Statistica Medica Prima Lezione

Scienze infermieristiche (Università degli Studi di Siena)

Scansiona per aprire su Studocu

STATISTICA MEDICA

STATISTICA MEDICA

La statistica serve in ambito sanitario perché per fare ricerca la produzione scientifica è orientata verso pubblicazione di articoli con ricerche fate con la statistica.
Anche i protocolli derivano dalle ricerche.

LA STATISTICA E' UNA SCIENZA? ED E' ESATTA?

La statistica non è nemmeno una scienza ma una disciplina in quanto non si hanno certezze concrete. La statistica tenta di interpretare un fenomeno attraverso le sue variabili e grandezze senza voler dire che quello che si afferma è l'assoluta certezza. Si fornisce una rappresentazione della realtà che sicuramente mi pone di fronte a punti interrogativi su quella che è la sua accuratezza e validità e precisione. Quindi il suo tentativo è quello di sintetizzare la realtà attraverso alcune variabili cercando di minimizzare gli errori. Per questo motivo non è esatta. Si deve quantificare questo errore. Ad esempio se dico che è probabile al 95% il mio errore è del 5%. La statistica è esatta entro certi errori.
La statistica è una serie di metodi che serve per poi sintetizzare la realtà per la presa in decisione.
La statistica ha varie definizioni che sono cambiate nel tempo:

- È la sola scienza che permette a differenti esperti, usando gli stessi dai, di arrivare a differenti conclusioni
- È la scienza di produrre fatti inaffidabili a partire da dati certi

In tutti i settori della ricerca è indispensabile la conoscenza dei concetti e dei metodi statistici, sia per i problemi di gestione, sia per quelli di indagine. Per pubblicare i risultati di una ricerca, per esempio, tutte le riviste scientifiche richiedono che la presentazione dei dati e la loro elaborazione seguano criteri ritenuti validi universalmente.

STATISTICA MODERNA

È la raccolta, la presentazione e la elaborazione numerica delle informazioni, per agevolare l'analisi dei dati ed i processi decisionali, ma anche l'insieme di metodi e norme atti a cogliere, analizzare ed interpretare dati numerici o numerabili.

La statistica moderna può essere distinta in 3 parti:

- STATISTICA DESCRITTIVA: spiega come dati raccolti devono essere riportati in tabella, rappresentati in grafici e sintetizzati in indici matematici, allo scopo di individuare le caratteristiche fondamentali del campione.
- STATISTICA MATEMATICA: presenta le distribuzioni teoriche sia per misure discrete sia per misure continue, allo scopo di illustrarne le caratteristiche fondamentali, le relazioni che esistono tra esse, gli usi possibili.
- INTERFERENZA STATISTICA: serve per la verifica delle ipotesi e può essere distinta in vari capitoli, in rapporto alle caratteristiche dei dati (se permettono o meno il ricorso alla distribuzione normale: statistica parametrica e non parametrica) e al numero di variabili (se una, due o più: statistica univariata, bivariata, multivariata).

Dalla definizione: la statistica che è l'insieme di metodi e norme che colgono ed analizzano ed interpretano dati numerici o numerabili →al problema della misurazione: il dato

statistico è rappresentato dall'assegnazione di numeri a variabili o più in generale a concetti. Esistono diversi modi per rilevare variabili o concetti e anche la loro misurazione può rappresentare diversi gradi di difficoltà.
Nell'analisi statistica, occorre porre sempre molta attenzione alle caratteristiche dei dati. Già la fase dell'esperimento che conduce alla raccolta delle informazioni è punto fondamentale, poiché da essa dipendono sia i metodi di descrizione, sia i test da applicare. Quando si compie una rilevazione di dati, che in seguito saranno elaborati mediante tecniche descrittive, in realtà si rilevano i valori o le modalità con le quali si presentano determinate VARIABILI.
Sesso, età, altezza, colore degli occhi ed altre ancora non sono che variabili le cui determinazioni possono essere oggetto di una rilevazione statistica. Ognuna di queste presenta particolari caratteristiche che permettono di classificarla in una ben precisa categoria.

LE CATEGORIE DI VARIABILI

Le variabili QUALITATIVE sono caratterizzate da attributi e si dividono in:

- NOMINALI: l'attributo rappresenta la modalità con la quale si presenta il dato.
- ORDINALI: come sopra con la differenza che le modalità sono organizzate in una scala logica.

Le variabili QUANTITATIVE sono caratterizzate da numeri e si dividono in:

- DISCRETE: numeri non necessariamente interi, collocabili in una scala di tipo ordinale, ma capaci di assumere solo certi valori puntuali.
- CONTINUE: numeri che possono assumere tutti gli infiniti valori all'interno o meno di un determinato intervallo.

È necessario porre estrema attenzione al reale significato da attribuire ai valori numerici che vengono utilizzati. Si possono avere numeri che apparentemente hanno le stesse caratteristiche, ma che in realtà richiedono elaborazioni diverse ed impongono il ricorso a metodi differenti per rispondere ai medesimi quesiti di ricerca.

Esempio:

VARIABILI	DESCRIZIONE	DETERMINAZIONI	CATEGORIA
X	Numero di prestazioni radiologiche dell'U.O. nel mese	(103;250;125;..)	QUANTITATIVA DISCRETA
Y	Tipo di prestazione ricevuta	(RX, TAC, RSM, …)	QUALITATIVA NOMINALE
Z	Costo mensile delle prestazioni(migliaia di €)	(53,1; 125,7; 87,9)	QUANTITATIVA CONTINUA
K	Classe d'età degli utenti	(meno di 30 anni; tra 30 e 55 anni)	QUALITATIVA ORDINALE
Z	Grado di	(scarso; buono; …)	QUALITATIVA

	soddisfazione del servizio ricevuto		ORDINALE

LA RAPPRESENTAZIONE DELLE VARIABILI

Le misure che vengono effettuate su più soggetti in relazione alla stessa variabile sono rappresentate in forma vettoriale.

- N: numero di soggetti su cui viene effettuata la misurazione (N se il gruppo coincide con la popolazione oggetto dello studio)
- X: è la variabile misurata
- X_i: sono le misurazioni della variabile realizzate sui soggetti con i=1, 2, …, n

$$X = (X_1, X_2, \dots , X_n)$$

Se per esempio fosse stata realizzata un'indagine su n=504 soggetti e fossero state misurate le seguenti variabili:

- X: età
- Y: sesso
- K: pressione sistolica

$X = (X_1, X_2, \dots , X_{504}) = (32, 64, …., 56)$
$Y = (Y_1, Y_2, \dots , Y_{504}) = (maschio, femmina, ….., femmina)$
$K = (K_1, K_2, \dots , K_{504}) = (135, 109, ……., 151)$

Un insieme di misure è detto SERIE STATISTICA o SERIE DEI DATI.
Una sua prima ed elementare elaborazione può essere una distribuzione ordinata di tutti i valori, in modo crescente o decrescente, detta SERIAZIONE.
Il valore minimo e il valore massimo insieme permettono di individuare immediatamente il campo (o intervallo) di variazione. Successivamente, la serie può essere raggruppata in classi, contano quanti valori od unità statistiche appartengono ad ogni gruppo o categoria. Si ottiene una distribuzione di frequenza o di intensità, detta anche semplicemente distribuzione.

RAPPRESENTARE I DATI: TABELLE E GRAFICI

Consideriamo per esempio una variabile ottenuta dal conteggio del numero di giorni di incubazione del virus influenzale trascorsi dal momento di inoculazione del virus in una cavia.

5	6	3	4	7	2	3	2	3	2	6	4	3	9	3
2	0	3	3	4	6	5	4	2	3	6	7	3	4	2
5	1	3	4	3	7	0	2	1	3	1	5	0	4	5

n=45
$X = (X_1, X_2, \dots , X_{45}) = (5, 6, 3, 4, ….., 5, 0, 4, 5)$

Il primo passaggio, quindi intuitivo in una distribuzione discreta, consiste nel definire le classi:

- È sufficiente identificare il valore minimo (0, nei dati della tabella) e quello massimo (9)
- Contando quante volte compare ogni modalità di espressione (cioè quanti sono le cavie con uguale numero di giorni di incubazione del virus)

Queste informazioni di norma sono presentate in una tabella impostata come la seguente:

CLASSE X	FREQUENZA ASSOLUTA n
0	3
1	3
2	7
3	12
4	7
5	5
6	4
7	3
8	0
9	1
TOTALE	45

La distribuzione di frequenza offre una lettura rapida delle caratteristiche più importanti della serie di dati. Nella tabella precedente, la cavia "tipica" sviluppa il virus in 3 giorni. Se dovessimo sintetizzare con un solo valore il numero di giorni che occorrono per sviluppare l'influenza diremmo 3, che rappresenta la TENDENZA CENTRALE.
Altra caratteristica importante è il numero minimo e il numero massico, 0 e 9, che insieme forniscono il campo di variazione, una indicazione della VARIABILITA' o DISPERSIONE.
La distribuzione del numero di giorni di incubazione tende a diminuire in modo simile allontanandosi da 3, seppure mantenga frequenze più alte nelle classi con un numero maggiore di giorni: sono indicazioni sulla FORMA DELLA DISTRIBUZIONE, che in questo esempio non è simmetrica (ma asimmetrica) rispetto alla tendenza centrale, a causa di un eccesso di valori più alti.

Sulla base della distribuzione di frequenza è possibile ricavare:

- FREQUENZA RELATIVA: ottenuta dividendo la frequenza assoluta per il numero totale
- FREQUENZA CUMULATA: è la somma di tutte le frequenze relative delle classi minori con quella classe stessa.

Entrambe sono rappresentate da percentuali.

CLASSE X	FREQUENZA ASSOLUTA n	FREQUENZA RELATIVA	FREQUENZA RELATIVA CUMULATA
0	3	6,7%	6,7%

1	3	6,7%	13,3%
2	7	15,6%	28,9%
3	12	26,7%	55,6%
4	7	15,6%	71,1%
5	5	11,1%	82,2%
6	4	8,9%	91,1%
7	3	6,7%	97,8%
8	0	0,0%	97,8%
9	1	2,2%	100%
TOTALE	45	100%	

Le frequenze relative sono utili quando si vogliono confrontare due o più distribuzioni, che hanno un differente numero complessivo di osservazioni. La frequenza cumulata offre informazioni importanti quando si intende stimare il numero totale di osservazioni inferiore (o superiore) ad un valore prefissato (ad esempio il 71% delle cavie sviluppa il virus in meno di 5 giorni, il 56% al massimo in 3 giorni).

Consideriamo adesso l'altezza di un gruppo di pazienti visitati in un ambulatorio ortopedico:

107	83	100	128	143	127	117	125	64	119
98	111	119	130	170	143	156	126	113	127
130	120	108	95	192	124	129	143	198	131
163	152	104	119	161	178	135	146	158	176

n=40

$X = (X_1, X_2, \dots, X_{40}) = (107, 83, 100, \dots\dots, 146, 158, 176)$

È evidente come non sia conveniente fare una classe per ogni centimetro, in analogia a quanto fatto con i dati dell'esempio precedente. In questo caso, il numero di modalità sarebbe nettamente superiore al numero di osservazioni, anche se il campione avesse un numero di osservazioni doppio o triplo. Di conseguenza, si impone la necessità di un raggruppamento in classi, che comprendano più modalità di espressione.

Classe	X_i	60-79	80-99	100-19	120-39	140-59	160-79	180-99
Freq. Assoluta	n_i	1	3	10	12	7	5	2
Freq. Relativa %	f_i	2,5	7,5	25,0	30,0	17,5	12,5	5,0
Freq. Cumulata	---	2,5	10,0	35,0	65,0	82,5	95,0	100,0

RAPPRESENTARE GRAFICAMENTE I DATI:

Le rappresentazioni grafiche servono per evidenziare in modo semplice, a colpo d'occhio, le quattro caratteristiche fondamentali di una distribuzione di frequenza. Le rappresentazioni

grafiche sono numerose e debbono essere scelte in rapporto al tipo di dati e quindi alla scala utilizzata.
Per dati QUANTITATIVI riferiti a variabili continue misurate su scale ad intervalli o di rapporti, di norma si ricorre ad ISTOGRAMMI o poligoni. Gli istogrammi sono grafici a barre verticali nei quali:

- Le misure della variabile sono ripotate lungo l'asse orizzontale
- Nell'asse verticale si ha il numero assoluto, oppure la frequenza relativa o quella percentuale, con cui compaiono i valori di ogni classe

Un istogramma deve essere inteso come una rappresentazione areale: sono le superfici dei vari rettangoli che devono essere proporzionali alle frequenza corrispondenti. Quando le classi hanno la stessa ampiezza, le basi dei rettangoli sono uguali, di conseguenza le loro altezze risultano proporzionali alle frequenze che rappresentano. Solo quando le basi sono uguali, è indifferente ragionare in termini di altezze o di aree di ogni rettangolo. Ma se le ampiezze delle classi sono diverse, bisogna ricordare il concetto generale che le frequenze sono rappresentate dalle superfici e quindi è necessario rendere l'altezza proporzionale.
I POLIGONI sono figure simili agli istogrammi e sono utilizzati di norma per la rappresentazione di valori relativi o di percentuali, n quanto è implicito che l'area totale sottesa sia uguale a 1 o 100%. Come nel caso degli istogrammi, l'asse orizzontale rappresenta il fenomeno, mentre l'asse verticale rappresenta la proporzione o percentuale di ogni classe. Un poligono può essere ottenuto a partire dal vostro istogramma, unendo con una linea spezzata i punti centrali di ogni classe.

Per la distribuzione di dati QUALITATIVI, le rappresentazioni grafiche più frequenti sono:

- Diagrammi a rettangoli distanziai (barre): detti anche grafici a colonne, sono formati da rettangoli con basi uguali ed altezze proporzionali alle intensità (o frequenze) dei vari gruppi considerati. A differenza degli istogrammi, i rettangoli non sono tra loro contigui, ma distaccati. Di conseguenza sull'asse delle x non vengono riportati misure ordinate ma nomi, etichette o simboli, propri delle classificazioni qualitative. Con dati quantitativi o nominali, le basi dei rettangoli sono sempre identiche avendo un solo significato simbolico.
- Gli ortogrammi (nastri): sono uguali ai rettangoli distanziati. L'unica differenza è che gli assi sono scambiati per una lettura più facile.
- Gli areogrammi o grafici a torta: sono grafici in cui le frequenze o le quantità di una variabile qualitativa sono rappresentate da superfici di figure piane, come quadrati, rettangoli o più frequentemente cerchi oppure loro parti. Nel caso dei diagrammi circolari o a torta, si divide un cerchio in parti proporzionali alle classi di frequenza. Gli areogrammi vengono usati soprattutto per rappresentare frequenze percentuali. Hanno il vantaggio di fare capire con immediatezza che la somma di tutte le classi è uguale all'unità (1 0 100%). I diagrammi circolari sono utilizzati per le distribuzioni di variabili nominali, al fine di evitare di stabilire anche involontariamente un ordine, che non esiste tra variabili qualitative. Mettono in evidenza come sono distribuite le singole parti, rispetto all'interno: il cerchio rappresenta l'intero fenomeno ed i componenti sono rappresentati da settori che sono distinti da tratteggi colori o gradazioni di colori differenti. Hanno l'inconveniente che evidenziano con estrema

difficoltà le differenze che non sono molto marcate. Per differenze piccole si dimostrano meno efficaci degli ortogrammi.

- I diagrammi polari (radar): serve per rappresentare le variabili cliniche (mensili, settimanali, giornaliere), come le quantità di pioggia e la temperatura media mensile oppure la quantità di inquinanti presenti nell'aria in un ciclo di 24 ore. A partire da un punto centrale, chiamato polo, si traccia una serie di cerchi concentrici, la cui distanza dal centro misura l'intensità del fenomeno.

LE MISURE DI TENDENZA CENTRALE

Per i CARATTERI QUALITATIVI, la tabella e le rappresentazioni grafiche esauriscono quasi completamente gli aspetti descrittivi, quando sia possibile leggere con esattezza le frequenze delle varie classi.

Per i CARATTERI QUANTITATIVI, si pone il problemi di sintesi oggettive che possano essere elaborate matematicamente e quindi che siano numeriche, al fine di un'analisi obiettiva che deve condurre tutti i ricercatori, con gli stessi dati, alle medesime conclusioni.

Una serie di dati numerici è compiutamente descritta da 3 proprietà principali:

1) TENDENZA CENTRALE o POSIZIONE
2) DISPERSIONE o VARIABILITA'
3) FORMA

1) TENDENZA CENTRALE

Le misure i tendenza centrale o posizione servono per individuare il valore intorno al quale i dati sono raggruppati. La tendenza centrale è la misura più appropriata per sintetizzare l'insieme di osservazioni, se una distribuzione di dati dovesse essere descritta con un solo valore. È la prima indicazione della dimensione di un fenomeno.

Le misure proposte sono 3:

A) MEDIA
B) MODA
C) MEDIANA

La scelta della misura di tendenza centrale di una serie di dati dipende dalle caratteristiche della distribuzione e dal tipo di scala.

A) MEDIA ARITMETICA

La media aritmetica è la misura di tendenza centrale più comunemente utilizzata. È definita come la somma del valore di tutte le osservazioni, diviso il numero di unità.

tutte le osservazioni, diviso il numero di unità.

$$m = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad \text{ovvero...} \qquad m = \frac{\sum_{i=1}^{n} x_i}{n}$$

dove ...

- m = media del campione
- x_i = **i**-esima osservazione della variabile X
- n = numero di osservazioni del campione
- $\sum_{i=1}^{n}$ = sommatoria di tutti gli x_i del campione

La media può essere vista come il baricentro della distribuzione campionaria, quando ogni singola osservazione è rappresentata da un peso convenzionale, identico per tutte, lungo l'asse che riporta i valori su una scala i intervalli o di rapporti. La rappresentazione grafica dei dati e della media, riportata nella figura, mostra otticamente come la somma della distanza dalla media dei valori collocati prima sia uguale alla somma dei valori collocati dopo.

MEDIA ARITMETICA PONDERATA

In una distribuzione di frequenza raggruppata in classi, come valore rappresentativo di ogni classe è preso il dato centrale, nell'assunzione che, entro ogni classe, i dati siano distribuiti in modo uniforme. La media aritmetica di distribuzioni di frequenza raggruppate in classi, detta media aritmetica ponderata, è calcolata più rapidamente secondo la seguente formula:

$$m = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

dove ...

- $m_{\bar{x}}$ = media della distribuzione in classi,
- x_i = valore medio della **i**-esima classe di intervallo,
- f_i = numero di osservazioni della classe **i**-esima classe,
- n = numero di classi,
- $\sum$ = sommatoria per tutte le **n** classi.

MEDIA GEOMETRICA

In genere la media aritmetica viene utilizzata quando tra le grandezze in esame esiste un legame di natura additiva. Esistono altre medie (appartenenti ancora al gruppo delle medie analitiche) da utilizzare quando tra le grandezze esistono altri tipi di legami. La media geometrica (che si indica con m_g) si applica di fronte a grandezze in cui vi è un comportamento di tipo moltiplicativo.

Per il calcolo della media geometrica, è condizione necessaria che la quantità siano tutte positive. Se alcune fossero negative, si deve ricorrere al valore assoluto:

$$m_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \qquad \text{ovvero...} \qquad m_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Conviene ricorrere alla media geometrica quando è necessario uno strumento più sensibile ad una variazione assoluta dei termini bassi piuttosto che ad una eguale variazione dei termini alti.

MEDIA ARMONICA

La media armonica è la stima più corretta della tendenza centrale, per distribuzioni di dati in cui devono essere usati gli inversi. È utilizzata quando i valori di X sono espressi come rapporti di un totale costante od in misure di relazione.

La media armonica è data da:

$$m_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

C) MEDIANA

La mediana è il valore che occupa la posizione centrale in un insieme ordinato di dati. È una misura robusta, in quanto poco influenzata dalla presenza di dati anomali. La sua utilizzazione è indispensabile nel caso di scale ordinali o di ranghi. Le sue caratteristiche più importanti sono 2:

- È calcolata sul numero di osservazioni. Si ricorre al suo uso quando si vuole attenuare l'effetto di valori estremi o comunque prendere in considerazione solo l'informazione fornita dai ranghi
- In una distribuzione o serie di dati, ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.

Per calcolare la mediana di un gruppo di dati, occorre prima di tutto disporre i valori in una fila ordinata in modo crescente oppure decrescente e contare il numero totale n di dati. Se n è:

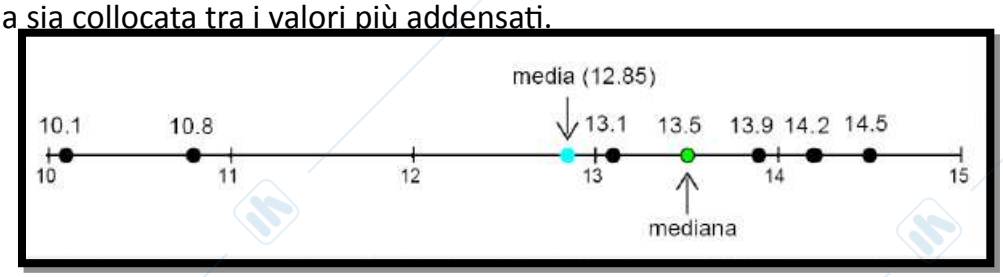
- DISPARI: la mediana corrisponde al valore numerico dato dal centrale, quello che occupa la posizione (n+1)/2
- PARI: la mediana è stimata utilizzando i due valori centrali che occupano le posizioni n/2 e (n/ 2)+1. Con poche osservazioni, come mediana viene assunta la media aritmetica di queste due osservazioni intermedie. Con molte osservazioni raggruppate in classi si ricorre talvolta alle proporzioni.

ESEMPIO: calcolare la mediana nella serie di 6 dati: 10,1; 10,8; 13,1; 13,9; 14,2; 14,5. Essendo il numero di osservazione pari (6) i due valori centrali sono 13,1 e 13,9. La mediana è quindi individuata dalla loro media aritmetica e quindi uguale a 13,5.

Per meglio comprendere le differenze tra media aritmetica e mediana, con la stessa serie di 6 dati (10,1; 10,8; 13,1; 13,9; 14,2; 14,5) in cui:

- La media è 12,85
- La mediana 13,5

La rappresentazione grafica evidenzia come la media sia il baricentro della distribuzione e la mediana sia collocata tra i valori più addensati.



Il calcolo della mediana in una distribuzione di frequenza è leggermente più complesso. Se per esempio la variabile è discreta si deve utilizzare la distribuzione cumulata delle frequenze e trovare con i criteri sopra indicati il valore mediano, il quale indicherà la classe a cui corrisponde la mediana.
Nel caso in cui la distribuzione di frequenza sia sitata fatta su una variabile in classi è necessario ricorrere a specifiche formule per il calcolo della Me.

X (giorni di incubazione)	*Y (Frequenze)*	*Frequenze cumulate*
1	1	1
2	3	4
3	3	7
4	7	14
5	11	25
6	8	33
7	4	37
8	2	39
9	1	40
Totale	*40*	

B) MODA
La moda (detta più raramente anche dato prevalente) è il valore più frequente di una distribuzione. Essa non è influenzata dalla presenza di nessun valore estremo, tuttavia viene utilizzata solamente a scopi descrittivi, perché è meno stabile e meno soggettiva delle altre misure di tendenza centrale. Può infatti differire nella stessa serie di dati, quando si formano classi di distribuzione con ampiezza differente. Per individuare la moda entro una classe di frequenza, non conoscendo come i dati sono distribuiti, si ricorre all'ipotesi della uniforme ripartizione. Oltre alle distribuzioni di frequenza che hanno una sola moda e che si chiamano distribuzioni unimodali, si trovano distribuzioni di frequenza che presentano due o più mode e sono denominate distribuzioni bimodali o plurimodali.
La moda corrisponde alla DETERMINAZIONE DELLA VARIABILE A CUI A' ASSOCIATA LA MAGGIORE FREQUENZA.

X= giorni di incubazione della malattia ABC
n=40
$X = (X_1, X_2, \dots, X_{40}) = (5,6,\dots\dots,5)$

X (giorni di incubazione)	Y (Frequenze)
1	1
2	3
3	3
4	7
5	11
6	8
7	4
8	2
9	1
Totale	40

La moda della variabile X è 5 perché a questa determinazione è associata la frequenza più alta (11).

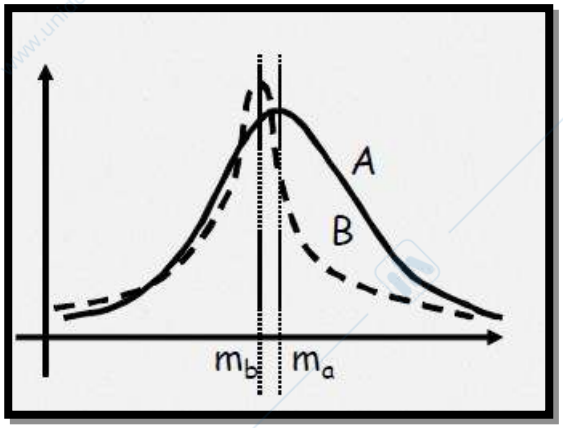
Più problematica e meno precisa è la determinazione del valore modale nel caso in cui la variabile è organizzata in classi la moda non corrisponde più ad una singola determinazione, ma ad una classe di valori della variabile.

Classi (X)	Frequenze (Y)
100-199	857
200-299	4311
300-399	4127
400-499	2109
500-599	1864
Totale	13268

Il valore modale è contenuto nella classe 200-299, ma non se ne conosce il valore preciso (occorrerebbe applicare una specifica formula).
La moda è un efficace indicatore di tendenza centrale solo quando si verifica un'elevata concentrazione di frequenze nei pressi della classe modale (non nel caso della precedente tabella).

2) LE MISURE DI DISPERSIONE O VARIABILITA'

La dispersione o variabilità è la seconda importante caratteristica di una distribuzione di dati. Essa definisce la forma più o meno raccolta della distribuzione intorno al valore centrale.



A) IL CAMPO DI VARIAZIONE

La prima misura ad esser stata storicamente utilizzata per descrivere la dispersione o variabilità dei dati è il campo o intervallo di variazione, definito come la differenza tra il valore massimo e quello minimo.

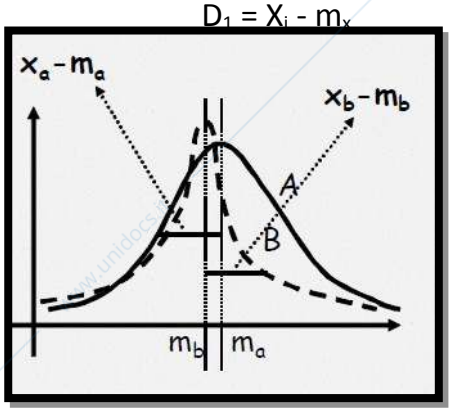
Nell'esempio della variabile X (giorno di incubazione della malattia ABC) il minimo è 1 mentre il massimo è 9 e quindi il CAMPO DI VARIAZIONE è 8.

Tale indicatore, dipendendo solo dal primo e dall'ultimo termine della successione di variazioni, non tiene conte dei termini intermedi e dunque della loro concentrazione attorno al valore medio.

IL CONCETTO DI BASE NEL CALCOLO DEGLI INDICI DI VARIABILITA'

Tutti gli indici di variabilità sono calcolati sulla distanza tra le singole osservazioni ed il rispettivo valore medio della variabile.

$$D_1 = X_i - m_x$$



Nota bene: la distanza non può essere una misura negativa. Questo concetto trova uno specifico riscontro all'interno delle formule.

B) LO SCOSTAMENTO SEMPLICE MEDIO (S_m)

La formula per lo scostamento semplice medio è la somma delle distanze prese in valore assoluto tra le osservazioni e la media aritmetica:

$$S_m = \frac{\sum_{i=1}^{n} |x_i - m|}{n}$$

Oppure nel caso delle distribuzioni di frequenza:

$$S_m = \frac{\sum_{i=1}^{n} |x_i - m| \cdot y_i}{\sum_{i=1}^{n} y_i}$$

C) DEVIAZIONE STANDARD (S)

La formula è molto simile a quello dello S_m, ma al posto della norma troviamo un elevamento alla seconda potenza e l'estrazione della radice quadrata per ridimensionare l'effetto dell'elevamento a potenza.

$$S = \sqrt[2]{\frac{\sum_{i=1}^{n} (x_i - m)^2}{n}}$$

Oppure nel caso delle distribuzioni di frequenza:

$$S = \sqrt[2]{\frac{\sum_{i=1}^{n} (x_i - m)^2 \cdot y_i}{\sum_{i=1}^{n} y_i}}$$

D) LA VARIANZA (S^2)

Il quadrato di S genera un ulteriore indicatore di variabilità che prende il nome di varianza.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n}$$

Oppure nel caso delle distribuzioni di frequenza:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2 \cdot y_i}{\sum_{i=1}^{n} y_i}$$

ESEMPIO DI CALCOLO DELLE MISURE PER UNA SERIE DI DATI
X= punteggio di 15 studenti al test di ammissione di infermieristica

STUDENTI	PUNTEGGIO (X)	X-m	(X-m)^2	|X-m|
1	27	-11,1	122,5	11,1
2	35	-3,1	9,4	3,1
3	38	-0,1	0,0	0,1
4	19	-19,1	363,5	19,1
5	45	6,9	48,1	6,9
6	41	2,9	8,6	2,9
7	40	1,9	3,7	1,9
8	36	-2,1	4,3	2,1
9	51	12,9	167,3	12,9
10	56	17,9	321,6	17,9
11	39	0,9	0,9	0,9
12	40	1,9	3,7	1,9
13	25	-13,1	170,7	13,1
14	27	-11,1	122,5	11,1
15	52	13,9	194,1	13,9
TOTALE			1540,9	118,9

Per prima cosa, prima di calcolare dalla 3 colonna in poi, dobbiamo calcolare la media.
N= 15
Media = 38,067
Una volta calcolata la media calcolo la 3,4 e 5 colonna, dove vengono calcolate le differenze dal valore medio (semplici, al quadrato e in valore assoluto). Successivamente sommo tutti i valori di ogni colonna tra di loro per ottenere il numeratore che mi servirà poi per il calcolo degli indici di variabilità.
Scostamento semplice medio = 7,929
Deviazione standard = 10,136
Varianza = 102,729

E) IL COEFFICIENTE DI VARIAZIONE (CV)

Si ricorre a questo indice quando è necessario confrontare la variabilità di due grandezze che differiscono non solo per ordine di grandezza ma anche l'unità di misura delle rispettive metriche. È un indice di variabilità relativa poiché ottenuto rapportando un indice assoluto (la deviazione standard) con la consistenza media del fenomeno. Il risultato è un numero adimensionale che permette paragoni tra distribuzioni con caratteristiche differenti.

CV =

ESEMPIO: in un ambulatorio di diabetologia viene rilevato il peso di due gruppi di soggetti: il primo di numerosità n_1=5 composto da ragazzi con età inferiore a 10 anni ed il secondo di numerosità n_2=8 composto da anziani con età superiore a 65 anni. I dati sono riportati nella seguente tabella:

SOGGETTO	PESO IN KG GRUPPO 1 (ragazzi)	PESO IN KG GRUPPO 2 (anziani)
1	35	75
2	27	58
3	32	63
4	38	80
5	25	58
6		68
7		69
8		72

È possibile affermare che le due distribuzioni di valori abbiano la stessa variabilità? È possibile dire che i valori di peso nei due gruppi si addensano in modo simile attorno al valore medio?

È difficile poterlo affermare, i due gruppi sono misurati su scale di misura non omogenee, e noi sappiamo quanto il livello di grandezza del valore influisca sul valore della variabilità.

Se calcoliamo il valore degli indici di tendenza centrale e di quelli di variabilità il risultato è il seguente:

MEDIA GRUPPO 1: 31,4

DEVIAZIONE STANDARD GRUPPO 1: 5,413

MEDIA GRUPPO 2: 67,9

DEVIAZIONE STANDARD GRUPPO 2: 10,134

Nel gruppo degli anziani sembra esserci una variabilità doppia rispetto a quella riscontrata con i ragazzi.

Se invece calcoliamo il CV:

CV gruppo 1: 0,172

CV gruppo 2: 0,149

Si nota che la variabilità degli anziani sia addirittura inferiore a quella dei ragazzi.

INDICATORI DELLE CARATTERISTICHE DISTRIBUTIVE

Insieme agli indici di tendenza centrale e agli indicatori di variabilità, è possibile utilizzare alcuni indicatori per completare la conoscenza delle caratteristiche distributive di una variabile. Si tratta di:

A) PERCENTILI
B) INDICE DI ASIMMETRIA
C) INDICE DI CURTOSI

Mentre nel primo caso si può parlare di un indicatore di posizione, nel secondo e terzo caso gli indicatori forniscono informazioni sulla morfologia delle distribuzioni.

A) PERCENTILI

Hanno la funzione di dividere la distribuzione in parti ugualmente numerose, così come la mediana divideva in due sezioni esattamente uguali la distribuzioni. Si classificano in:

- QUARTILI (Q): dividono la distribuzione dei casi (posta in un ordinamento non decrescente) in 4 parti ugualmente numerose
- DECILI (D): divide in 10 parti ugualmente numerose
- CENTILI (C): divisi in 100 parti ugualmente numerose

I quartili quindi si dividono in 3 parti:
Q1: quel valore della distribuzione al di sotto del quale si trovano ¼ dei termini considerati e al di sopra del quale si trovano i rimanenti ¾.
Q2: quel valore della distribuzione al di sotto del quale si trova ½ dei termini considerati e al di sopra del quale si trovano i rimanenti 1/2
Q3: quel valore della distribuzione al di sotto del quale si trovano ¾ dei termini considerati e al di sopra del quale si trovano i rimanenti ¼.

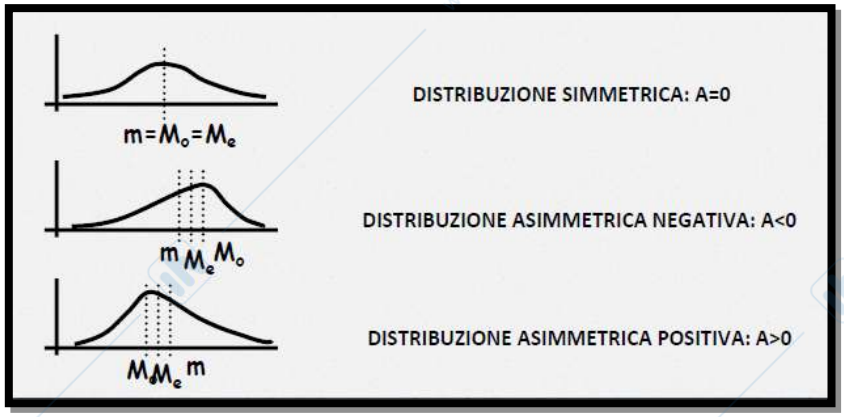
Un'importante caratteristica è che il 2 quartile, il quinto decile ed il cinquantesimo centile coincidono esattamente con la mediana della distribuzione.
Q2=D5=C50=Me

B) INDICE DI ASIMMETRIA

Una qualsiasi curva si dice simmetrica se innalzato verticalmente un asse in corrispondenza del valore medio, la distribuzione risulta divisa in due parti specularmente identiche. In una curva PERFETTAMENTE SIMMETRICA $M_0 = M_e = m$.
Invece la curva ASIMMETRICA POSITIVA se $M_0 < M_e < m$.
Invece è ASIMMETRICA NEGATIVA se $M_0 > M_e > m$.

Per calcolare l'indice di asimmetria si utilizzano i quartili. La formula è:

$$A = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$



C) INDICE DI CURTOSI

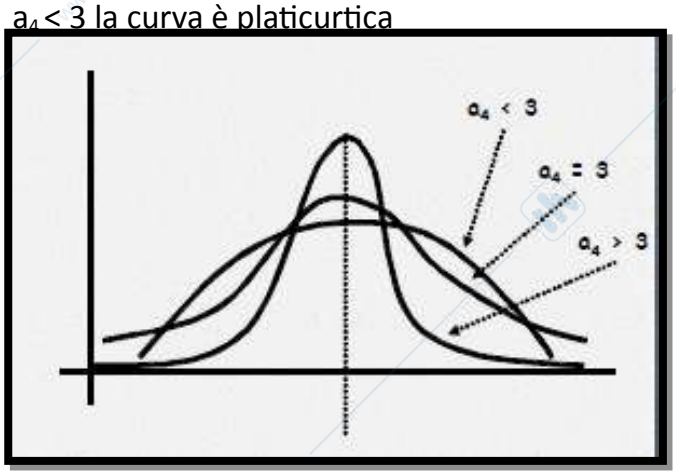
La curtosi descrive le caratteristiche morfologiche in termini di lunghezza delle code della distribuzione e della larghezza della parte sommitale. La formula è:

$$a_4 = \frac{\overline{m}_4}{S^4}$$

Dove : $\overline{m}_4 = \frac{\sum_{i=1}^{n}(x_i - m)^4}{n}$ m_4 è il quarto momento centrale ed S^4 è l'elevamento alla quarta potenza della deviazione standard.

L'indice a_4 è sensibile all'altezza ed alla lunghezza delle code della distribuzione ed in misura inferiore alla "gibbosità" della curva intorno al suo massimo. Se le code sono lunghe ed il vertice della distribuzione è piuttosto appuntito l'indice di curtosi è elevato, viceversa se le code sono corte e la sommità è piatta l'indice risulta basso.

$a_4 > 3$ la curva è leptocutica
$a_4 = 3$ la curva è mesocurtica (normale)
$a_4 < 3$ la curva è platicurtica



CENNI SULLE DISTRIBUZIONI BIVARIATE

Finora abbiamo parlato di distribuzioni di frequenza facendo riferimento alle determinazioni di una sola variabile. In realtà si possono realizzare distribuzioni di

frequenza facendo riferimento alle determinazioni combinate di due grandezze. Ali strumenti vanno sotto il nome di TAVOLE DI CONTINGENZA.
Si tratta di tavole a doppia entrata dove in ogni riga e colonna si trovano le determinazioni assunte dalle variabili, mentre all'interno della tavola si distribuiscono le frequenze corrispondenti alle varie combinazioni assunte dalle determinazioni delle variabili.
All'interno delle celle della tavola di contingenza si possono trovare i valori di:

- Frequenze assolute
- Percentuali di riga
- Percentuali di colonna
- Percentuali riferite al totale

A seconda del numero di righe e colonne le tavole di contingenza possono essere 2x2, 2x3, 3x3, etc.
La tavola di contingenza è con valori assoluti.
ESEMPIO:

- X = sesso
- Y = sigarette fumate al giorno
- K = peso corporeo in classi

TAVOLA DI CONTINGENZA CON PERCENTUALI DI RIGA:

Tavola di contingenza Sigarette fumate al giorno * Sesso

			Sesso		
			Maschio	Femmina	Totale
Sigarette fumate al giorno	1-10	Conteggio	12	8	20
		% entro Sigarette fumate al giorno	60,0%	40,0%	100,0%
	11-20	Conteggio	5	9	14
		% entro Sigarette fumate al giorno	35,7%	64,3%	100,0%
	Oltre 20	Conteggio	3	3	6
		% entro Sigarette fumate al giorno	50,0%	50,0%	100,0%
Totale		Conteggio	20	20	40
		% entro Sigarette fumate al giorno	50,0%	50,0%	100,0%

Tavola di contingenza Peso corporeo in classi * Sesso

			Sesso		
			Maschio	Femmina	Totale
Peso corporeo in classi	Meno di 60 Kg	Conteggio	6	4	10
		% entro Sesso	30,0%	20,0%	25,0%
	61-80 Kg	Conteggio	9	8	17
		% entro Sesso	45,0%	40,0%	42,5%
	81-100 Kg	Conteggio	4	5	9
		% entro Sesso	20,0%	25,0%	22,5%
	Più di 100 Kg	Conteggio	1	3	4
		% entro Sesso	5,0%	15,0%	10,0%
Totale		Conteggio	20	20	40
		% entro Sesso	100,0%	100,0%	100,0%

PROBABILITA'

Tutta la teoria probabilistica poggia su tre concetti elementari:

- PROVA O ESPERIMENTO

- EVENTO O RISULTATO
- PROBABILITA'

La prova genera l'evento con una certa probabilità.
ESEMPIO: lancio della moneta.

- Prova: lancio della moneta
- Evento: uscita della faccia testa
- Probabilità: la probabilità associata

La probabilità è rappresentata da un numero, il quale esprime il livello di fiducia che vogliamo assegnare alla realizzazione di un risultato, ovvero al verificarsi di un evento nel contesto di una prova.
Sempre nel contesto della moneta, supponiamo effettuare N lanci nelle medesime condizioni e supponiamo che durante queste prove l'evento "testa" si sia presentato un numero n di volte. Sotto queste condizioni è ovvio che:

$$0 \leq n \leq N$$

La quantità di eventi favorevoli (testa) viene anche detta FREQUENZA ASSOLUTA. Se dividiamo la precedente disuguaglianza per il numero complessivo di prove effettuate otteniamo la cosiddetta FREQUENZA RELATIVA dell'evento considerato, cioè:

$$0 \leq \leq 1$$

Nell'impostazione frequentista delle teorie probabilistiche il riprodursi di un evento ripetibile un numero di volte molto elevato, porta la frequenza relativa verso una stabilità che cresce con il numero delle prove, fino ad arrivare ad un valore costante. Questa è una definizione abbastanza approssimativa di probabilità che può anche essere riformulata nel seguente modo:

$$\Pr(E) = \lim_{N\to\infty} \frac{n}{N} \qquad 0 \leq \Pr(E) \leq 1$$

E = evento favorevole che attendiamo dall'esperimento, N il numero complessivo di volte in cui viene ripetuta tale prova, n è il numero di volte in cui si verifica E, infine n/N è la frequenza relativa dell'evento E.

VARIABILI CASUALI: è rappresentata da un insieme di coppie di valori del tipo:

$$(X_1, p_1); (X_2, p_2); (X_3, p_3); \ldots\ldots; (X_n, p_n);$$

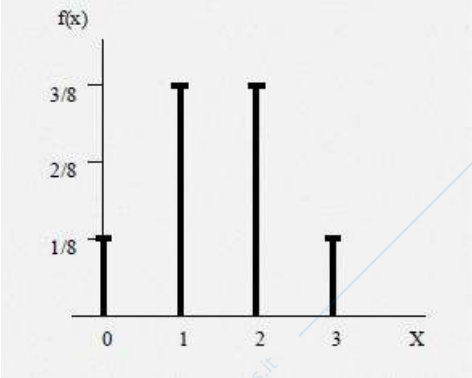
Dove il primo valore indica la determinazione della grandezza X, mentre il secondo la probabilità che viene associata al verificarsi dell'evento.
La variabile X, che descrive l'esperimento, esprime il numero di "croci" che possono apparire in 3 lanci di una moneta. Tale numero non è noto a priori, ma possiamo con certezza affermare che i possibili risultati dell'esperimento variano in uno spazio discreto tra 0 e 3. In questo spazio si trovano le determinazioni X_i della variabile X. Per individuare le probabilità associate ad ogni singolo evento è opportuno costruire lo SPAZIO CAMPIONARIO illustrato nella seguente tabella:

Probabilità	*Eventi*
(1/2)*(1/2)*(1/2)=1/8	C,C,C
(1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8	C,C,T C,T,C T,C,C
(1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8	C,T,T T,C,T T,T,C
(1/2)*(1/2)*(1/2)=1/8	T,T,T

Da essa si ricava lo SPAZIO CAMPIONARIO RIDOTTO, nel quale sono contenuti i valori assunti dalla variabile e le probabilità ad essi associate:

Spazio campionario originario		*Spazio campionario ridotto*	
Probabilità	*Eventi*	*X*	*P*
(1/2)*(1/2)*(1/2)=1/8	C,C,C	3	(1/8)=0,125
(1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8	C,C,T C,T,C T,C,C	2	(3/8)=0,375
(1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8	C,T,T T,C,T T,T,C	1	(3/8)=0,375
(1/2)*(1/2)*(1/2)=1/8	T,T,T	0	(1/8)=0,125
		Totale	1

Le determinazioni di una variabile casuale discreta come quella dell'esempio precedente generano una FUNZIONE DI MASSA, che graficamente è rappresentata da un diagramma a bastoni, in cui la somma delle asticelle che vengono innalzate in corrispondenza delle determinazioni della variabile deve essere pari ad 1.
Parliamo di funzione, al posto del simbolo p_i possiamo sostituire quello di $f(X_i)$, in quanto valore associato alla variabile X.
Nell'esempio precedente:



$$\sum_{i=1}^{4} f(x_i) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = \frac{8}{8} = 1$$

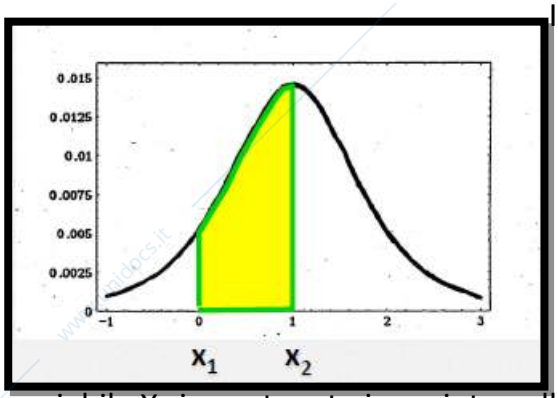
DISTRIBUZIONE DI VARIABILI CASUALI CONTINUE

Se ampliamo il concetto alle variabili casuali continue allora dovremmo definire la forma distributiva di queste variabili come FUNZIONE DI DENSITA'.
Immaginiamo che, a determinati intervalli di una variabile casuale continua, venga associata una certa probabilità di realizzazione dei valori in essi contenuti.

Siccome ogni rettangolo contiene un certo numero di frequenze n_i, la probabilità che una determinazione della variabile X sia contenuta nell'intervallo (X_1, X_2) è data da:

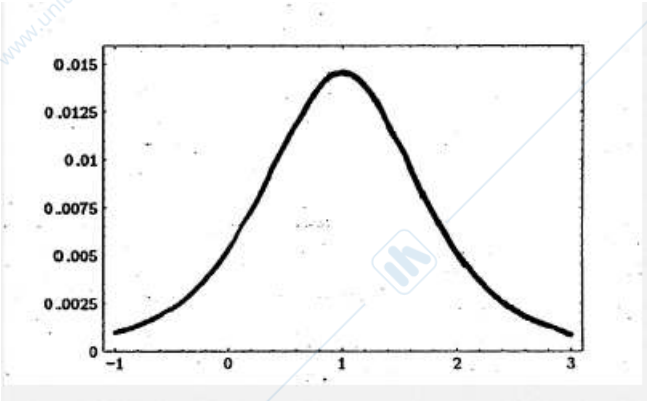
$$\Pr(x_1 < X < x_2) = \frac{n_i}{N}$$

Se decidiamo di stringere l'intervallo (X_1, X_2) fino a dimensioni molto piccole, allora vedremmo ridurre l'ampiezza dei rettangoli dell'istogramma fino a divenire un continuo che descrive una vera e propria funzione di densità, in cui L'AREA SOTTOSTANTE LA CURVA E' PARI ALL'UNITA'.



In questa funzione la probabilità che un valore della variabile X sia contenuto in un intervallo (X_1, X_2) è data dall'area racchiusa dalla curva in quell'intorno, ovvero dall'integrale della forma funzionale calcolato nell'intervallo X_1, X_2.

La classica rappresentazione di una variabile casuale X è quella in cui la distribuzione assume una curva dal caratteristico andamento campanulare. Nella teoria statistica questa classe di funzioni prende il nome di DISTRIBUZIONE NORMALE (GAUSSIANA):



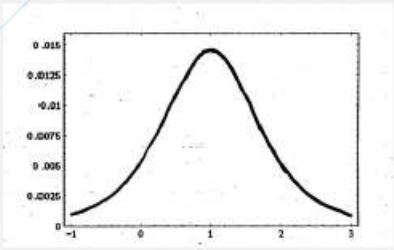
In sintesi si può scrivere:

$$X \sim N(\mu, \sigma)$$

X è la stessa variabile nel campione e nella popolazione, ma con distribuzioni diverse:

Campione	Popolazione

Classe X	Frequenza assoluta n
0	3
1	3
2	7
3	12
4	7
5	5
6	4
7	3
8	0
9	1
Totale	*45*



Esistono tanti diversi tipi di distribuzioni di variabili casuali continue, ognuna con una specifica forma della funzione di densità:

- Normale
- Normale standardizzata
- T-student
- F di Snedecor
- Chi-quadro

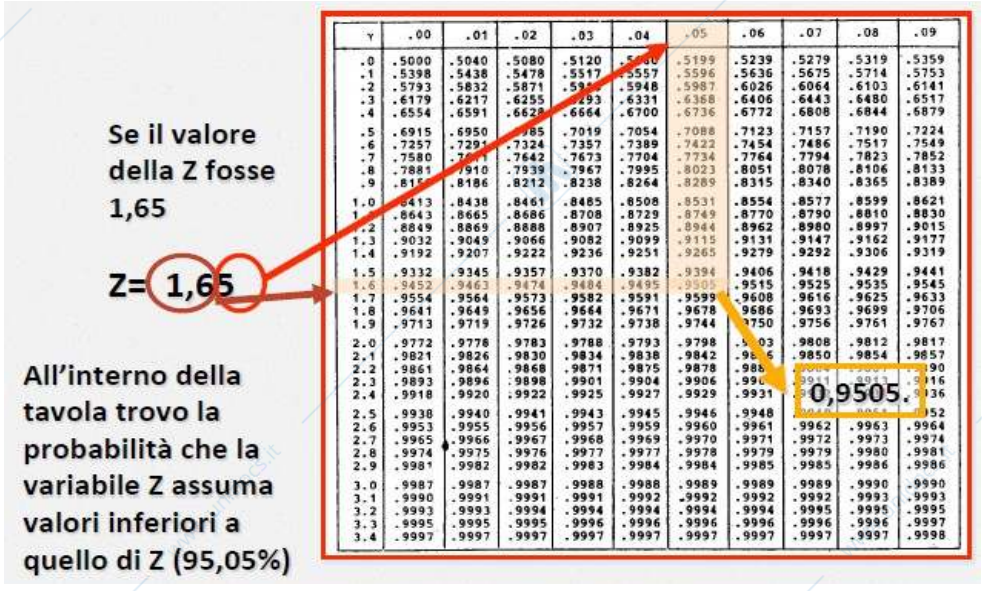
Una di queste vale la pena analizzarla con più attenzione: LA NORMALE STANDARDIZZATA.

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

La tavola della normale standardizzata

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Se il valore della Z fosse 1,65, dividendo 1,65 in 1,6 (nelle righe) e 0,05 (nelle colonne) e vado ad incrociare i due all'interno della tabella trovo la probabilità che la variabile Z assuma valori inferiori a quello di Z. essa è del 95,05%.



ALCUNE RIFLESSIONI PRIMA DI ANDARE AVANTI: abbiamo visto come sia necessario studiare un campione per trarre conclusioni sull'intera popolazione. Abbiamo visto, inoltre, come per la selezione del campione sono necessarie tecniche probabilistiche. Infine abbiamo visto come nella popolazione esistano distribuzioni di variabili governate da assunti probabilistici. A questo punto mi chiedo:
Ma tutto questo come si riconduce al problema della stima, ovvero al processo di generalizzazione dei parametri campionari ai valori della popolazione (INFERENZA STATISTICA)?
Supponiamo che in una popolazione P di numerosità N sia stata misurata la variabile X e che

$$X \sim N(\mu, \sigma)$$

Supponiamo inoltre che dalla popolazione P si estragga un campione C1 di numerosità n e su questo siano calcolati m (media campionaria) ed S (deviazione standard).
Cosa succede a tali parametri se dalla stessa popolazione potessi estrarre un secondo campione C2 della stessa numerosità del primo e poi un terzo C3 e così via?
Consideriamo la solita popolazione di N= 10.000 individui. Come abbiamo detto, da questa è possibile estrarre numerosissimi campioni ognuno di numerosità n, ma con media campionaria e deviazione standard differenti. Supponiamo di estrarne solo 8 di numerosità n=1.000, in cui i valori di media e deviazione standard calcolati per una ipotetica VARIABILE CASUALE X sono i seguenti:

	Campioni								*Popolazione*
	C1	C2	C3	C4	C5	C6	C7	C8	
Media	9,2	11	10	8,7	11	10	9,5	9,8	10
Deviazione Stan.	4,3	4,5	2,7	5,3	4,9	2,9	3,1	2,7	3,8

Prendendo in considerazione la media: queste medie campionarie possono essere considerate e DETERMINAZIONI DI UNA VARIABILE e questi sono 8 dei possibili valori che questa può assumere (gli altri sono dati dal resto dei campioni che non sono stati estratti ma che sarebbe possibile estrarre dalla popolazione). Questa nuova variabile indica la DISTRIBUZIONE DELLE MEDIE CAMPIONARIE la cui simbologia è:

$$\overline{X}$$

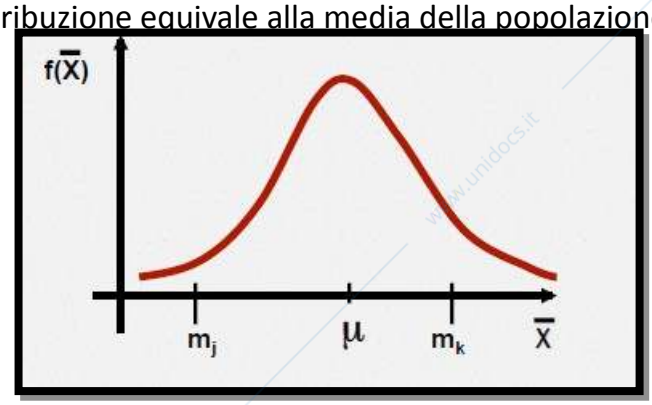
Se una variabile X si distribuisce $X \sim N(\mu, \sigma)$,allora la variabile $\overline{X}$ si distribuisce $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$.

Nell'esempio la distribuzione delle medie campionarie ha una media molto vicina alla media della popolazione, ma i campioni sono solo 8!

	Campioni								
	C1	C2	C3	C4	C5	C6	C7	C8	Popolazione
Media	9,2	11	10	8,7	11	10	9,5	9,8	10
Deviazione Stan.	4,3	4,5	2,7	5,3	4,9	2,9	3,1	2,7	3,8

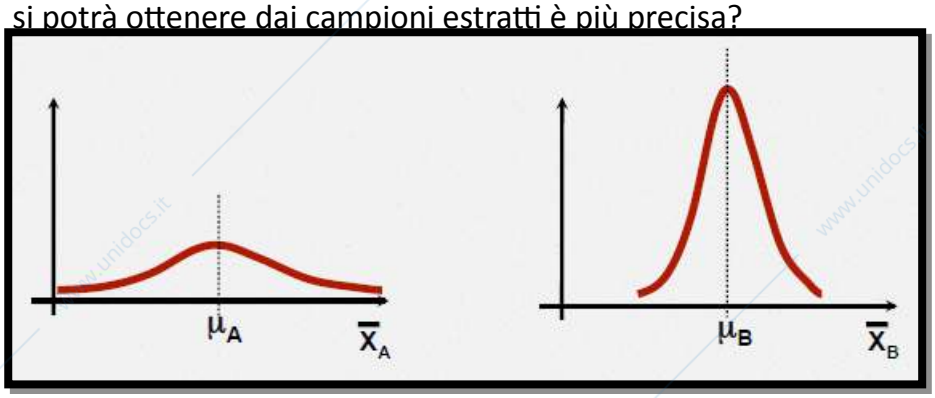
Media delle medie dei campioni=9,95

La variabile $\overline{X}$ ha dunque la stessa distribuzione normale della variabile originaria X, sull'asse orizzontale si trovano tutte le possibili medie dei campioni che è possibile estrarre dalla popolazione, sull'asse verticale si trovano le probabilità che questi hanno di verificarsi. La media della distribuzione equivale alla media della popolazione μ.

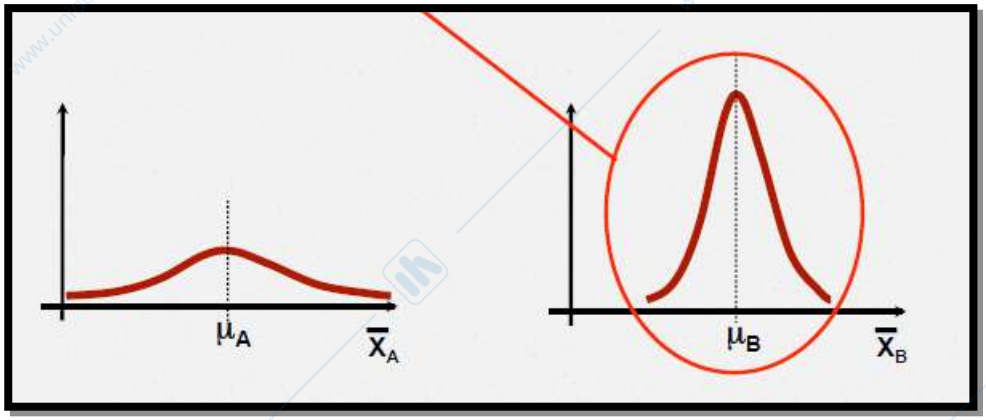


IL PROBLEMA DELLA STIMA

Se per le variabili X_A e X_B, le cui distribuzioni delle medie campionarie sono illustrate sotto, si estraggono due campioni di uguale numerosità n_A e n_B, in quale dei due casi la stima che si potrà ottenere dai campioni estratti è più precisa?



La distribuzione delle medie campionarie della variabile X_B fornirà delle stime più precise, in quanto c'è una maggiore probabilità di ottenere medie campionarie più vicine alla media della popolazione.



IL CAMPIONAMENTO

Se la POPOLAZIONE E' PICCOLA, come il numero di allevamenti zootecnici di una provincia, il numero di ammalati di una malattia rara o residenti su un territorio circoscritto, può essere conveniente raccogliere dati su tutta la popolazione.

Tuttavia spesso la POPOLAZIONE E' GRANDE, anche se finita. Quindi a causa del tempo e delle risorse disponibili, che in tutte le ricerche sono limitate, è possibile misurarne solamente una parte.

IL PROBLEMA DIVENTA: come costruire questo campione, in modo che esso fornisca informazioni corrette su tutta la popolazione?

NUMEROSITA' CAMPIONARIA

Prendiamo in considerazione ora un universo formato da N elementi e supponiamo di voler estrarre un campione di n elementi.

Quanto deve essere grande n? per rispondere a questa domanda, occorre considerare 4 fattori:

- A: dimensione dell'universo
- B: l'errore di campionamento
- C: il livello fiduciario
- D: il grado di eterogeneità dell'universo

A) DIMENSIONE DEL'UNIVERSO

All'aumentare di N tende a crescere anche n, e viceversa. L'aumento (o la diminuzione) dell'ampiezza del campione è però meno che proporzionale rispetto all'aumento (o alla diminuzione) della dimensione dell'universo. Del resto, nell'ambito di un universo di dimensioni limitate, un campione altrettanto ridotto è fortemente esposto a rischi di distorsione. In altri termini, se il campione è formato da pochi elementi, sono piuttosto elevati i rischi che questi elementi siano scarsamente rappresentativi dell'universo. Al crescere di N, la dimensione del campione può aumentare in misura meno che proporzionale, in quanto si riduce il rischio che gran parte del campione sia formata da ELEMENTI NON RAPPRESENTATIVI, dal momento che stiamo lavorando su grandi numeri.

La quantificazione di N non è sempre facile. Si pensi al caso di un'azienda di servizi, che può avere utenti saltuari o molto saltuari, e utenti sistematici che ricorrono al servizio più volte al meno o più volte alla settimana.

L'universo potrebbe corrispondere al numero di utenti dell'ultimo mese, ma questo numero potrebbe escludere parte degli utilizzatori più saltuari e, al contrario, potrebbe

comprendere una certa quota (ignota) di sovrapposizioni, ossia di utenti che nell'ultimo mese si sono rivolti molte vole al servizio in questione.

B) ERRORE DI CAMPIONAMENTO

Il valore rilevato con l'indagine campionaria non corrisponderà perfettamente al valore dell'universo, ma sarà caratterizzato da un errore per eccesso o per difetto (indichiamo questo errore con e).
Il valore dell'universo (che non conosciamo, ma che vogliamo stimare) sarà quindi compreso entro un certo intervallo rispetto al valore del campione. Questo intervallo è noto con il nome di INTERVALLO DI CONFIDENZA.
Di norma, viene deciso in fase di impostazione della ricerca quale è l'errore di campionamento che si è disposti ad accettare.
All'aumentare dell'errore accettato, si riduce l'ampiezza del campione e viceversa: del resto, per avere risultati più precisi, è intuitivo che occorra lavorare su un campione più grande.

C) LIVELLO FIDUCIARIO

Quando proiettiamo i risultati di un indagine campionaria sull'intero universo (incognito) sia effettivamente compreso entro l'intervallo di confidenza rispetto al valore del campione. Abbiamo infatti una certa probabilità che il valore reale dell'universo sia esterno a questo intervallo.
Il GRADO DI CERTEZZA CHE IL VALORE DELL'UNIVERSO SIA INTERNO ALL'INTERVALLO DI CONFIDENZA viene chiamato LIVELLO FIDUCIARIO o LIVELLO DI CONFIDENZA. Solitamente, anche il livello fiduciario che si è disponibili ad accettare viene deciso in fase di impostazione della ricerca. Maggiore è il grado di certezza che si vuole, più grande dovrà essere il campione su cui lavorare. Nella maggior parte delle indagini, il livello fiduciario accettato è pari al 95%.
Ciò significa che, se invece di estrarre un solo campione con una certa dimensione, ne venissero estratti 100, tutti con quella dimensione, 95 porterebbero a una stima corretta (con il valore incognito dell'universo compreso nell'intervallo di confidenza), mentre 5 porterebbero a una stima non corretta. In realtà, le ricerche vengono svolte su un solo campione: si ha il 95% di probabilità che questo campioni porti a stime corrette, 5% di probabilità che ciò non si verifichi.
Il grado di fiducia sulla correttezza della nostra stima è legato non solo alla dimensione del campione, ma anche all'intervallo di confidenza. A parità di ampiezza del campione, se vogliamo ridurre l'errore, dobbiamo accontentarci di un più basso grado di certezza. Al contrario se intendiamo accettare un errore più grande, possiamo contare su un livello fiduciario maggiore.

D) ETEROGENEITA' DELL'UNIVERSO

All'aumentare del grado di eterogeneità dell'universo, crescono i rischi connessi al campionamento. Più l'universo è eterogeneo, infatti, maggiori sono le probabilità di fornire

(tramite l'indagine su campione) una stima distante dal corrispondente valore dell'universo, in quanto sono maggiori anche le probabilità di lavorare su un campione meno rappresentativo. Per questo motivo, di fronte a una forte eterogeneità dell'universo occorre cautelarsi dai maggiori rischi di distorsione, utilizzando un campione più ampio. Si ha la situazione di massima omogeneità quando tutti i componenti dell'universo si comportano nello stesso modo, in riferimento al fenomeno oggetto di indagine. Al contrario, si ha la condizione di massima eterogeneità quando gli elementi dell'universo si presentano tra loro molto differenziati.
Il problema che a questo punto si incontra è dato dal fatto che, prima di fare un'indagine, non si conoscono le caratteristiche dell'universo e la sua eterogeneità. Se si disponesse di queste informazioni, si potrebbe evitare di effettuare l'indagine. Occorre allora ricorrere a un espediente, stimando il grado di eterogeneità con le informazioni che a diverso titolo possiamo avere a disposizione: per esempio, utilizzando i risultati di altre indagini condotte sull'argomento, oppure attraverso un piccolo sondaggio preliminare. Se queste strade non sono percorribili, si assume l'ipotesi più pessimistica, si considera cioè che il fenomeno abbia la massima eterogeneità possibile. Questa scelta prudenziale consente di evitare un campione sottodimensionato, ma può costare molto, in termini di numero di in termini di numero di interviste da effettuare.

POPOLAZIONE E CAMPIONE

La parola popolazione identifica un gruppo di persone appartenenti ad una certa collettività, non necessariamente fatta di persone. Raramente in uno studio statistico è possibile esaminare ogni singolo soggetto dell'intera popolazione a causa del grande numero i componenti. L'esame di un campione, ossia di un numero ridotto di osservazioni, invece dell'intera popolazione, consente di superare questo problema. Un campione non è altro che un insieme di elementi tratti da una popolazione (o universo).

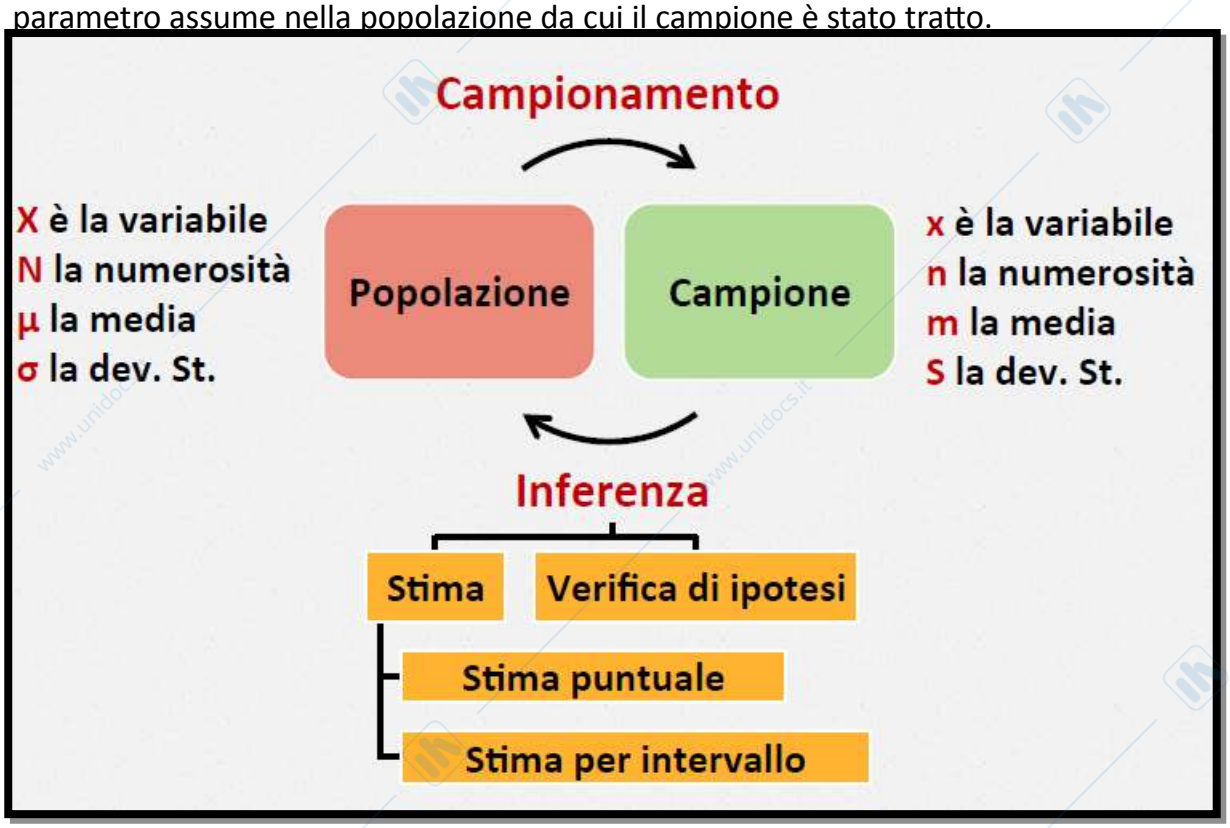
Il campione è sempre un sottoinsieme della popolazione $C \subset U$.
Se misuro una variabile X nel campione e nella popolazione i simboli che uso sono i seguenti:

Popolazione		Campione	
Numerosità dei casi	N	Numerosità dei casi	n
Media	μ	Media	m
Deviazione S.	σ	Deviazione S.	S

Lo studio dei fenomeni si attua spesso attraverso dei campioni, cioè una parte di tutta la popolazione dallo studio del campione, attraverso tecniche statistiche, si cerca di avere informazioni sull'intera popolazione, ovvero devo studiare in modo opportuno una popolazione attraverso un gruppo ristretto di oggetti, affinché le stime provenienti da quest'ultimo possano essere stese, con un determinato margine di errore, alla popolazione di riferimento.

INFERENZA STATISTICA

È un processo di generalizzazione dei parametri campionari ai corrispondenti parametri della popolazione ovvero STIMARE attraverso un dato campionario il valore che tale parametro assume nella popolazione da cui il campione è stato tratto.



L'indirizzo di un campione consiste nell'esaminare un numero n di soggetti selezionati tra gli N individui appartenenti alla popolazione. Gli n elementi estratti rappresentano quello che comunemente chiamiamo CAMPIONE CASUALE DELLA POPOLAZIONE oggetto di studio. È IMPORTANTE CHE IL CAMPIONE SIA PROBABILISTICO (ovvero scelto con tecniche probabilistiche) altrimenti le tecniche di inferenza statistica non possono essere utilizzate.

IL PROBLEMA DELLA STIMA

Il processo di generalizzazione si regge su una batteria di strumenti che sostanzialmente possiamo così riassumere:

A) STIMATORI PUNTIALI (medie, varianze, ….)
B) STIMATORI PER INTERVALLO (intervalli di confidenza)
C) TESTS DI SIGNIFICATIVITA'

Attraverso questi strumenti è possibile stabilire una corrispondenza tra i parametri del campione e quelli che caratterizzano la popolazione.

B) STIMATORI PER INTERVALLO O INTERVALLI DI CONFIDENZA

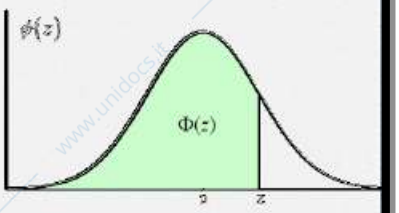
Per ridurre il grado di incertezza della stima è preferibile prendere in considerazione, piuttosto che un singolo valore della media, un intervallo di valori e vedere se, dato un certo livello di probabilità, il valore vero del parametro ignoto μ. Il livello di probabilità è

dato da 1-α, dove α rappresenta un numero compreso tra 0 e 1 e che generalmente è abbastanza piccolo. Di solito vengono adottati valori di α pari a 0,10 0,05 e 0,01 a cui corrispondono valori di 1-α rispettivamente di 0,90 0,95 e 0,99. Lo scopo degli intervalli di confidenza è quello di individuare, sulla base dei valori campionari di media e deviazione standard, un intervallo all'interno del quale è contenuto il valore vero della popolazione μ, dato un certo livello di probabilità 1-α.
Come abbiamo visto, se

$$X \sim N(\mu, \sigma) \Rightarrow \bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

utilizzando la normale standardizzata

$$P[u_{\alpha L} < Z < u_{\alpha H}] = 1 - \alpha$$



$$P\left[u_{\alpha L} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < u_{\alpha H}\right] = 1 - \alpha$$

La determinazione campionaria della media è contenuta in un intorno simmetrico della media della popolazione, dove $U_{\alpha L}$ e $U_{\alpha H}$ sono i due valori della distribuzione che isolano, il primo alla propria sinistra ed il secondo alla sua destra, due aree di probabilità pari ad α/2.
Dopo pochi passaggi aritmetici è facile arrivare alla FORMULA degli INTERVALLI DI CONFIDENZA:

$$P\left[\overline{X} - u_{\alpha}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{\alpha}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

In cui è facile verificare che il valore medio della popolazione contenuto nell'intervallo costruito sulla base delle determinazioni campionarie, infatti all'interno della formula è possibile sostituire i valori dati dagli stimatori:

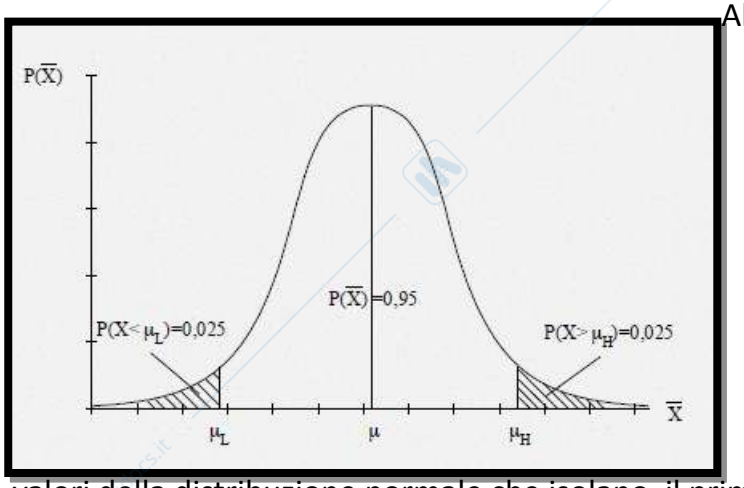
$$m = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

ESEMPIO: disponete di una variabile X che si distribuisce come una normale standardizzata. Avete estratto un campione di numerosità n=16, le caratteristiche di questo campione sono media pari a 5 e varianza pari a 4 ed infine abbiamo scelto di utilizzare un livello $\alpha = 0{,}05$. Come costruire l'intervallo di confidenza?

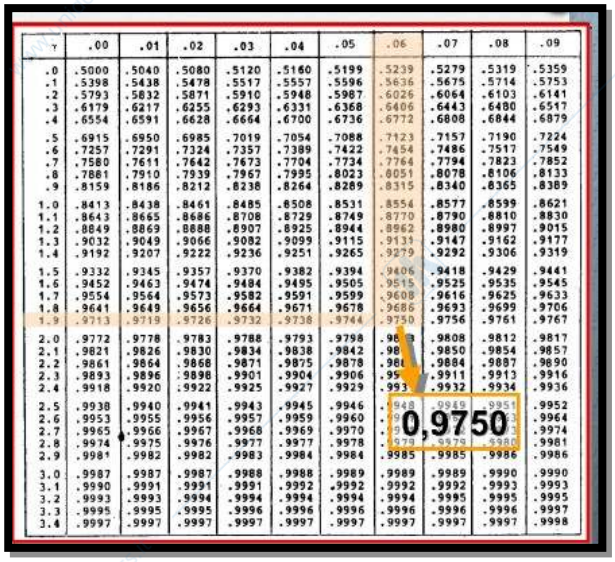
$$P\left[\bar{X} - u_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

Siccome disponiamo di tutti i dati, l'unico problema da affrontare è rispondere al quesito: quanto vale U_α e come calcolarlo?



Abbiamo già detto che questi sono i due valori della distribuzione normale che isolano, il primo alla propria sinistra ed il secondo alla sua destra, due aree di probabilità pari ad $\alpha/2$. Occorrono le tavole della normale standardizzata.

Se $\alpha=0{,}05$, il valore da cercare nelle tavole sarà $\alpha/2= 0{,}025$, ovvero il valore complementare dell'unità. Il valore di U_α sarà $U_\alpha= 1{,}96$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.98	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.98	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.98	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.99	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.99	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.99	.99	.99	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.99	.99	.99	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

0,9750

Se dunque: _ n=16

_ m = 5

_S = 4

_ $U_\alpha = 1,96$

L'intervallo di confidenza sarà

$$\Pr\left[5 - 1,96\frac{2}{\sqrt{16}} < \mu < 5 + 1,96\frac{2}{\sqrt{16}}\right] = 0,95$$

Da cui ...

$$\Pr[4,02 < \mu < 5,98] = 0,95$$

Il valore vero della media nella popolazione è compreso tra 4,02 e 5,98 con una probabilità del 95%.

La generalizzazione dei parametri campionari ai parametri dell'universo è pur sempre una stima e come tale affetta da errore: α, p, sign., diversi modi di esprimere l'incertezza della stima.
Abbiamo visto che per generalizzare le stime dal campione alla popolazione ho bisogno di campioni probabilistici, ma l'importanza della probabilità nell'inferenza sta solo nella possibilità di generare campioni rappresentativi?
Supponiamo di aver estratto un campione da una popolazione di N individui, supponiamo che la variabile X che intendiamo studiare si distribuisce normalmente nella popolazione con media μ e deviazione standard σ, supponiamo di aver calcolato il valore medio m delle osservazioni campionarie.

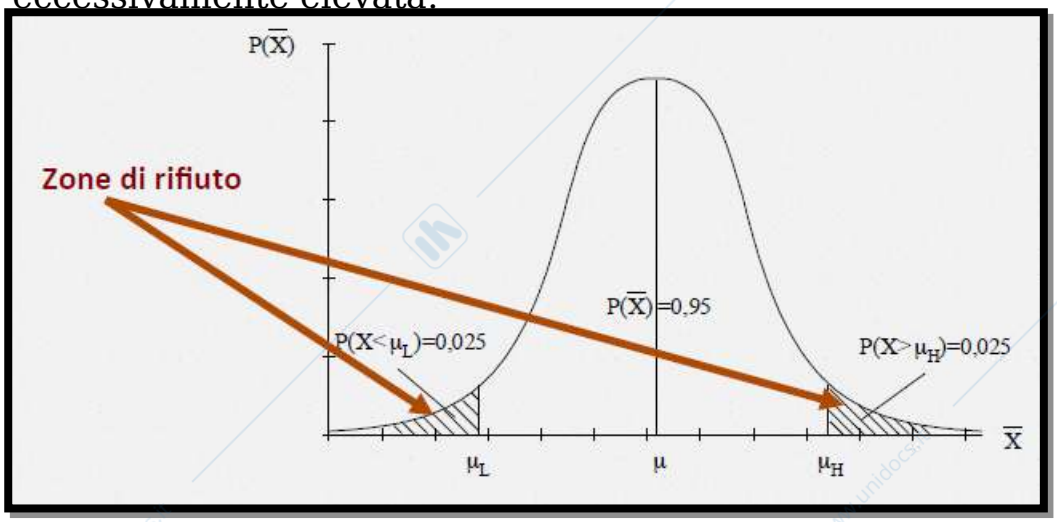
Potremmo sottoporre a test il valore medio campionario per sapere se questo è uguale a quello della popolazione formulando:

- IPOTESI NULLA: $H_o : m=\mu$ rappresenta l'ipotesi che viene sottoposta a est

- IPOTESI ALTERANTIVA: $H_A : m \neq \mu$ rappresenta il comportamento del test qualora l'ipotesi nulla debba essere rigettata.

IPOTESI NULLA CONTRO IPOTESI ALTERNATIVA: la media osservata m non è altro che una determinazione della variabile casuale media campionaria $\overline{X}$, la quale di distribuisce nuovamente come una normale
N(m, s/). Nell'ipotesi H_0, come si può osservare, si confronta il valore campionario con quello della popolazione. Per sua natura (essendo un parametro relativo ad un campione), può assumere anche valori molto diversi (distanti) da m.
È nostro compito individuare dei limiti oltre i quali la differenza tra m e μ è eccessivamente elevata.

C) I TEST E LE ZONE DI RIFIUTO
È nostro compito individuare dei limiti oltre i quali la differenza tra m e μ è eccessivamente elevata:

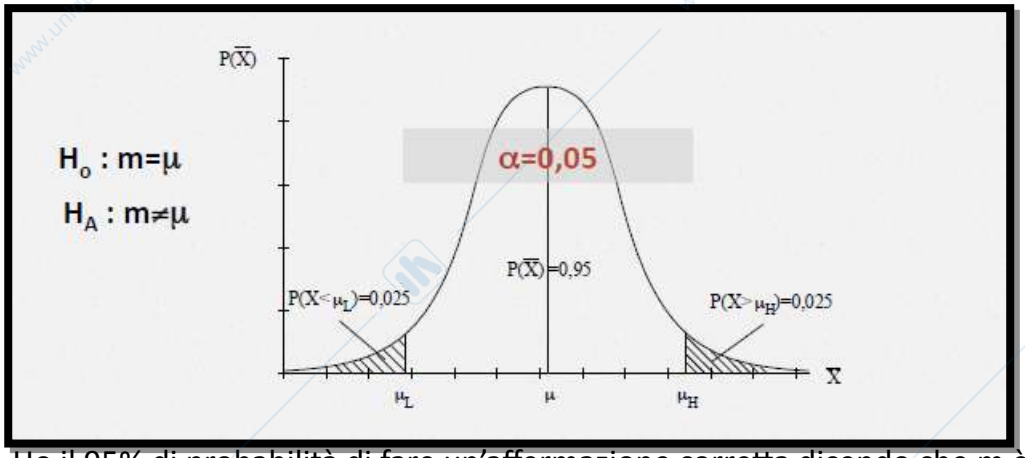


LA SIGNIFICATIVITA' DI UN TEST
In ogni test statistico per verificare l'ipotesi nulla è necessario stabilire un valore di probabilità (α) che è uguale o minore ad un livello prescelto. Il livello di probabilità = 1- α (dove α rappresenta un numero compreso tra 0 e 1 e che generalmente è abbastanza piccolo). Di solito vengono adottati valori di α pari a 0,10 0,05 e 0,01 a cui corrispondono valori di 1-α rispettivamente di 0,90 0,95 e 0,99. La scelta di un livello di significatività poco severo ci può aiutare a rifiutare l'ipotesi nulla, così come un livello di significatività troppo rigoroso può portare ad una sua accettazione arbitraria.
Nella pratica cosa significano questi termini?

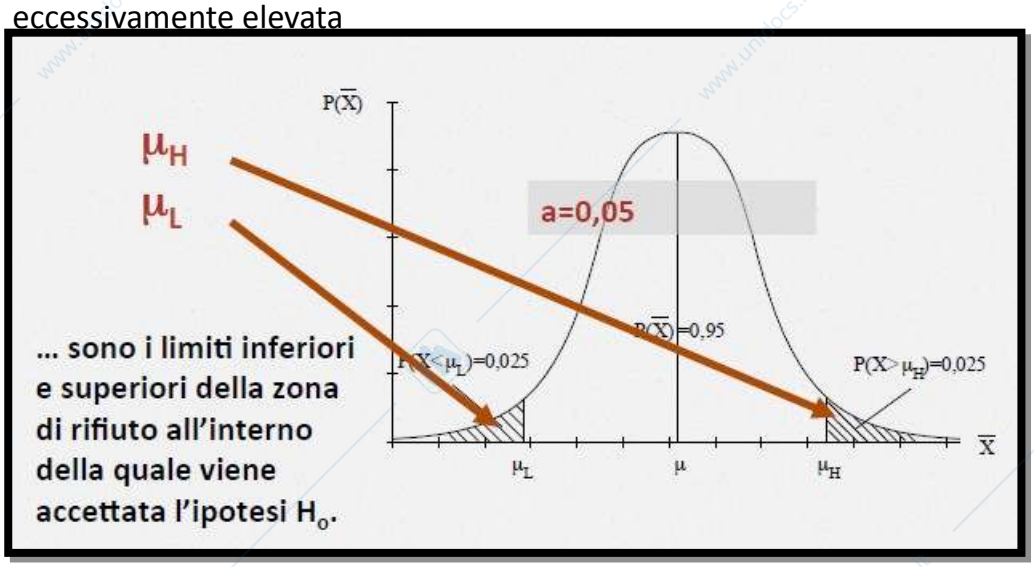
- 1 – α: livello di probabilità che abbiamo stabilito affinché la nostra affermazione possa essere ritenuta corretta.
- α: rischio di sbagliare che sono disposto a correre nella stima.

Se α = 0,05 e 1- α = 0,95 ho il 95% di probabilità di fare un'affermazione correttamente il 5% rappresenta il rischio che sono disposto a correre nella stima.

Nel caso precedente:



Ho il 95% di probabilità di fare un'affermazione corretta dicendo che m è uguale a μ, mentre sono disposto a tollerare un errore del 5% ad affermare quanto sopra.
Se è nostro compito individuare dei limiti oltre i quali la differenza tra m e μ è eccessivamente elevata

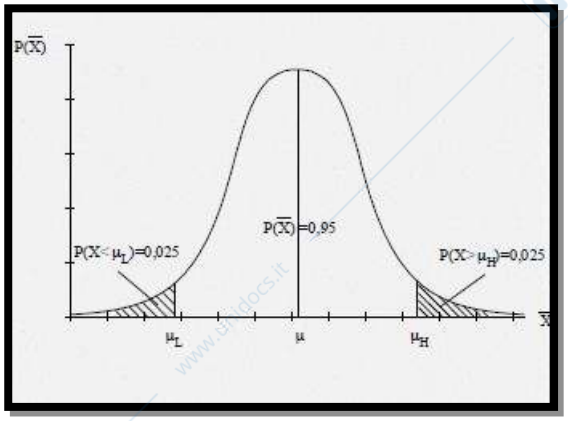


Se α = 0,05 dobbiamo scegliere due valori μ_L e μ_H in modo che

$$P(\overline{X} < \mu_L) + P(\overline{X} > \mu_H) = 0,05$$

Dalla simmetria della distribuzione avremmo che

$$P(\overline{X} < \mu_L) = 0,025$$
$$P(\overline{X} > \mu_H) = 0,025$$



Per effettuare il test è necessario ricorrere alle distribuzione della variabile Z, ovvero la normale standardizzata, dove Z_L e Z_H sono equivalenti ai limiti μ_L e μ_H traslati però nella variabile normale standardizzata.

L'ipotesi nulla è rigettata se $\frac{(m-\mu)\sqrt{n}}{\sigma} < Z_L = -1{,}96$ $\frac{(m-\mu)\sqrt{n}}{\sigma} > Z_H = 1{,}96$ In tal caso si accetta l'ipotesi alternativa.

L'ipotesi nulla non può essere rigettata se $Z_L < \frac{(m-\mu)\sqrt{n}}{\sigma} < Z_L$

Tutte le volte che il valore di α è inferiore a 0,05 la differenza messa sotto esame con il test è statisticamente significativa. Posso rigettare l'ipotesi nulla a favore di quella alternativa.

LA VERIFICA DI IPOTESI CON DATI NOMINALI: IL TEST CHI QUADRATO

Lo strumento fondamentale per l'applicazione di questo test è la cosiddetta TAVOLA DI CONTINGENZA (nel nostro caso 2X2).

Nelle tavole di contingenza le frequenze si distribuiscono in una tabella a doppia entrata a cui sono state associate una coppia di variabili. Si applica a coppie di variabili ed utilizza le tavole di contingenza sulla quale l'applicazione del test ha lo scopo di verificare l'esistenza e la forza di relazioni tra le variabili. Lo scopo del test è quello di determinare il grado di significatività dei dati con l'ipotesi nulla:

- $H_0 : p_1=p_2$ ovvero la probabilità di appartenere alla categoria 1 (successi) è la stessa nei due gruppi.

Da questa tabella con le frequenze

	Successi	*Insuccessi*	Totale
Gruppo 1	O 11	O 12	R1
Gruppo 2	O 21	O 22	R2
Totale	C1	C2	*N*

osservate si può passare a quella delle frequenze attese attraverso la formula ottenendo quindi

$$e_{ij} = \frac{(R_i \cdot C_j)}{N}$$

	Successi	*Insuccessi*	Totale
Gruppo 1 Gruppo 2	e 11 e 21	e 12 e 22	R1 R2
Totale	C1	C2	*N*

Se la tavola delle frequenze osservate sarà molto diversa da quella delle frequenze attese il grado di significatività dei dati rispetto ad H_0 sarà basso, presentando valori di significatività (p) pari a 0,05 o inferiori. La formula per il calcolo del test è:

$$\chi^2 = \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Per calcolare il grado di significatività del test per un certo gruppo di osservazioni, si deve fare ricorso ad apposite tavole nelle quali è possibile individuare il livello fi significatività del valore del test ottenuto.

$$\Pr\left(\chi^2 \geq \chi_0^2\right)$$

Per applicare in modo efficace questo test occorre tenere presente una regola empirica che assicura una accurata approssimazione, ovvero bisogna fare attenzione a verificare che tutte le frequenze attese e_{ij} siano maggiori di 5. Una regola più compiacente si accontenta che questi valori attesi siano maggiori di 2. In tutti questi casi sarebbe più opportuno ricorrere al test di Fisher. QUESTO TEST SI APPLICA SOLO SU VARIABILI QUALITATIVE.

ESEMPIO: se questa è la tavola delle frequenze osservate calcolare il test chi-quadro.

Esito *Farmaco*	*Guargione*	*Non guarigione*	Totale
Farmaco tipo A Farmaco tipo B	7 2	3 7	10 9
Totale	9	10	19

Le frequenze attese vengono calcolate come segue:

$$e_{1,1} = \frac{R_1 * C_1}{N} = \frac{10*9}{19} = 4,74 \qquad e_{1,2} = \frac{R_1 * C_2}{N} = \frac{10*10}{19} = 5,26$$

$$e_{2,1} = \frac{R_2 * C_1}{N} = \frac{9*9}{19} = 4,26 \qquad e_{2,2} = \frac{R_2 * C_2}{N} = \frac{9*10}{19} = 4,74$$

La tavola delle frequenze attese sarà:

Esito / *Farmaco*	*Guargione*	*Non guarigione*	Totale
Farmaco tipo A	4,74	5,26	10
Farmaco tipo B	4,26	4,74	9
Totale	9	10	19

Il confronto fra le due tavole ci porta a calcolare il test:

$$\chi^2 = \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$\chi^2 = \frac{(7-4,74)^2}{4,74} + \frac{(3-5,26)^2}{5,26} + \frac{(2-4,26)^2}{4,26} + \frac{(7-4,74)^2}{4,74} = 4,325$$

Visto che il valore del test è $X^2 = 4,325$, è necessario andare a trovare sulle apposite tavole della distribuzione del chi-quadro il valore critico sapendo che i g.d.l. si trovano secondo la seguente formula:

g.d.l. = (righe-1)*(colonne-1) = 1

Il valore critico sarà il seguente: $X^2 = 3,841$ con p=0,05

Rigetto l'ipotesi nulla con una probabilità del 95%.

LA TAVOLA DEI VALORI CRITICI (X^2)

Gradi di libertà	Area della coda di destra									
	0.995	0.99	0.975	0.95	0.90	0.1	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672

LA FORMULA DEL CHI-QUADRO CON LA CORREZIONE DI YATES

Una possibile fonte di errore nel test del chi-quadro è rappresentata dal fatto che mentre la distribuzione del chi-quadro è continua, i dati contenuti nelle tavole di contingenza sono sempre discreti, il che porta ad ottenere dei valori del test troppo elevati. Per questo si utilizza una formula leggermente modificata ma comunque il procedimento per il calcolo del test rimane esattamente lo stesso.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{[(o_{ij} - e_{ij}) - 0{,}5]^2}{e_{ij}}$$

Il test del chi-quadro può essere effettuato anche su tavole di contingenza con più di 2 righe e colonne. Il procedimento per il calcolo del test rimane esattamente lo stesso e le formule da utilizzare sono le stesse ma aumentano i gradi di libertà.

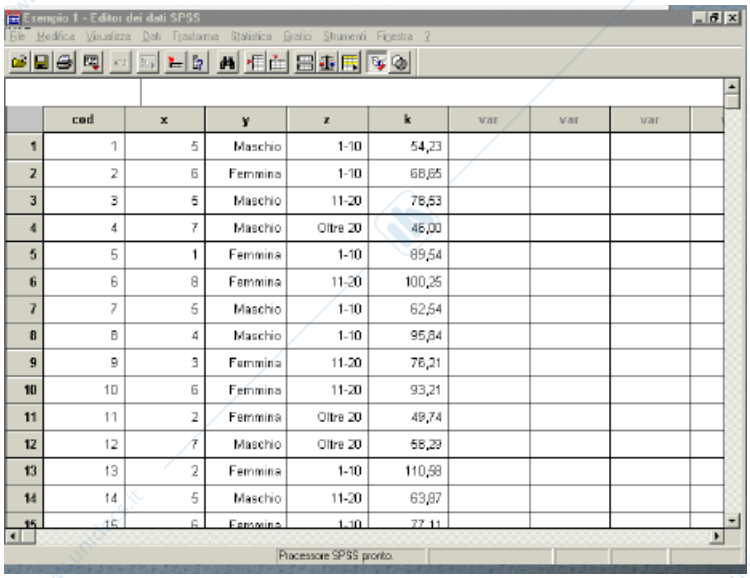
COME VERIFICARE IL LEGAME TRA DUE VARIABILI

Per analizzare i legami che esistono tra le grandezze prese in esame occorrono strumenti diversi dai test illustrati nella precedente lezione, il cui utilizzo, ancora una volta varia in funzione della categoria della variabile sottoposta ad analisi. In questa lezione parleremo di:

A) CORRELAZIONE/ASSOCIAZIONE TRA VARIABILI
B) REGRESSIONE

PARTIAMO DA UN ESEMPIO: abbiamo analizzato su un campione di 40 soggetti il periodo di incubazione di una malattia infettiva, rilevando le seguenti variabili:

- X: giorni di incubazione
- Y: sesso
- K: peso corporeo

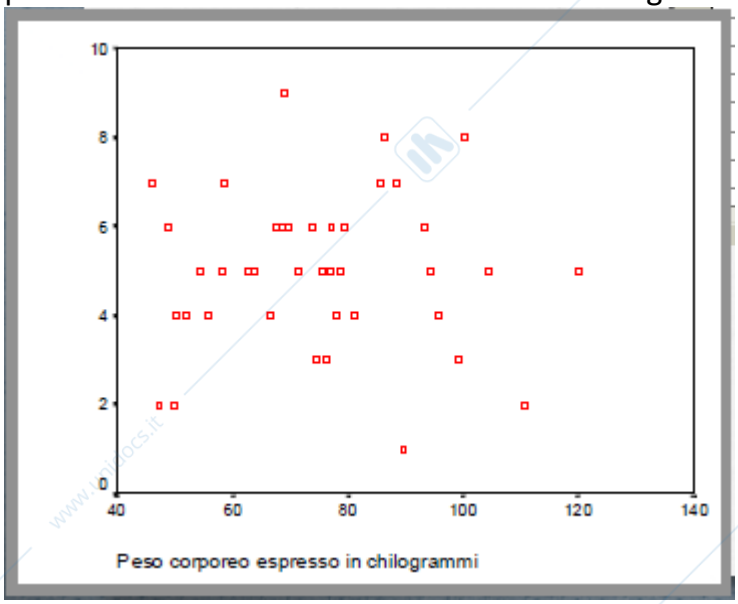

Esempio 1 - Editor dei dati SPSS

File Modifica Visualizza Dati Trasforma Statistica Grafici Strumenti Finestra ?

	cod	x	y	z	k	var	var	var
1	1	5	Maschio	1-10	54,23			
2	2	6	Femmina	1-10	68,65			
3	3	5	Maschio	11-20	78,53			
4	4	7	Maschio	Oltre 20	46,00			
5	5	1	Femmina	1-10	89,54			
6	6	8	Femmina	11-20	100,35			
7	7	5	Maschio	1-10	62,54			
8	8	4	Maschio	1-10	95,84			
9	9	3	Femmina	11-20	76,21			
10	10	6	Femmina	11-20	93,21			
11	11	2	Femmina	Oltre 20	49,74			
12	12	7	Maschio	Oltre 20	58,29			
13	13	2	Femmina	1-10	110,98			
14	14	5	Maschio	11-20	63,87			
15	15	6	Femmina	1-10	77,11			

Processore SPSS pronto.

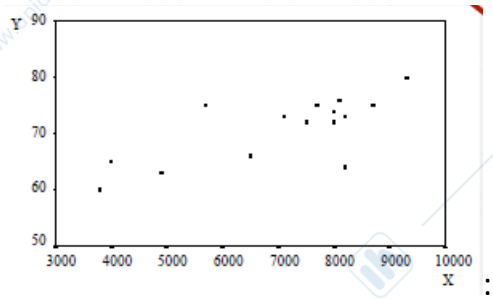
Una delle ipotesi di ricerca potrebbe essere: esiste una relazione trai i giorni di incubazione ed il peso dei soggetti? Se rappresentassimo i valori della X e della nel grafico sotto, si potrebbe affermare che tra le due esiste un legame?



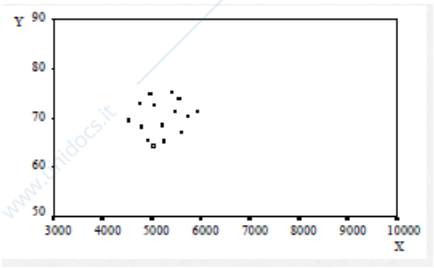
A) IL CONCETTO DI CORRELAZIONE

Correlare significa stabilire una reciproca relazione tra due ordini di grandezze, ovvero verificare se le variazioni che si verificano in una variabile possono o meno determinare delle corrispondenti variazioni nella seconda. In questa parte parleremo delle misure di correlazione, proponendo alcune metodologie che, applicate sui dati osservati, indicheranno se esista una relazione tra le variabili, quanto essa è forte ed in quale modo si esplica questa relazione.

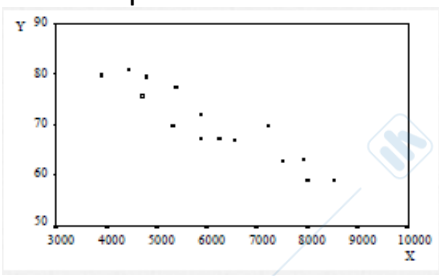
Date due grandezze X e Y si possono verificare 3 casi:



_ : CORRELAZIONE POSITIVA: ad un aumento della variabile X corrisponde un aumento anche della variabile Y, vi è un legame diretto tra le due grandezze. r≈ 1



_ : ASSENZA DI CORRELAZIONE: le osservazioni non rappresentano una direzione ben precisa, ma piuttosto tendono a disporsi come una "nuvola di punti" attorno ad un valore centrale. r≈0



_ : CORRELAZIONE NEGATIVA: le osservazioni assumono una tendenza decrescente e lasciano ipotizzare che tra la variabile X e Y vi sia un rapporto inverso. r≈-1

Il valore della correlazione lineare si ottiene con il coefficiente r di Pearson. Si calcola solo su VARIABILI QUANTITATIVE e serve a valutare l'INTENSITA' DEL LEGAME TRA DUE VARIABILI.
Il coefficiente r ha un campo di variazione che varia tra -1 e 1. Se r=0 non vie è legame tra X e Y, ovvero le variazioni della Y non possono essere spiegate dalla X. Quanto più il suo valore si differenzia da 0 e si avvicina agli estremi del campo di variazione, tanto più è evidente la presenza di un legame tra le suddette grandezze. In particolare, se r tende verso 1 (per esempio r=0,85, r=0,92 o superiori) allora vi è una relazione diretta tra le variabili, cioè al crescere dell'una si verifica un aumento anche nell'altra. Se invece tende verso -1, allora la relazione è di tipo inverso, cioè al crescere di X la Y decresce.

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

Dove:

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)\cdot(y_i - m_y)$$ (covarianza)

S_x deviazione standard della variabile X

S_y deviazione standard della variabile Y

TORNANDO ALL'ESEMPIO: il valore del coefficiente lineare di correlazione r di Pearson calcolato tra i giorni di incubazioni ed il peso corporeo ha fato il seguente risultato: 0,008.

Correlazioni

		Giorni di incubazione della malattia ABC
Peso corporeo espresso in chilogrammi	Correlazione di Pearson	,008
	N	40

Il valore della r è praticamente 0: tra le due variabili non esiste legame (ovvero le variazioni del peso non determinano le variazioni del numero di giorni di incubazione della malattia).

B) LA REGRESSIONE LINEARE

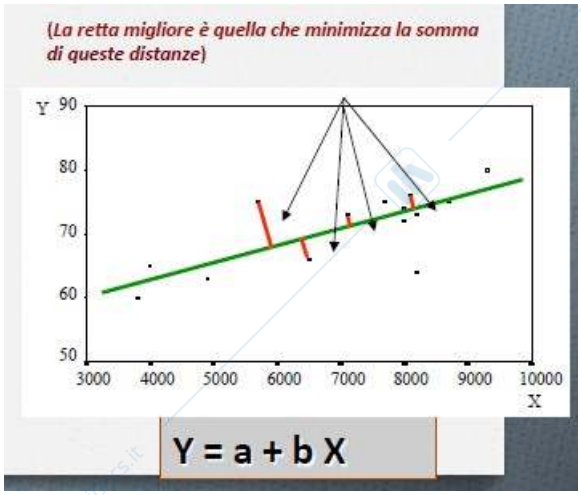
Se il coefficiente r indica l'intensità del legame tra due variabili, il calcolo della regressione restituisce la forma di questo legame. Potremmo, infatti, scegliere di descrivere l'andamento delle osservazioni mediante una funzione matematica che, interpolando i dati, sia in grado di rappresentarne la tendenza e di conservarne le informazioni principali. Tale metodo si chiama REGRESSIONE LINEARE.

Il metodo della regressione lineare consiste appunto nell'individuare una retta che sia capace di rappresentare la distribuzione dei punti nel piano a 2 dimensioni. Come è facile immaginare, se i punti corrispondenti alle osservazioni si dispongono vicino alla retta, allora il metodo scelto sarà in grado di descrivere con efficacia il legame tra le variabili. In teoria esistono un numero infinito di rette che potrebbero essere tracciate per interpolare le osservazioni, ma in pratica vi è un solo modello matematico che ottimizza la rappresentazione dei dati.

Il termine lineare sta proprio ad indicare che il legame che vogliamo studiare è di tipo lineare ed è rappresentato da una retta Y=a+bX. Esso è il modello matematico attraverso cui vogliamo rappresentare la forma della relazione tra le variabili. Questo modello utilizza la X come variabile indipendente e la Y come variabile dipendente. I parametri a e b (che rappresentano rispettivamente l'intercetta con l'asse delle Y ed il coefficiente angolare, ovvero la pendenza della retta), dovranno essere stimati sulla base delle osservazioni raccolte per le due variabili incluse nel modello. I valori osservati delle variabili X e Y vengono utilizzati per trovare la retta che fornisce la migliore rappresentazione dei dati sul piano cartesiano.

Il metodo di stima dei parametri va sotto il nome di MINIMI QUADRATI. Tra le rette che si possono tracciare per interpolare i dati la più adatta a rappresentarne l'andamento sarà

certamente quella che minimizza la somma delle distanze tra i valori stimati mediante il modello.



La stima dei due parametri si ottiene nel seguente modo:

$$\begin{cases} a = m_y - b \cdot m_x = m_y - \dfrac{S_{xy}}{S_x^2} \cdot m_x \\ b = \dfrac{S_{xy}}{S_x^2} \end{cases}$$

Dove:

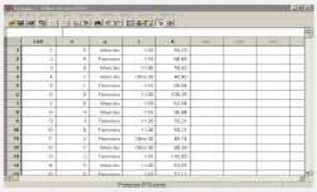
m_y = media della variabile Y	S_{xy} = covarianza
m_x = media della variabile X	S_x^2 = varianza della variabile X

TORNANDO ALL'ESEMPIO:

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati		
		B	Errore std.	Beta	t	Sig.
1	(Costante)	4,946	1,191		4,154	,000
	Peso corporeo espresso in chilogrammi	7,236E-04	,015	,008	,047	,963

a. Variabile dipendente: Giorni di incubazione della malattia ABC



Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	,008[a]	,000	-,026	1,78

a. Stimatori: (Costante), Peso corporeo espresso in chilogrammi

Come può essere scritto il modello ?

Y = a + b X

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	4,946	1,191		4,154	,000
	Peso corporeo espresso in chilogrammi	7,236E-04	,015	,008	,047	,963

a. Variabile dipendente: Giorni di incubazione della malattia ABC

Y = 4,946 + 0,00007236 X

Il coefficiente angolare del modello (parametro b) è praticamente 0, in termini matematici significa che qualsiasi modifica nei valori della variabile X non influenzano i valori della Y (confermando l'assenza di correlazione).