizione



cm	n.	%	% cumulativa
15	28	2.8	2.8
16	49	4.9	7.7
17	93	9.3	170
18	136	13.6	30.6
19	159	15.9	46.5
20	212	21.2	67.7
21	120	12.0	79.7
22	99	9.9	89.6
23	66	6.6	96.2
24	33	3.3	99.5
25	5	0.5	100

moda

distribuzione percentuale: è in forma "standardizzata" e quindi facilita il confronto con altri dati

percentuale cumulativa: utile per il calcolo dei percentili

[dati fittizi]

Gli indici di variabilità e di forma

Un indicatore di posizione non è in grado di fornire informazioni esaurienti su una distribuzione.
Occorre anche capire quanto le modalità assunte dalle varie unità statistiche siano *disperse* intorno all'indice di posizione.
Questa dispersione viene detta *variabilità* per caratteri quantitativi e *mutabilità* per caratteri qualitativi.

- Es. 1. Un reparto produce in serie pezzi meccanici che dovrebbero avere uno spessore prefissato. Conoscere la variabilità delle dimensione dei pezzi dà un'idea della qualità della produzione. (**imprecisione**).
- Es. 2. Ditta di ristorazione che necessita una previsione sul numero di pasti da preparare (**incertezza**).
- Es. 3. Collettivo di studenti su cui rileviamo il numero di esami superati ad una certa data (**disomogeneità**).
- Es. 4. Distribuzione della ricchezza (carattere trasferibile) in una popolazione (**concentrazione**).

Occorrono allora indicatori della variabilità che abbiano come obiettivo quello di quantificare il grado di dispersione di un carattere.

Gli indici di variabilità e di forma

Esistono due criteri per misurare la dispersione:

- **Dispersione intorno a un valore medio (μ, Me, …)**
- **Dispersione tra le diverse modalità**

Nel primo caso l'indice di variabilità rappresenta una media degli scarti delle modalità osservate rispetto ad una media.

L'indice più importante è certamente la varianza σ^2 di una distribuzione definita come il *quadrato della media quadratica degli scarti dalla media aritmetica.*

In formula si ha

$$\sigma^2 = \Sigma_i (x_i - \mu)^2 / n$$

oppure, nel caso di distribuzione per frequenze

$$\sigma^2 = \Sigma_i (x_i - \mu)^2 \, n_i / n$$

Gli indici di variabilità e di forma

1. Esempio: Una ginnasta è esaminata da una giuria di 5 persone e i voti che riporta (in trentesimi) sono:

25 25 26 27 29

La media aritmetica vale μ=26.2 e la varianza è

$$(25-26)^2+(25-26)^2+(26-26)^2+(27-26)^2+(29-26)^2/\ 5 = 2.24$$

2. Esempio con dati organizzati per frequenza:

Es. *distribuzione delle partite di calcio dello scorso campionato per numero di gol segnati*

N. gol (x_i)	F_i	$(x_i-\mu)$	$(x_i-\mu)^2$	$(x_i-\mu)^2\ n_i$
0	36	-2.65	7.0225	323.035
1	51	-1.65	2.7225	166.073
2	80	-0.65	0.4225	35.0675
3	52	+0.35	0.1225	6.3700
4	36	+1.35	1.8225	65.6100
5	22	+2.35	5.5225	121.495
6	18	+3.35	11.2225	202.005
7	7	+4.35	18.9225	264.915
8	4	+5.35	28.6225	114.490
Totale	306			1299.06

La media vale
(0*36+1*51+2*80+3*52+4*36+
5*22+6*18+7*7+8*4)/306 = 2.65 (μ)
e la varianza è

$$\sigma^2=1299.06/306 = 4.245$$

Proprietà della varianza

- Come tutti gli indici di variabilità, σ^2 vale 0 quando tutte le unità assumono la stessa modalità.
- Per un carattere trasferibile, il massimo della varianza, fissata la media, si ha quando tutte le unità assumono il valore 0 e una sola detiene il totale nμ In questo caso si ha:

$$\sigma 2 = (0- \mu)2 * (n-1)/n + (n\mu - \mu)2 *1/n =$$
$$\mu 2(n-1)/n + \mu 2\ (n-1)2/n = \mu 2n\ (n-1)/n = \mu 2\ (n-1)$$

La varianza non è espressa nella stessa unità di misura delle osservazioni e per questo viene spesso preferito calcolare lo **scarto quadratico medio (s.q.m.)** o **deviazione standard** che non è altro che la radice quadrata della varianza ovvero σ.

Gli indici di variabilità e di forma

Deviazione standard σ (o **s** se riferita ad un campione)

misura della distanza media dei dati dalla media

$\sigma = \sqrt{\text{varianza}}$

popol. 1	popol. 2
2, 3, 4, 5, 6, 7, 8, 9, 10	5, 6, 6, 6, 6, 6, 6, 6, 7
media = 6	media = 6
σ = 2.6	σ = 0.4

- La deviazione standard è utile per conoscere quanto i dati sono "dispersi" rispetto alla loro media
- Conoscere soltanto la MEDIA non è sufficiente per farsi un'idea della distribuzione dei dati. Esempio
- Nella presentazione dei dati, oltre alla media bisogna sempre specificare almeno la deviazione standard (se i dati hanno distribuzione normale) o un altro indice di variazione (es. mediana)

Per riassumere:

dati a distribuzione normale: specificare media e deviazione standard;

dati a distribuzione deformata: specificare media e percentili.

É importante ricordare che, in ogni caso, la conoscenza della sola media (non accompagnata dalla deviazione standard o dai percentili) non è sufficiente a fornire un'idea della distribuzione dei dati di partenza.

Gli indici di variabilità e di forma

La varianza o lo s.q.m. sono indici assoluti. Spesso è necessario rapportare la dispersione di un fenomeno alla sua entità media.

Esempio:

X_i	2	4	6	Totale
N_i	20	10	20	50

In questo caso la varianza vale

$$\sigma^2 = [(2-4)^2\ 20 + (4-4)^2\ 10 + (6-4)^2\ 20\ ]/50=3.2$$

Se invece

X_i	2000002	2000004	2000006	Totale
N_i	20	10	20	50

La varianza è esattamente la stessa anche se il fenomeno appare molto meno variabile …

Conviene allora considerare il Coefficiente di Variazione, definito come $C.V.=\sigma/\mu$ che nei due casi vale, rispettivamente

3.2/4= 0.8 e 3.2/2000004=0.00000159.

Gli indici di variabilità e di forma

Il *Coefficiente di variazione* viene utilizzato quando si vuole confrontare la variabilità di caratteri espressi con diverse unità di misura o o con un diverso ordine di grandezza.

Tale indice viene calcolato effettuando il rapporto tra l'indice di variabilità e la media. Il suo risultato (un numero puro, cioè un valore indipendente dall'unità di misura) ci dice quante volte la media è contenuta nello scarto quadratico medio.

Esempio: Si vuole conoscere fra i caratteri in studio della tabella quello che presenta minore variabilità

Carattere	μ	σ
Statura	168,59 cm	6,49 cm
Diametro trasverso del torace	28,58 cm	2,35 cm
Peso	64,48 Kg	7,27 cm

C.V. (Statura) = (6,49/168,59)*100 = **3,85**

C.V. (Diametro torace) = 2,35/28,58)*100 = **8,22**

C.V. (Peso) = (7,27/64,48)*100 = **11,27**

Indici di variabilità basati sui quantili

Così come la media aritmetica, tra gli indici di posizione può essere fuorviante in presenza di valori anomali, così la varianza può essere *gonfiata* da valori particolarmente distanti dalla media.

Per evitare tali inconvenienti sono stati proposti indici robusti di variabilità, tra cui ricordiamo il **Range Interquartile**: differenza tra il terzo e il primo quartile (75-esimo e 25-esimo percentile)

$Q_3 - Q_1$

Es.: Classi di statura

Classi	f_i	F_i
[155-164)	0.093	0.093
[164-169)	0.194	0.287
[169-174)	0.290	0.577
[174-179)	0.248	0.825
[179-184)	0.126	0.951
[184-194)	0.049	1.000
Totali	1.000	

Gli indici di variabilità e di forma

Es.: Classi di statura

Classi	f_i	F_i
[155-164)	0.093	0.093
[164-169)	0.194	0.287
[169-174)	0.290	0.577
[174-179)	0.248	0.825
[179-184)	0.126	0.951
[184-194)	0.049	1.000
Totali	1.000	

Abbiamo già calcolato la mediana (*Me = 172.67*). Allo stesso modo si calcolano Q_1 e Q_3.

***La classe di Q_1 è* [164-169) e, applicando la formula analoga a quella vista per la mediana,**

$Q_1 = 164 + (169-164) * (0.25-0.093)/(0.287-0.093) = 168.046$

La classe di Q_3 è [174-179) e, analogamente,

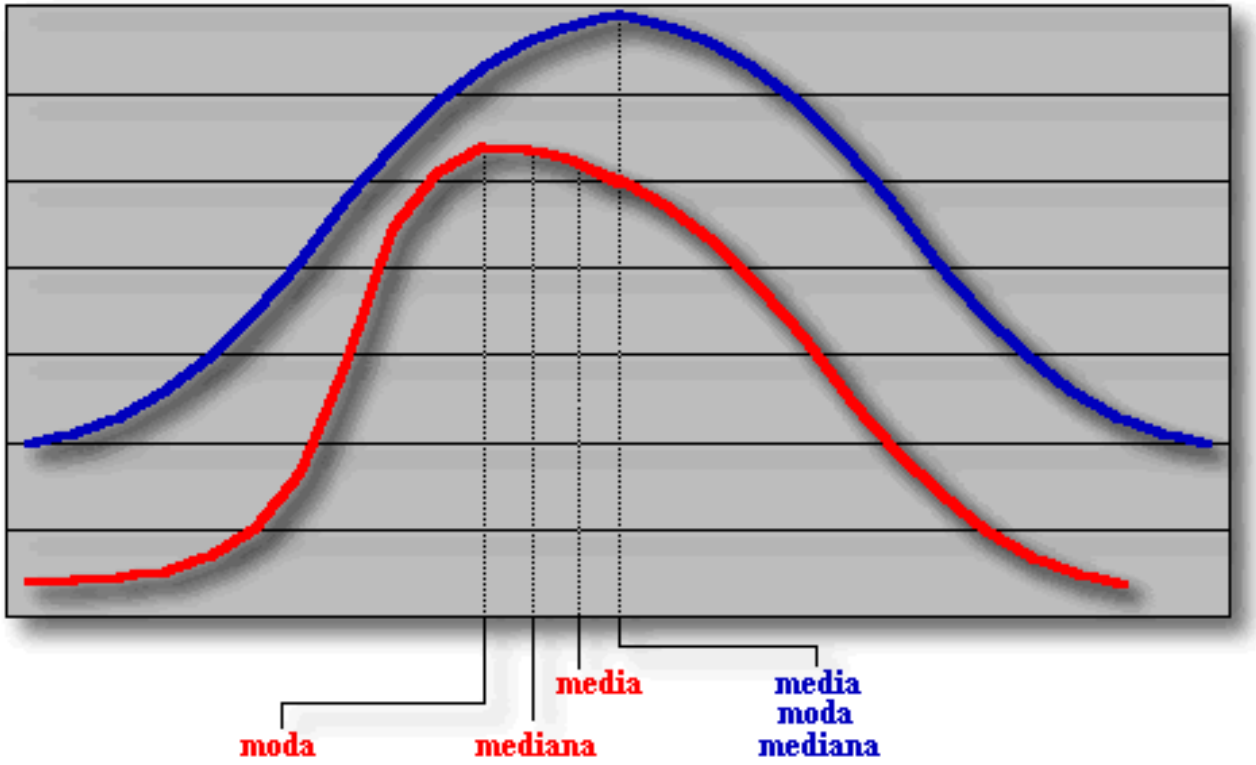
$Q_3 = 174 + (179-174) * (0.75-0.577)/(0.825-0.577) = 177.488$

da cui discende

$Q_3 - Q_1 = 177.488 - 168.046 = 9.442$

Gli indici di variabilità e di forma: l'assimetria

La rappresentazione grafica delle distribuzioni può fornire informazioni cruciali sul comportamento del carattere nella popolazione in esame. Non sempre i dati originano curve simmetriche; talvolta possono essere generate curve più o meno asimmetriche (eventualmente con andamento *bimodale* o trimodale ecc.). Fra le curve asimmetriche, una di quelle più tipiche originata da misurazioni biologiche assume un andamento simile alla curva rossa del sottostante grafico (o un andamento ad essa speculare).

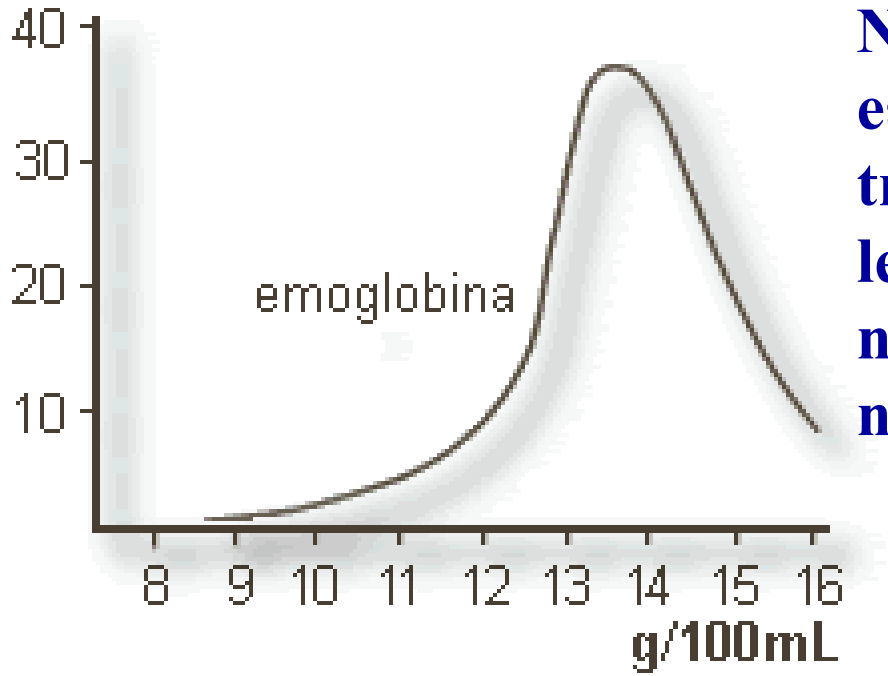


Nella distribuzione **simmetrica**, media, moda e mediana **coincidono**.

Nella distribuzione **deformata**, media, moda e mediana **non coincidono** e la **MEDIA** è l'indice che viene più **distorto** dai dati estremi.

Gli indici di variabilità e di forma: l'assimetria

Molti parametri ematologici hanno una distribuzione pressochè normale. Tuttavia, alcuni di essi mostrano, nell'uomo, una distribuzione con coda deformata positivamente (es. fosfatasi alcalina). Altri test evidenziano, invece, una coda verso sinistra e quindi la distribuzione è asimmetrica e deformata verso i valori negativi.



Nella figura è mostrata un buon esempio di distribuzione asimmetrica: si tratta della frequenza delle distribuzioni della concentrazione di emoglobina nel sangue umano.

Gli indici di variabilità e di forma: l'assimetria

- Ad esempio, quando la media μ coincide con la mediana *Me*, molto spesso questo vuol dire che la distribuzione è di tipo simmetrico unimodale.
- Se invece risulta $Me < \mu$, gran parte delle osservazioni si posiziona su valori bassi ma alcuni valori particolarmente alti spostano la media verso destra: si parla in tal caso di **asimmetria positiva**.
- Se poi risulta $Me > \mu$, gran parte delle osservazioni si posiziona su valori relativamente alti ma alcuni valori bassi spostano la media verso sinistra: si parla in tal caso di **asimmetria negativa**.

Pearson ha proposto come coefficiente di asimmetria:

$$A_2 = \frac{\mu - Me}{\sigma}$$

che varia tra -1 e 1.

Esso è anche detto **coefficiente di Skewness** (in inglese asimmetria).

Gli indici di variabilità e di forma: l'assimetria

Es.:

X_i	1	2	3	4	5	6	7
n_i	7	6	5	4	3	2	1

Qui si ha $\mu = Me = 4$ ma la asimmetria è decisamente positiva. Invece

Y_i	1	2	3	4	5	6	7
n_i	1	2	3	4	5	6	7

Qui $\mu = Me = 4$ ma la asimmetria è decisamente negativa.

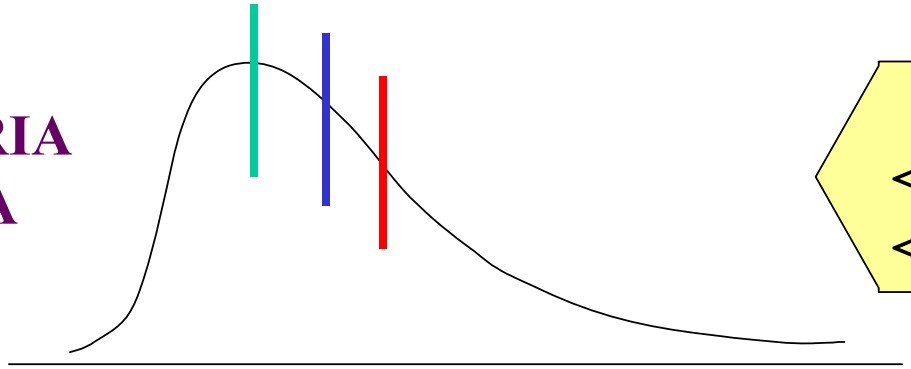
Un indice più sofisticato che risolve questi problemi è stato proposto da Fisher

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3$$

ovvero il cubo della media potenziata di ordine 3 della distribuzione standardizzata delle x_i.

Gli indici di variabilità e di forma: l'assimetria

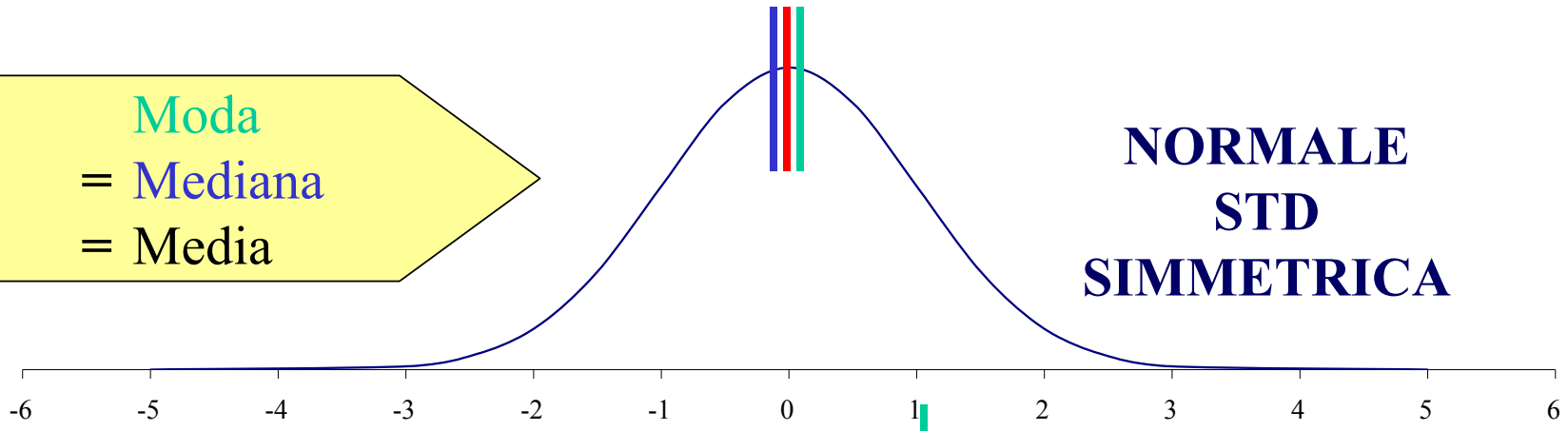
ASIMMETRIA POSITIVA



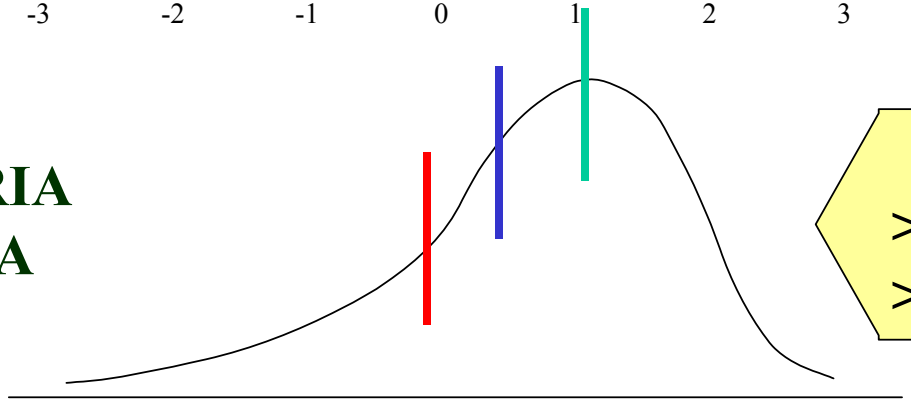
Moda
< Mediana
< Media

Moda
= Mediana
= Media

NORMALE STD SIMMETRICA



ASIMMETRIA NEGATIVA



Moda
> Mediana
> Media

Gli indici di variabilità e di forma: Curtosi

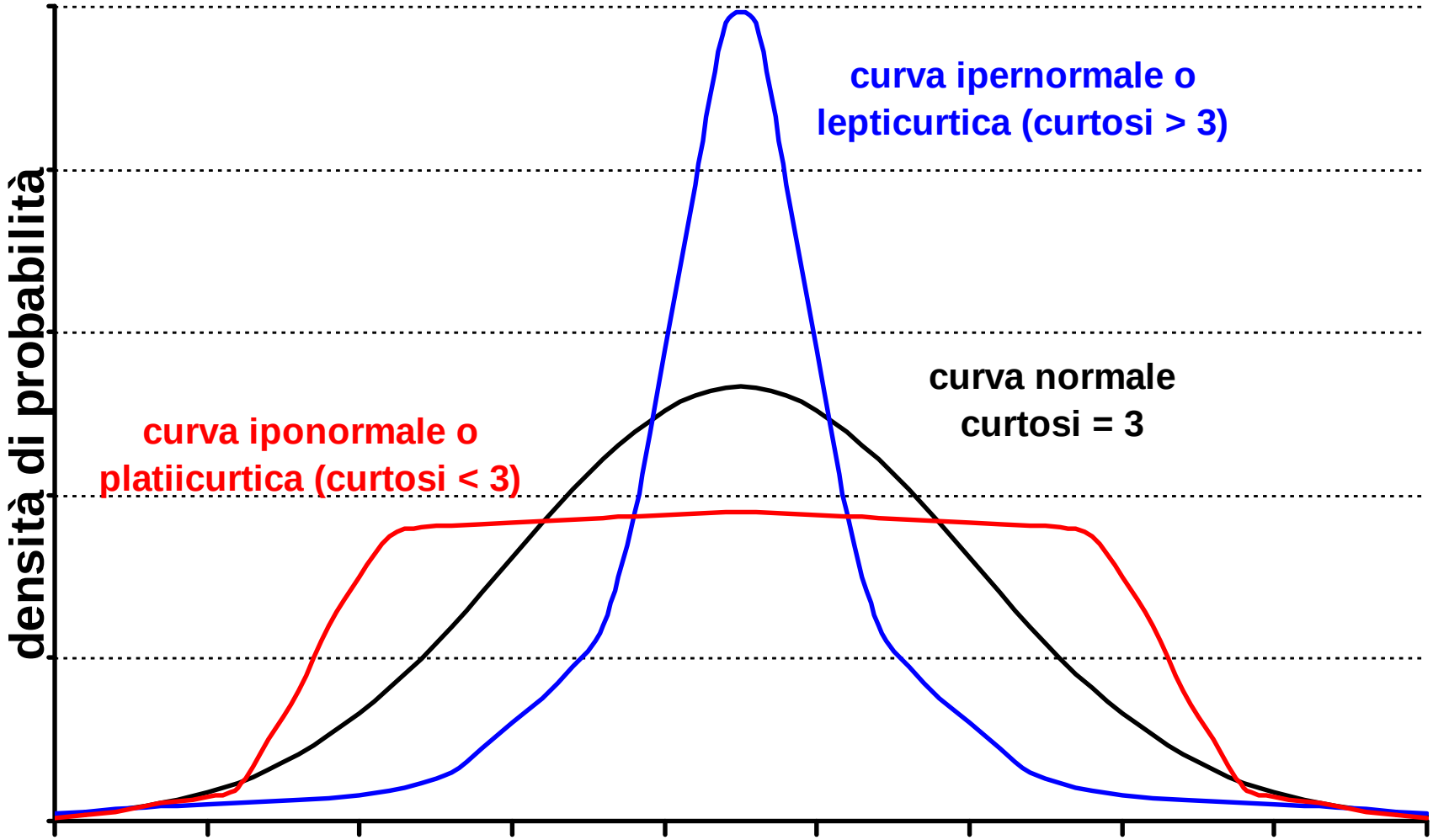
Quando, nelle applicazioni, si ha una distribuzione empirica che è ben rappresentata da una curva simmetrica e campanulare ci si può domandare: la distribuzione è normale? Oppure iponormale, cioè presenta rispetto alla normale una minore frequenza dei valori centrali e di quelli estremi e una maggiore frequenza dei valori intermedi? Oppure è ipernormale, cioè presenta, rispetto alla normale, maggiore frequenza dei valori centrali e dei valori estremi?

Una **misura della anormalità** o **curtosi** della distribuzione (cioè di quanto si discosta dalla curva normale) è il coefficiente di curtosi

$$\text{Indice di Curtosi} = [\Sigma(x - \bar{x})^4/n] \,/\, [\Sigma(x - \bar{x})^2]^2$$

per la curva normale si dimostra che è = 0
per la distribuzione ipernormale è > 0
per la distribuzione iponormale è < 0

Gli indici di variabilità e di forma: Curtosi



Confronti tra distribuzioni: la standardizzazione.

Per poter confrontare almeno qualitativamente due distribuzioni dobbiamo eliminare i fattori che potrebbero oscurare le differenze tra le due distribuzioni. Per questo motivo i confronti si effettuano a parità di media (convenzionalmente posta uguale a zero) e di scarto quadratico medio (posto uguale a 1).

Questa operazione si chiama *standardizzazione* e consiste nel trasformare un carattere *X* attraverso una trasformazione lineare

$$Y = \frac{X - \mu_x}{\sigma_x}$$

si ottiene subito che $\mu_y = \frac{\mu_x - \mu_x}{\sigma_x} = 0$ e $\sigma_y = \frac{\sigma_x}{\sigma_x} = 1$

In tal modo i confronti fra due diverse distribuzioni vengono depurati delle eventuali differenze in posizione e variabilità.

Variabilità biologica, deviazione standard e normalità

Fonti di variazione sono presenti in ogni misurazione di un carattere biologico. Tale variabilità non è tuttavia del tutto imprevedibile: infatti, molti fenomeni naturali seguono un modello teorico definito «curva di distribuzione normale» o «gaussiana».

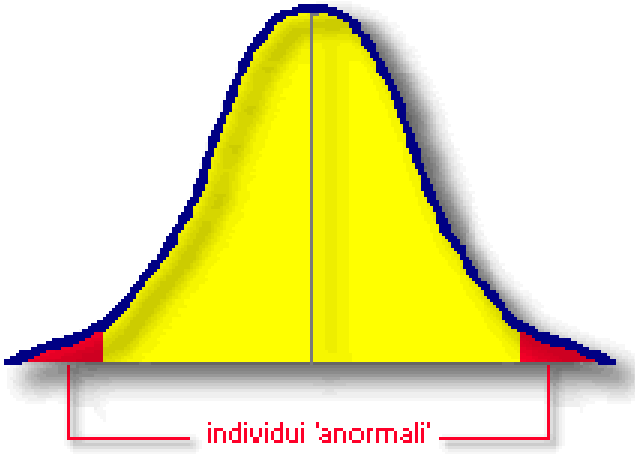
Questo modello è particolarmente utile, in quanto possiamo impiegarlo conoscendo soltanto la media e la deviazione standard.

Infatti, in una gaussiana il 95% dei dati cade nell'intervallo media ± 2 volte la deviazione standard.

Più precisamente, si può dimostrare che l'intervallo (media ± deviazione standard) comprende il 68% circa dei dati; l'intervallo (media ± 2 deviazioni standard) ne comprende il 95% e l'intervallo(media ± 3 deviazioni standard) comprende pressoché tutti i dati (99.7%).

Variabilità biologica, deviazione standard e normalità

La curva di distribuzione normale (o simmetrica o gaussiana) è un modello teorico che si adatta a molti fenomeni naturali. Ha aspetto "a campana" ed è simmetrica rispetto alla media



In una gaussiana

il 95% dei dati cade entro l'intervallo **media ± 2σ**

il 99.7% dei dati cade entro l'intervallo **media ± 3σ**

in medicina è criterio comune assumere come **LIMITI DELLA NORMALITA'** il 2.5° ed il 97.5° percentile della distribuzione dei dati di una popolazione sana. Cioè: **MEDIA + 2 σ**

Nella scienza medica una delle domande più frequenti che sorgono immediatamente quando si viene a conoscenza di un valore di una misura biologica eseguita su un individuo è:

«si tratta di un valore «normale»?

Variabilità biologica, deviazione standard e normalità

ESEMPI. Sono stati ottenuti i seguenti valori. Possono essere considerati "normali"?

- **Frequenza cardiaca: 120 pulsazioni al minuto;**
- **globuli bianchi: 6.000;**
- **Emoglobina: 12,5;**
- **Pressione arteriosa: 140/80.**

La definizione dei limiti della normalità è un processo complicato; per la variabilità biologica, qualsiasi valore potrebbe essere normale, almeno in linea teorica. La complessità del problema, anzi la impossibilità a risolverlo in maniera definitiva, è dimostrata indirettamente dal fatto che sono stati proposti diversi criteri per stabilire la «normalità», e che nessuno di essi è immune da critiche.

Tuttavia:

Variabilità biologica, deviazione standard e normalità

In medicina il criterio di 'normalità' accettato è quello di assumere come limiti il 2.5 ed il 97.5 percentile della distribuzione dei dati di una popolazione 'sana'

Quindi

normale=frequente e anormale=raro

Se tale distribuzione segue una curva gaussiana (come avviene spesso), allora il limite della normalità ora indicato corrisponde alla espressione (media ± 2 deviazioni standard).

Nel caso in cui la distribuzione sia asimmetrica, pur valendo sempre il principio del 2.5° e 97.5° percentile, il range di normalità non potrà essere calcolato semplicemente come (media ± 2 ds), ma dovrà essere accertato in altro modo (ad esempio individuando i percentili in un tracciato cumulativo di frequenze).

Variabilità biologica, deviazione standard e normalità

Una semplice critica che si può avanzare riguardo alla definizione di normalità ora esposta è la seguente: se vengono considerati anormali tutti gli individui che si trovano al di sotto del 2.5 percentile ed al di sopra del 97.5 percentile, allora la *prevalenza* (ossia la frequenza) di ogni malattia dovrebbe essere esattamente pari al 5%; cioè, in una popolazione sarà sempre ammalato il 5% degli individui.

Ciò evidentemente non è compatibile con il comune modo di intendere la frequenza di una malattia.

Attenzione, una critica alla critica: nell'obiezione ora esposta si assume (erroneamente!!!) che anormale sia sinonimo di ammalato.

La Statistica Inferenziale

Statistica Inferenziale

- **Metodo induttivo**
  (dal particolare al generale)
- **Rilevazioni parziali**
- **Stima dei parametri di popolazione**
- **Verifica delle ipotesi**
- **Previsione**

Probabilità: storia e definizione

Breve storia

- Il calcolo della probabilità costituisce uno dei principali strumenti della statistica inferenziale.
- Le sue origini risalgono al XVII secolo e sono abitualmente attribuite a Blaise Pascal (1623-1662) la cui attenzione fu richiamata su problematiche concernenti il gioco dei dadi dal Cav. De Mèrè.
- Oggi il calcolo della probabilità è inteso come un ramo della matematica che studia il concetto di probabilità di cui ancora non esiste una interpretazione univoca.

Definizione classica

Secondo tale impostazione la *probabilità* è il rapporto tra il numero *K* dei casi favorevoli all'evento e il numero *N* dei casi possibili, purchè questi ultimi siano tutti ugualmente probabili

Probabilità: definizione

Più propriamente, è bene parlare di «frequenza relativa»: infatti la probabilità (Pr) che si verifichi un evento aleatorio A è data dal rapporto tra il numero di casi favorevoli (quelli in cui A si verifica) ed il numero di casi possibili (cioè il numero di volte che A può verificarsi).

La probabilità (Pr) di un evento è l'espressione quantitativa delle frequenza con cui esso si verifica ("FREQUENZA RELATIVA")

La Pr di un evento A si esprime come:

$$Pr(A) = \frac{\text{numero delle volte che A si verifica}}{\text{numero delle volte che A può verificarsi}}$$

Esempio 1: lanciando un dado regolare, la probabilità di avere come risultato la faccia contrassegnata dal numero 6 è uguale a 1/6, dove 1 rappresenta il numero dei casi favorevoli e 6 il numero dei casi possibili.

Probabilità: esempi

ESEMPIO 2: Lanciando una moneta l'evento «testa» si verifica una volta ogni due lanci, quindi la probabilità di tale evento è 1/2, cioè 0.5.

ESEMPIO 3: Qual è la probabilità che una carta da gioco estratta a caso da un mazzo di 52 sia un asso? Poiché nel mazzo vi sono 4 assi, la probabilità è di 4/52. In questo il numero di casi favorevoli è pari a 4 (asso di cuori, quadri, fiori, picche), mentre ognuna delle 52 carte del mazzo rappresenta un *potenziale* evento favorevole.

ESEMPIO 4: Supponiamo che in un episodio di intossicazione alimentare si siano verificati 48 casi su un totale di 192 persone alimentate con il cibo contaminato. La probabilità di ammalare per un soggetto scelto a caso è stata pertanto: 48/192 = 0.25 ovvero 25%.
Notare che, diversamente dai tre esempi precedenti, in questo caso si tratta di una probabilità *a posteriori*, cioé valutata su un evento già accaduto.

Probabilità: definizione

Poiché *K* è minore o uguale ad *N*, tale quoziente assume sempre valori maggiori o uguali a 0 oppure minori o uguali a 1.

Dagli esempi si nota che la probabilità può venire espressa attraverso una frazione, un numero decimale o una percentuale. Il numero decimale assume sempre un valore compreso fra 1 (quando l'evento si verifica sempre, e quindi il numeratore è uguale al denominatore) e 0 (quando l'evento non si verifica mai, e quindi il numeratore è uguale a 0).

É intuitivo che: la probabilità che un certo carattere sia vero per un individuo scelto a caso da una popolazione equivale alla proporzione della popolazione che è provvista di quel carattere.

Esempio: se il 10% della popolazione è mancina, avremo una probabilità pari a 0.1 (10%) che un individuo preso a caso da quella popolazione sia mancino.

Probabilità: spazio degli eventi

L'insieme di tutti gli eventi rappresentanti i risultati della prova è definito *spazio degli eventi* ed è indicato con la lettera Ω (omega).

Es.1: lanciando un dado, lo spazio degli eventi è costituito dai numeri compresi tra 1 e 6 che possono aversi come esito finale dello esperimento, Ω = {1,2,3,4,5,6};

Es.2: sottoponendo un individuo ad un intervento chirurgico, Ω è costituito da tutti i possibili risultati dell'operazione, Ω = {guarigione del paziente, non guarigione del paziente};

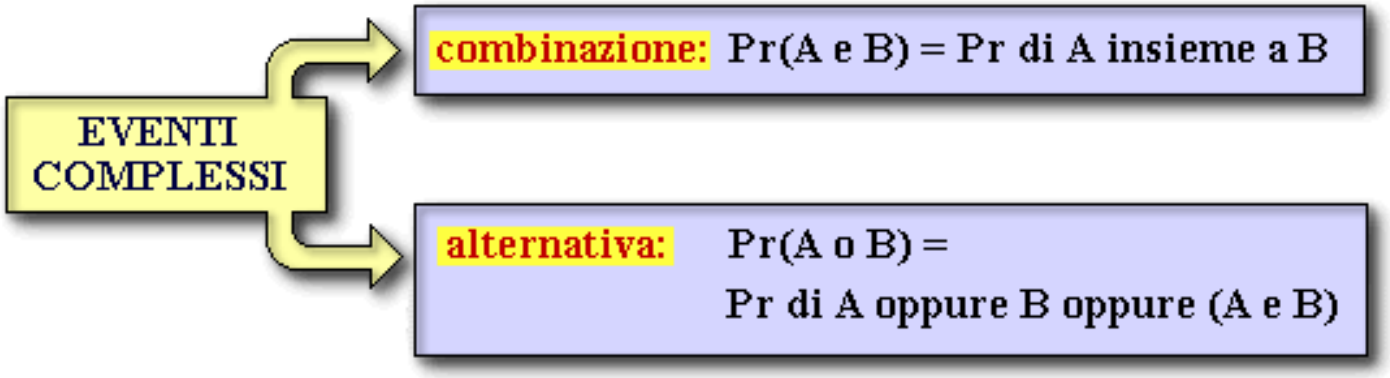
Es.3: prelevando un campione di sangue da un individuo per stabilirne il gruppo di appartenenza, lo spazio degli eventi è individuato dall'insieme di tutti i possibili gruppi sanguigni esistenti, Ω = {A, B,AB, 0}.

Probabilità: eventi complessi

Definizione:

Vi sono situazioni in cui occorre valutare la probabilità di eventi che si esprimono come *combinazioni* specifiche (es. evento A *e* evento B) oppure come *alternative* specifiche (es. eventi A *o* evento B). In questi casi si parla di "eventi complessi".

Gli eventi complessi si gestiscono attraverso due regole di base:



Gli eventi complessi si gestiscono attraverso due regole di base:

- **la regola della moltiplicazione**
- **la regola della addizione**

Probabilità: regola della moltiplicazione

La regola della moltiplicazione si applica ad una combinazione di eventi; essa stabilisce che la probabilità (Pr) che si verifichino contemporaneamente l'evento A e l'evento B equivale al prodotto delle probabilità di ciascun evento:

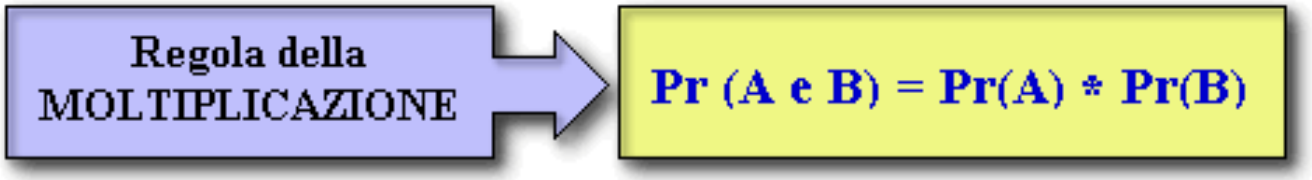
$$Pr(A \text{ e } B) = Pr(A) * Pr(B)$$

ed anche

$$Pr(A \text{ e } B \text{ e } C) = Pr(A) * Pr(B) * Pr(C)$$

e così via.

Questa regola vale soltanto se A e B sono indipendenti, cioè nel caso in cui il verificarsi di A non influenzi il verificarsi di B e viceversa.

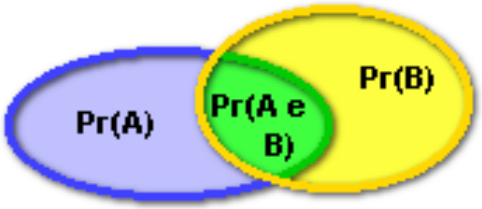


Probabilità: regola della addizione

La regola dell'addizione si applica, invece, ad una alternativa di eventi; essa stabilisce che la probabilità che si verifichi A oppure B oppure entrambi equivale alla somma delle probabilità dei singoli eventi. E' necessario però considerare se i due eventi si escludono reciprocamente (ossia il verificarsi di uno inibisce la possibilità del verificarsi dell'altro) oppure no.

Esempio: Se si lancia un dado, gli eventi "ottenimento di un 2" e "ottenimento di un 3" si escludono reciprocamente. Infatti, non è possibile ottenere contemporaneamente un 2 e un 3 nello stesso lancio. Il verificarsi di un evento esclude la possibilità dell'altro.

Nel caso in cui A e B non si escludano reciprocamente (cioè possa verificarsi A e B congiuntamente), a tale somma è necessario sottrarre la Pr (A e B).



$$Pr(A \text{ o } B) = Pr(A) + Pr(B) - Pr(A \text{ e } B)$$

Probabilità: regola della addizione

Regola della ADDIZIONE

$$Pr(A \text{ o } B) = Pr(A) + Pr(B) - Pr(A \text{ e } B)$$

ma se A e B si escludono reciprocamente:

$$Pr(A \text{ o } B) = Pr(A) + Pr(B)$$

Esempio:

Supponiamo che un certo farmaco induca nel 10% dei casi un effetto collaterale (EC). Se somministriamo il farmaco a 2 soggetti: A e B

qual'è la probabilità che entrambi presentino EC (eventi indipendenti: EC in A non influenza EC in B

$Pr(A \text{ e } B) = 0.1*0.1 = 0.01$ (1%)

qual'è la probabilità che almeno uno presenti EC (gli eventi non si escludono reciprocamente)

$Pr(A \text{ o } B) = 0.1+0.1-0.01 = 0.19$ (19%)

Probabilità: distribuzione

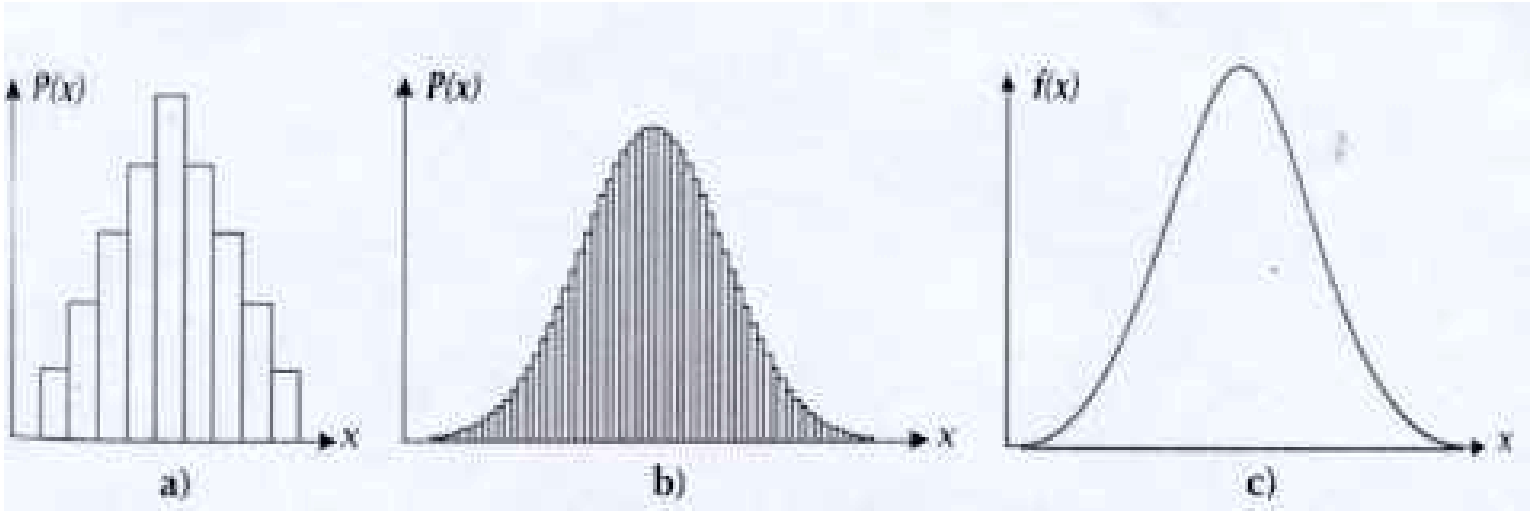
L'insieme delle probabilità a (somma unitaria) associate ai risultati di una prova costituisce la *distribuzione di probabilità*.

Esempio 1: Immaginiamo di sapere che su 100 individui sottoposti ad una determinata operazione chirurgica, 95 sono guariti. Possiamo determinare la probabilità di guarigione (95/100) = 0,95 e la probabilità di non guarigione [(100-95)/100] = (5/100) = 0,05 e, quindi, costruire una una distribuzione di probabilità.

Questa è costituita dai risultati dell'esperimento "intervento chirurgico" (guarigione e non guarigione del paziente) e dalle probabilità ad essi associate, la somma delle quali è (0,95 + 0,05) = 1.

Probabilità: distribuzione

Esempio 2: Immaginiamo di rilevare la statura di un gruppo di 60 donne. Poiché la variabile statura è continua, la rappresentazione grafica che ne deriva sarà la fig. a.

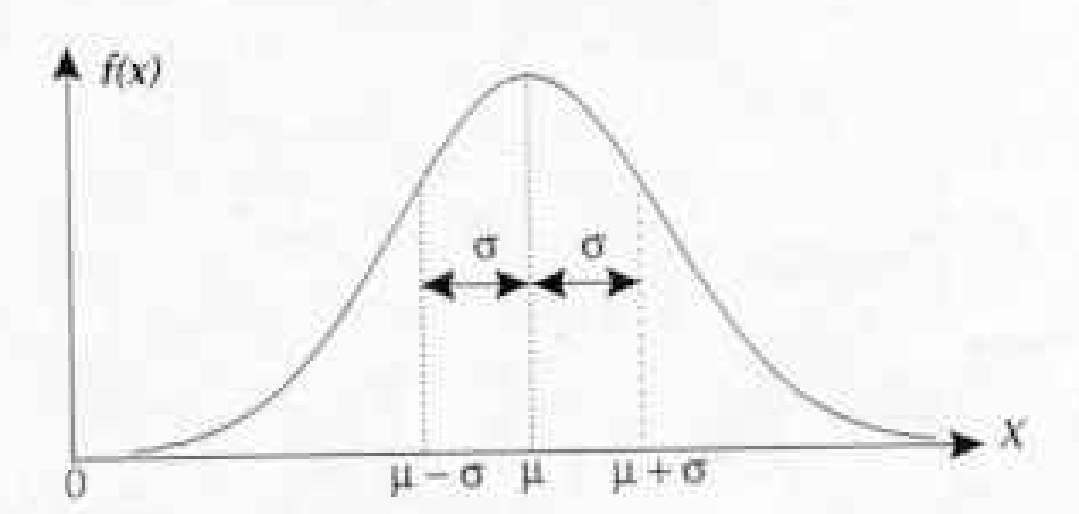


Supponiamo ora di voler rappresentare graficamente la distribuzione di frquenze della statura di un insieme di 10.000 donne. Aumentando il numero delle prove, aumentano i risultati (fig. b). Di conseguenza, gli istogrammi rappresentanti tale distribuzione sono una quantità maggiore, più sottili e meno distinguibili l'uno dall'altro, tanto da poter essere bene approssimati da una curva (fig. c).

Probabilità: distribuzione teorica

La *curva normale o di Gauss*. È una *distribuzione di probabilità teorica*, ossia una distribuzione di probabilità definita da una funzione matematica nota, della quale è conosciuta la rappresentazione grafica e mediante la quale possiamo descrivere alcuni fenomeni reali.

La sua importanza risiede nel fatto che, nella popolazione, molte delle misure biomediche (pressione sanguigna, statura, glicemia, temperatura corporea, peso, colesterolo, ecc.) si distribuiscono normalmente.



Probabilità: proprietà della curva di Gauss

La curva di gauss gode delle seguenti proprietà:

- **può assumere tutti i valori compresi tra - ∞ e + ∞;**
- **è a forma di campana, unimodale e simmetrica rispetto alla μ della distribuzione;**
- **è crescente da - ∞ a μ e decrescente a + ∞;**
- **è asintotica rispetto all'asse delle ascisse, ossia si avvicina molto senza toccarlo mai;**
- **le osservazioni si addensano attorno a μ mentre sono meno frequenti man mano che ci si allontana dalla stessa;**
- **l'area sottesa alla curva è pari a 1 e, a seguito della sua simmetria è posizionata metà a destra e metà a sinistra della stessa μ;**
- **il parametro μ determina la posizione della curva, mentre il parametro scarto quadratico medio (σ) la larghezza.**

  Se il valore assunto dalla μ aumenta (diminuisce), la curva si sposta verso destra (sinistra); se invece aumenta (diminuisce) lo σ, la curva diventa più bassa e più larga (più alta e più stretta).

Probabilità: curva di Gauss

La distribuzione di Gauss assume un ruolo di primaria importanza nelle tecniche tipiche della inferenza statistica e nella risoluzione di alcuni problemi pratici.

Esempio: sapendo che la pressione sistolica si distribuisce normalmente con media pari a 120 mmHg e σ = 10 mmHg, possiamo voler determinare la frequenza relativa degli individui con valori della pressione di 130 mmHg.

Il problema è risolvibile ricorrendo alla *distribuzione normale standardizzata*, che rappresenta una particolare distribuzione gaussiana con μ pari a 0 e varianza σ^2 uguale a 1, mediante la quale possono essere risolti i problemi inerenti tutte le distribuzioni normali con media non nulla e varianza maggiore di 1.

Per poter utilizzare la normale standardizzata, tuttavia, è necessario procedere ad un'opportuna operazione di trasformazione del valore x = 130 mmHg, denominata standardizzazione.

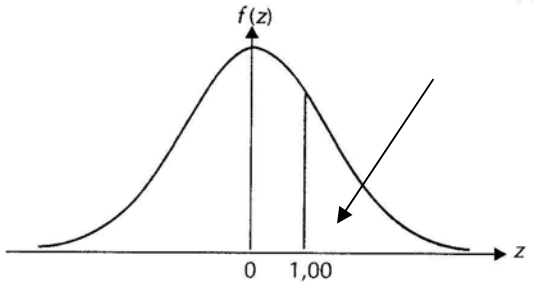
Probabilità: distribuzione standardizzata

Tale trasformazione è eseguita sottraendo ad x la media μ =120 mmHg e dividendo il risultato per σ = 10 mmHg, in modo da ottenere: $z = \{(x-\mu)/\sigma\} = \{(130-120)/10\} = 1$ in cui z rappresenta, appunto, la trasformata di x.

Una volta calcolato z attraverso l'utilizzo delle tavole della curva normale standardizzata possiamo determinare la porzione di area a destra di tale valore. Quest'ultima rappresentando la somma di un numero infinito di sottilissimi istogrammi, corrisponde alla frequenza relativa degli individui che hanno pressione > di 130 mmHg.

I valori di z sono letti con due decimali, il primo dei quali compare nella prima colonna della tabella mentre il secondo nella prima riga della stessa.

La frequenza relativa dei soggetti con pressione maggiore o uguale a 130 mmHg è pari a 0,1587, ossia il 15,9% circa della popolazione presenta una pressione sistolica > di 130 mHg



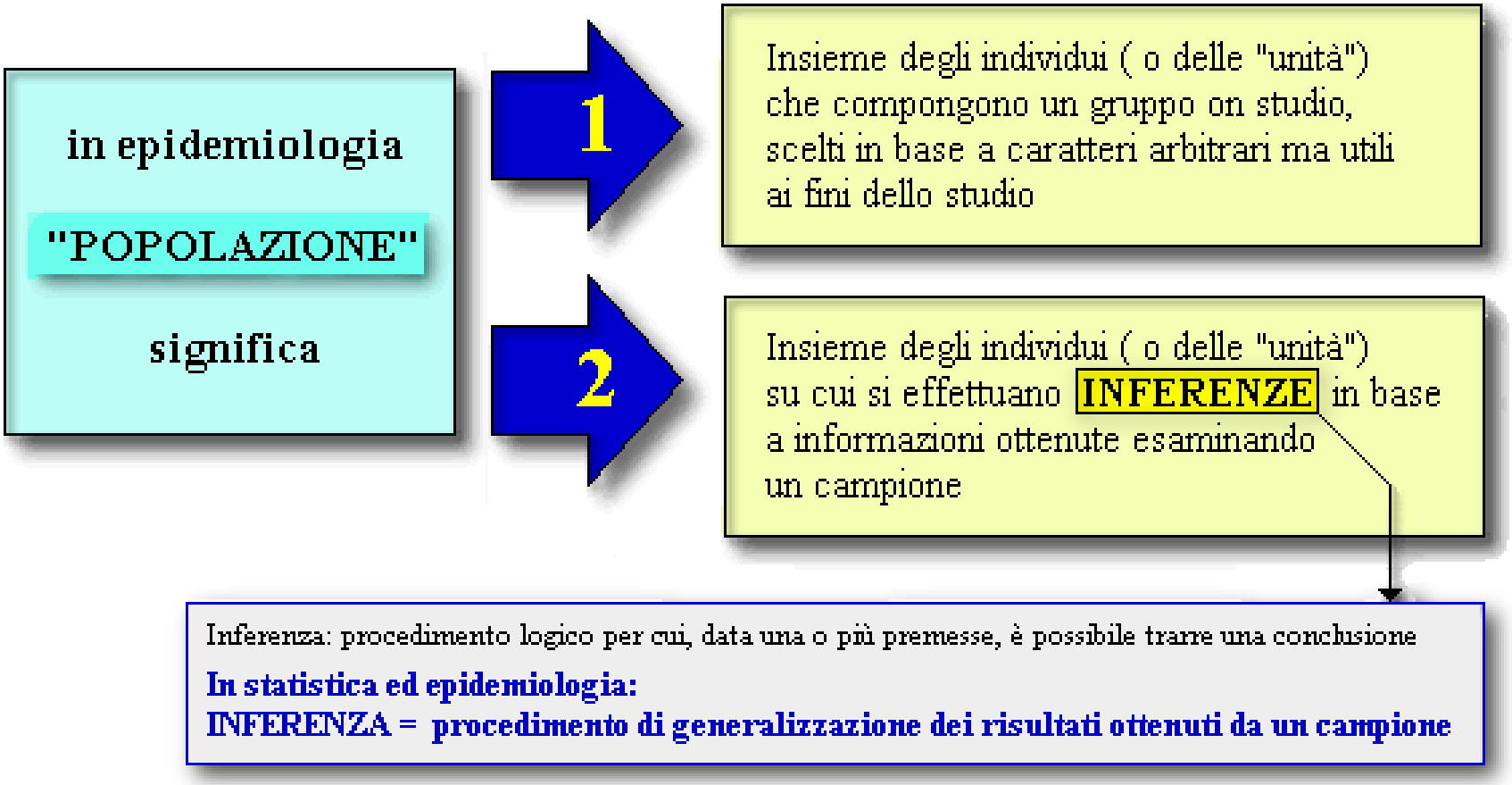
Probabilità: distribuzione standardizzata

Area compresa tra 0 e + ∞ nella curva normale standartizzata

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	**0,1587**	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	**0,1251**	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,8691	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0570	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	**0,0281**	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0238	0,0233
2,0	0,0227	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183

Il campionamento: scopi

Il principale obiettivo di un campionamento è quello di raccogliere dati che consentiranno di generalizzare all'intera popolazione i risultati ottenuti dal campione. Questo processo di generalizzazione è detto «inferenza».

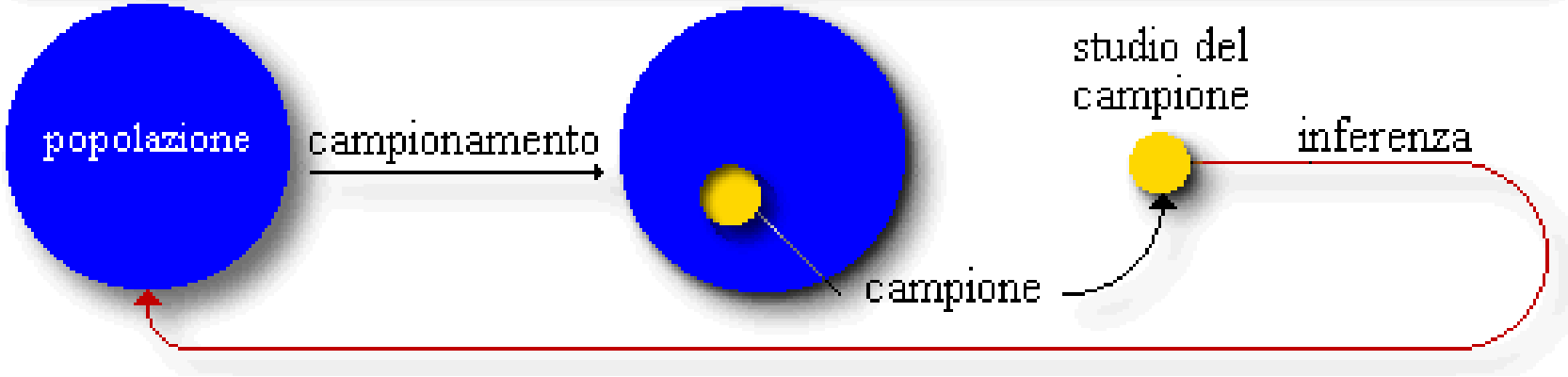


Il campionamento: obiettivi

esaminare tutta la popolazione → censimento

esaminare un campione → indagine o sondaggio o inchiesta (*survey*)

- Raramente possono essere studiate TUTTE le unità che compongono una popolazione
- Pertanto, si studia spesso soltanto una parte (CAMPIONE) della popolazione, per poi generalizzare i risultati (questo processo di generalizzazione si definisce 'inferenza')



Il campionamento: limiti

Quando si effettua uno studio per mezzo di un campione, è necessario tener presente che non si otterranno mai risultati del tutto affidabili. Per valutare la "*bontà*" di uno studio campionario è indispensabile tener conto di vari fattori, fra i quali i più importanti sono:

- **i criteri di scelta della popolazione in studio;**
- **il metodo con cui si è selezionato il campione;**
- **il periodo di osservazione;**
- **i metodi adottati per identificare i casi di malattia;**
- **le tecniche di analisi;**
- **la precisione delle misure effettuate.**

Il campionamento: caratteri del campione

In qualsiasi modo il campione venga scelto, i suoi caratteri non saranno mai IDENTICI a quelli della

popolazione di origine

detta **TARGET**

(insieme degli individui o unità su cui si intende effettuare una inferenza)

la differenza fra il risultato ottenuto dal campione e la vera caratteristica del target è detta

ERRORE DI CAMPIONAMENTO

che non può mai essere calcolato con esattezza ma....

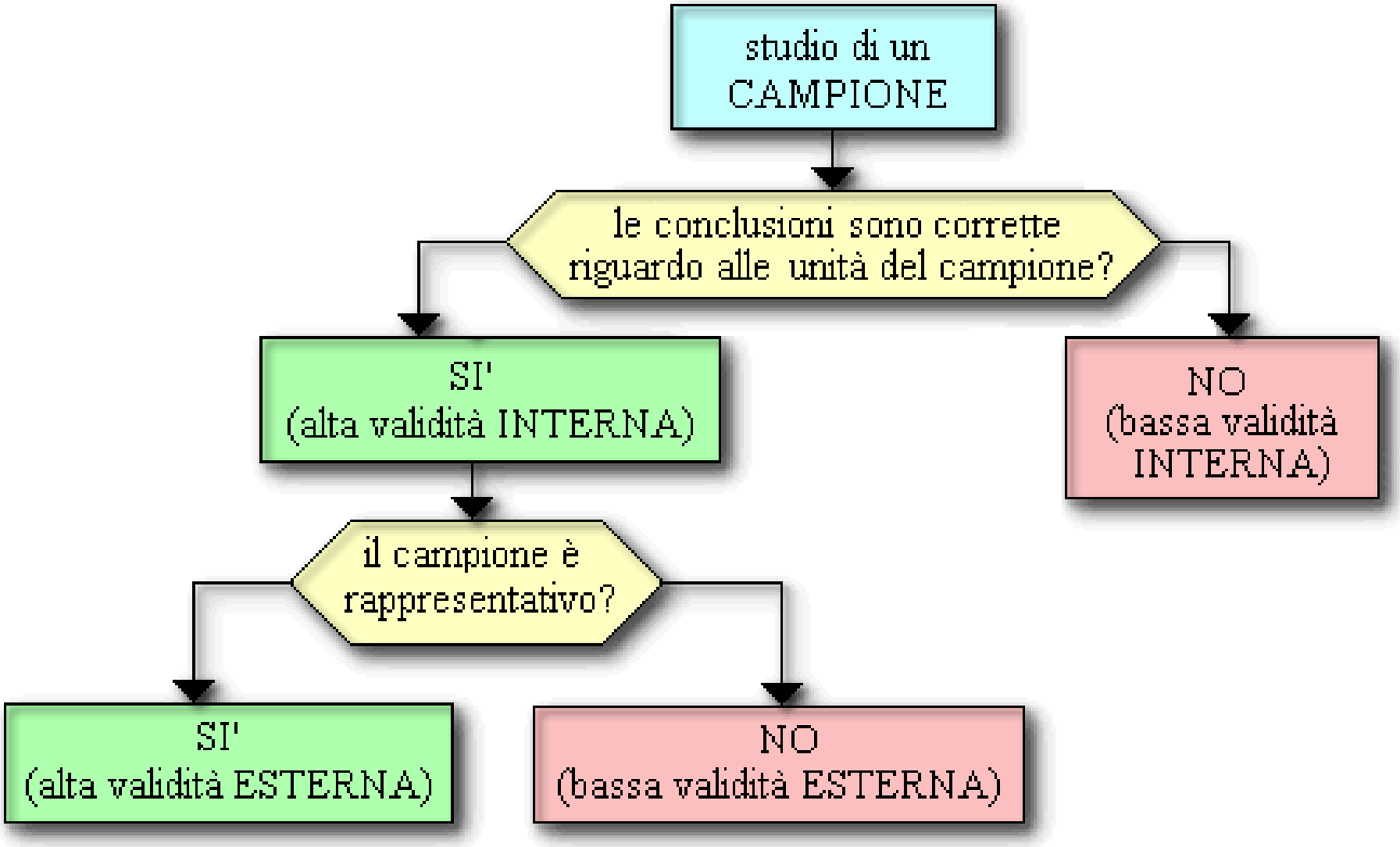
PUO ESSERE STIMATO

Immaginiamo di aver effettuato una indagine esaminando ciascuna unità che componeva il campione. A questo punto, esaminando i dati forniti dal campione al fine di trarne delle conclusioni, si pongono due domande fondamentali:

1. **Le conclusioni sono corrette per le unità che compongono il campione?**
2. **Se sì, il campione rappresenta bene la popolazione da cui è stato estratto?**

Il campionamento: validità del campione

La risposta alle domande è rappresentata da questo schema



Il campionamento: limiti

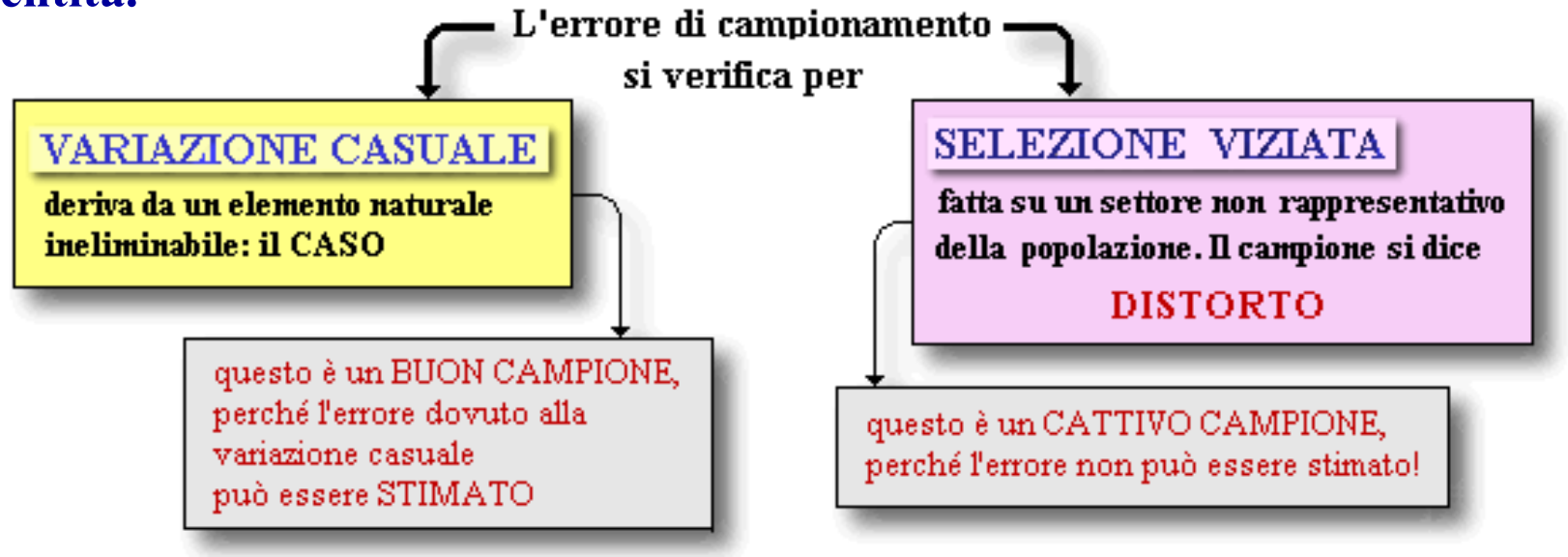
Attraverso lo studio di un campione, si può soltanto *stimare* (cioè determinare con un certo margine di errore) il carattere della popolazione da cui il campione deriva; tuttavia, tale carattere non potrà mai essere determinato con esattezza; la *accuratezza* della *stima* è direttamente correlata al numero di osservazioni che si compiono del fenomeno in studio.

Con qualunque *metodo* si effettui il campionamento, si otterranno dal campione dei risultati che quasi certamente si discostano (poco o tanto) dalla «vera» misura della popolazione. Ciò avviene perché non possiamo mai essere sicuri che il campione rappresenti una copia perfetta della popolazione da cui esso è stato estratto.

Il campionamento: errore

L'errore di campionamento è rappresentato dalla differenza tra i risultati ottenuti dal campione e la vera caratteristica della popolazione che vogliamo stimare.

L'errore di campionamento non può mai essere determinato con esattezza, in quanto la «vera» caratteristica della popolazione è (e resterà!) ignota. Esso tuttavia può essere contenuto entro limiti più o meno ristretti adottando appropriati metodi di campionamento. Inoltre, esso può essere *stimato*; ciò significa che, con adatti metodi statistici, si possono determinare i limiti probabili della sua entità.



Il campionamento: errore

- La *selezione viziata* è quella effettuata su un campione non rappresentativo

- *bias o distorsione*: differenza, causata da un errore sistematico, tra la stima ottenuta da un campione e la vera caratteristica della popolazione

Soltanto quando la scelta degli individui che compongono il campione è stata dettata dal puro e semplice caso, è possibile prevedere e calcolare l'entità della differenza tra campione e popolazione. In caso contrario, il campione si dice «distorto» o «biassato».
Con un campione distorto, non è possibile calcolare l'errore di campionamento ed i dati ottenuti saranno difficilmente utilizzabili.

Il campionamento: Metodi

Campionamento non probabilistico

- Non basato sulla randomizzazione ma su altri criteri (es. comodità, accessibilità ecc.)
- Soggetto a forte distorsione *(bias)*

Campionamento per randomizzazione stratificata

- la popolazione da cui selezionare il campione viene suddivisa in strati in base ad un determinato carattere
- all'interno di ciascuno strato si seleziona un campione con il metodo della randomizzazione semplice o sistematica

Campionamento per randomizzazione sistematica

- le unità del campione vengono selezionate dalla popolazione ad intervalli regolari
- è un metodo più pratico rispetto alla randomizzazione semplice
- può essere influenzato da una variabile esterna che si manifesta con ciclicità

Il campionamento: casuale semplice

Campionamento per randomizzazione semplice

- si effettua con un metodo che garantisce la casualità della selezione
- in genere si utilizza un calcolatore con apposito software oppure le tavole generatrici di numeri casuali

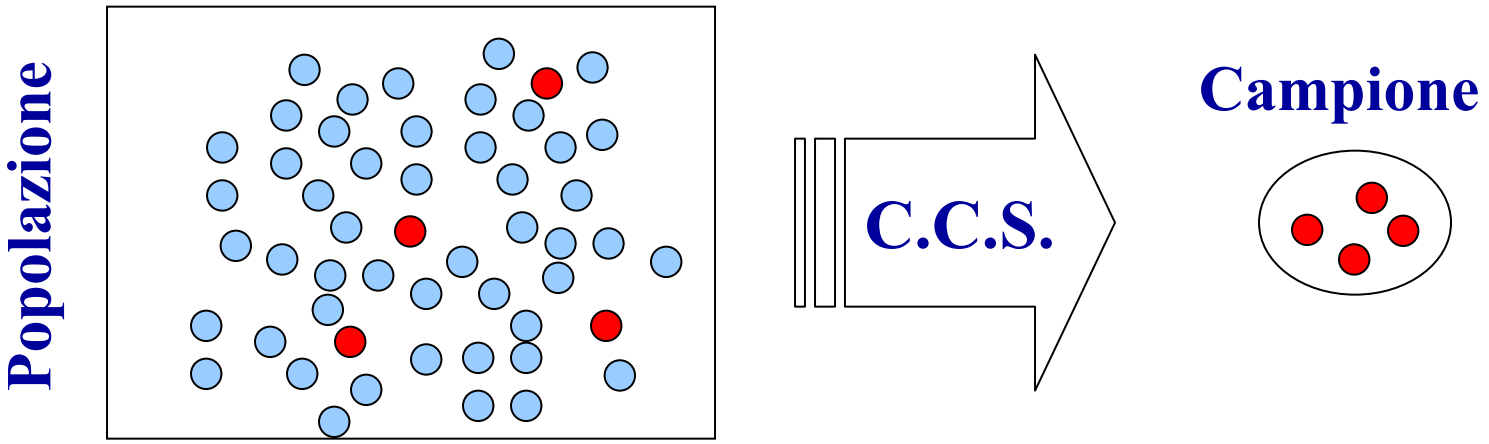
Il campionamento per randomizzazione semplice si effettua estraendo una certa quota di unità dalla popolazione attraverso un metodo che garantisce la casualità delle estrazioni. Questa viene ottenuta, ad esempio, con il classico sistema dell'estrazione di un numero, come avviene nel gioco della "tombola".

Più verosimilmente, nella pratica si utilizza un computer provvisto di apposito software oppure le cosiddette «tavole generatrici di numeri casuali».

TAVOLA NUMERI RANDOMIZZATI

Riga								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	83089	57067	50211	47487
....	.....							

Il campionamento: Casuale Semplice

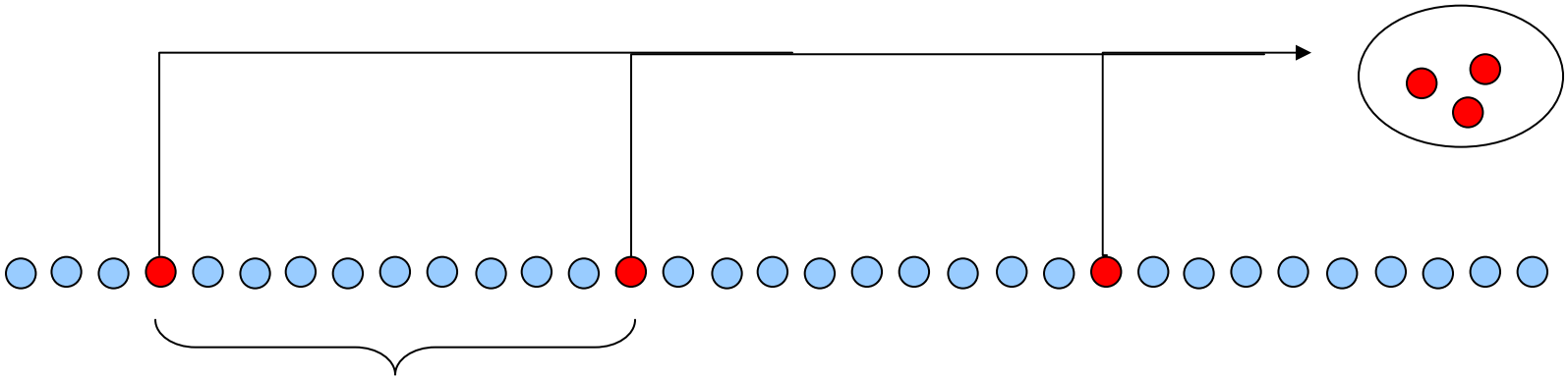


Tutte le *unità di rilevazione* della popolazione hanno la medesima probabilità di essere selezionate e costituire dunque le *unità di osservazione* del campione

Il numero di campioni possibili è dato da: $C_N^n = \frac{N!}{(N-n)! \cdot n!}$

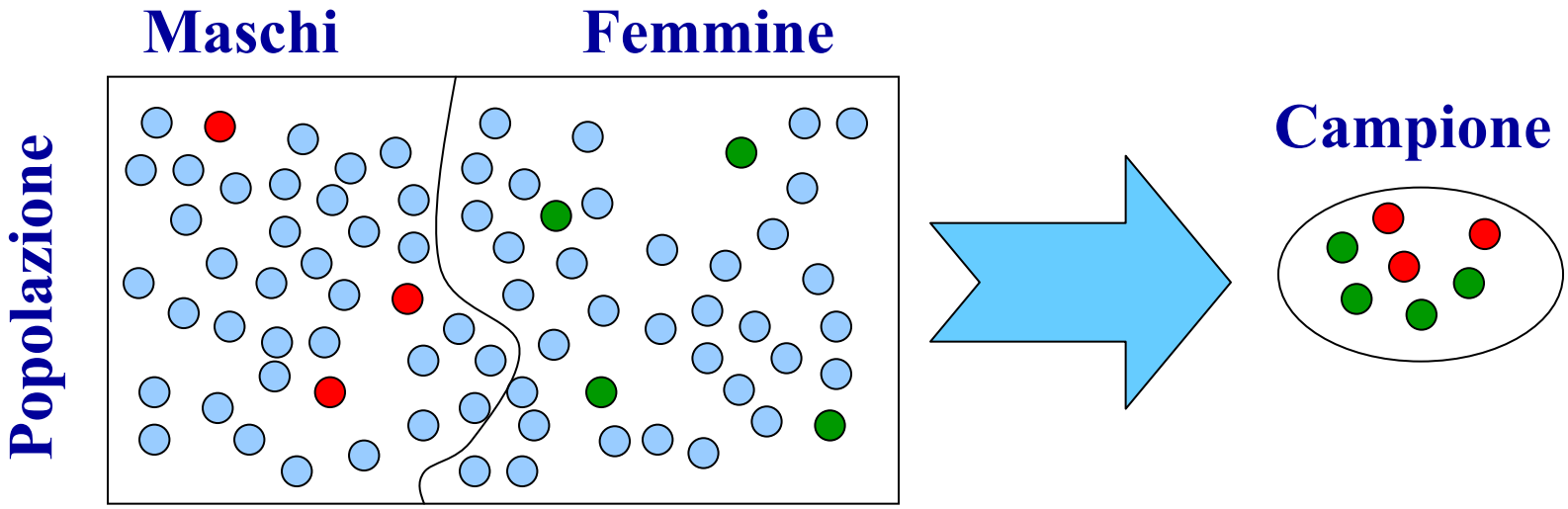
dove: n/N = frazione di campionamento

Il campionamento: Sistematico



- **Si dispone di una lista "*naturalmente*" ordinata delle unità di rilevazione**
- **Si stabilisce la numerosità campionaria *n***
- **Si determina il passo di campionamento $k = N/n$**
  (Inverso della frazione di campionamento)
- **Si estrae un numero a caso tra i primi *k* (*1ª unità campionata*) e si campionano successivamente le unità della popolazione, mantenendo un "passo" costante pari a *k***

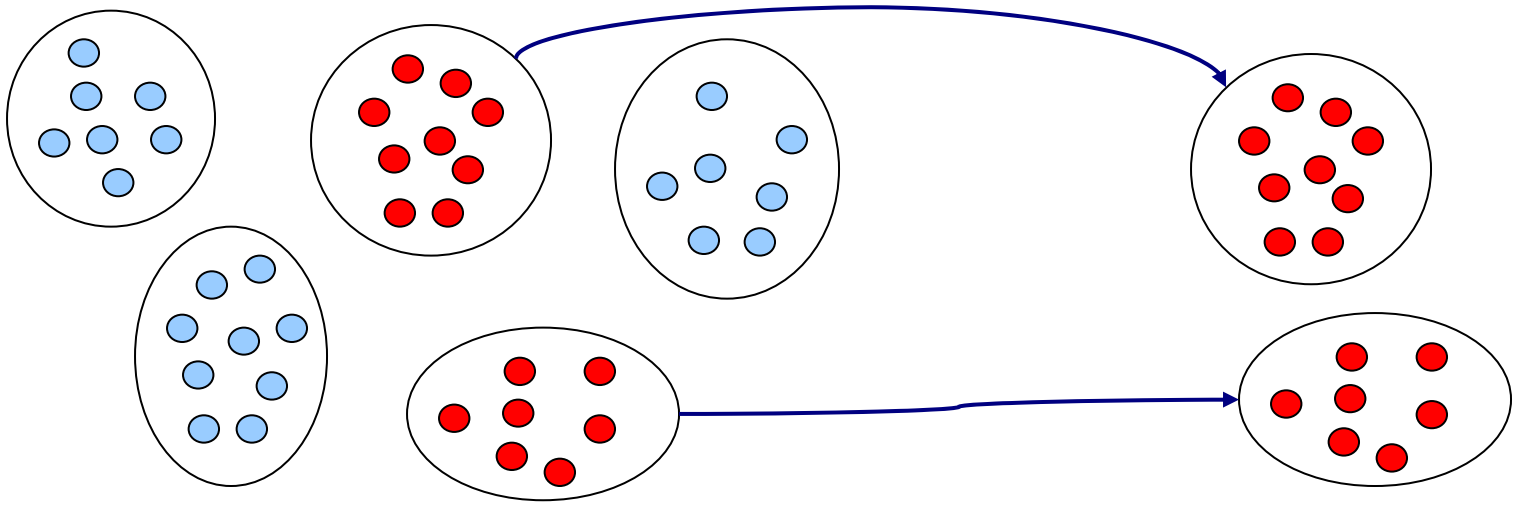
Il campionamento: Stratificato



La popolazione viene suddivisa in strati in base ad una o più *variabili ausiliarie* (e.g. maschi/femmine), da ciascuno dei quali si campionano i soggetti attraverso un C.C.S.

Vengono quindi condotte analisi separate per ciascuno strato, in seguito combinate e ponderate appropriatamente

Il campionamento: a Grappolo



- **Dalla popolazione, divisa in M gruppi, si estrae con un C.C.S.s.r. un numero di grappoli m (m=M);**
- **successivamente tutte le unità appartenenti ai gruppi (grappoli) selezionati vengono campionate**

esempio

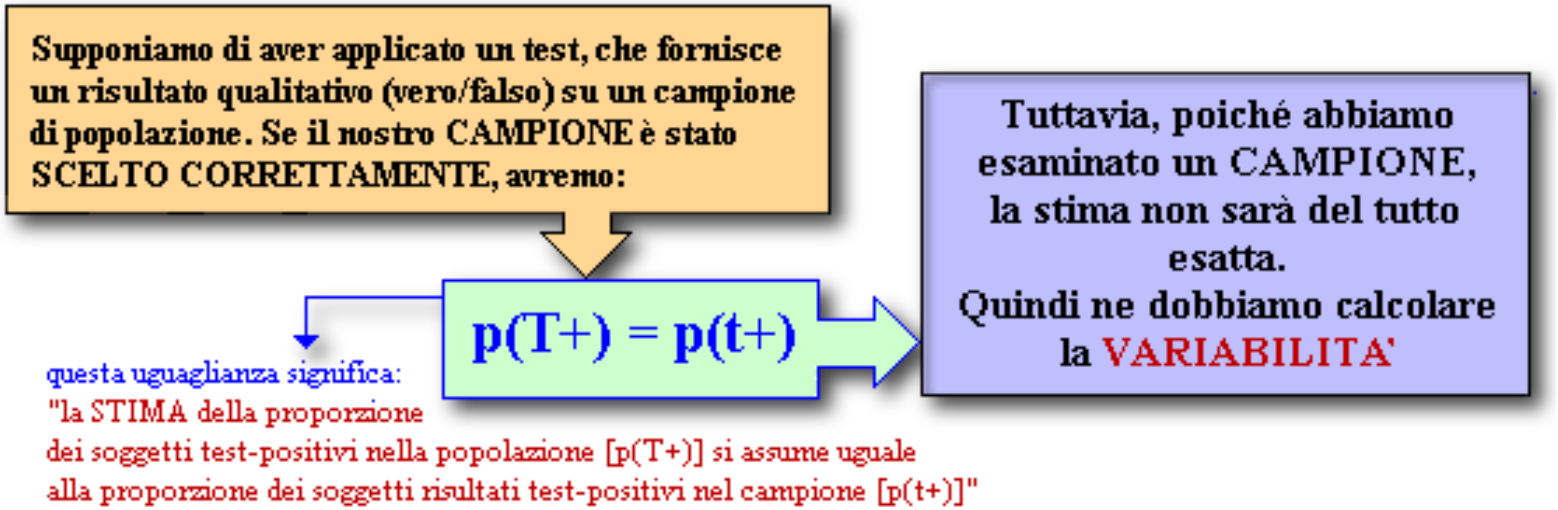
i gruppi possono essere gli ospedali di una determinata area e le unità elementari tutti i pazienti che hanno subito un certo intervento

Variabilità di una stima

Dopo aver estratto un campione di unità da una popolazione, si procede alla misurazione del parametro che interessa su tutti gli individui (o le unità di interesse) che compongono il campione.
Alla fine della nostra indagine, noi conosciamo esattamente lo stato dei soggetti che compongono il campione, ma possiamo soltanto stimare lo stato di tutti i soggetti della popolazione da cui essi provengono.

ESEMPIO:
E' necessario valutare la copertura anticorpale per il virus X in un gruppo di 8000 soggetti. Viene prelevato un campione di sangue da 20 soggetti presi a caso, e poi, sul siero, si effettua la titolazione degli anticorpi. Dei 20 soggetti, 18 (90%) risultano provviste di un titolo tale da farli ritenere «protetti». Pertanto, *stimiamo* che il 90% dei soggetti del gruppo siano protetti.

Variabilità di una stima (segue)



Le caratteristiche del campione ci interessano soltanto in quanto applicabili all'intera popolazione.

Questo processo di generalizzazione, detto inferenza, porta sempre con sé una certa quota di errore, in quanto il campione non potrà mai essere perfettamente rappresentativo della popolazione da cui proviene. Pertanto, attraverso la misura ottenuta dal campione potremo soltanto ottenere una stima della «vera» misura della popolazione.

- **Campioni di grandi dimensioni permettono stime più precise**
- **Studiando un BUON CAMPIONE possiamo ottenere una BUONA STIMA della "vera" misura della popolazione**

Calcolo della variabilità di una stima

In un campione randomizzato di **40** [redacted], **14** sono positivi ad un test

$$p(t+) = \frac{14}{40} = 0.35 = 35\,\%$$

e quindi la STIMA della proporzione di positivi nella popolazione è

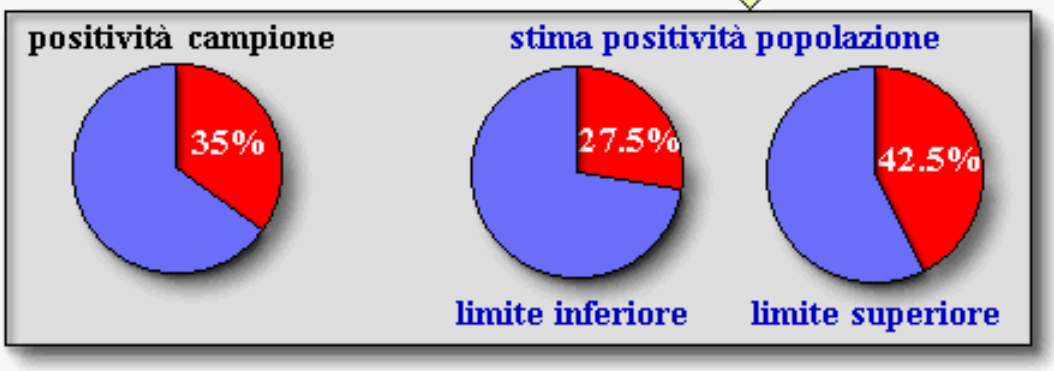
$$P(t+) = \frac{14}{40} = 0.35 = 35\,\%$$

Ora, calcoliamo la VARIABILITA' della stima:

$$\text{varianza}(p) = \frac{p(1-p)}{n} = \frac{0.35 * 0.65}{40} = 0.00569$$

$$\text{errore standard} = \sqrt{\text{varianza}(p)} = \sqrt{0.00569} = 0.0754 = 7.54\%$$

la positività nella popolazione viene STIMATA come: 0.35 ± 0.0754 oppure $35\% \pm 7.54$



Errore standard e limiti fiduciali

La variabilità di una MISURA è caratterizzata da → DEVIAZIONE STANDARD

La variabilità di un VALORE STATISTICO (es. media o percentuale) è caratterizzata da → ERRORE STANDARD

L'errore standard rappresenta un parametro fondamentale, che viene comunemente impiegato per il calcolo dei limiti fiduciali o intervalli fiduciali o intervalli di confidenza. Il limite fiduciale è molto utile per avere un'idea della vera caratteristica della popolazione in studio tramite un campione.

Quanto più piccolo è l'errore standard tanto più attendibile è la statistica

Per campioni ragionevolmente ampi, il limite fiduciale 95% viene espresso come: valore statistico ±2 ES

→ indica che questo intervallo include con probabilità 95% la VERA caratteristica della popolazione

Errore standard e limiti fiduciali (segue)

Il limite fiduciale può essere collocato al livello di probabilità da noi desiderato; comunemente si utilizza un limite fiduciale pari a 0.95 o 0.99. Ci si può esprimere anche in probabilità percentuale, ed allora si dirà «limite fiduciale 95%» o «limite fiduciale 99%».

Che cosa significa l'espressione «confidenza 95%» o «limite fiduciale 95%»?

Significa che vi è una probabilità del 95% che l'intervallo trovato includa la vera caratteristica della popolazione.

In altre parole, se ripetessimo la stessa indagine per 100 volte con gli stessi metodi (ma su 100 campioni diversi), probabilmente otterremmo ogni volta una stima diversa; tuttavia, il vero valore della popolazione sarebbe all'interno del nostro intervallo di confidenza 95 volte su 100.

Per campioni ragionevolmente ampi (almeno 50-60 osservaz.), i limiti fiduciali 95% possono essere espressi come: valore statistico ± 2 volte l'errore standard.

Errore standard e limiti fiduciali (segue)

Nell' esempio già visto di calcolo della variabilità di una stima sul gruppo in studio, il valore statistico trovato era pari a 0.35 e l'errore standard era di 0.0754. Pertanto, la vera positività nella popolazione è compresa fra [0.35 - 0.0754] e [0.35+0.0754]. Adottando i limiti fiduciali al 95%, tali valori diventano pari a 0.202 e 0.498 (20.2% e 49.8%).

In conclusione, possiamo affermare che abbiamo una probabilità del 95% che la percentuale di positività nella popolazione sia compresa fra 20.2 e 49.8.

Il limite fiduciale 95% di una proporzione può essere facilmente stimato con la seguente formula, in cui *p* è la proporzione osservata e N è il numero di unità del campione (per una maggior precisione, moltiplicare per 1.96 anziché per 2).

$$p \pm 2\sqrt{\frac{p(1-p)}{N}}$$

questo è l'errore standard

Nell'esempio:

$$0.35 \pm 2\sqrt{\frac{0.35(1-0.35)}{40}} = 0.35 \pm 2\sqrt{\frac{0.0056875}{40}} = 0.35 \pm 0.150831031$$

Errore standard e limiti fiduciali (segue)

Come si fa a calcolare l'errore standard di una media?

Basta dividere la deviazione standard s per la radice quadrata della numerosità del campione: $ES = \frac{s}{\sqrt{n}}$

Notare che - anche in questo caso - l'errore standard è influenzato dalla numerosità della popolazione studiata: piú grande è la dimensione dello studio, piú grande sarà l'attendibilità della media.

ESEMPIO. Abbiamo misurato il peso di un campione di 80 soggetti. La media è risultata pari a 82.5 kg, con una deviazione standard di 3.5 kg. L'errore standard della media sarà: $ES = \frac{3.5}{\sqrt{80}} = 0.39$

INTERVALLO DI CONFIDENZA
(o intervallo fiduciale) ⇨ stima ± margine di errore

Dimensione o numerosità del campione

Prima di intraprendere un'indagine epidemiologica, è bene conoscere quante «unità di interesse» dovranno essere esaminate per raggiungere con sufficiente attendibilità l'obiettivo desiderato.

Questa è una delle parti più delicate nella pianificazione di una indagine. Ovviamente, più grande sarà il campione e più precisi e attendibili saranno i risultati; tuttavia, indagini su campioni di grandi dimensioni sono costose e richiedono più tempo.

Dimensione o numerosità del campione (segue)

La **varianza** è una misura del grado di variazioni o oscillazioni presenti, relativamente al parametro che vogliamo stimare, nella popolazione.

L'entità di queste variazioni può essere derivata, almeno approssimativamente, dai risultati di altre analoghe indagini effettuate in precedenza o dalla conoscenza della storia naturale della malattia, o da altri fattori.

Una popolazione in cui il parametro da misurare presenta ampie oscillazioni ha una varianza elevata; una popolazione in cui le oscillazioni sono scarse ha una varianza bassa.

L'**intervallo di confidenza** rappresenta una misura della bontà di una stima. Un intervallo di confidenza molto ampio suggerisce che non siamo molto sicuri del punto in cui si trova il «vero» valore. Viceversa, un intervallo ristretto indica che siamo abbastanza sicuri che il valore trovato è piuttosto vicino al valore vero della popolazione; in questo caso la stima sarà, quindi, più precisa.

Il livello di confidenza è una misura della sicurezza della stima: ad esempio, con un livello di confidenza 95% siamo sicuri al 95% che il valore vero cade nell'intervallo trovato. Cioè, se ripetessimo lo studio 20 volte, in media sbaglieremmo 1 volta ma saremmo nel giusto 19 volte.

Confronto fra popolazioni

- Uno scopo della statistica è determinare se le caratteristiche di due popolazioni sono differenti o meno.
- Si traggono cioè conclusioni sulla popolazione, determinando un'inferenza statistica.
- Possiamo confrontare campioni o popolazioni attraverso le medie o le varianze.
- Per effettuare un confronto si ricorre al test statistico.
- Il test statistico è il procedimento che consente di rifiutare o non rifiutare (accettare) un'ipotesi sulla popolazione.
- Il test assegna un certo valore di probabilità all'ipotesi che viene formulata.

Significatività statistica e casualità

La significatività statistica viene determinata attraverso test (es. χ^2, t di Student ecc.) che sono INDISPENSABILI per interpretare i risultati di un confronto

Supponiamo che, in una piccola serie di animali, in media la comparsa di malattia sia meno frequente dopo vaccinazione con vaccino A che con vaccino B.

DOMANDA: le differenze osservate sono estrapolabili all'intera popolazione di pazienti di quel tipo?

Ecco 3 IPOTESI per giustificare la diversità osservata:

1. **il confronto è viziato perché qualche fattore (es. immunità passiva ecc.) non considerato è responsabile della differenza**
2. **la differenza è dovuta al caso (variabilità biologica)**
3. **A è effettivamente più attivo di B**

l'ipotesi 1 può essere esclusa disegnando bene l'esperimento

l'ipotesi 2 può essere esclusa applicando un test di SIGNIFICATIVITA' STATISTICA

soltanto dopo aver escluso le ipotesi 1 e 2 possiamo concludere che A è migliore di B

Prove di significatività: l'ipotesi nulla

Tutti i test statistici di significatività assumono inizialmente la cosiddetta «ipotesi zero» (o «ipotesi nulla»).

Quando si effettua il confronto fra due o più gruppi di dati, l'ipotesi zero prevede sempre che non esista alcuna differenza tra i gruppi riguardo al parametro considerato. In altre parole, secondo l'ipotesi zero i gruppi sono fra loro uguali e le eventuali differenze osservate vanno attribuite al solo caso.

Ovviamente l'ipotesi zero può essere accettata o respinta, ma in che modo?

Si procede applicando un test statistico di significatività, il cui risultato in genere va confrontato con un valore critico tabulato in apposite tabelle. Se il risultato del test di significatività supera il valore critico, allora la differenza fra i gruppi viene dichiarata statisticamente significativa e, quindi, l'ipotesi zero viene respinta. In caso contrario l'ipotesi zero viene accettata.

Prove di significatività: l'ipotesi nulla (segue)

Come sempre avviene, i risultati di un test statistico non hanno un valore di assoluta e matematica certezza, ma soltanto di **probabilità**. Pertanto, una decisione di respingere l'ipotesi zero (presa sulla base del test statistico) è **probabilmente** giusta, ma potrebbe essere errata. La misura di questo rischio di cadere in errore si chiama «**livello di significatività**» del test.

Il livello di significatività di una prova può essere scelto a piacere dallo sperimentatore. Tuttavia, di solito si sceglie un livello di probabilità di 0.05 (5%) o di 0.01 (1%). Questa probabilità (detta valore p) rappresenta una stima quantitativa della probabilità che le differenze osservate siano dovute al caso.

C'è sempre il rischio che la DECISIONE di RIFIUTARE l'ipotesi zero SIA ERRATA

la misura di questo rischio si chiama...

LIVELLO DI SIGNIFICATIVITA' (LS) del test

- LS può essere scelto arbitrariamente
- in genere si sceglie LS=0.05 oppure LS=0.01
- il valore LS più basso al quale l'ipotesi zero può essere respinta è il "VALORE P" o più semplicemente "P"

Prove di significatività: l'ipotesi nulla (segue)

ESEMPIO: Abbiamo effettuato una sperimentazione su due gruppi di soggetti affetti da una determinata malattia. Uno dei due gruppi è stato trattato con il farmaco A e l'altro con il farmaco B; i soggetti con A sembrano guarire con maggiore frequenza di quelli trattati con B.

Calcolando il valore *p* otteniamo una stima quantitativa della probabilità che le differenze osservate siano dovute al caso. In altre parole, *p* è la risposta alla seguente domanda: «se non ci fossero differenze fra A e B, e se la sperimentazione fossa eseguita molte volte, quale proporzione di sperimentazioni condurrebbe alla conclusione che A è migliore di B?»

Il livello di significatività 5% viene adottato molto frequentemente in quanto si ritiene che il rapporto 1/20 (cioè 0.05) sia sufficientemente piccolo da poter concludere che sia «piuttosto improbabile» che la differenza osservata sia dovuta al semplice caso. In effetti, la differenza *potrebbe* essere dovuta al caso, e lo sarà 1 volta su 20. Tuttavia, questo evento è «improbabile». Ovviamente, se si vuole escludere con maggiore probabilità l'effetto del caso, si adotterà un livello di significatività inferiore (es. 1%).

Prove di significatività: l'ipotesi nulla (segue)

Esempio: Immaginiamo che il tasso medio di acido urico rilevato in 52 uomini adulti sia risultato pari a $\overline{x} = 7,3$ mg/100 ml. Sapendo che il livello medio di acido urico in una popolazione maschile adulta e sana è pari a $\mu = 5,7$ mg / 100 ml con una varianza $\sigma^2 = 1$ mg / 100 ml ci si domanda:

se il valore riscontrato nei 52 soggetti campionati (diverso da quello dell'intero collettivo) sia da attribuire a semplici errori campionari oppure ad una qualche patologia che ne altera con sistematicità la concentrazione.

Supponiamo che la concentrazione media di uremia rilevata nei 52 soggetti appartenenti al campione (pari a 7,3 mg / 100ml) si differenzia dal tasso generico che è $\mu = 5,7$ mg / 100 ml a causa della non completezza delle osservazioni sulle quali abbiamo stimato il parametro d'interesse.

Quanto appena asserito rappresenta l'ipotesi nulla $H_{(0)}$ viceversa potremmo pensare che i soggetti costituenti il campione

Prove di significatività: l'ipotesi nulla (segue)

Non siano stati selezionati dalla popolazione di individui sani, bensì da una popolazione di oggetti affetti da qualche malattia che altera sistematicamene il livello di acido urico.

Formuliamo pertanto l'ipotesi alternativa $H_{(1)}$, secondo cui $\bar{x}$ è diversa da μ perché esiste un fattore che altera l'uremia nei 52 soggetti selezionati.

Stimata μ e definite le due ipotesi, dobbiamo identificare la **statistica test**, o **funzione test**, definita dal rapporto avente come numeratore la differenza tra il valore assunto dalla stima del parametro e il parametro stesso, nel nostro caso $(x - \mu)$, e come denominatore $(\sigma / \sqrt{n})$, denominato **errore standard**.

Nel caso da noi preso in esame, quindi, il valore assunto dalla statistica test è pari a

$$z = \frac{x - \mu}{\sigma / \sqrt{n}} = \frac{(7{,}5 - 5{,}7)}{1 / \sqrt{52}} = \frac{1{,}8}{0{,}14} = 12{,}9$$

che rappresenta uno degli infiniti valori che la distribuzione di probabilità della variabile normale standardizzata può assumere.

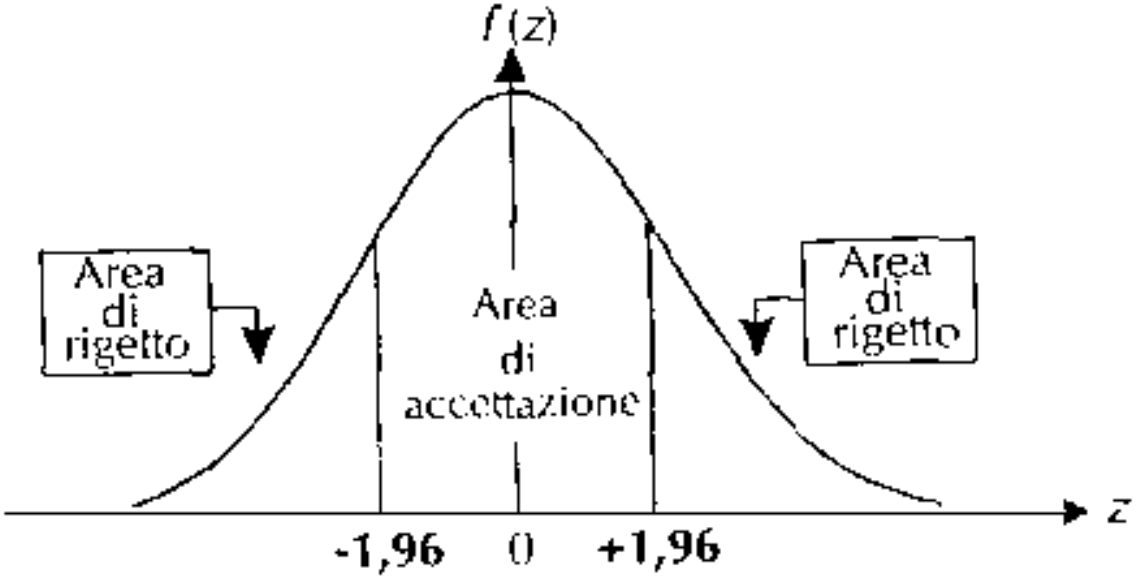
Prove di significatività: l'ipotesi nulla (segue)

Stabilita la distribuzione campionaria della statistica test, dobbiamo definire il **livello di significatività** α (solitamente posto uguale a 0,05, 0,01, 0,001), definito come quella probabilità che consente di dividere la s uddetta distribuzione in due aree: l'**area di accettazione dell'ipotesi nulla**, e l'**area di rigetto dell'ipotesi nulla**.

Fissato α ad un livello dello 0,05, dalle tavole della curva normale standardizzata desumiamo $z_1 = -1,96$ e $z_2 = +1,96$.

Tutta l'area a Sx e a Dx di tali quantità costituisce l'area di rigetto di $H_{(0)}$, mentre l'area interna all'intervallo [- 1,96; 1,96] rappresenta l'area di accettazione della medesima.

Fasi dell'indagine statistica: esercitazione



Poiché z = 12,9 "cade" nella regine di rifiuto dell'ipotesi nulla, possiamo concludere che $H_{(0)}$ è falsa, i 52 individui sorteggiati che presentano un livello medio di acido urico pari a 7,5 mg / 100 ml non appartengono alla popolazione di individui sani bensì ad una popolazioe di soggetti affetti da qualche patologia che ne altera la concentrazione.

Area compresa tra 0 e + ∞ nella curva normale standartizzata

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	**0,1587**	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	**0,1251**	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,8691	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0570	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	**0,0281**	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0238	0,0233
2,0	0,0227	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183

Prove di significatività: l'ipotesi nulla (segue)

"Significativo" non è sinonimo di "importante"

"Significativo" = difficilmente dovuto al puro caso

Numerosi test statistici vengono usati per determinare con un certo grado di probabilità l'esistenza (o l'assenza) di differenze significative nei dati in esame o meglio, più in generale, di accettare o rigettare l'ipotesi zero.

Test statistici di comune impiego in medicina

SIGNIFICATIVITA' STATISTICA DI UNA DIFFERENZA	
Test	*Confronto fra:*
Test del chi quadrato	due o più proporzioni o percentuali (con molte osservazioni)
Test esatto di Fisher	due o più proporzioni o percentuali (con poche osservazioni)
Test U di Mann-Whitney	due mediane
Test *t* di Student	due medie
Test *t* di Bonferroni	più medie
Test *f*	2 o più medie

DESCRIZIONE DEL GRADO DI UNA ASSOCIAZIONE	
Test	*Confronto fra:*
Coefficiente di regressione	una variabile indipendente ed una variabile dipendente
Test *r* di Pearson	due variabili

Il metodo del chi-quadrato nel confronto tra 2 percentuali

Esempio di applicazione del test più comune, il «chi-quadrato» per confrontare due gruppi o popolazioni riguardo ad un parametro.

L'esempio prende in considerazione due percentuali; lo scopo è quello di verificare se la loro differenza è dovuta al caso oppure no, cioè sia «significativa».
Il metodo del chi-quadrato è utilizzabile quando il numero complessivo di osservazioni è >40; altrimenti, occorre usare altri test.

Supponiamo di voler raffrontare l'efficacia nella terapia di una malattia infettiva di un nuovo antibiotico (che chiameremo con un nome di fantasia: xmicina) con un antibiotico ben noto (streptomicina).

Dati ottenuti

	guariti	non guariti	*totali*
xmicina	a = 60	b = 2	62
streptomicina	c = 37	d = 11	48
totali	97	13	110

I dati grezzi sembrano indicare che la xmicina è più efficace della streptomicina. Tuttavia, prima di giungere ad una conclusione affrettata, occorre rispondere alla seguente domanda:

Ammesso che, in realtà, non esistano differenze nell'efficacia dei due trattamenti, che probabilità c'è di osservare - in uno studio di dimensioni simili a questo - differenze uguali o superiori a quelle che abbiamo osservato?

Il metodo del chi-quadrato nel confronto tra 2 percentuali

La risposta a questa domanda dipende da *quanto i dati ottenuti si discostano dai dati che «sarebbe lecito attendersi se i trattamenti avessero la stessa efficacia e se i dati fossero influenzati soltanto dalla variazione casuale»*.

I nostri dati dimostrano che complessivamente (cioè indipendentemente dal tipo di antibiotico) il trattamento è risultato efficace nell'88% dei casi (97 guariti su 110 trattati). Allora, applichiamo questa percentuale di successo a ciascuno dei due gruppi di soggetti in esame e ricaviamo la corrispondente tabella:

Dati attesi (valori approssimati)

	guariti	non guariti	*totali*
xmicina	a = 55	b = 7	62
streptomicina	c = 42	d = 6	48
totali	97	13	110

Nella tabella soprastante, il valore a=55 è stato ottenuto assumendo una percentuale di guarigione dell'88% nei 62 casi trattati con xmicina: 62*88/100=54.56, cioè, approssimando all'unità, a=55. Lo stesso calcolo è stato eseguito per ottenere il valore c=42 (48*88/100). I valori delle celle b e d possono essere facilmente ottenuti per differenza.

Il test chi-quadrato, che quantifica la differenza fra i numero osservati e quelli attesi, è la somma delle quattro celle a, b, c e d, per ciascuna delle quali si calcola il valore della frazione:

$$\frac{(\text{numero osservato} - \text{numero atteso})^2}{\text{numero atteso}}$$

Il metodo del chi-quadrato nel confronto tra 2 percentuali

La magnitudine del chi-quadrato è determinata dalla differenza fra i numeri osservati ed i numeri attesi nel caso in cui i due trattamenti avessero avuto lo stesso effetto. La differenza al numeratore della frazione viene elevata al quadrato; ciò elimina i numeri negativi che possono comparire quando il numero osservato è minore di quello atteso. Quindi, il quadrato della differenza viene diviso per il numero atteso; in questo modo la differenza per ogni cella viene aggiustata in rapporto al numero di individui della stessa cella.

Pertanto, il chi-quadrato viene calcolato come segue:

$$\chi^2 = \frac{(60-55)^2}{55} + \frac{(2-7)^2}{7} + \frac{(37-42)^2}{42} + \frac{(11-6)^2}{6} = 8.79$$

É evidente che il chi-quadrato aumenta con l'aumentare della differenza dei dati posti a raffronto. Se esso supera certi valori (tabella «valori di chi-quadrato»), la differenza viene ritenuta significativa; in caso contrario, non si può affermare l'esistenza di una significativa differenza tra i due fenomeni considerati.

Non ci resta quindi che confrontare il valore ottenuto con una Tabella dei valori di chi-quadrato Nel nostro caso, il valore ottenuto è un chi-quadrato con «1 grado di libertà» (il grado di libertà è pari al numero di osservazioni-1; esso, confronto tra due percentuali - come nel nostro caso - si pone sempre =1).

Il metodo del chi-quadrato nel confronto tra 2 percentuali

Valori di χ^2

Gradi di libertà	Probabilità	
	5%	1%
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.475
...	...	...

Il nostro valore è >6.635, e quindi la differenza fra i due gruppi è da ritenere significativa al livello di probabilità 1%. nel nostro caso, la differenza tra soggetti trattati con xmicina e quelli trattati con streptomicina è statisticamente significativa al livello di probabilità 1%.

In altre parole: ammettendo che i due antibiotici abbiano pari efficacia e ripetendo l'esperimento infinite volte, potremo osservare molto raramente (ossia meno di 1 volta su 100!) dati simili a quelli ottenuti oppure ancor più favorevoli al farmaco A.

In sostanza: in base ai risultati del test del chi-quadrato, l'affermazione «xmicina è più efficace di streptomicina» ha il 99% di probabilità di essere vera (e quindi ha il 1% di probabilità di essere falsa).

Il fatto che la differenza fra i due gruppi in studio sia risultata statisticamente significativa non implica necessariamente che, nella pratica clinica, la xmicina avrebbe sostituito la streptomicina nella terapia della infezione. Ad esempio, la xmicina potrebbe essere molto più tossica, oppure dotata di gravi effetti collaterali, oppure molto più costosa ecc.

L'indice chi-quadrato si può estendere al confronto di più gruppi, ma in tal caso il calcolo è diverso da quello dell'esempio.

Tavola dei valori critici del chi-quadrato

α g.l.	0,05	0,025	0,01	0,005	0,001	0,0005
1	3,84	5,02	6,63	7,88	10,80	12,10
2	5,99	7,38	9,21	10,60	13,80	15,20
3	7,81	9,35	11,30	12,80	16,30	17,70
4	9,49	11,10	13,30	14,90	18,50	20,00
5	11,10	12,80	15,10	16,70	20,50	22,10
6	12,60	14,40	16,80	18,50	22,50	24,10
7	14,10	16,00	18,50	20,30	24,30	26,00
8	15,50	17,50	20,10	22,00	26,10	27,90
9	16,90	19,00	21,70	23,60	27,90	29,70
10	18,30	20,50	23,20	25,20	29,60	31,40
11	19,70	21,90	24,70	26,80	31,30	33,10
12	21,00	23,30	26,20	28,30	32,90	34,80
13	22,40	24,70	27,70	29,80	34,50	36,50
14	23,70	26,10	29,10	31,30	36,10	38,10
15	25,00	27,50	30,60	32,80	37,70	39,70
20	31,40	34,20	37,60	40,00	45,30	47,50
25	37,70	40,60	44,30	46,90	52,60	54,90
30	43,80	47,00	50,90	53,70	59,70	62,20
40	55,80	59,30	63,70	66,80	73,40	76,10

Parametro t di Student (confronto fra medie)

Esempio di applicazione del test t di student nel caso di un campione composto da poche unità e di cui non è nota la varianza della popolazione.

Supponiamo che il tasso di colesterolo (espresso in mg/100 ml) rilevato su 12 individui apparentemente sani sia il seguente: 195, 253, 206, 221, 222, 189, 218, 228, 234, 284, 259, 268 con media $\bar{x}$ pari a 231,42 mg 100 ml.

Sapendo che il livello medio di colesterolo in soggetti sani è di 210 mg/100 ml, il maggiore valore osservato negli individui in esame è attribuibile all'aver selezionato gli individui da una popolazione con ipercolesterolemia?

Per rispondere a questa domanda procediamo attraverso la teoria dei test d'ipotesi, definendo anzitutto l'ipotesi nulla $H_{(0)}$ (secondo cui la media campionaria si discosta da quella della popolazione per puro effetto del caso) e l'ipotesi alternativa $H_{(1)}$ (per

Parametro t di Student (confronto fra medie)

la quale $H_{(1)}$ $\bar{x}$ è diversa da μ perché i soggetti selezionati sono malati).

Non essendo nota la varianza della popolazione, il calcolo del test statistico definito come rapporto tra la differenza di $(\bar{x} - \mu)$ e il quoziente $(\sigma^2/\sqrt{n})$, richiede la preliminare stima di σ^2:

$$S^2 = \frac{(195 - 231{,}42)^2 + (253 - 231{,}42)^2 + \ldots (268 - 231{,}42)^2}{(12 - 1)} = 836{,}36$$

A questo punto, sottraendo μ da $\bar{x}$ e dividendo il tutto per l'errore standard, definito ora dal rapporto $(S/\sqrt{n - 1})$, determiniamo la funzione test:

$$t = \frac{(231{,}42 - 210)}{(29{,}38) / (\sqrt{12 - 1})} = 2{,}42$$

che rappresenta uno dei possibili valori che la variabile *t di student* può assumere.

Parametro t di Student (confronto fra medie)

Quest'ultima è una variabile continua, può assumere i valori compresi tra - ∞ e + ∞ ed è simmetrica rispetto alla media. Graficamente è molto simile alla curva di Gauss, rispetto alla quale è tuttavia più larga e più piatta, a causa della maggiore imprecisione delle stime $\bar{x}$ e s^2, determinate su un campione con ampiezza minore delle 30 unità.

Decidiamo di accettare o rifiutare $H_{(0)}$ confrontando il valore t da noi determinato con il valore teorico t desumibile dalla **tavola t di Student,** per il cui utilizzo è necessario determinare preliminarmente un fattore addizionale, i *gradi di libertà (g.l.)*, pari al denominatore della stima della varianza, nel nostro caso (n - 1) = (12 - 1) = 11.

Fissato $\alpha = 0{,}05$ e calcolato g.l. = 11, dobbiamo ricavare dalle tavole della distribuzione della t di Student il valore t.

Poiché tale distribuzione è simmetrica, per individuare t dobbiamo dimezzare il livello di significatività 0,05 e cercare sulla tavola

Parametro t di Student (confronto fra medie)

In corrispondenza di $\alpha/2 = 0{,}025$ e g.l. = 11.

Una volta ricavato t = 2,20, e quindi $t_1 = -2{,}20$ e $t_2 = +2{,}20$, dobbiamo valutare se t = 2,42 appartiene all'area di accettazione o di rigetto della ipotesi nulla.

Essendo 2,42 esterno all'intervallo [-2,20; + 2,20], rifiutiamo $H_{(0)}$, a conferma di una anomalia dei dati attribuibile, appunto, ad una ipercolesterolemia.

Tavola dei valori critici della t di Student

α/2 *g.l.*	0,05	0,025	0,01	0,005	0,001	0,0005
1	6,314	12,710	31,820	63,660	318,300	636,600
2	2,920	4,303	6,965	9,925	22,330	31,600
3	2,353	3,182	4,541	5,841	10,220	12,940
4	2,132	2,776	3,747	4,604	7,173	8,610
5	2,015	2,571	3,365	4,032	5,893	6,859
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,405
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	1,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140
15	1,753	2,131	2,602	2,947	3,733	4,073
20	1,725	2,086	2,528	2,845	3,552	3,850
30	1,697	2,042	2,457	2,750	3,385	3,646
40	1,684	2,021	2,423	2,704	3,307	3,551
50	1,676	2,009	2,403	2,678	3,262	3,495

I Test Statisici

Permettono di verificare su campioni ipotesi riferite a popolazioni statistiche

parametrici

Tests

non parametrici

Parametrici: (si applicano a seriazioni)

- t di Student
- ANOVA

I principi fondamentali della verifica delle ipotesi

Per poter utilizzare correttamente un test parametrico è necessario che siano soddisfatte le seguenti condizioni:

1- indipendenza dei dati

2- normalità delle distribuzioni campionarie

3- omogeneità delle varianze campionarie

I principi fondamentali della verifica delle ipotesi

Tests non parametrici:

- prescindono dalle assunzioni sulle distribuzioni campionarie;
- sono di semplice attuazione e possono essere applicati anche a serie (distribuzioni con caratteri qualitativi)

- chi quadrato (χ_2) di Pearson
- test di Wilcoxon
- test di Kolmogorov-Smirnov
- test U di Mann-Whitney

I principi fondamentali della verifica delle ipotesi

I principali vantaggi offerti dai metodi non parametrici dipendono essenzialmente dalla loro indipendenza dalla distribuzione dei dati; essi sono:

1. Il livello di probabilità individuato è esatto qualunque sia il tipo di distribuzione da cui sono stati estratti i campioni
2. L'elaborazione dei dati è molto semplice, basandosi essenzialmente sul calcolo combinatorio e i risultati sono di facile interpretazione
3. In generale i metodi non parametrici consentono un'ampia libertà nella formulazione delle ipotesi
4. Per quanto riguarda la potenza dei test, questi metodi sono tanto più efficaci quanto maggiormente la distribuzione dei dati si discosta da quella normale

D'altro canto i metodi non parametrici presentano indubbi svantaggi per quanto riguarda il calcolo delle stime e l'utilizzazione completa dell'informazione

I principi fondamentali della verifica delle ipotesi

I tests parametrici si applicano a seriazioni, cioè a distribuzione di dati con modalità quantitative, utilizzando indici statistici quali:

- la media
- la devianza
- la varianza

Perché possano essere utilizzati è necessario che siano soddisfatte alcune condizioni:

1. Indipendenza dei dati
2. Normalità delle distribuzioni campionarie
3. Omogeneità delle varianze campionarie

I principi fondamentali della verifica delle ipotesi

1- L'indipendenza dei dati è verificata quando le modalità assunte dai diversi elementi che costituiscono il campione non dipendono in alcun modo da quelle degli elementi di un altro campione

" prima e dopo il trattamento "

chiaramente gli elementi che costituiscono i due gruppi sono gli stessi, e la modalità assunta dopo il trattamento non è determinata solo dal trattamento, ma anche dalla modalità rilevata prima

Per dati dipendenti, nel caso di due gruppi è possibile applicare il " t paired test " o t per dati appaiati

I principi fondamentali della verifica delle ipotesi

2- Le distribuzioni campionarie non devono differire in maniera significativa dalla distribuzione normale

La normalità di una distribuzione può essere verificata standardizzando le singole modalità, ricercando nella tabella l'integrale di probabilità relativo ad ogni singola modalità, moltiplicando ciascun integrale per il numero totale delle frequenze e confrontando con il $\chi 2$ queste frequenze teoriche con quelle osservate

I principi fondamentali della verifica delle ipotesi

I gradi di libertà sono uguali al numero delle modalità considerate meno 3, essendo tre (x, s, n) i legami introdotti con il calcolo

Se il valore ottenuto è più basso del valore tabulare del $\chi 2$, **per i gradi di libertà calcolati ed al livello di probabilità scelto**,

la distribuzione campionaria può essere considerata simile alla normale in quanto non differisce in maniera significativa da quest'ultima

I principi fondamentali della verifica delle ipotesi

3- La variabilità all'interno dei singoli campioni deve essere omogenea, nel senso che le varianze campionarie non possono differire tra loro in maniera significativa

Nel caso di 2 campioni l'omogeneità delle varianze si verifica calcolando il rapporto tra s^2 maggiore e s^2 minore; il valore ottenuto va confrontato, **al livello di probabilità scelto e per i gradi di libertà delle due s^2, nella tabella dell' F (rapporto tra varianze)**

Se il valore ottenuto è minore del valore tabulare, le varianze possono essere considerate omogenee ed il test parametrico può essere usato

Per varianze non omogenee si può applicare

l' approssimazione di Cochran

Nel confronto tra più di due gruppi, l'omogeneità delle varianze deve essere valutata con il

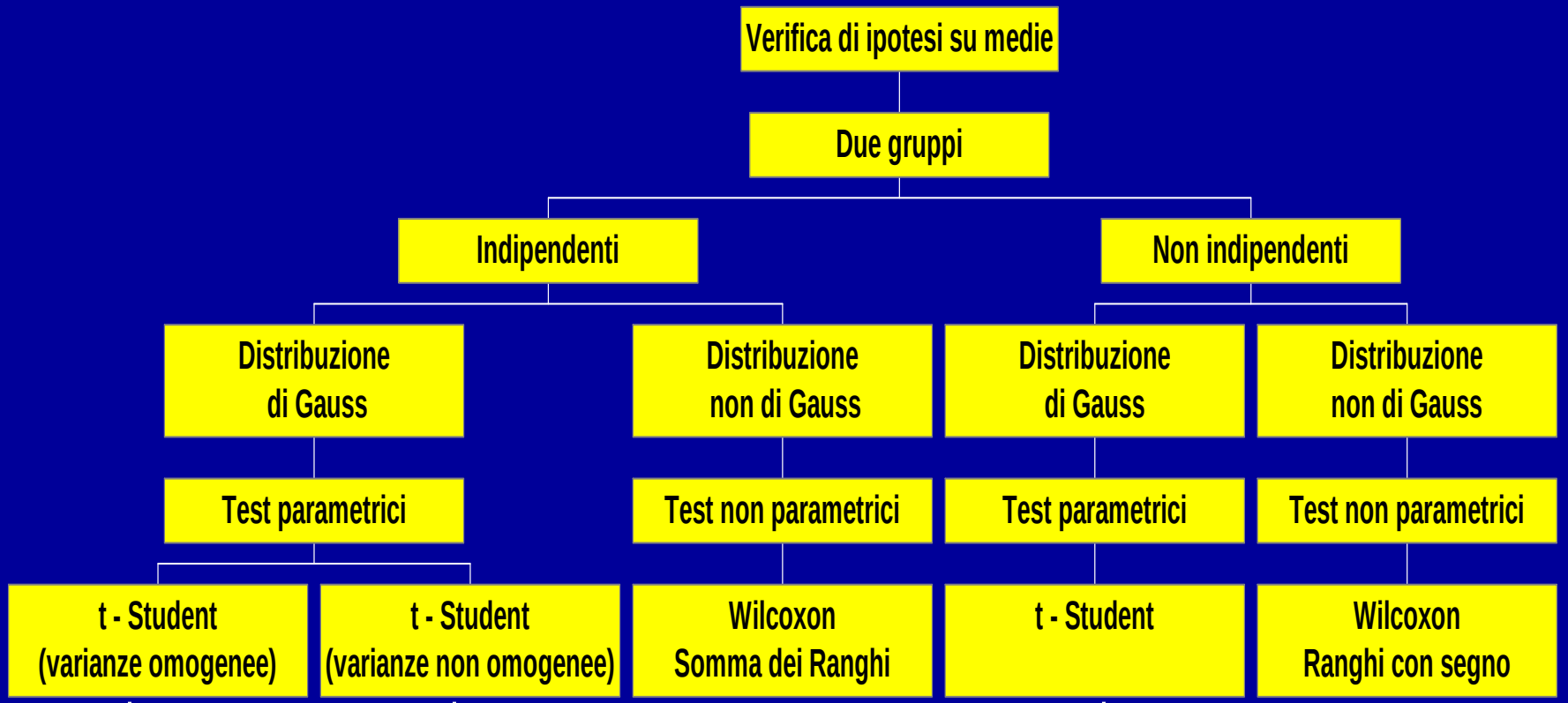
test di Bartlett

I principi fondamentali della verifica delle ipotesi

VERIFICA DELLE IPOTESI

METODI PARAMETRICI Esistenza di distribuzioni		METODI NON PARAMETRICI Non esistenza di distribuzioni
TESTS t-student Analisi della varianza Confronto proporzioni	CONFRONTO TRA GRUPPI	TESTS sui RANGHI (Wilcoxon) Analisi della varianza non parametrica
Correlazione Coefficiente di Pearson Regressione	LEGAMI TRA VARIABILI	Correlazione Coefficiente di Spearman Tabelle di contingenza

I principi fondamentali della verifica delle ipotesi

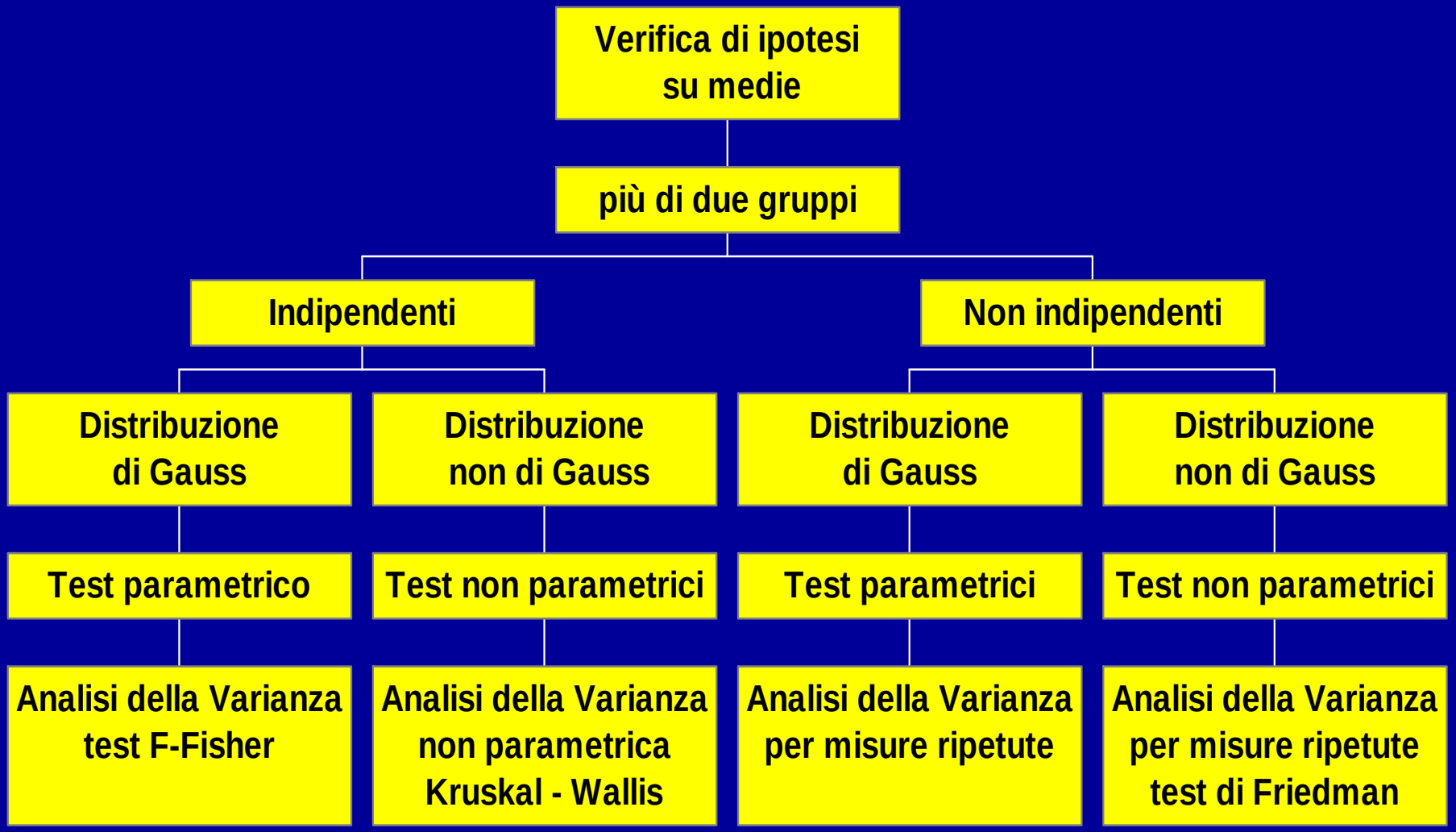


$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

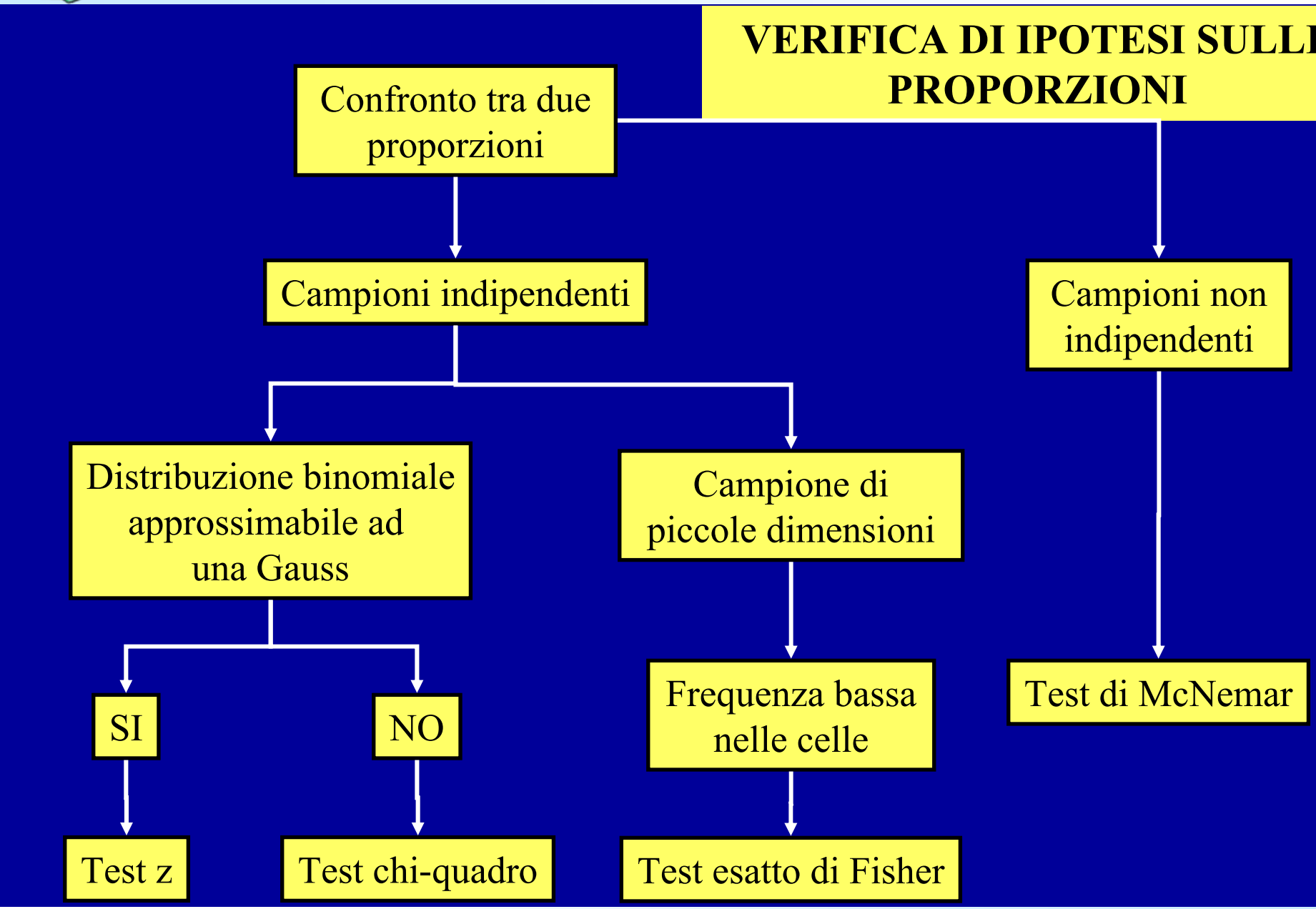
$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}}$$

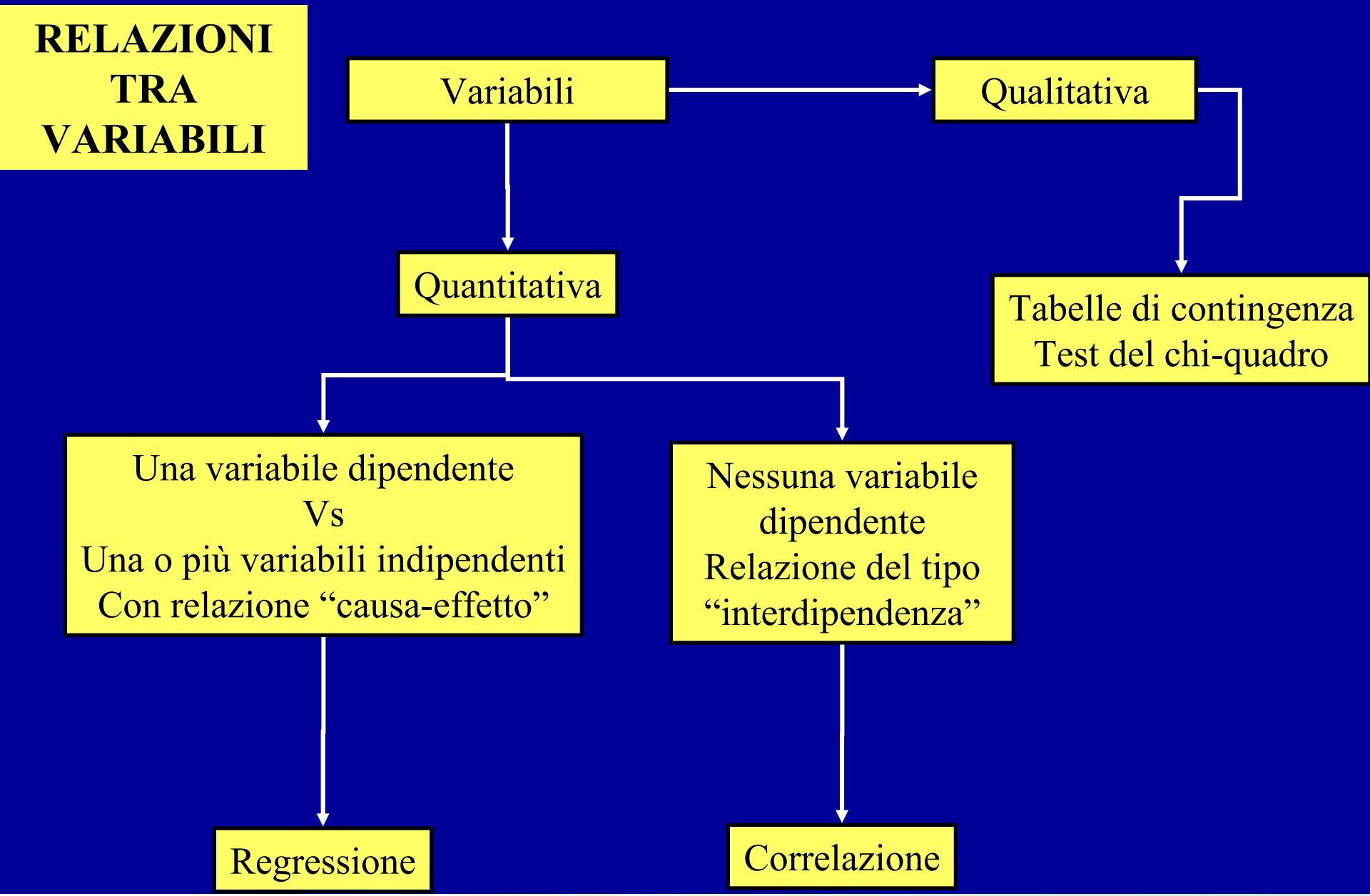
I principi fondamentali della verifica delle ipotesi



I principi fondamentali della verifica delle ipotesi

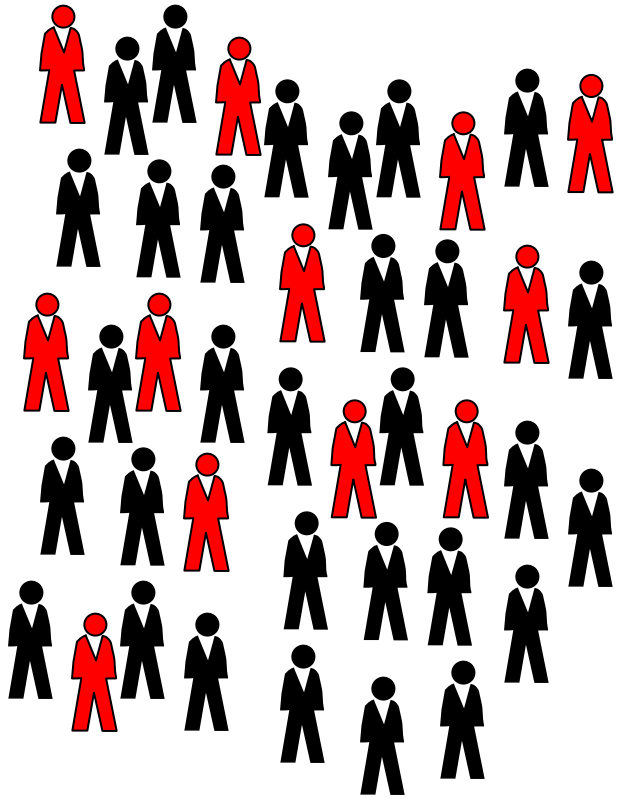


I principi fondamentali della verifica delle ipotesi



MISURE DI RISCHIO E ASSOCIAZIONE

Misurare l'incidenza di malattia in un gruppo di popolazione...



12/42 * 100 => **28,6%**

Misurare l'incidenza di malattia equivale a valutare la probabilità (ossia il rischio) di avere nuovi casi di malattia nella popolazione nel periodo prescelto

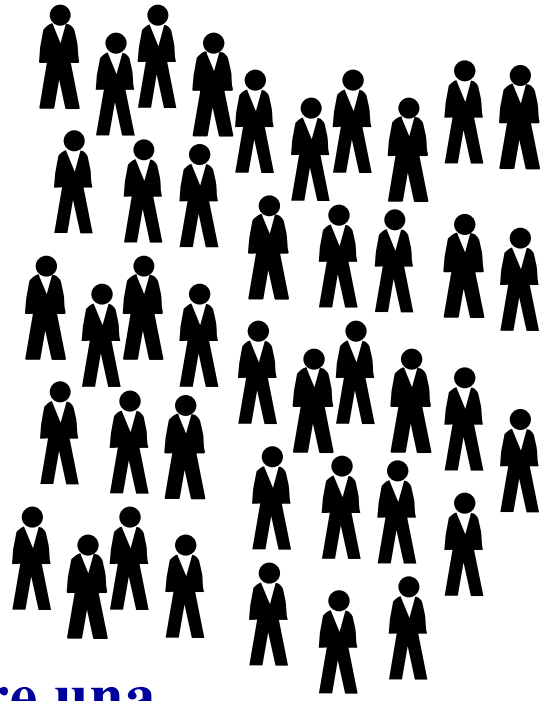
Misurare l'incidenza di malattia in un gruppo di popolazione...

Se l'incidenza cumulativa risulta pari a 28,6%, questo indica che la **probabilità**, per ciascun individuo appartenente alla popolazione, di sviluppare la malattia nel periodo considerato é pari al 28,6%

Valutare il rischio individuale...

Ma gli individui che compongono questa popolazione sono tutti uguali?

La probabilità che ha ciascun individuo é effettivamente pari all'incidenza cumulativa calcolata su tutta la popolazione?



La probabilità di contrarre una malattia per un individuo può aumentare o diminuire in base alla presenza/assenza di particolari condizioni

Cosa é un fattore di rischio?

E’ un fattore esponendosi al quale aumenta la probabilità di contrarre una determinata malattia.

Può essere rappresentato da una condizione geneticamente determinata, da una abitudine personale, da un particolare stile di vita, da un evento accidentale, ecc.

L'epidemiologia analitica...

- Indaga le cause o i fattori che determinano l'insorgere di una malattia o ne influenzano la diffusione
- ...ovvero indaga la **relazione causa-effetto** fra fattori di rischio e patologie

Indagare una relazione causa-effetto

Fra un evento variabile (*causa o fattore di rischio*) ed una malattia (*o una determinata condizione in studio*) può esistere una certa associazione statistica

Per *associazione* intendiamo il grado di dipendenza statistica tra due o più eventi variabili

Associazione fra due eventi

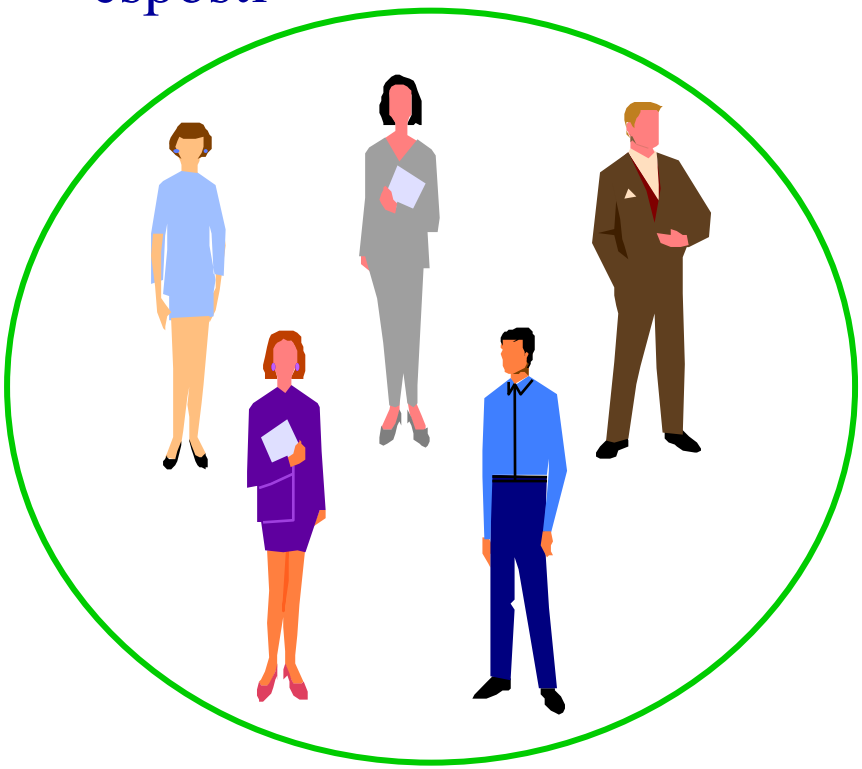
- Causale (o eziologica)
- Indiretta (o secondaria)
- Spuria (o non causale)

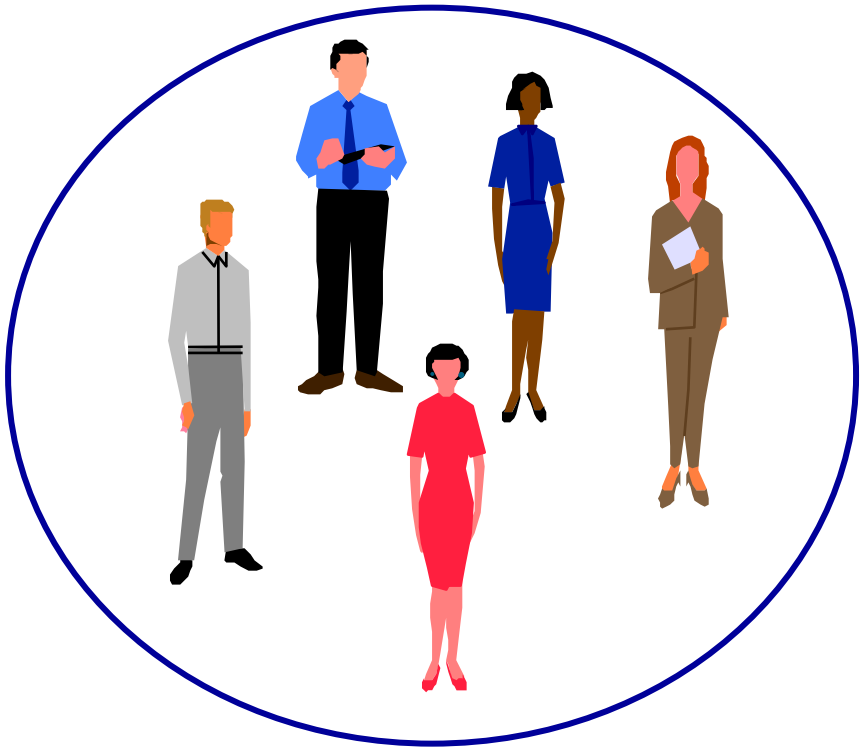
Come valutare un'associazione causale

- Forza dell'associazione
- Consistenza dell'osservazione
- Relazione temporale
- Plausibilità biologica
- Relazione di dipendenza dose-risposta

Schema generale di uno studio analitico

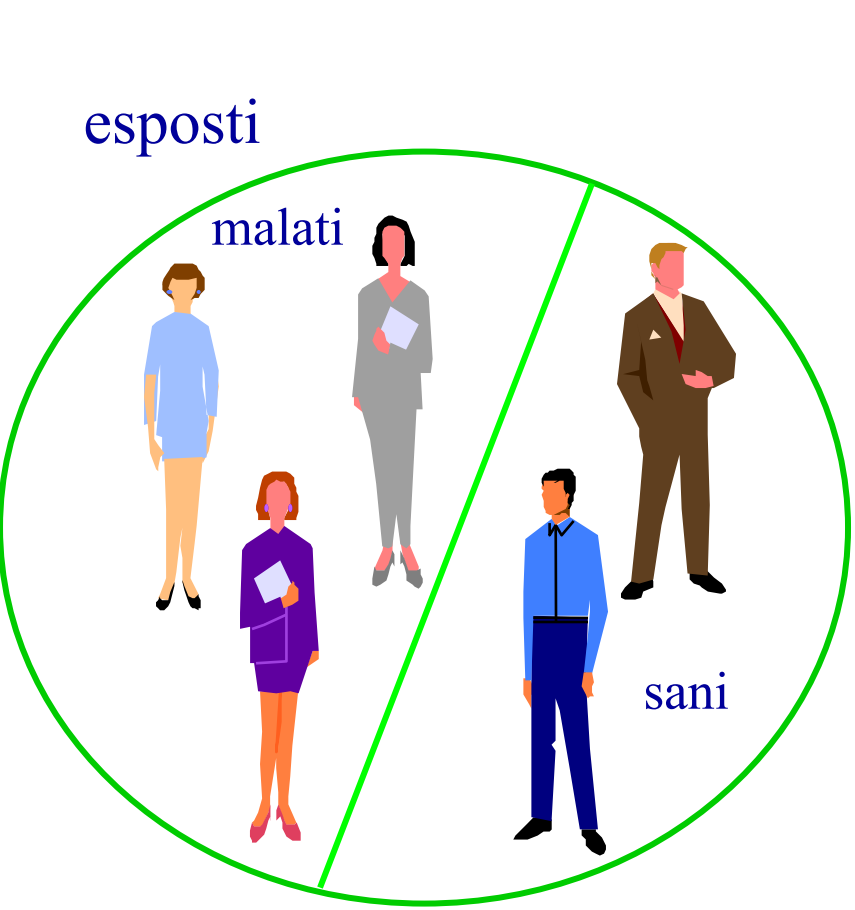
esposti

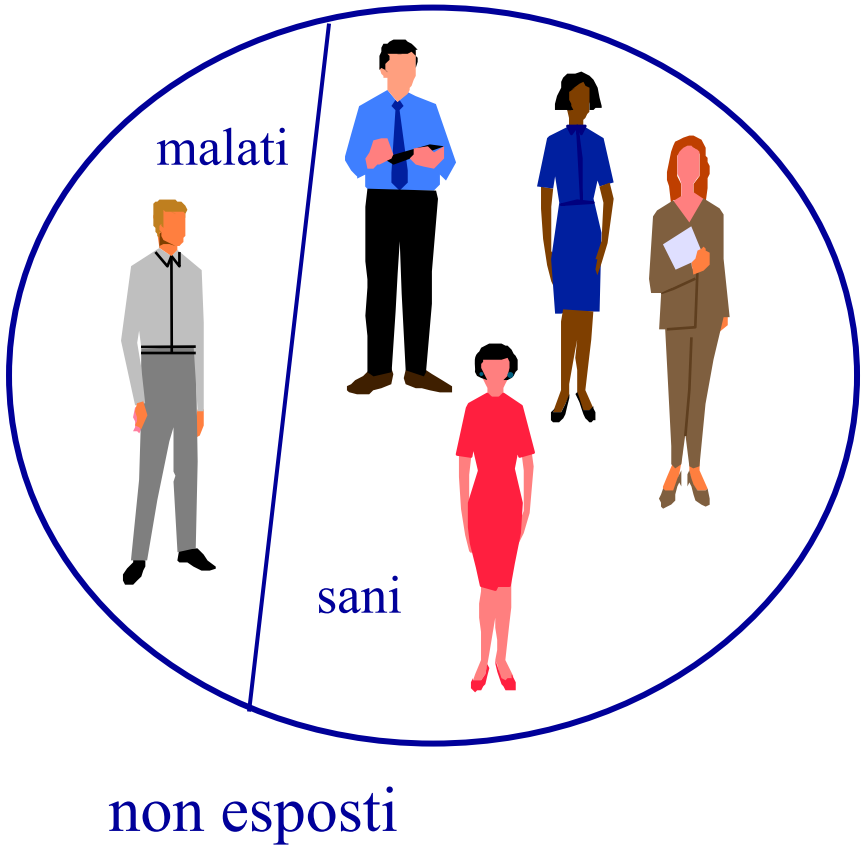




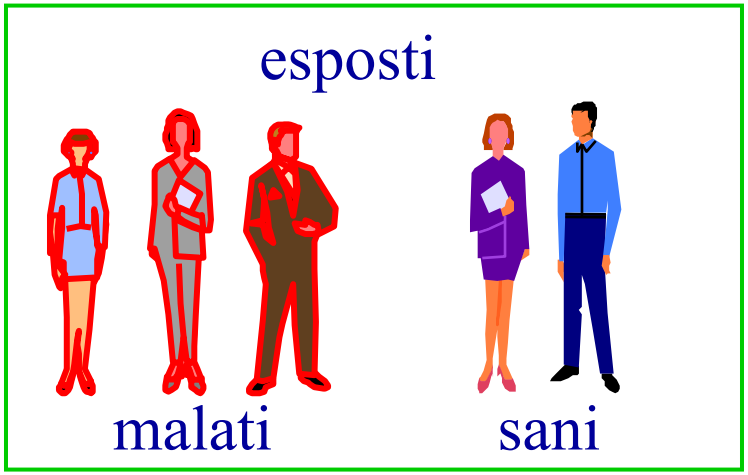
non esposti

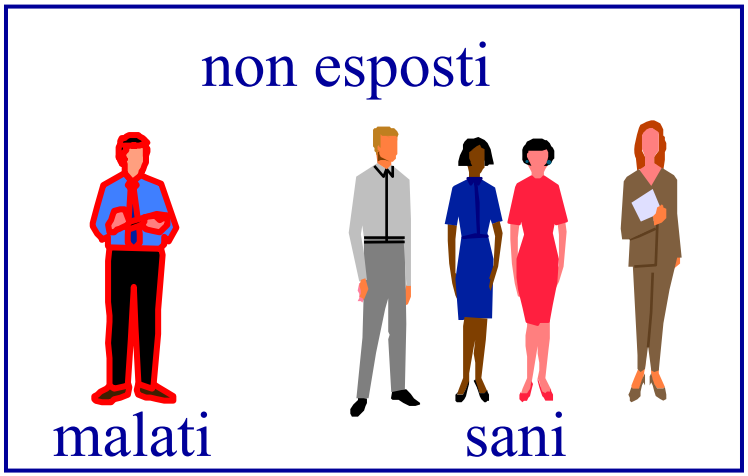
Schema generale di uno studio analitico





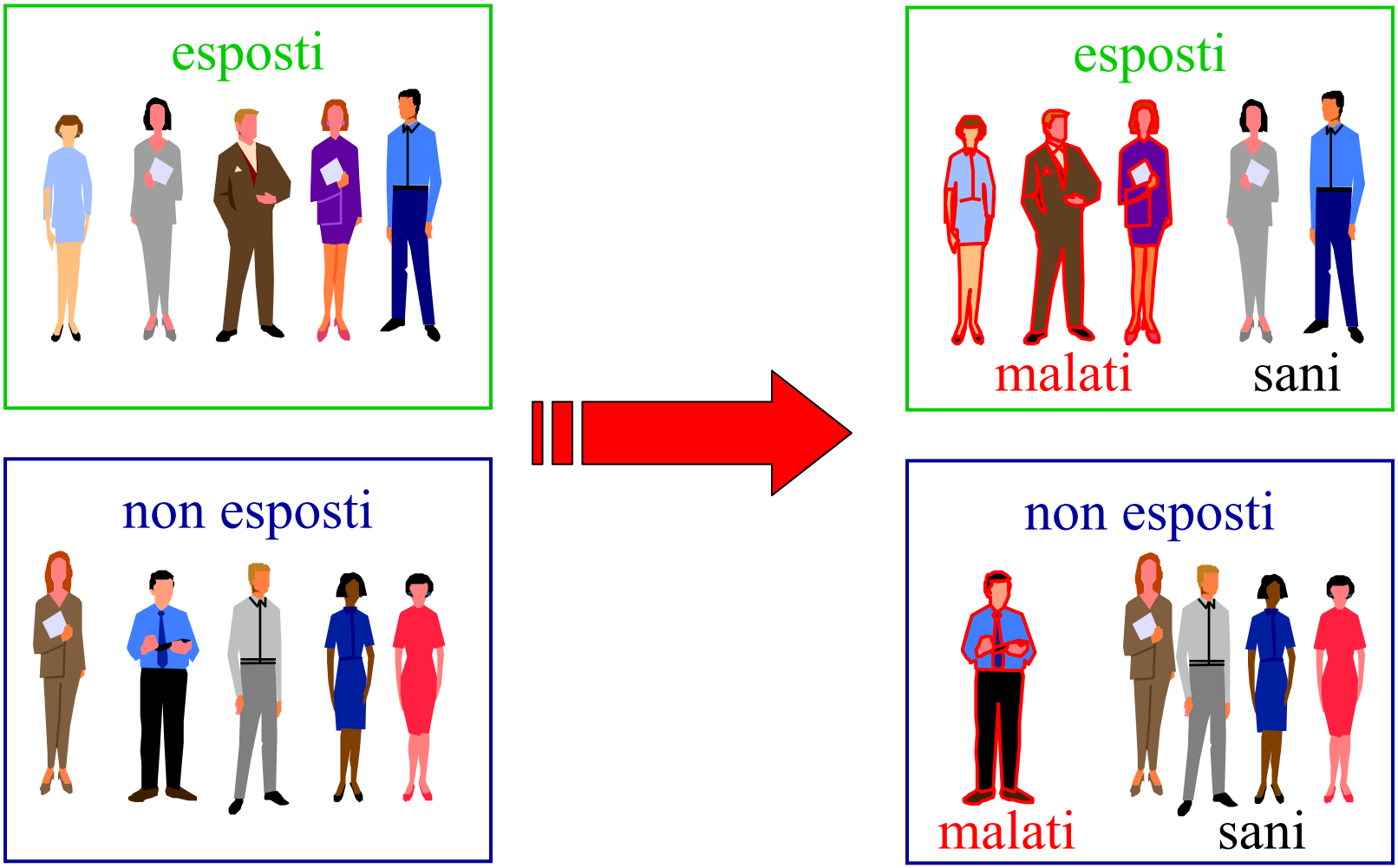
La tabella 2x2



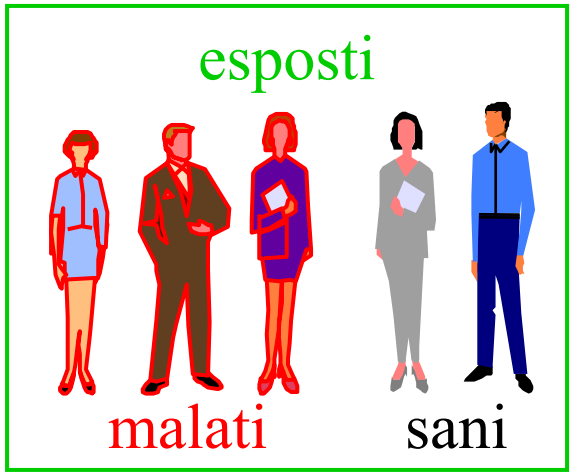


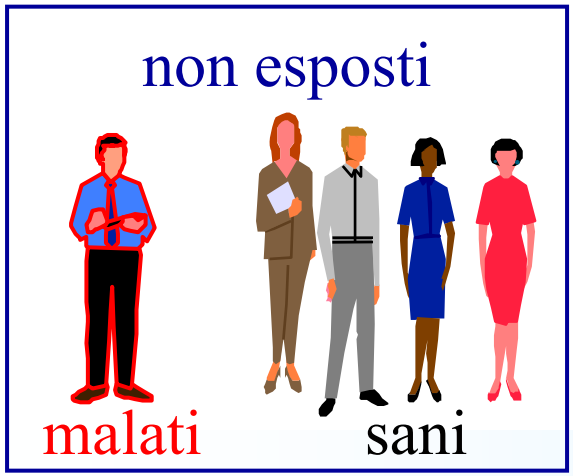
	M+	M-	
E+	3 *a*	2 *b*	*5*
E-	1 *c*	4 *d*	*5*
	4	*6*	***10***

Un modello di studio: lo studio di coorte



Un modello di studio: lo studio di coorte





	M+	M-	
E+	3 *a*	2 *b*	*5*
E-	1 *c*	4 *d*	*5*
	4	*6*	***10***

$$\text{Rischio Relativo} = \frac{I_{E+}}{I_{E-}} = \frac{a/(a+b)}{c/(c+d)} = \frac{3/5}{1/5} = 3$$

Interpretare il RR

- **RR >1**
- **RR =1**
- **RR <1**

Significatività del RR

	M+	M-	
E+	200 *a*	150 *b*	*350*
E-	170 *c*	250 *d*	*420*
	370	*400*	***770***

RR = 1,41

IC 95%= (1,22 - 1,64)

	M+	M-	
E+	20 *a*	15 *b*	*35*
E-	17 *c*	25 *d*	*42*
	37	*40*	*77*

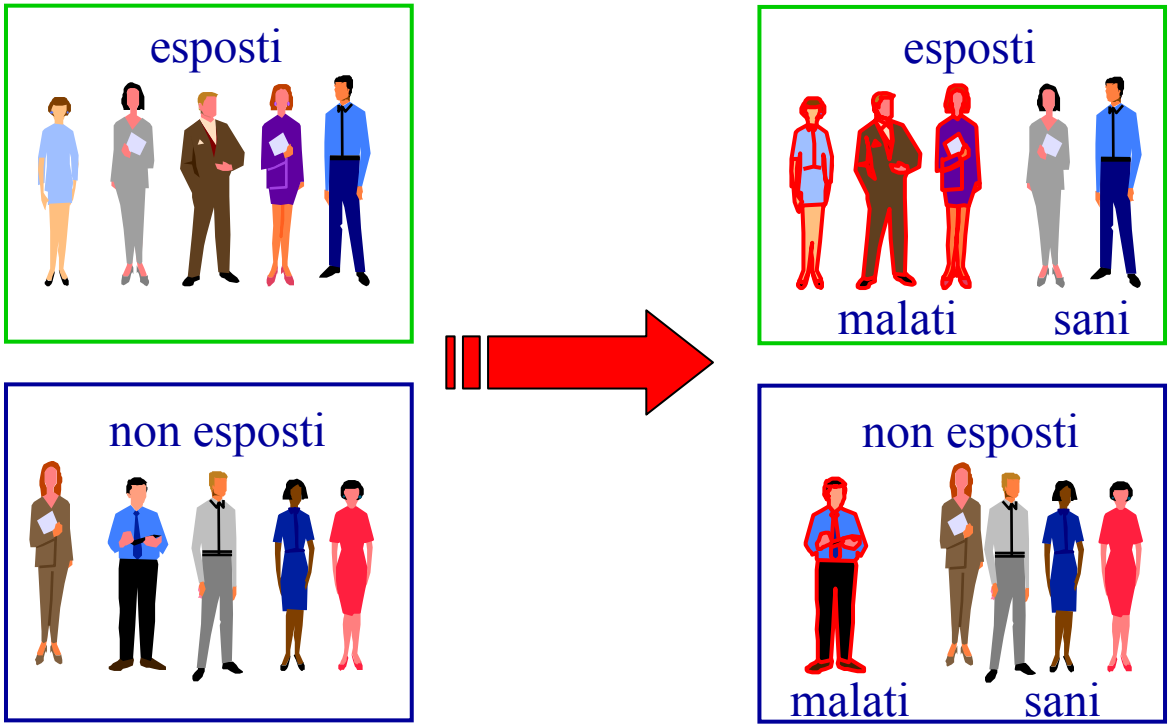
RR = 1,41

IC 95%= (0,89 - 2,25)

La relazione causa-effetto

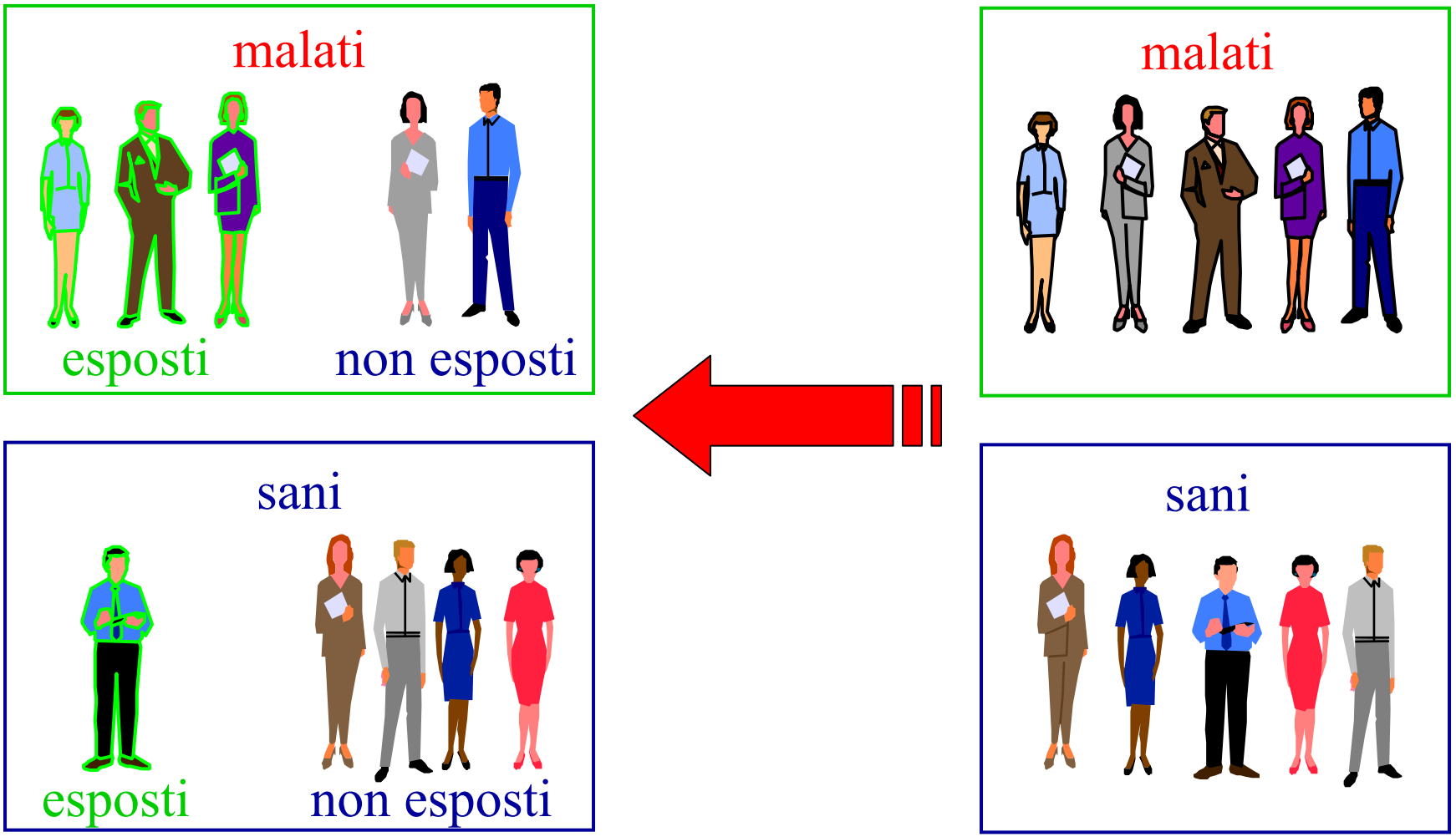
- Gli studi *caso-controllo* ed altri modelli di studio
- Vantaggi e limiti degli studi analitici
- Misure di impatto sulla popolazione

Limiti di uno studio di coorte

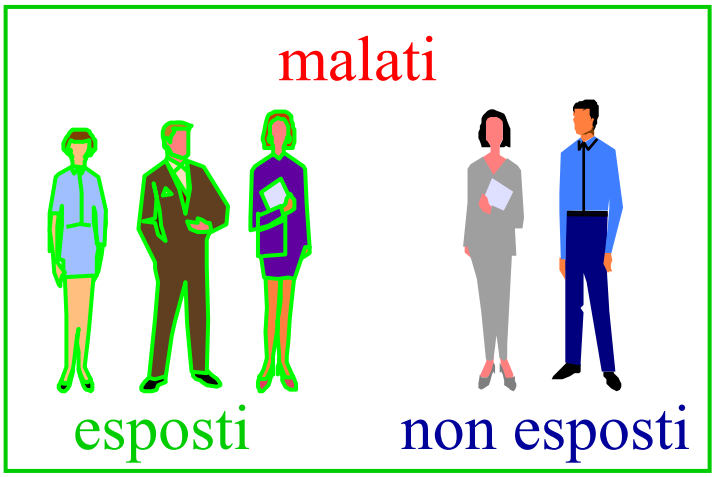


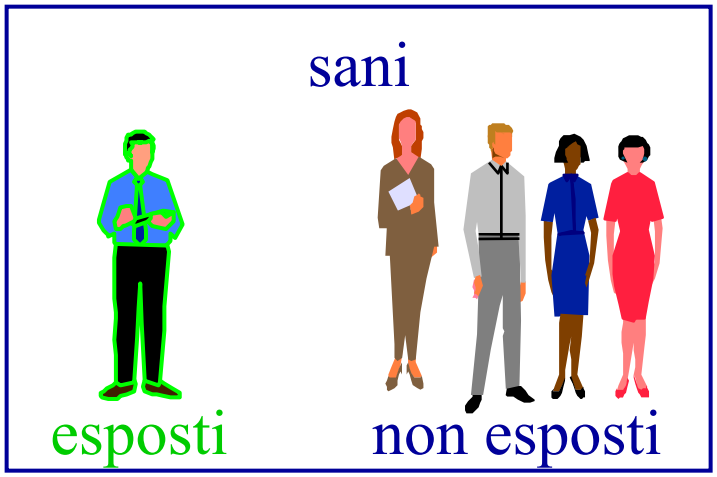
nello studio di coorte si seleziona una popolazione sana e si valuta l'incidenza di malattia in funzione dell'esposizione ad un fattore di rischio

Lo studio caso-controllo



Lo studio caso-controllo



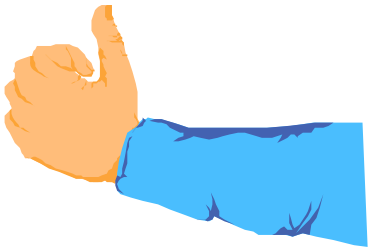


	M+	M-	
E+	3 *a*	1 *b*	*4*
E-	2 *c*	4 *d*	*6*
	5	*5*	***10***

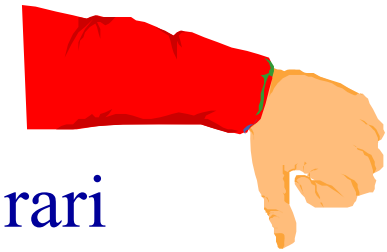
$$\text{Odds Ratio} = \frac{\text{odds}_{M+}}{\text{odds}_{M-}} = \frac{3/2}{1/4} = 6$$

$$\textit{Odds Ratio} = \frac{a/c}{b/d} = \frac{a \times d}{b \times c}$$

Vantaggi e limiti degli studi caso-controllo

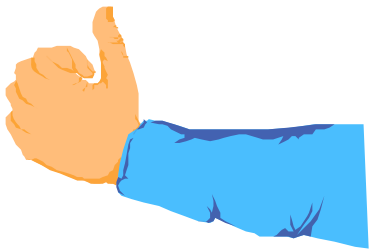


- Costo
- Possibilità di valutare fattori di rischio multipli per un'unica patologia
- Possibilità di studiare patologie rare



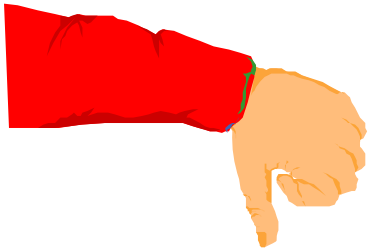
- Maggiore intervento di bias
- Non adatti allo studio di fattori di rischio rari
- Forniscono solo la stima del rischio nella popolazione

Vantaggi e limiti degli studi di coorte



- Misura diretta dell'incidenza
- Possibilità di valutare effetti multipli di un unico fattore di rischio
- Possibilità di studiare fattori di rischio rari

- Costo
- Non adatti allo studio di patologie rare



Indagine di una epidemia di tossinfezione alimentare

Approccio 1: studio di coorte

- Intervista a tutti i partecipanti
- Calcolo dei tassi d'attacco specifici per alimento
- Valutazione del rischio relativo

Approccio 2: studio caso-controllo

- Intervista ad un campione di partecipanti
- Valutazione degli alimenti consumati da casi e controlli
- Valutazione dell'odds ratio

Indagine di una epidemia di tossinfezione alimentare

Nome	sesso	età	Data sintomi	Ora sintomi	Antipasto	Primo	Secondo	Contorno	Dolce
Intervistato 1	M	20	12/12/97	12:00	X	X	X	X	X
Intervistato 2	M	24	13/12/97	09:00	X	X			
Intervistato 3	F	35	13/12/97	11:30	X	X	X	X	X
Intervistato 4	M	12	12/12/97	10:30	X		X	X	X
Intervistato 5	F	45	13/12/97	08:15		X	X	X	X
Intervistato 6	F	48	13/12/97	15:00		X			X
Intervistato 7	F	10			X		X	X	X
Intervistato 8	M	25				X			X
Intervistato 9	M	41			X		X	X	
Intervistato 10	F	10			X	X	X	X	X
Intervistato 11	F	5				X			X
Intervistato 12	M	49			X		X	X	X

coorte	$RR_{primo} = \frac{5/8}{1/4} = 2,5$	$RR_{secondo} = \frac{4/8}{2/4} = 1,0$
caso controllo	$OR_{primo} = \frac{5/1}{3/3} = 5,0$	$OR_{secondo} = \frac{4/2}{4/2} = 1,0$

Rischio attribuibile individuale

(rischio attribuibile per gli esposti)

$$RA = I_{Exp+} - I_{Exp-}$$

quantità di rischio supplementare (eccesso di rischio) attribuibile al fattore di rischio considerato

é una misura di incidenza

Frazione attribuibile

(rischio attribuibile fra gli esposti)

$$FA = \frac{I_{Exp+} - I_{Exp-}}{I_{Exp+}}$$

la proporzione di malati fra gli esposti che eviterebbe la malattia se fosse rimosso il fattore di rischio

é una proporzione

Rischio attribuibile di popolazione

$$RAP = RA \times P$$

dove P rappresenta la prevalenza del fattore di rischio nella popolazione

la quantità di casi nell'intera popolazione che non ammalerebbero se fosse rimosso il fattore di rischio considerato

é una misura di incidenza

Frazione eziologica nella popolazione

$$FE = \frac{RAP}{I_{Tot}} = \frac{I_{Tot} - I_{Exp-}}{I_{Tot}}$$

la proporzione di casi nell'intera popolazione dovuti al fattore di rischio

é una proporzione

Misure di rischio e associazione

Per poter identificare i fattori di rischio delle Malattie occorre agire secondo due fasi in Sequenza:

I FASE → Ricercare l'associazione statistica Fra fattore di rischio e malattia;

II FASE → risalire alla possibilità che il fattore di rischio associato abbia in realtà un vero e proprio ruolo favorente lo sviluppo della malattia

Misure di rischio e associazione

Il rischio può essere valutato in vari modi:

- ASSOLUTO
- RELATIVO
- ATTRIBUIBILE

RISCHIO ASSOLUTO

Rappresenta l'incidenza della malattia tra gli esposti al fattore di rischio, ossia la proporzione di soggetti che durante il periodo di osservazione sviluppa la malattia.

Tale misura non fornisce tuttavia alcuna informazione Su quanto quel fattore di rischio influisca realmente Sullo sviluppo della malattia, poiché l'incidenza Potrebbe essere uguale (o addirittura superiore) anche In coloro che non risultano esposti; per ottenere questa Informazione deve essere considerato il **RISCHIO RELATIVO**

RISCHIO RELATIVO

E' definito dal rapporto fra incidenza negli esposti e quella nei non esposti allo stesso fattore di rischio:

$$RR = \frac{I\ exp +}{I\ exp -}$$

Esprime di quanto maggiore è il rischio di coloro che sono esposti al fattore rispetto ai non esposti.

RR costituisce una misura statistica della forza della associazione tra fattore di rischio e malattia e dovrebbe risultare pari a 1 se il fattore di rischio considerato non ha influenza nello sviluppo della malattia

RISCHIO ATTRIBUIBILE

Rappresenta la quota di rischio supplementare attribuibile al fattore di rischio considerato, ossia la quota di malati che eviterebbero la malattia se fosse completamente rimosso dalla popolazione il detto fattore di rischio.

$$RA = (I\ exp\ +) - (I\ exp\ -)$$

RA è dato dalla differenza tra incidenza negli esposti ed Incidenza nei no esposti.
Il valore RA esprime quante volte è maggiore il rischio di ammalare negli esposti rispetto ai non esposti.

A parità di RR, RA può essere molto diverso indicando un ben differente impatto assoluto della presenza del fattore di rischio.

Misure di rischio e associazione

$$RA = (I\ exp +) - (I\ exp -)$$

Talvolta si preferisce esprimere lo stesso concetto in termini di rischio attribuibile negli esposti (RAE) (detto anche attributable proportion nella terminologia anglosassone) che rappresenta la proporzione di malati in una popolazione esposta che può essere evitata rimuovendo il fattore di Rischio. E' dato dalla differenza tra incidenza negli esposti Ed incidenza nei non esposti diviso l'incidenza degli esposti

$$RAE = \frac{(I\ exp +) - I\ exp -)}{(I\ exp +)}$$

Misure di rischio e associazione

I due esempi (**A** e **B**), riportati in due Tabelle 2x2, riguardano due ipotetici studi a coorte in cui si è valutato il ruolo dell'esposizione ad un inquinante chimico in relazione all'incidenza di due malattie.

Es. A	M+	M-	Totale
Exp +	5	495	500
Exp -	1	499	500
Totale	**6**	**994**	**1000**

$$RR = \frac{5/500}{1/500} = \frac{0,01}{0,002} = 5$$

RA = 0,01 – 0,002 = 0,008 = 0,8%

Es. B	M+	M-	Totale
Exp +	100	300	400
Exp -	30	570	600
Totale	**130**	**870**	**1000**

$$RR = \frac{100/400}{30/600} = \frac{0,25}{0,05} = 5$$

RA = 0,25 – 0,05 = 0,20 = 20%

Si può notare come, a parità di RR, un RA più alto indica che una percentuale più alta di esposti si ammala a causa del fattore di rischio (4 su 500 nell'es. A pari allo 0,8% e 80 su 400 nell'es. B pari al 20%). Questi casi non si sarebbero quindi verificati se fosse stato rimosso il fattore di rischio. Pertanto, a parità di RR, il RA è tanto più alto quanto maggiore è l'incidenza della malattia.

Dividendo il RA ottenuto per l'incidenza negli esposti si ottiene la percentuale di casi attribuibili al fattore o rischio attribuibile negli esposti (RAE); esso rappresenta la percentuale di casi di malattia teoricamente prevenibile nella popolazione degli esposti dopo rimozione del fattore di rischio considerato.

Misure di rischio e associazione

$RAE_A = 0{,}008/0{,}01 = 80\%$

$RAE_B = 0{,}20/0{,}25 = 80\%$

La misura è direttamente equivalente al rischio relativo; infatti nell'esempio specifico i due risultati sono uguali

Dall'esempio si evince come il RR rappresenti una misura prettamente etiologica, mentre il RA una misura di impatto nella popolazione. Da notare che abbiamo parlato di rapporto tra due rischi (incidenze cumulative) ma le stesse misure possono essere calcolate anche considerando i tassi di incidenza persona/tempo o quelli di mortalità.

Associazione tra caratteri

1) studio dei possibili fattori che provocano o facilitano l'instaurarsi di una malattia (studio delle associazioni);
2) misure di rischio per valutare le possibili associazioni tra esposizione e malattia;
3) fattori di confondimento e distorsioni del metodo statistico.

L'associazione e' il grado di dipendenza statistica tra 2 o piu' eventi variabili;

Infatti l'associazione puo' essere:

- causale o eziologica (il fumo di tabacco provoca il cancro);
- secondaria o indiretta (la bronchite cronica, causata dal fumo, e' associata al cancro);
- non causale o spuria o artificiale: e' determinata da una circostanza esterna: o un fattore di confondimento o una distorsione della metodologia statistica usata.

Misure di rischio

Facciamo l'esempio di due gruppi di soggetti (ad es. quelli con colesterolo alto e quelli con colesterolo basso), inizialmente sani, che esposti ad un fattore di rischio (colesterolemia alta) dopo un certo tempo sviluppano una malattia (cardiopatia).

Al termine del periodo di follow-up si avranno 4 categorie di soggetti:

malati esposti (a),

malati non esposti (c),

non malati esposti (b)

***non malati non esposti (d)*:**

Si consideri uno studio prospettico (1)

	Malato (M+)	Non malato (M-)	Totale
Esposto (E+)	a=50	b=450	500
Non esposto (E-)	c=25	d=475	500

La probabilità che un soggetto esposto sia malato è detta Incidenza o rischio assoluto:

a/a+b, cioe' 50/500

… oppure i risultati di un Trial (2)

	Morti	Non Morti	Totale
Terapia tradizionale (TT)	35	41	76
Terapia Sperimentale (TS)	49	26	75

Rischio attribuibile individuale (RA) o Riduzione del Rischio Assoluto (RRA)

Rappresenta la quantita' di rischio supplementare attribuibile al fattore di rischio (o alla terapia tradizionale):

(1) $RA = I_{E+} - I_{E-} = 0.10 - 0.05 = 0.05$

(il fattore di rischio aumenta il rischio del 5%)

(2) $RA = I_{(TT)} - I_{(TS)} = 0.46 - 0.65 = -0.19$

(la terapia sperimentale aumenta il rischio di morte del 19%: si noti il segno negativo di RA)

Rischio Relativo (RR o risk ratio)

Rapporto fra incidenza negli esposti e incidenza nei non esposti, cioe':

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{50/500}{25/500} = \frac{0.10}{0.05} = 2 \quad (1)$$

(cioe' gli esposti hanno un rischio doppio dei non esposti).

Se il valore e' attorno a 1 indica che il fattore non ha influenza nello sviluppo della malattia;
se e' <1 indica che il fattore ha un ruolo protettivo,
se e' >1 indica che esiste un'associazione tra fattore e malattia.

Rischio Relativo (RR o risk ratio)

Rapporto fra incidenza negli esposti e incidenza nei non esposti, cioe':

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{35/76}{49/75} = \frac{0.46}{0.65} = 0.71 \quad (2)$$

(cioe' i pazienti trattati con terapia tradizionale hanno un rischio minore rispetto ai pazienti trattati con terapia sperimentale)

Se il valore e' attorno a 1 indica che le due terapie sono equivalenti;

se e' <1 indica che la terapia al numerat. è più efficace

se e' >1 indica che è meno efficace

Riduzione del Rischio Relativo (RRR)

Rapporto fra incidenza negli esposti e incidenza nei non esposti, cioe':

RRR = 1-RR = 1- 0.71=0.29 **(2)**

(cioe' i pazienti trattati con terapia sperimentale hanno un rischio del 29% più alto dei pazienti trattati con terapia tradizionale)

Rischio Relativo (RR o risk ratio)

$$ES(\ln RR) = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

$$\ln RR \pm z_{\alpha/2} ES(\ln RR)$$

ODDS RATIO O RISCHIO RELATIVO STIMATO

Poiché il rischio relativo può essere correttamente calcolato per mezzo di studi di coorte, si può fare ricorso ad esempio nel caso di studi caso-controllo ad una stima del rischio relativo:

$$OR \approx RR$$

OR valuta l'entità dell'associazione che si considera presente quando il suo valore è significativamente > 1

ODDS RATIO O RISCHIO RELATIVO STIMATO

	Malati (casi)	Non malati (controlli)
Esposti (E+)	a	b
Non esposti (E-)	c	d
Totale	a + c	b + d

$$OR = \frac{a\,d}{b\,c}$$

ODDS RATIO

Cerchiamo di stimare l'intensità dell'associazione tra due variabili nominali. In una tabella 2x2, tale stima è l'**odds ratio**.

Se un evento si verifica con probabilità p, l'odds in favore dell'evento è: $p/(1-p) = p:(1-p)$.

Definizione di Odds(e) e di P(e)

Se un evento (**e**) si verifica con probabilità **P(e)=1/2**, l'odds in favore dell'evento è:

Odds(e) = (1/2)/(1-1/2) = (1/2)/(1/2)=1:1.

Se un evento (**e**) ha un odds in favore **Odds(e)=a:b**,

la probabilità che l'evento si verifichi è **p(e) = a/(a + b).**

Esempio: Se due variabili casuali dicotomiche associate indicano ***malattia*** (m) ed ***esposizione (E) a fattore di rischio***, allora ***l'odds ratio*** è il rapporto tra l'odds di malattia tra i soggetti esposti e l'odds di malattia tra i soggetti non esposti, ovvero:

$$OR = \left[\frac{P(M|E)}{1 - P(M|E)}\right] \Big/ \left[\frac{P(M|\bar{E})}{1 - P(M|\bar{E})}\right]$$

Misure di rischio e associazione

Un campione generico di n soggetti (malati, non malati) (esposti, non esposti) può dare origine ad una tabella di contingenza 2x2 di seguito riportata.

	esposizione		
malattia	Si	No	Totale
Si	a	b	a+b
No	c	d	c+d
Totale	a+c	b+d	n=a+b+c+d

$$p(\text{malattia} \mid \text{esposti}) = \frac{a}{a+c} \qquad 1\text{-}p(\text{malattia} \mid \text{esposti}) = 1 - \frac{a}{a+c} = \frac{c}{a+c}$$

$$P(\text{malattia} \mid \overline{\text{esposti}}) = \frac{b}{b+d} \qquad 1\text{-}p(\text{malattia} \mid \overline{\text{esposti}}) = 1 - \frac{b}{b+d} = \frac{d}{b+d}$$

Con questi rapporti esprimiamo lo stimatore dell'odds ratio:

$$OR = \left(\frac{a}{a+c} \Big/ \frac{c}{a+c} \right) \Big/ \left(\frac{b}{b+d} \Big/ \frac{d}{b+d} \right) = [a/c] \,/\, [b/d] = ad \,/\, bc$$

Misure di rischio e associazione

Il rischio relativo (RR) è il rapporto della probabilità di malattia tra i soggetti esposti e la probabilità di malattia tra i soggetti non esposti.

$$RR = \frac{p(\text{malati} \mid \text{esposti})}{p(\text{malati} \mid \overline{\text{esposti}})} = \left(\frac{a}{a+c} \Big/ \frac{b}{b+d} \right) = \frac{a(b+d)}{b(a+c)}$$

L'odds ratio non è uguale al rischio relativo; ma ...

In presenza di eventi rari (i valori di a e b sono tali da rendere il prodotto $ab \cong 0$) il rischio relativo è approssimato dall'odds ratio:

$$RR = (ab + ad) / (ba + bc) \cong ad / bc = OR$$

In genere si preferisce utilizzare l'odds ratio, poiché la distribuzione campionarla di OR presenta minori difficoltà di quella di RR.

esempio

Scopo: determinare se il monitoraggio elettronico fetale durante il parto faciliti la decisione di parto cesareo,

Metodi: uno studio ha incluso 5.824 neonati e di questi ne ha sottoposti 2.850 a monitoraggio e 2.974 no.

Monitoraggio elettronico fetale			
Parto cesareo	Si	No	Totale
Si	358	229	587
No	2492	2745	5237
Totale	2850	2974	5824

L'odds ratio di parto cesareo nel gruppo sottoposto a monitoraggio fetale verso il gruppo non sottoposto è:

$$OR = [(358)\ (2.745)] / [(229)\ (2.492)] = 1{,}72.$$

intervallo di confidenza di OR

Sembra esistere una moderata associazione tra l'utilizzo del monitoraggio ed il tipo di parto.

Nota: ciò non implica, tuttavia, che il monitoraggio elettronico causi un parto cesareo; E' possibile che i feti a maggior rischio di parto cesareo siano quelli sottoposti a monitoraggio.

L'incertezza di questa stima è riflessa dall'ampiezza del intervallo di confidenza (IC) di OR;

Si ricordi che l'espressione per l'I.C. al 95% per una media

$$(\overline{x} - 1{,}96 \times es(\overline{x}) \ , \ \overline{x} + 1{,}96 \times es(\overline{x}))$$

si basa sulla assunzione che i valori della popolazione originaria siano normalmente distribuiti.

poniamo per comodità $\hat{y}=\ln(OR)$

La distribuzione di probabilità dell'odds ratio e asimmetrica a destra; Infatti l'odds ratio assume solo valori positivi tra 0 ed infinito. Al contrario, la distribuzione di probabilità del logaritmo naturale dell'odds ratio e più simmetrica ed approssimativamente normale.

$$[\ln(OR) - 1.96\, es\{\ln(OR)\}, \ln(OR) + 1.96\, es\{\ln(OR)\}]$$

Pertanto, per calcolare un intervallo di confidenza per ln(OR) prima di tutto dobbiamo conoscere l'errore standard (es) di questa quantità. Per una tabella 2x2 rappresentata nel mode seguente :

L'errore standard di log(OR) è stimato da

$$es(\log(OR)) = [(1/a + 1/b + 1/c + 1/d)]^{1/2}$$

intervallo di confidenza di OR

Se uno dei valori della tabella è uguale a zero , l'errore standard non è definito. In questo caso, aggiungendo 0,5 ad ogni valore correggeremo la situazione. Pertanto, la stima modificata dell'errore standard è:

$$\sqrt{\frac{1}{a+0.5}+\frac{1}{b+0.5}+\frac{1}{c+0.5}+\frac{1}{d+0.5}}$$

La stima appropriata può essere sostituita nell'espressione precedente.

Per trovare L'IC al 95% per l'odds ratio, calcoliamo l'anti-logaritmo dei limiti inferiore e superiore per ottenere:

$$\left(e^{\ln(OR) - 1{,}96\ es\ [\ln(OR)]}\ ;\ e^{\ln(OR) + 1{,}96\ es\ [\ln(OR)]}\right).$$

Per la relazione tra monitoraggio elettronico fetale e tipo di parto, il logaritmo dell'odds ratio stimato e:

L'espressione dell'I.C. al 95% per il logaritmo naturale dell'odds ratio

Misure di rischio e associazione

$$\ln(OR) - 1{,}96 \times es[\ln(OR)] \; ; \; \ln(OR) + 1{,}96 \times es[\ln(OR)]$$

La stima dell'errore standard di ln(OR) è:

$$es\,[\ln(OR)] = [(1/a + 1/b + 1/c + 1/d)\,]^{1/2}$$

$$= (1/358 + 1/229 + 1/2492 + 1/2745)^{1/2} = 0{,}089$$

L'IC al 95% per il logaritmo dell'odds ratio è:

$$(0{,}542 - 1{,}96 \times (0{,}089) \;,\; 0{,}542 + 1{,}96 \times (0{,}089)) = (0.368\,,0.716).$$

e l'IC al 95% per l'odds ratio è: [(exp(0, 368), exp(0,716)] = (1.44 , 2.05).

Siamo confidenti al 95% che l'odds di parto cesareo tra i feti sottoposti a monitoraggio durante il travaglio è da 1,44 a 2,05 volte maggiore dell'odds dei feti non sottoposti a monitoraggio. Si noti che questo intervallo non include il valore 1.

esempio

Viene condotto uno studio sulle cause del tumore dell'esofago

Si intervistano:
435 ***casi*** di tumore
451 ***controlli***

Si vuole verificare il ruolo del fumo e dell'abitudine all'alcool come cause del tumore

L'indagine evidenzia che non assumevano alcool:
107 casi
193 controlli

Misure di rischio e associazione

E' possibile riassumere i dati in una tabella 2x2

	Casi	Controlli	Totale
Alcool si	328	258	586
Alcool no	107	193	300
Totale	435	451	886

1. C'è una associazione significativa tra l'assunzione di alcool e il tumore dell'esofago?
2. Qual è il rischio di avere tumore dell'esofago in seguito all'assunzione di alcool?
3. Il rischio di avere tumore dell'esofago in seguito all'assunzione di alcool è statisticamente significativo?
4. E' possibile determinare un intervallo di confidenza per il rischio?

Misure di rischio e associazione

1. Associazione tra le variabili

$$X^2 = \frac{N\,(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} =$$

$$= \frac{886\,(328*193 - 258*107)^2}{586*300*435*451} = 32{,}74$$

$p<0{,}0001$; l'associazione è significativa

2. Determinazione del rischio

$$OR = \frac{p_t / q_t}{p_c / q_c} = \frac{ad}{bc} = \frac{328*193}{258*107} = 2{,}29$$

3. Verifica di ipotesi su OR

Il test per la verifica dell'ipotesi H_0 OR=1 (se non il rischio non è significativo l'OR è diverso da 1 solo per effetto del caso)

$$X^2 = \frac{(|a - E(a)| - 0{,}5)^2}{Var(a_i)} = (|328\text{-}287{,}71|\text{-}0{,}5)^2/(49{,}64) = 31{,}89$$

$$E(a) = [(a+b)\,(a+c)] / N = 586 * 435 / 886 = 287{,}71$$

$$Var(a) = [(a+b)(c+d)(a+c)(b+d)] / N^2\,(N-1) =$$
$$= (586*435*451*300)/(886*886*885) = 49{,}64$$

$p<0{,}0001$; l'OR è statisticamente significativo

Misure di rischio e associazione

Determinazione dell'intervallo di confidenza. E' possibile risolvere questo problema con due metodi diversi

Metodo di Miettinen. Questo si basa sull'assunzione che sia noto l'errore dell'OR.

- L'OR non segue una distribuzione normale
- Il logaritmo dell'OR segue approssimativamente la distribuzione di Gauss

➢ Un test per la verifica di ipotesi può essere: z=ln(OR) / ES[ln(OR)]

L'intervallo di confidenza si basa sul test per cui:

$$\ln OR \pm \left[z_{1-\alpha/2} \times ES(\ln OR) \right]$$

$$\ln OR \pm \left[z_{1-\alpha/2} \times \frac{\ln OR}{\sqrt{\chi^2}} \right] \Rightarrow \ln OR \times \left(1 \pm \frac{z_{1-\alpha/2}}{\sqrt{\chi^2}} \right)$$

Misure di rischio e associazione

Trasformando con l'esponenziale avremo:

$$OR^{\left(1 \pm \frac{z_{1-\alpha/2}}{\sqrt{\chi^2}}\right)}$$

E' il valore connesso al livello di confidenza: 1,96 se al 95%

E' la radice quadrata del test eseguito per la verifica di ipotesi dell'OR. Si evidenzia che La radice quadrata del X^2 si distribuisce approssimativamente come una z

Con questo metodo l'intervallo di confidenza al 95% è:

$2{,}29^{(1 - 1{,}96 / 5{,}65)} = 1{,}71$

$2{,}29^{(1 + 1{,}96 / 5{,}65)} = 3{,}06$

Misure di rischio e associazione

Nella stessa indagine sono state rilevate informazioni anche sull'abitudine al fumo:
come cambia il rischio associato al consumo di alcool considerando che un individuo può essere fumatore?
309 casi sono anche fumatori
di questi 265 assumono alcool
208 controlli sono fumatori
di questi 151 assumono alcool
Sono anche bevitori:
63 casi che sono anche non fumatori
107 controlli non fumatori

Misure di rischio e associazione

Fumatori (f)			
	Casi	Controlli	Totale
Alcool si	265	151	416
Alcool no	44	57	101
Totale	309	208	517
Non fumatori (nf)			
Alcool si	63	107	170
Alcool no	63	136	243
Totale	126	243	369

ORf= 265*57 / 151*44=
=2,27

ORnf= 63*136 / 107*63=
=1,27

E' possibile determinare un unico OR che ci dia informazioni sul rischio di essere bevitori e anche fumatori?

Misure di rischio e associazione

L'odds ratio di Mantel-Haenszel è un metodo per risolvere il problema

$$OR_{MH} = \frac{\Sigma(a_i d_i / N_i)}{\Sigma(b_i c_i / N_i)}$$

$$OR_{MH} = \frac{[(265*57)/517] + [(63*136)/369]}{[(44*151)/517] + [(63*107)/369]} = \frac{52,44}{31,12} = 1,69$$

Misure di rischio e associazione

Il test di significatività per verificare l'ipotesi H_0 OR=1 è sempre un test X^2

$$X^2 = \frac{[\Sigma a_i - \Sigma E(a_i)]^2}{\Sigma Var(a_i)}$$

I valori attesi e la varianza si calcolano come al solito:

$\Sigma E(a_i)=E(a_1)+E(a_2)=$

$=[(a_1+b_1)(a_1+c_1)/N_1]+[(a_2+b_2)(a_2+c_2)/N_2]=$

$=248,63+58,05=306,68$

$\Sigma Var(a_i)=Var(a_1)+Var(a_2)=$

$=\{[(a_1+b_1)(c_1+d_1)(a_1+c_1)(b_1+d_1)] / N_1^2 (N_1-1)\}+$

$+\{[(a_2+b_2)(c_2+d_2)(a_2+c_2)(b_2+d_2)] / N_2^2 (N_2-1)\} =$

$=19,58+20,67=40,25$

Misure di rischio e associazione

Applicando i dati del nostro esempio al metodo di Mantel-Haenszel si ha:

$$X^2 = \frac{[\Sigma a_i - \Sigma E(a_i)]^2}{\Sigma Var(a_i)} = \frac{(328\text{-}306{,}68)^2}{40{,}25} = 10{,}77$$

$p<0{,}001$; l'OR è statisticamente significativo

L'intervallo di confidenza può essere calcolato con il metodo di Miettinen:

$$OR^{\left(1 \pm \frac{z_{1-\alpha/2}}{\sqrt{\chi^2}}\right)}$$

Limiti di confidenza al 95%

$$1{,}69^{(1 - 1{,}96 / 3{,}28)} = 1{,}23$$

$$1{,}69^{(1 + 1{,}96 / 3{,}28)} = 2{,}32$$

Misure di rischio e associazione

ln(***p*** avere tumore / ***q*** non avere tumore)=

b_0(parametro che indica il rischio di tumore senza altri fattori)+

+b_1(parametro che "pesa" la variabile)*fumo+

+b_2*alcool

Alcool e fumo non sono variabili quantitative!

Si trasformano in variabili apparentemente quantitative assegnando arbitrariamente ***valore 1*** alla ***presenza*** del fattore di rischio e *valore 0* all'*assenza* del fattore di rischio.
Queste variabili vengono comunemente chiamate **variabili "dummy".**

Misure di rischio e associazione

Il rischio di tumore in relazione all'abitudine all'alcool e al fumo può essere valutata con la ***regressione logistica***

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_nx_n$$

Dove $y = \ln\left(\frac{p}{q}\right)$ → Trasformazione logit

Questo rapporto è l'***odd***

Semplificando con una sola variabile: $y=b_0+b_1x_1$

$$p = \frac{\exp(b_0+b_1x_1)}{1+\exp(b_0+b_1x_1)} \Longrightarrow OR = \exp b_1$$

Misure di rischio e associazione

Si voglia valutare la relazione tra due variabili dello studio osservazionale sui ricoveri per infarto miocardico:

Decesso

Sesso

Pregresso infarto

Pregresso ictus

Età

…

E' una variabile qualitativa e non consente la valutazione della relazione con un modello di regressione in cui

Y=decesso

X=sesso

X=pregresso infarto

X=pregresso ictus

X=età

Per applicare la regressione posso effettuare una trasformazione della variabile decesso

Misure di rischio e associazione

La variabile decesso è classificata come presenza o assenza

Vogliamo conoscere la probabilità p di avere il risultato
Decesso=presenza
Quando l'età assume un certo valore

p=(numero di decessi) / (totale delle osservazioni)
q=(numero di non decessi) / (totale delle osservazioni) = 1-p

p / q = odd di decesso ed è un numero che assume valori
→0 se prevalgono i sopravvissuti
→1 se i due gruppi si equivalgono
→ $+\infty$ se prevalgono i deceduti

log (p / q) = log (odd) di decesso ed è un numero che assume valori
→ $-\infty$ se prevalgono i sopravvissuti
→ 0 se i due gruppi si equivalgono
→ $+\infty$ se prevalgono i deceduti

Misure di rischio e associazione

log (p / q) è la variabile trasformata da “presenza di decesso”
Può essere utilizzata come variabile dipendente di un modello di regressione.

Questa trasformazione prende il nome di
trasformazione logit
per cui la regressione si chiama
regressione logistica

Esemplificando con una sola variabile: $\log (p / q) = y = b_0 + b_1 x_1$

$$p = \frac{\exp(b_0 + b_1 x_1)}{1 + \exp(b_0 + b_1 x_1)} \Longrightarrow OR = \exp b_1$$

Sensibilità, specificità e valori predittivi

Le tabelle di frequenza cosi' come descritte per il calcolo del Rischio Relativo (o del CHI^2) possono servire per il calcolo della sensibilita' e della specificita' di un test, cioe' due modi per descrivere l'accuratezza di un test.

Test ideale

Un test ideale dovrebbe essere affidabile e valido.

Per **affidabilità** generalmente s'intende la capacità di un test di offrire sempre lo stesso risultato nel corso di misurazioni ripetute. Questa è pertanto una caratteristica intrinseca al test e dipendente dalla bontà dello strumento e/o dell'operatore.

Esiste però un altro parametro importante da valutare, rappresentato dalla **validità**: ovvero dalla capacità di un test di distinguere in una popolazione i soggetti sani da quelli malati.

Utilizzando un test ideale, pertanto, tutti i soggetti sani dovrebbero risultare negativi al test e analogamente tutti i malati dovrebbero risultare positivi.

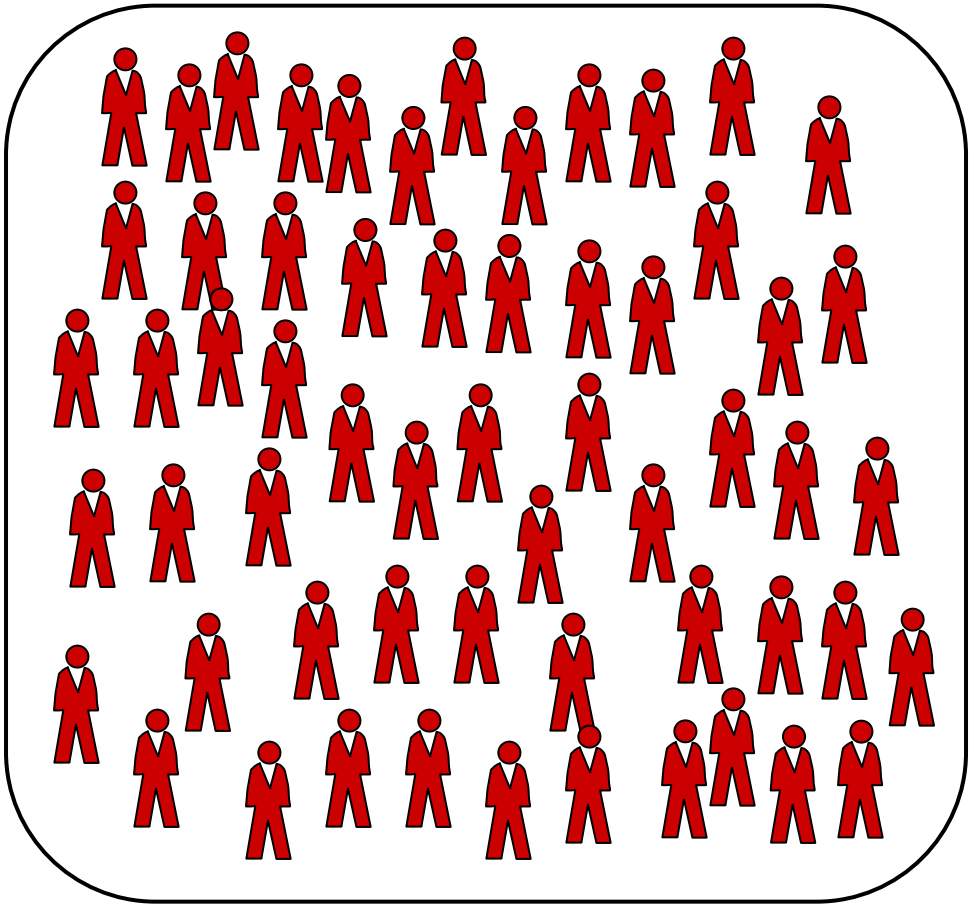
Sensibilità, specificità e valori predittivi

Test ideale

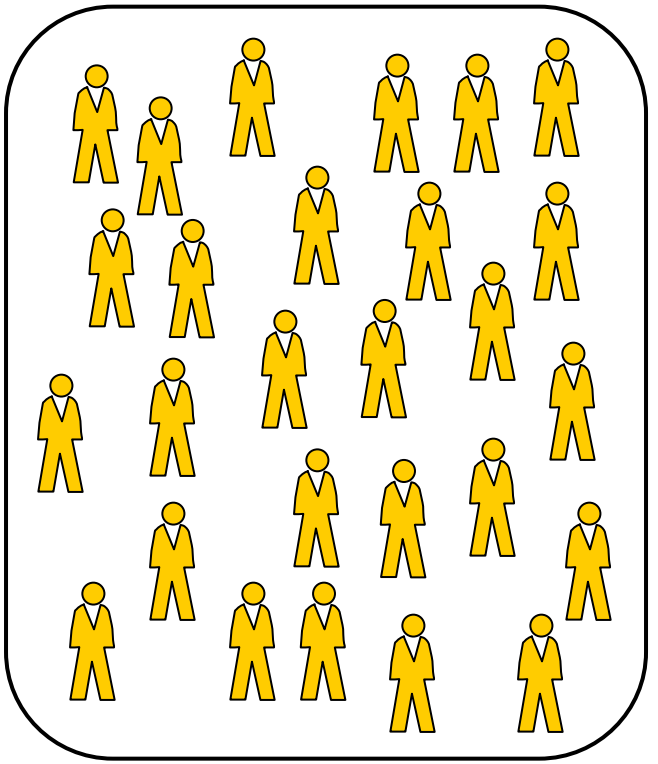
sani

malati

negativi al test



positivi al test

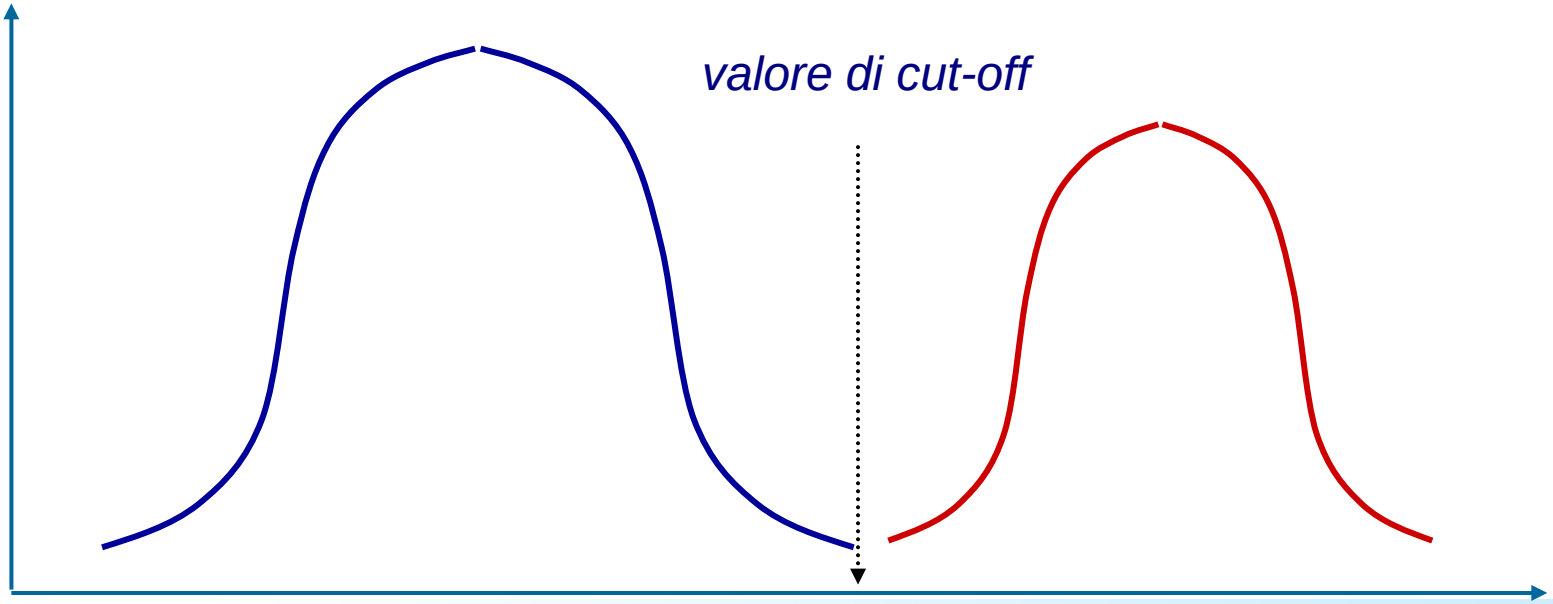


Sensibilità, specificità e valori predittivi

Il valore della variabile, discriminante per assegnare un soggetto al gruppo dei sani o dei malati, viene chiamato **valore di cut-off**.

Ipotizzando che la variabile misurata sia la glicemia basale e la patologia indagata sia il diabete mellito, potremmo dire che (se questo test fosse un test ideale!) dato il valore di 110 mg/ml, tutti i diabetici risulterebbero avere una glicemia basale >110 e tutti i soggetti non diabetici un valore inferiore a 110.

Sfortunatamente in medicina la realtà è notevolmente differente...



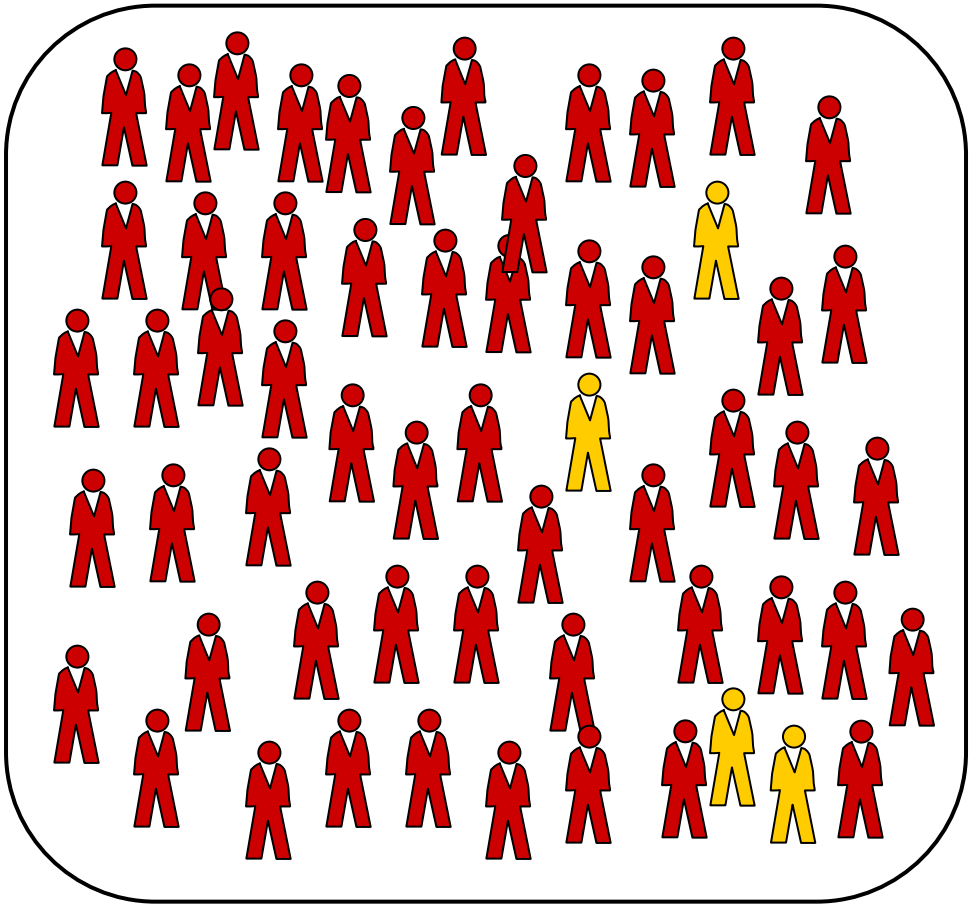
Sensibilità, specificità e valori predittivi

Test reale...

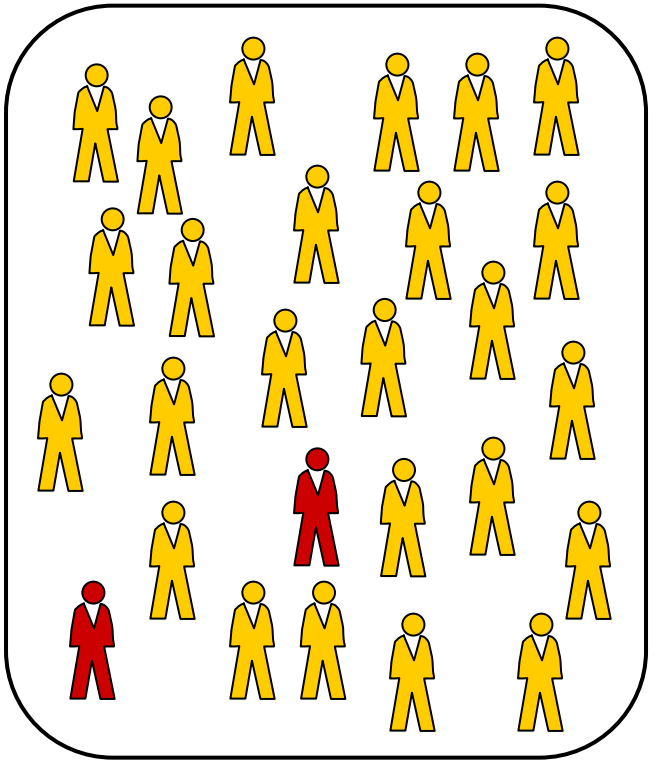
sani

malati

negativi al test

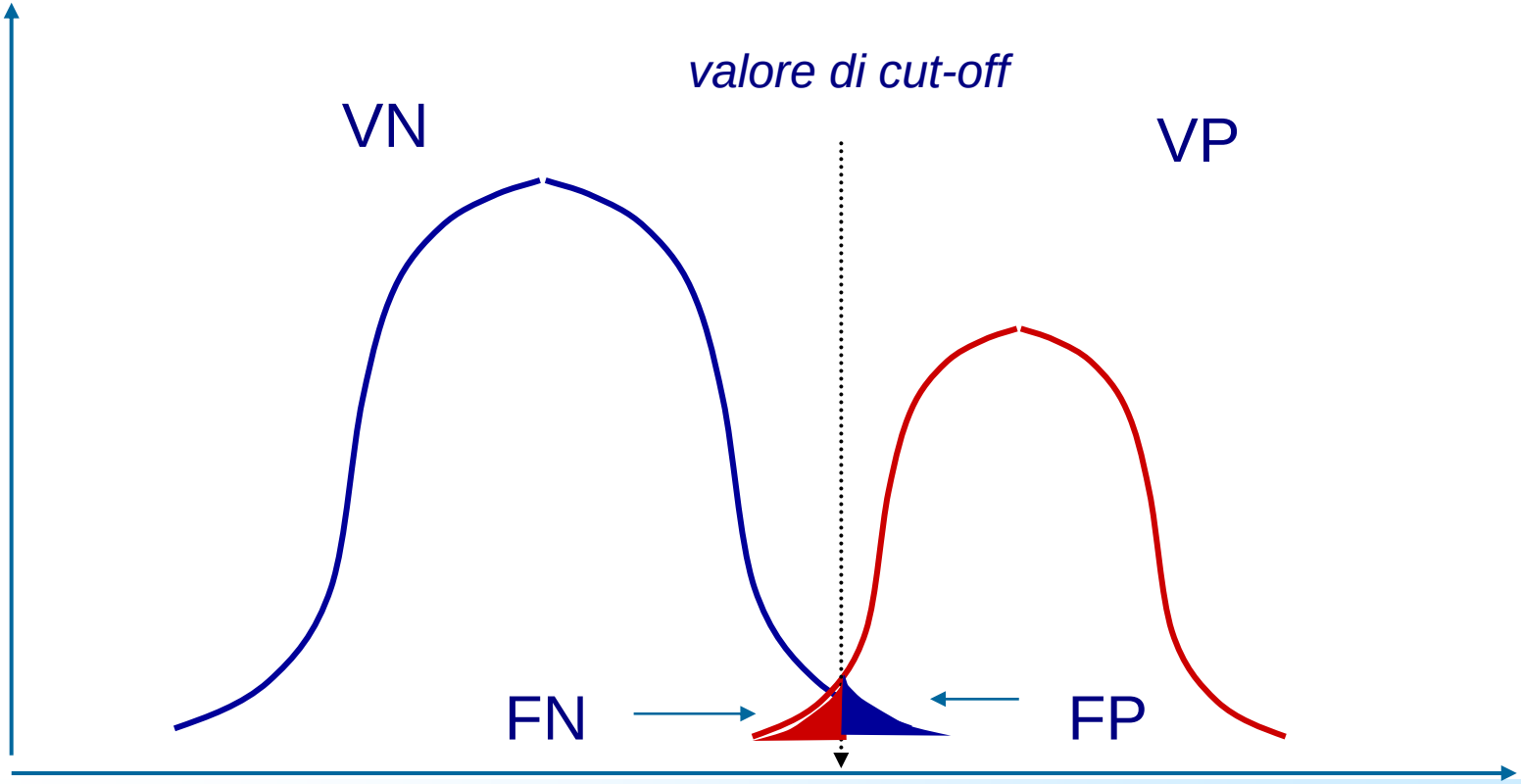


positivi al test



Sensibilità, specificità e valori predittivi

Infatti, sottoponendo una qualsiasi popolazione a un test di screening, purtroppo (dato un certo valore di cut-off) avremo sempre un certo numero di soggetti sani che risulteranno positivi al test e, simmetricamente, un certo numero di soggetti malati che il test non riuscirà a identificare come tali, e pertanto saranno erroneamente classificati come "sani".



Distribuzione della popolazione in relazione al test

Un ulteriore modo di rappresentare la distribuzione di un'ipotetica popolazione in funzione della presenza/assenza di malattia e dei risultati di un test può esser dato dalla classica tabella 2x2, a noi già familiare.

Le colonne rappresentano la distinzione dei soggetti in malati e sani; nelle righe invece i pazienti sono distribuiti in funzione del risultato al test.

Tanto più basse saranno le quote di falsi positivi e falsi negativi, tanto più il test sarà valido.

Distribuzione della popolazione in relazione al test

	M+	M-	
T+	VP	FP	T_P
T-	FN	VN	T_N
	T_{M+}	T_{M-}	N

Sensibilità

Per sensibilità si intende la capacità di un test di individuare in una popolazione i soggetti malati. Essa è data dalla proporzione dei soggetti realmente malati e positivi al test (veri positivi) rispetto all'intera popolazione dei malati.

	M+	M-	
T+	VP	FP	TP
T-	FN	VN	TN
	TM+	TM-	N

capacità del test di individuare in una popolazione i soggetti malati

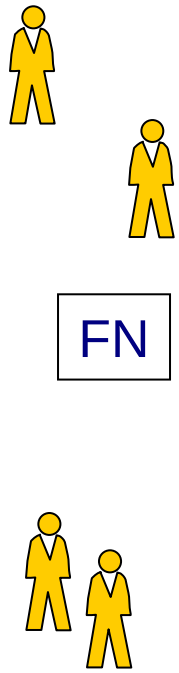
$$\frac{VP}{TM+} = \frac{VP}{VP + FN}$$

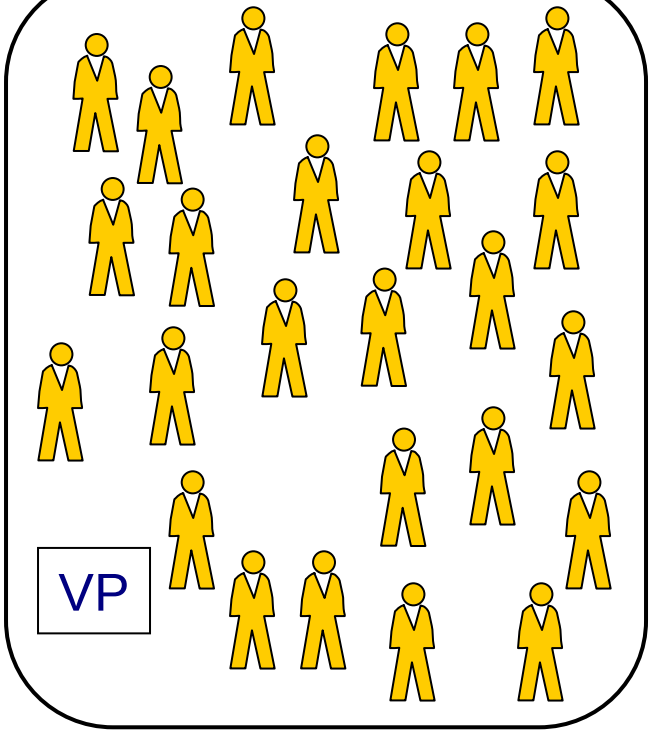
Sensibilità

malati

negativi al test | 25/29= 86,2% | positivi al test

Il calcolo della sensibilità tiene quindi in conto esclusivamente la popolazione dei malati, in funzione dell'identificazione o meno come positivi o negativi al test.





Specificità

Per specificità si intende la capacità di un test di identificare come negativi i soggetti sani. Un test molto specifico, ci consente di limitare la possibilità che un soggetto sano risulti positivo al test.

	M+	M-	
T+	VP	FP	TP
T-	FN	VN	TN
	TM+	TM-	N

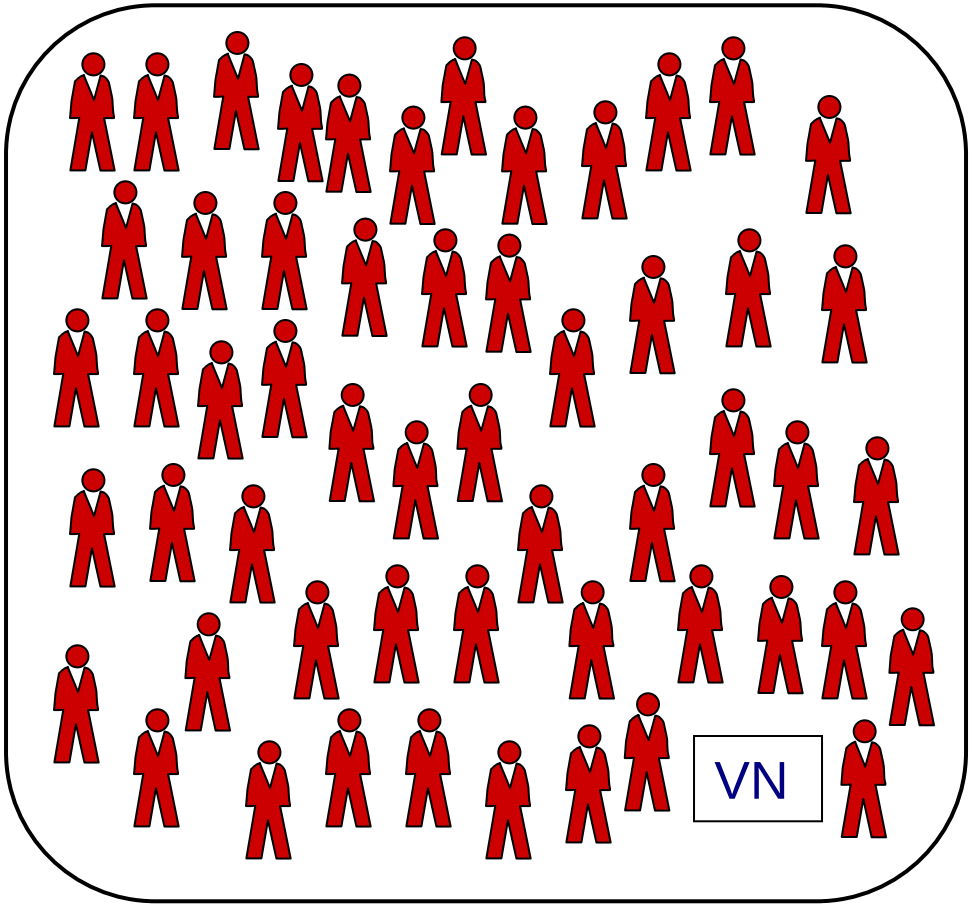
capacità del test di individuare come negativi i soggetti sani

$$\frac{VN}{TM-} = \frac{VN}{VN + FP}$$

Specificità

sani

negativi al test | 55/57= 96,5% | positivi al test



FP

la specificità sarà data dalla proporzione di sani identificati come tali dal test (55) e il totale dei sani (57), quindi sarà pari al 96,5%.

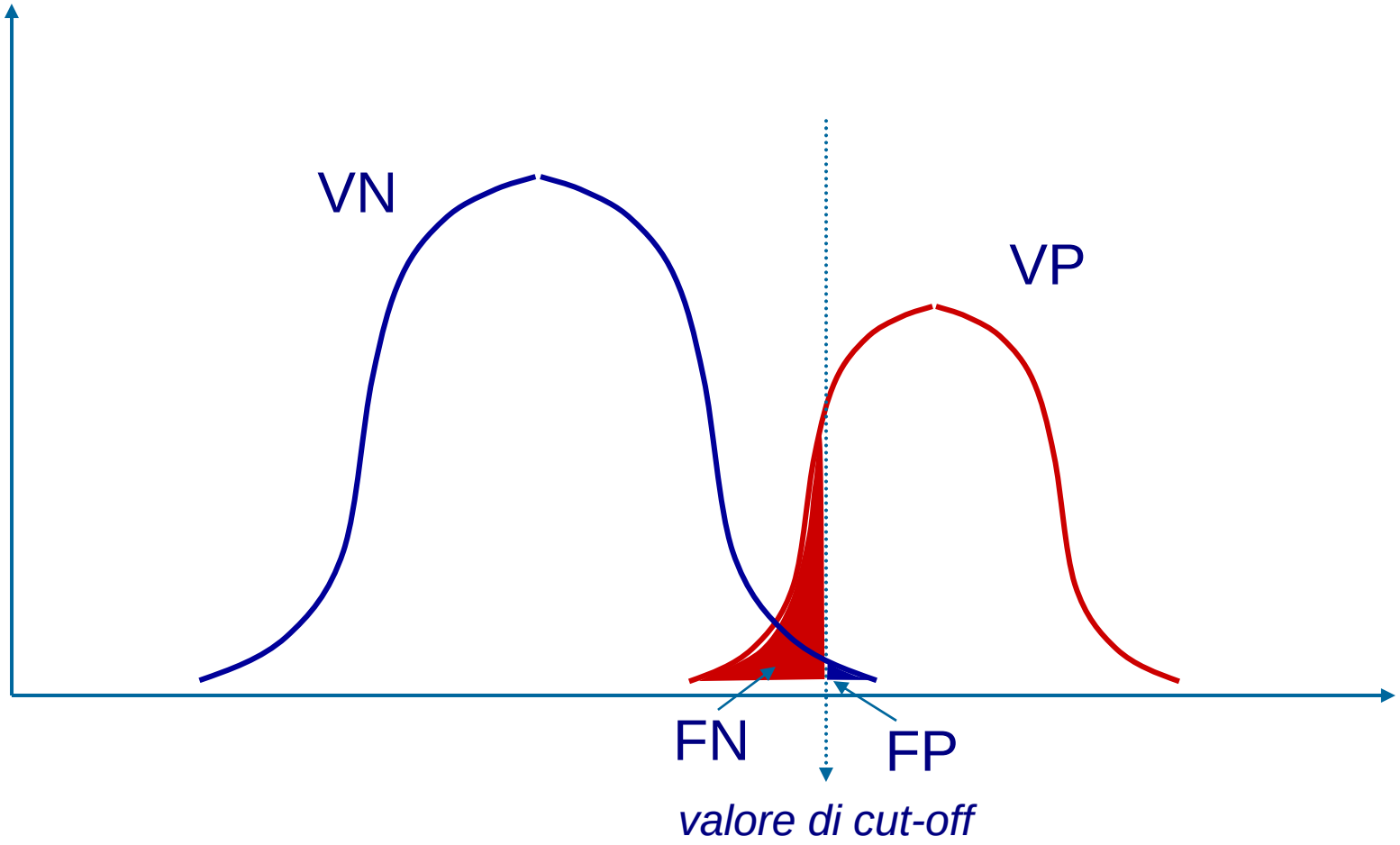
Aumentando la specificità...

Un test altamente specifico sarà dunque un test che produrrà una bassa quota di falsi positivi. Se il test in questione fosse rappresentato dalla misurazione di una variabile continua (per esempio, la glicemia dell'esempio precedente), una maniera per aumentarne la specificità sarebbe quello di aumentare il limite di cut-off, ovvero il livello al di sopra del quale "etichettare" un soggetto come malato.

In ogni caso, ovviamente, la reale distribuzione della popolazione fra malati e sani in funzione della variabile misurata non cambierebbe! Pertanto spostando la linea a destra avremmo una riduzione globale dei soggetti positivi al test con un conseguente aumento della quota di falsi negativi, cioè di soggetti realmente malati che andremmo ad identificare come sani.

Essendo aumentata la quota di falsi negativi, diminuirebbe quindi la sensibilità.

Aumentando la specificità...

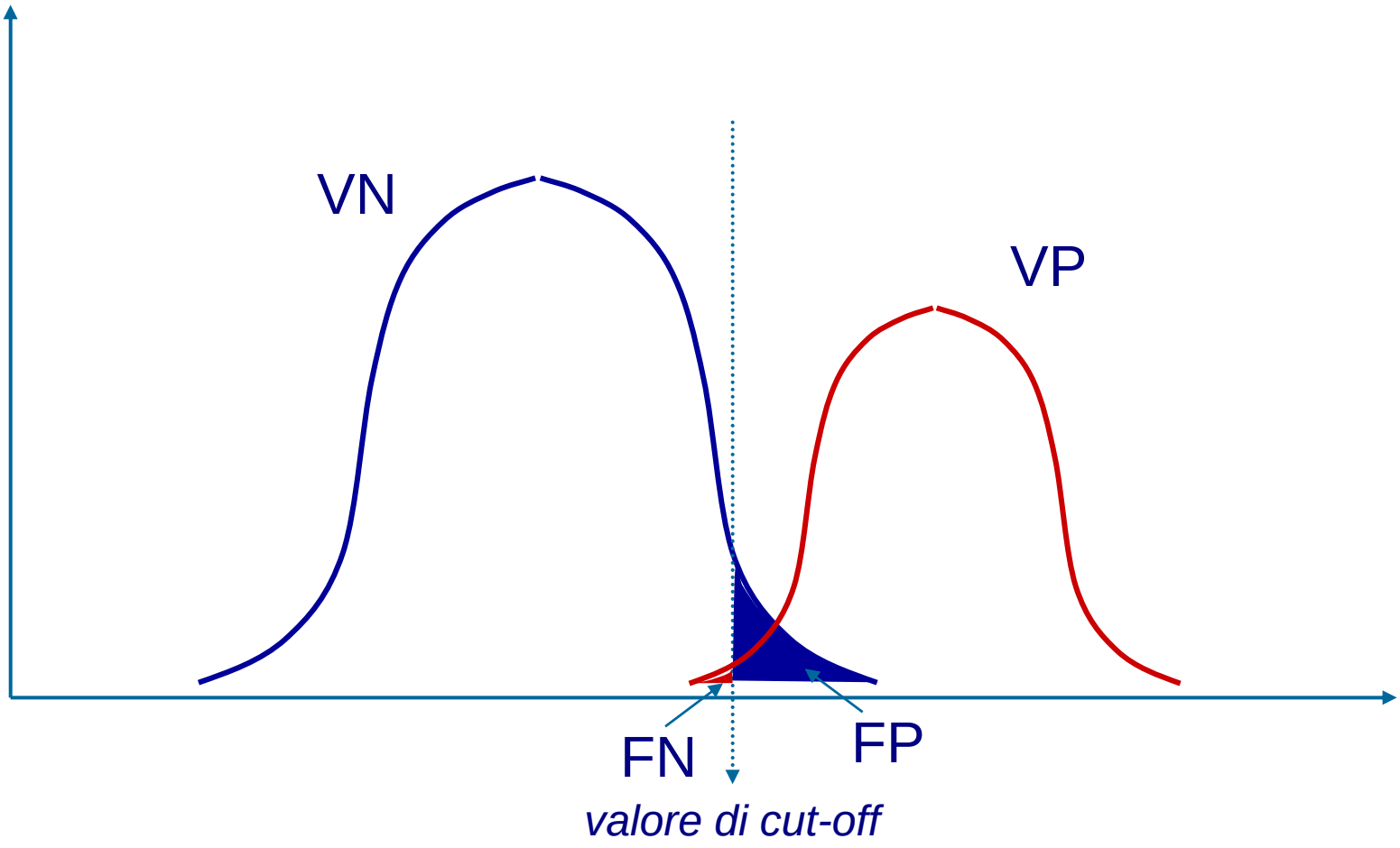


Aumentando la sensibilità...

In maniera analoga, potremmo invece aumentare la sensibilità del test (se, per esempio, volessimo garantirci di poter riconoscere la quota più alta possibile di soggetti malati). In tal caso, inevitabilmente, abbassando il livello di cut-off, includeremmo nel gruppo dei positivi un certo numero di sani (la coda destra della curva dei sani) che rappresenterebbero i falsi positivi: diminuirebbe pertanto la specificità del test.

Sensibilità e specificità sono, quindi, due parametri reciprocamente dipendenti.

Aumentando la sensibilità...



Valore predittivo positivo

Finora abbiamo trattato di parametri che, in un certo senso, sono definibili a priori: sensibilità e specificità sono caratteristiche intrinseche di un test. Esse ci informano su quale sia la probabilità di reclutare soggetti malati o sani da una certa popolazione di partenza (di malati o di sani), mentre nulla ci dicono sulla probabilità che abbiamo, di fronte ad un singolo risultato positivo, che quel soggetto sia realmente malato.

Per rispondere a questo interrogativo dobbiamo poter calcolare un nuovo parametro: il **valore predittivo positivo** (**VPP**). Esso esprime proprio la probabilità che ha un soggetto, risultato positivo al test, di essere realmente malato.

Il VPP si calcola come quota di soggetti veri positivi sul totale dei positivi (veri e falsi positivi).

Valore predittivo positivo

	M+	M-	
T+	VP	FP	T_P
T-	FN	VN	T_N
	T_{M+}	T_{M-}	N

la probabilità che un soggetto positivo al test sia effettivamente malato

$$\frac{VP}{T_P} = \frac{VP}{VP + FP}$$

Valore predittivo positivo

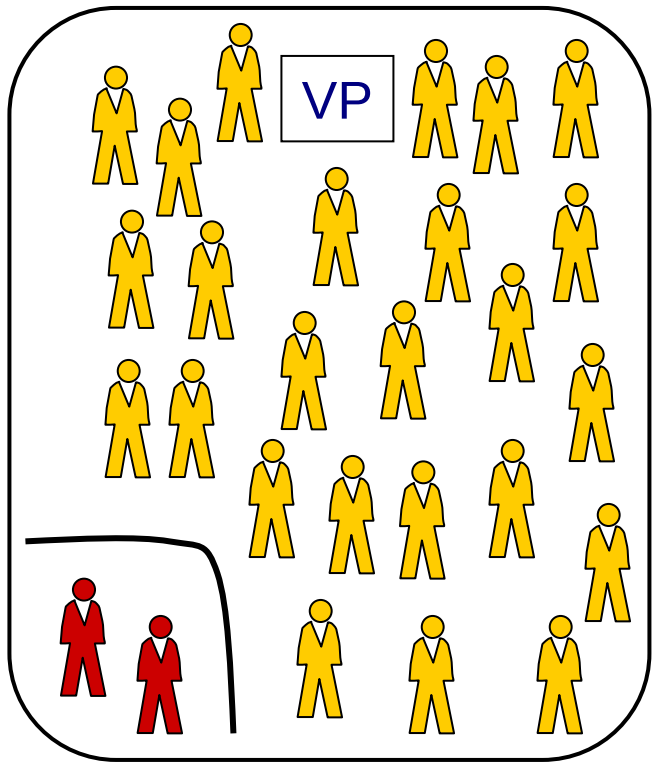
25/27 = 92,6%

sani

malati

Ritornando allo schema della popolazione, valutando tutti i soggetti identificati come positivi al test, il VPP sarà dato dal numero di soggetti realmente malati (veri positivi), cioè 25, su tutti quelli risultati positivi (veri e falsi), cioè 27.

Il valore risultato (92,6%) indica la probabilità per un soggetto con un test positivo di essere realmente malato.



Valore predittivo negativo

In maniera del tutto speculare possiamo calcolare il valore predittivo negativo, come la quota di veri negativi sul totale dei negativi.

	M+	M-	
T+	VP	FP	T_P
T-	FN	VN	T_N
	T_{M+}	T_{M-}	N

la probabilità che un soggetto negativo al test sia effettivamente sano

$$\frac{VN}{T_N} = \frac{VN}{VN + FN}$$

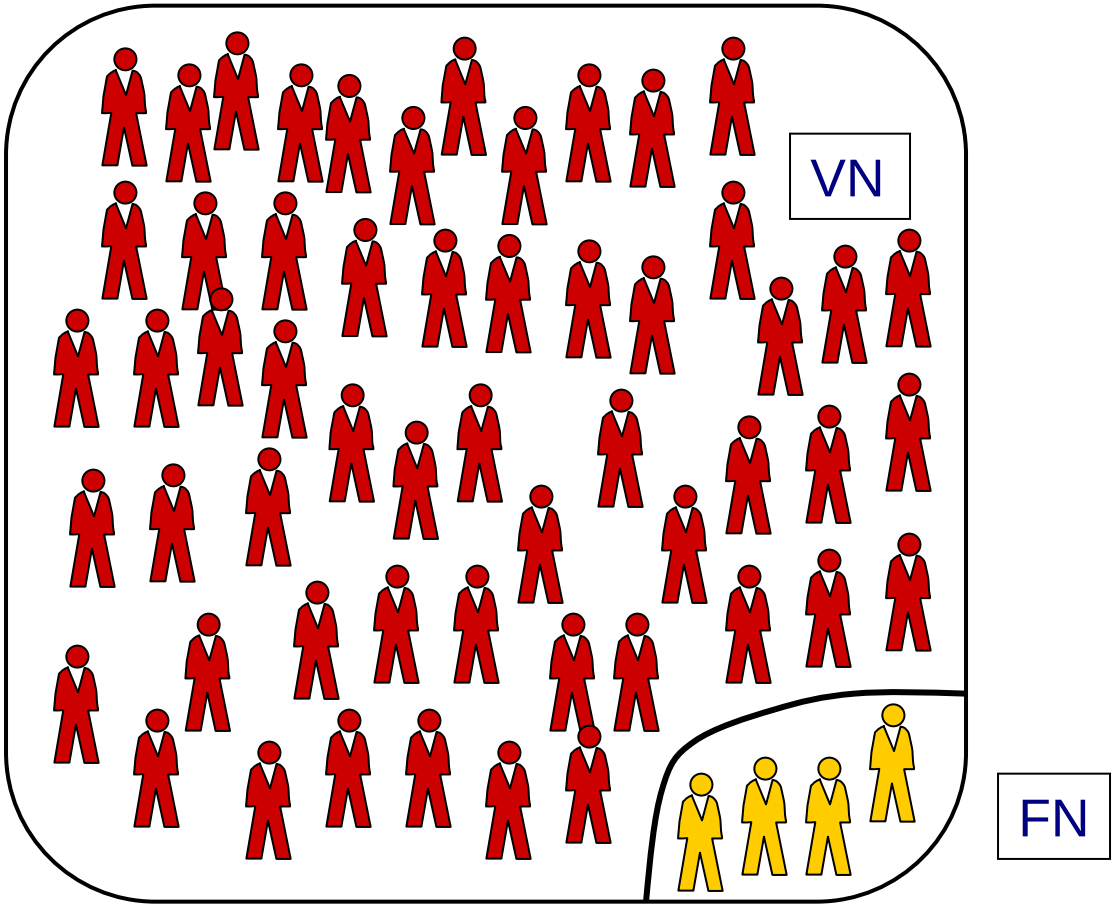
Valore predittivo negativo

sani

malati

negativi al test

55/59 = 93,2%



Nell'esempio in figura esso sarà pari a 55 (veri negativi) sul totale dei negativi (che include anche 4 falsi negativi). Il 93,2% indicherà la probabilità che ha un soggetto risultato negativo al test di essere effettivamente sano.

Influenza della prevalenza nei valori predittivi

I valori predittivi di un test sono influenzati pesantemente dalla prevalenza della condizione in esame.

Un test con un valore predittivo positivo molto vicino al 100% sarà comunque poco utile se la prevalenza della condizione che vogliamo studiare è molto bassa.

In altre parole quanto più la prevalenza della condizione in esame è elevata, tanto migliore sarà la performance di un test con un elevato valore predittivo.

La conseguenza diretta di questa osservazione è che lo stesso test diagnostico potrà funzionare in modo diverso secondo la popolazione che viene ad esso sottoposto.

Influenza della prevalenza nei valori predittivi

Se vogliamo applicare un test di screening alla popolazione generale, la probabilità di incontrare una determinata condizione patologica sarà uguale alla prevalenza.

Se invece vogliamo applicare il test diagnostico ai pazienti afferenti a un ambulatorio specialistico, la prevalenza di questa popolazione sarà notevolmente maggiore di quella della popolazione generale.

Questa "prevalenza", o meglio, la probabilità di incontrare un paziente con una certa malattia si definisce come "**probabilità pre-test**".

Tale probabilità può variare secondo la prevalenza nella popolazione generale, il gruppo di età, il sesso, la presenza di sintomi clinici, e, appunto, lo scenario nel quale il paziente viene osservato.

Likelihood ratio (rapporto di verosimiglianza)

Il likelihood ratio, o rapporto di verosimiglianza, **esprime di quante volte la probabilità di una determinata diagnosi di malattia è modificata per effetto del test**.

Per il rapporto di verosimiglianza positivo, valori superiori a 10 indicano che il test è molto efficace nell'aumentare la nostra probabilità pre-test. Allo stesso modo, per il rapporto di verosimiglianza negativo, valori minori di 0,1 sono da considerare tipici di test particolarmente attendibili.

L'uso di questo parametro permette di eseguire valutazioni della performance di un test diagnostico del tutto indipendenti dalla prevalenza della condizione in esame e di verificarne l'utilità secondo la propria realtà specifica.

Likelihood ratio (rapporto di verosimiglianza)

	M+	M-	
T+	VP	FP	TP
T-	FN	VN	TN
	TM+	TM-	N

LR+ proporzione di veri positivi rispetto alla proporzione di falsi positivi

$$\frac{VP}{FP} = \frac{\text{sensibilità}}{\text{1-specificità}}$$

LR- proporzione di falsi negativi rispetto alla proporzione di veri negativi

$$\frac{FN}{VN} = \frac{\text{1-sensibilità}}{\text{specificità}}$$