Statistica Descrittiva

Scopo: descrivere il campione (dati) in modo sintetico ed efficace mediante tabelle, grafici, numeri

Premessa: Le caratteristiche che osserviamo sul campione variano da un'unità di osservazione all'altra → variabili

Le variabili possono essere **discrete** o **continue**

Variabili discrete: assumono un numero finito o un'infinità numerabile di valori

Variabili continue: possono assumere qualsiasi valore

Tabelle e Grafici di Frequenza (1)

Un primo utile sistema per riassumere i dati è la costruzione di tabelle e grafici di frequenza

Esempio nel discreto: lancio di un dado

Il risultato del lancio è una variabile discreta (può assumere uno dei seguenti valori 1 2 3 4 5 6)

50 lanci → ottengo una sequenza di 50 (n) numeri

Sintetizzo i dati costruendo una tabella

Tabelle e Grafici di Frequenza (2)

Esempio nel discreto *(continua)*

risultato del lancio	frequenza (f)	frequenza relativa (f/n)
1	9	0.18
2	12	0.24
3	6	0.12
4	8	0.16
5	10	0.20
6	5	0.10

$\sum f = n$ $\sum (f/n) = 1.00$

prima colonna: possibili risultati del lancio

seconda colonna: numero totale di volte in cui è stato ottenuto quel risultato (**frequenza assoluta**)

terza colonna: frequenza assoluta del risultato divisa per il numero totale (n) di osservazioni (**frequenza relativa**)

Tabelle e Grafici di Frequenza (3)

Esempio nel discreto *(continua)*

risultato del lancio	frequenza (f)	frequenza relativa (f/n)
1	9	0.18
2	12	0.24
3	6	0.12
4	8	0.16
5	10	0.20
6	5	0.10

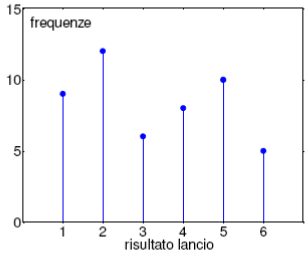
seconda colonna: **distribuzione di frequenze**

terza colonna: **distribuzione di frequenze relative**

Tabelle e Grafici di Frequenza (4)

Esempio nel discreto *(continua)*

La distribuzione di frequenza e la distribuzione di frequenza relativa possono essere rappresentate graficamente mediante **diagrammi a bastoncino**



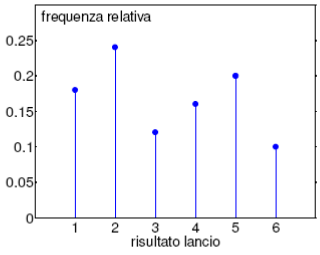


Tabelle e Grafici di Frequenza (5)

Esempio nel continuo

Campione di 200 uomini (n=200) estratto da una certa popolazione. Rileviamo l'altezza in cm.

La variabile osservata è continua. Non ha senso parlare di frequenza del singolo valore poiché non c'è alcuna possibilità di osservare due stature esattamente uguali.

L'intervallo che contiene tutti i valori osservati viene suddiviso in un certo numero di sottointervalli (classi) e si contano quante osservazioni cadono nei diversi sottointervalli.

Tabelle e Grafici di Frequenza (6)

Esempio nel continuo (continua)

Limiti intervallo (cm)	Valore centrale (cm)	frequenza (*f*)	frequenza relativa (*f*/*n*)
141.5-148.5	145	2	0.01
148.5-155.5	152	7	0.035
155.5-162.5	159	22	0.11
162.5-169.5	166	13	0.065
169.5-176.5	173	44	0.22
176.5-183.5	180	36	0.18
183.5-190.5	187	32	0.16
190.5-197.5	194	13	0.065
197.5-204.5	201	21	0.105
204.5-211.5	208	10	0.05

$\sum f = n$ $\quad \sum (f/n) = 1.00$

Tabelle e Grafici di Frequenza (7)

Quante classi (sottointervalli) vi devono essere?

- compromesso ragionevole tra una distribuzione troppo dettagliata ed una troppo sintetica

Le classi vengono in genere scelte in modo che il valore centrale sia un numero intero

In quale classe viene posizionata una osservazione che cade al limite tra due classi?

In genere la si pone nella classe superiore

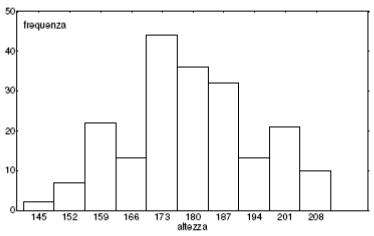
Ad esempio il valore 162.5 viene posto nella classe 162.5-169.5.

Si tratta quindi di sottointervalli del tipo [ )

Tabelle e Grafici di Frequenza (8)

Esempio nel continuo (continua)

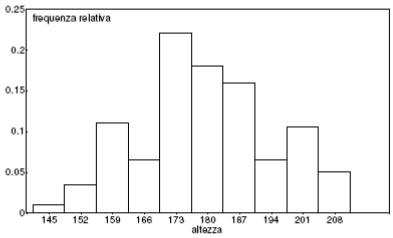
I dati raggruppati in sottointervalli possono essere rappresentati graficamente mediante **istogrammi**



istogramma della distribuzione di frequenze

base = ampiezza della classe

altezza = frequenza assoluta

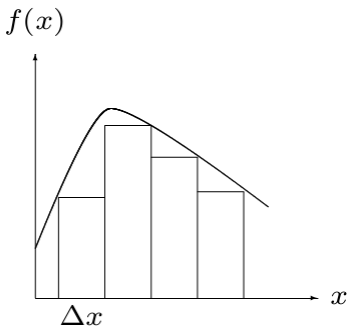


istogramma della distribuzione di frequenze relative

base = ampiezza della classe

altezza = frequenza relativa

Obiettivo: Descrizione di un Istogramma



Indici di descrizione statistica di un campione

INDICI

- di posizione
  - MODA
  - MEDIANA
  - MEDIA
- di dispersione
  - SCARTO QUADRATICO MEDIO
  - VARIANZA
  - RANGE
- di forma
  - ASIMMETRIA (SKEWNESS)
  - CURTOSI (KURTOSIS)

Misure di posizione : La Media Aritmetica (1)

La media aritmetica (m) è la più comune misura di posizione

Le osservazioni (x_1, x_2,...x_n) vengono sommate tra di loro, quindi la somma divisa per n (cioè per il numero di osservazioni):

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

Per calcolare l'altezza media del nostro campione di 200 individui dobbiamo sommare le 200 osservazioni e dividere la somma per 200

dato un insieme di m elementi $\{x_1, x_2, \ldots x_m\}$, e

dato un insieme di m di numeri reali $\{p_1, p_2, \ldots p_m\}$

- Si dice **media aritmetica pesata**

$$\bar{x} = \frac{x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots + x_m \cdot p_m}{p_1 + p_2 + \ldots + p_m}$$

che utilizza un peso p_j o la frequenza di ogni dato x_j per j=1,…,m

Misure di posizione : La Media Aritmetica (2)

Se è nota solo la distribuzione di frequenza per sottointervalli (non le singole osservazioni)

→ si calcola una **media approssimata**

Sia f_1 il numero di osservazioni che cadono nel sottointervallo 1;

tali osservazioni vengono approssimate dal valore centrale del sottointervallo (x_1)

analogamente per tutti gli altri sottointervalli

$$m \cong \frac{1}{n} \cdot \sum_{i=1}^{k} f_i x_i = \sum_{i=1}^{k} x_i \left(\frac{f_i}{n} \right)$$

f_i / n: frequenza relativa del sottointervallo *i*-esimo

k : numero dei sottointervalli

Esempio di media pesata

La media della lunghezza di un gruppo di $\mathbf{f_1}$= 7 neonati ⇨ $\mathbf{m_1}$=48.0 cm
e di altri $\mathbf{f_2}$= 3 neonati ⇨ $\mathbf{m_2}$=49.5 cm.

Per calcolare la media delle lunghezze dell'insieme totale di **10 neonati** pur senza avere la conoscenza dei valori delle lunghezze individuali, si utilizzano le proprietà della media aritmetica :

la somma delle lunghezze dei primi 7 è	48.0 × 7 = 336.0
la somma delle lunghezze dei secondi 3 è	49.5 × 3 = 148.5
la somma delle lunghezze di tutti i 10 è	= 484 .5

La media della lunghezza di tutti i 10 neonati è = 484.5/10 = 48.45

Ovvero $\mathbf{Media = (f_1 \times m_1 + f_2 \times m_2)/(f_1 + f_2)}$
$\mathbf{Media = (7 \times 48.0 + 3 \times 49.5)/(7+3)}$

esempio di media aritmetica

Lunghezza(cm) in un campione di 60 neonati.

51.0	49.4	49.0	52.5	51.5	51.8
46.5	47.8	49.7	44.5	49.8	53.0
48.7	50.0	52.9	50.8	46.2	48.9
54.5	48.2	48.9	51.2	49.5	56.3
46.0	52.2	47.0	50.8	50.0	52.5
51.2	51.1	54.7	52.3	48.2	50.8
55.0	50.2	50.3	47.7	48.5	53.8
50.2	53.4	47.4	50.5	51.7	49.5
44.4	49.2	50.5	49.5	52.9	50.5
54.0	46.5	51.5	50.9	51.6	52.7

la media aritmetica dei primi 6 valori di lunghezza di 6 neonati è:
$\bar{x}$ = (51.0+49.4+49.0+52.5+51.5+51.8)/6 = 305.2/6 = 50.87

la media aritmetica di tutti i 60 valori di lunghezza è:
= (55.9+51.3+53.0+50.5+54.9+53.4+…+53.8)/60 = 3021.8 /60
$\bar{x}$ = 50.363

La media aritmetica di N dati distinti è …

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

MEDIA per dati raggruppati in classi

ALTEZZA(cm) di un campione di 60 neonati.

limiti di classe	x_i	$f(x_j)$	$x_i f(x_j)$
44.25- 45.75	45.0	2	90.0
45.75- 47.25	46.5	5	232.5
47.25- 48.75	48.0	7	336.0
48.75- 50.25	49.5	14	693.0
50.25- 51.75	51.0	16	816.0
51.75- 53.25	52.5	9	472.5
53.25- 54.75	54.0	5	270.0
54.75- 56.25	55.5	1	55.5
56.25- 57.75	57.0	1	57.0
Σ		60	3022.5

Nell'*esempio* del campione di 60 misure di lunghezza dei neonati:

$$\bar{x} = \frac{45.0\times 2 + 46.5\times 5 + \ \ldots\ 57.0\times 1}{2+3+\ \ldots\ 1} = \frac{3022.5}{60} = 50.375$$

La media per dati raggruppati in m classi è ...

$$\bar{x} = \frac{\sum_{j=1}^{m} x_j \cdot f(x_j)}{\sum_{j=1}^{m} f(x_j)}$$

dove **m** è il numero di classi e ,

$\sum_{j=1}^{m} f(x_j) = N$ se $f(x_i)$ indica le frequenze assolute,

oppure $\sum_{j=1}^{m} f(x_j) = 1$ se $f(x_i)$ indica le frequenze relative.

media aritmetica e mediana

Si consideri un campione di valori di VES (*velocità di eritrosedimen-tazione*, mm/ora) misurati in 7 pazienti

{8, 5, 7, 6, 35, 5, 4}

In questo caso, **la media** ($\bar{x}$ = 10 mm/ora) **non è** un valore **tipico** della distribuzione: soltanto un valore su 7 è superiore alla media!

Conviene usare come indice del centro **la mediana**, definita come quel valore che divide a metà la distribuzione, sicché **l'insieme dei valori è per metà minore e per metà maggiore della mediana**.

Per **calcolare la mediana** si dispongono i dati in ordine crescente:
ordine originale: {8, 5, 7, 6, 35, 5, 4}
ordine crescente: {4, 5, 5, 6, 7, 8, 35}

mediana

Se n è **dispari**, la mediana è il valore che occupa la posizione (n+1)/2 nell'insieme ordinato.

Nell'*esempio*, poiché (n+1)/2=4, la mediana è 6 mm/ora, ed è tipica nel senso che si avvicina a buona parte dei valori del campione.

Se n è **pari**, la mediana è la media dei valori che occupano le posizioni (n/2) ed [(n/2)+1] nell'insieme ordinato dei numeri.

Se, nell'***esempio***, si esclude il valore più alto, si ottiene l'insieme ordinato {4, 5, 5, 6, 7, 8},

(n/2)=3 e [(n/2)+1]=4,

e la mediana vale (5+6)/2=5.5.

Frattili di una distribuzione

Una distribuzione può essere descritta per mezzo dei suoi **frattili**.

Si dice frattile (sinonimi: **centile, percentile e quantile**) p-*esimo* di una distribuzione quel valore xp tale che la frequenza relativa cumulata $F(x_p) = p$.

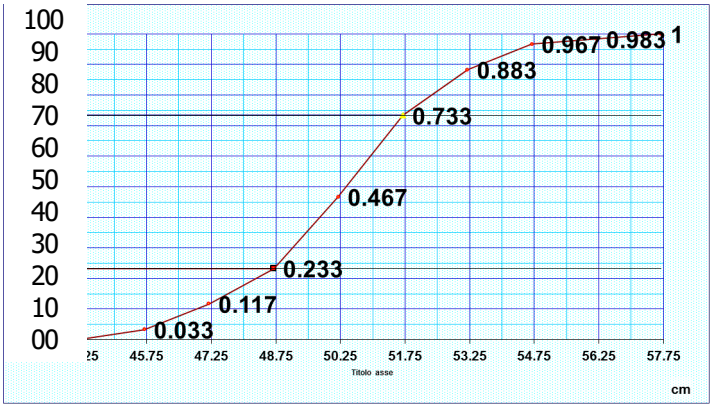
Ad esempio, **il 50° centile** di una distribuzione è il valore che, sull'asse dei numeri reali, ha alla sua sinistra il 50% dei valori della distribuzione, e **coincide con la mediana**.

Il 10° centile è il valore che ha alla sinistra il 10% della distribuzione.

ALTEZZA(cm) di un campione di 60 neonati.

limiti di classe	x_i	$f(x_j)$	$x_i f(x_j)$
44.25- 45.75	45.0	2	90.0
45.75- 47.25	46.5	5	232.5
47.25- 48.75	48.0	7	336.0
48.75- 50.25	49.5	14	693.0
50.25- 51.75	51.0	16	816.0
51.75- 53.25	52.5	9	472.5
53.25- 54.75	54.0	5	270.0
54.75- 56.25	55.5	1	55.5
56.25- 57.75	57.0	1	57.0
Σ		60	3022.5

Nei **grafici cumulati**, i valori riportati sull'asse verticale indicano la **frequenza** delle rilevazioni con **valore pari o minore** ai valori in corrispondenza sull'asse orizzontale



La Moda

Più di rado si incontra una terza misura di posizione, la moda; è il *valore che si verifica più spesso (frequenza assoluta più elevata);* la modalità della variabile in cui si registra il maggior numero di casi.

Quanto sono usualmente lunghi i bimbi alla nascita?

Guardando i dati a nostra disposizione, è subito evidente maggior numero (16) di bimbi è lungo tra i 50.3 cm e i 51.7 cm.

la classe modale è dunque 50.25-51.75.

Se la distribuzione ha più di due valori massimi o se la frequenza più alta riscontrata nell'insieme considerato non supera di molto le altre la moda non è un buon indicatore di tendenza centrale.

La moda

Lunghezza supina (cm) in un campione di 60 neonati.
Valori ottenuti con l'infantometro Harpenden.

Estremi di classe	Valore Centrale	Freq Semplici n	Freq Semplici %	Freq cumulate n	Freq cumulate %
44.3-45.7	45.0	2	0.033333	2	0.033333
45.8-47.2	46.5	5	0.083333	7	0.116667
47.3-48.7	48.0	7	0.116667	14	0.233333
48.8-50.2	49.5	14	0.233333	28	0.466667
50.3-51.7	51.0	16	0.266667	44	0.733333
51.8-53.2	52.5	9	0.15	53	0.883333
53.3-54.7	54.0	5	0.083333	58	0.966667
54.8-56.2	55.5	1	0.016667	59	0.983333
56.3-57.7	57.0	1	0.016667	60	1

Nella classe 50.3-51.7 , piu' vicino alla casse con freq=14

$$50{,}25+\frac{1{,}5\times|16-14|}{|16\text{-}14|+|16\text{-}9|}=50{,}583$$

quale misura di posizione usare?

A quale misura di tendenza centrale ci riferiamo?

- Il proprietario di una ditta afferma "Lo stipendio medio nella nostra ditta è 2.700 euro"
- Il sindacato dei lavoratori dice che "lo stipendio mensile è di 1.700 euro".
- L'agente delle tasse dice che "lo stipendio è stato quasi sempre di 2.200 euro".

Queste risposte diverse sono state ottenute tutte dai dati della seguente tabella.

Stipendio mensile	N° di lavoratori
1.300	2
1.700	22
2.200	19
2.600	3
6.500	2
9.400	1
23.000	1

Media aritmetica= lire 2.700
Mediana = lire 2.200
Moda = lire 1.700

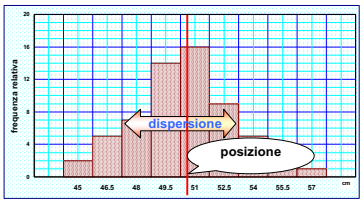
interpretazione delle misure di posizione

- La **media aritmetica** indica che, se il denaro fosse distribuito in modo che ciascuno ricevesse la stessa somma, ciascun dipendente avrebbe avuto 2.700 euro
- La **moda** ci dice che la paga mensile più comune è di 1.700.euro
- La moda si considera spesso come il valore tipico dell'insieme di dati poiché è quello che si presenta più spesso. **Non tiene però conto degli altri valori** e spesso in un insieme di dati vi è **più di un valore** che corrisponde alla definizione di moda.
- La **mediana** indica che circa metà degli addetti percepiscono meno di 2.200.euro, e metà di più.
- La mediana **non è influenzata dai valori estremi** eventualmente presenti ma solo dal fatto che essi siano sotto o sopra il centro dell'insieme dei dati.

Statistica Descrittiva

- Intervallo di variazione
- Devianza
- Varianza
- Deviazione Standard
- Intervallo interquartile

dispersione di una distribuzione

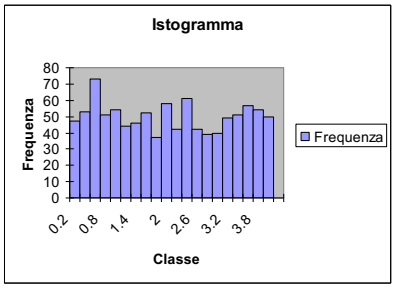


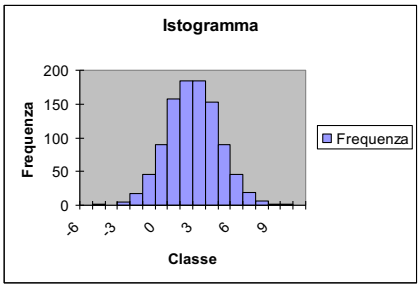
35

Media e varianza:

Media uguale

Deviazione Standard Diversa





Media=2
Varianza=1.33

Media=2
Varianza=4

36

dispersione di una distribuzione

Numero di ore di sonno	frequenza Maschi	Femmine
1	1	3
2	3	6
3	3	7
4	7	8
5	11	5
6	8	3
7	4	1
8	2	1
9	1	1
10	-	-
11	-	1
12	-	1
13	-	1
14	-	1
15	-	1

Diamo un'occhiata alla distribuzione di frequenza delle ORE DI SONNO indotte da un sonnifero, dormite da **40 maschi** e **40 femmine**.

37

dispersione di una distribuzione

- La **misura della variabilità**, permette di descrivere in modo più completo la distribuzione di una variabile.
- Le misure di tendenza centrale: **media**, **mediana e moda** individuano l'elemento "centrale" della distribuzione.
- Diamo, di nuovo, un'occhiata alla distribuzione di frequenza delle **ORE DI SONNO** dei 40 soggetti.
  - ✓ La media è di **5 ore** ma uno sguardo alla tabella mostra che un buon numero di pazienti sono molto diversi tra loro.
  - ✓ Alcuni presentano un periodo di sonno **più breve** ed altri **più lungo della media**.
- La media **non dice** in che misura i dati siano dispersi attorno al valore centrale.

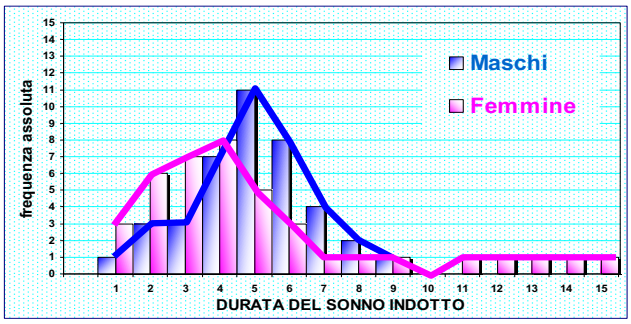
38

dispersione di una distribuzione

Il numero medio di "letture" risulta di 5 ore in entrambe i sessi

Uguale *durata del sonno indotto* ?

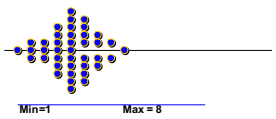
Per facilitare i confronti riportiamo i dati in grafico.

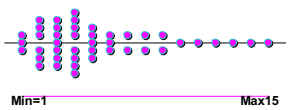


39

L'intervallo di variazione

- Mentre **in media** le femmine presentano un durata del sonno uguale ai maschi, alcune di loro hanno un durata del sonno ancora superiore ai tempi più elevati dei maschi.
- Quindi le medie non sono insufficienti: per completare il quadro occorrono alcune misure di variabilità.
- L'**intervallo di variazione** o range consiste semplicemente nella differenza tra il valore massimo e il valore minimo della distribuzione.





40

L'intervallo di variazione

Esempio:

Gli insiemi di valori di VES

{A}: { 8, 5, 7, 6, 35, 5, 4}
{B}: { 11, 8, 10, 9, 17, 8, 7}

hanno la stessa media ($\bar{x}$=10),

ma in {A} i valori sono più dispersi che in {B}:

in {A} i valori sono inclusi tra 4 e 35

in {B} i valori sono inclusi tra 7 e 17

La differenza tra il massimo e il minimo valore di un insieme di dati è detto **intervallo di variazione** (o **range**).

il **range** di {A} è $R_A = 35 - 4 = 31$

il **range** di {B} è $R_B = 17 - 7 = 10$

Il **range** è il più ***intuitivo*** fra gli indici di dispersione, ha però il difetto di basarsi solo sui due valori estremi, nei quali si manifesta maggiormente la variabilità di campionamento e l'errore di misura.

41

La devianza

Gli indici di dispersione di più largo uso sono basati sugli **scarti dalla media**: per un campione di dimensione n, $\{x_1, x_2, \ldots x_n\}$, sono così definiti

Devianza: $D = \sum (x_i - \bar{x})^2$

Varianza campionaria: $s^2 = \frac{D}{n-1}$

Deviazione standard: $s = \sqrt{s^2}$

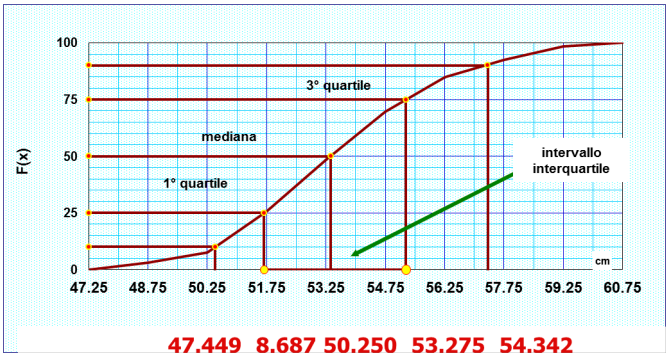
Coefficiente di variazione: $CV\% = 100 \times \frac{s}{\bar{x}}$

La **devianza** è la somma dei quadrati degli scarti tra ogni elemento del campione (x_i) e la media campionaria ($\bar{x}$).

42

l'intervallo interquartile

Un indice di dispersione di uso comune è l'**intervallo interquartile**, dato dalla **differenza tra 3° e 1° quartile** (cioè tra 75° e 25° centile): tale intervallo contiene la metà dei valori inclusi nel campione, indipendentemente dalla forma della distribuzione della variabile.



Principali indici statistici

I grafici finora analizzati ci danno informazioni qualitative; possiamo quantificarle ricorrendo ai seguenti indici.
Siano $x_1, x_2, \ldots, x_n$ n osservazioni numeriche

- di posizione
  - MODA
  - MEDIANA
  - MEDIA
- INDICI
  - di dispersione
    - SCARTO QUADRATICO MEDIO
    - VARIANZA
    - RANGE
  - di forma
    - ASIMMETRIA (SKEWNESS)
    - CURTOSI (KURTOSIS)

50

Inferenza Statistica

Come è possibile estendere a tutta la popolazione le informazioni contenute in un campione?

Per rispondere a questa domanda abbiamo bisogno di nozioni di **teoria della probabilità**

Richiami di Teoria della Probabilità

- **probabilità**
- **funzione densità di probabilità**
- **densità di probabilità normale (gaussiana)**

Probabilità (1)

Che cosa è la probabilità?

È una misura che associa una valutazione numerica (numero compreso tra 0 ed 1) al verificarsi di un evento

Consideriamo un **esperimento** (es: lancio di un dado)

Un esperimento può dar luogo a diversi risultati: ciascuno di essi è detto **evento**

A ciascun evento possiamo far corrispondere un numero compreso tra 0 e 1, che rappresenta la probabilità che l'evento ha di verificarsi

Formal definition of Probability

A probability measure P on the countable sample space Ω is a set function

$$P : \mathcal{F} \to [0, 1],$$

satisfying the following conditions

- $P(\Omega) = 1$.
- $P(\omega_i) = p_i$.
- *If* $A_1, A_2, A_3, \ldots \in \mathcal{F}$ *are mutually disjoint, then*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Probabilità (2)

Definizione di probabilità **nel caso di eventi equiprobabili!**

Dato un *esperimento* ed un *evento* A, la probabilità che l'evento A si verifichi è data da

$$Pr\{A\} = \lim_{n \to \infty} \frac{n_A}{n}$$

n = numero totale di prove effettuate

n_A = numero di volte in cui si è verificato l'evento A

$Pr\{A\}$ è il limite della frequenza relativa dell'evento al tendere all'infinito del numero di prove

Probabilità (3)

$Pr\{A\}$ è un numero compreso tra 0 ed 1

S = insieme di tutti i possibili risultati di un esperimento (**evento certo**)

$Pr\{S\} = 1$

Descrizione di una POPOLAZIONE Mediante descrittori statistici

Funzione densità di probabilità (1)

Supponiamo ora di associare ad ogni possibile risultato di un esperimento un numero X

X è una **variabile casuale** (aleatoria): non conosciamo a priori il valore di X, solo dopo aver eseguito l'esperimento X assume un valore ben preciso x

N.B.: D'ora in avanti

- la lettera maiuscola (X) rappresenta la variabile casuale
- la lettera minuscola (x) rappresenta un valore specifico che la v.c. può assumere, cioè una possibile realizzazione di X

$$P(X \in A) = \int_A p_X(x)\, \mathrm{d}\, x$$

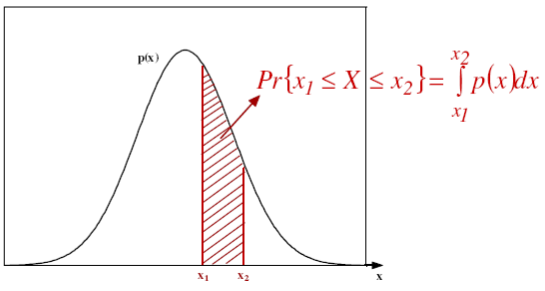
Definizione:
Qualsiasi funzione f(x) si può definire una funzione densità di probabilità se e solo se:
l'integrale su tutto lo spazio di p(x) deve essere 1.

Di conseguenza ogni funzione non negativa, integrabile secondo Lebesgue, con integrale su tutto lo spazio uguale a 1, è la funzione densità di probabilità di una ben definita distribuzione di probabilità.

Funzione densità di probabilità (2)

Si definisce **funzione densità di probabilità** di X (e la si indica con $p(x)$), la funzione

$$p(x) = \lim_{\Delta x \to 0} \frac{Pr\{x \le X \le x + \Delta x\}}{\Delta x}$$



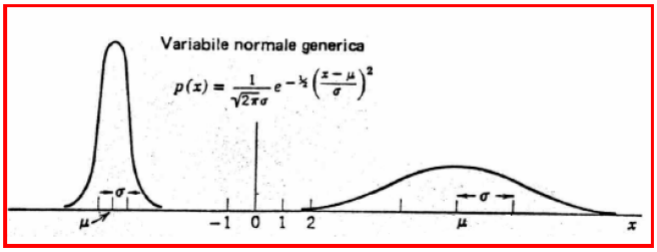
La Generica Distribuzione Normale

Una variabile casuale X è normale con media μ e varianza σ^2 se la sua funzione densità di probabilità ha la seguente espressione

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

assume un massimo in corrispondenza a μ

è simmetrica rispetto a μ



La Distribuzione Normale Standardizzata (1)

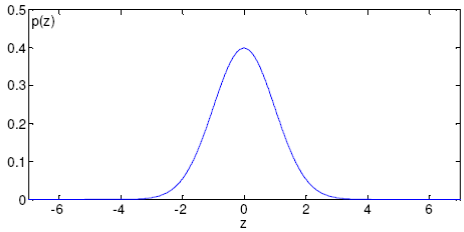
La variabile normale standardizzata Z ha valore medio nullo e varianza unitaria: $\mu_Z = 0$

$$\sigma_Z^2 = 1$$

La sua funzione di probabilità è $p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

assume un massimo in corrispondenza a 0

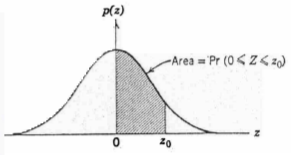
è simmetrica rispetto all'origine



La Distribuzione Normale Standardizzata (2)

Le probabilità che corrispondono alle superfici racchiuse dalla curva normale standardizzata tra il valore medio (che è uguale a zero) ed un qualsiasi valore z_0 sono state tabulate

	Seconda cifra decimale di z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389



La Distribuzione Normale Standardizzata (3)

Esempi per illustrare l'uso della tabella

$$Pr\{0 \le Z \le 1.96\} = 0.475 = 47.5\%$$

$$Pr\{-1.96 \le Z \le 1.96\} = 0.95 = 95\%$$

$$Pr\{1.2 \le Z \le 2.3\} = Pr\{0 \le Z \le 2.3\} - Pr\{0 \le Z \le 1.2\} =$$
$$= 0.4893 - 0.3849 = 0.1044$$

La Distribuzione Normale Standardizzata (4)

Nel seguito indicheremo con

z_P = valore che lascia una probabilità pari a P nelle due code della distribuzione normale standardizzata

$z_{0.05}$ = valore che lascia nelle due code una probabilità pari al 5% (cioè in ciascuna delle due code una probabilità del 2.5%) = 1.96

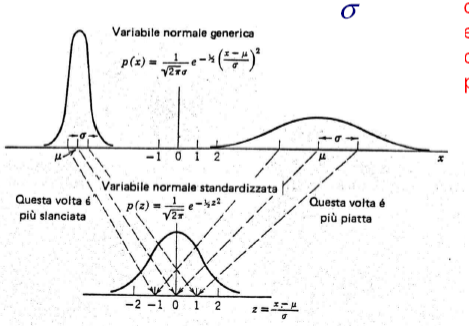
$z_{0.02}$ = valore che lascia nelle due code una probabilità pari al 2% (cioè in ciascuna delle due code una probabilità dell' 1%) ≈ 2.33

$z_{0.01}$ = valore che lascia nelle due code una probabilità pari all' 1% (cioè in ciascuna delle due code una probabilità dello 0.5%) ≈ 2.58

Standardizzare una generica var. norm. (1)

Una generica variabile normale X (con media μ e varianza σ^2) essere trasformata nella forma standardizzata ponendo

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



"standardizzare" una variabile normale consiste nello spostare l'origine in modo da farla coincidere con il valore medio, e nel cambiare la scala in modo che la deviazione standard risulti pari ad 1

Standardizzare una generica var. norm. (2)

Data una generica variabile normale X (con media μ e varianza σ^2) ci interessa determinare

$$Pr\{x_1 \le X \le x_2\} = \int_{x_1}^{x_2} p(x)dx$$

Queste probabilità possono essere calcolate ricorrendo alla tavola della variabile normale standardizzata

Standardizzare una generica var. norm. (3)

Consideriamo una variabile normale X con media $\mu = 100$ e dev. st. $\sigma = 5$. Vogliamo determinare

$$Pr\{X \ge 110\}$$

$$\Updownarrow$$

$$Pr\left\{\frac{X-100}{5} \ge \frac{110-100}{5}\right\}$$

$$\Updownarrow$$

$$Pr\{Z \ge 2\}$$

Standardizzare una generica var. norm. (4)

$$Pr\{Z \geq 2\} = 1 - Pr\{Z \leq 2\} =$$
$$= 1 - Pr\{Z \leq 0\} - Pr\{0 \leq Z \leq 2\} =$$
$$= 1 - 0.5 - 0.4772 = 0.0228$$

$$Pr\{X \geq 110\} = 0.0228$$

Perché abbiamo richiamato questi concetti di teoria della probabilità?

Perché la teoria della probabilità applicata al campione estratto dalla popolazione ci permette di compiere un' inferenza statistica sull'intera popolazione

Vediamo meglio che cosa comporta estrarre un campione da una popolazione

Campionamento (1)

In generale consideriamo una popolazione

Il manifestarsi del fenomeno oggetto di studio in ciascuna unità viene descritto mediante una v.c. X.

Questa v.c. X ha un valore medio (μ, media della popolazione) e una varianza (σ^2, varianza della popolazione)

La popolazione oggetto del nostro interesse è fissa → μ e σ^2 sono delle costanti (incognite) che prendono il nome di **parametri della popolazione**

Campionamento (2)

Dalla popolazione estraiamo un campione casuale di n unità

Sul campione estratto calcoliamo valore medio m e varianza s^2 (sono valori ben precisi)

Una volta estratto un campione occorre tenere presente che è solo uno dei possibili campioni che avremmo potuto estrarre.

I risultati osservati sarebbero stati diversi su un altro campione

Siamo in presenza di **variabilità campionaria**

Campionamento (3)

Pertanto i valori di media (m) e di varianza (s^2) ottenuti sul campione estratto non sono altro che particolari realizzazioni di due v.c.: la media campionaria M e la varianza campionaria S^2 (dette **statistiche campionarie**)

La media campionaria M e la varianza campionaria S^2 variano al mutare del campione secondo una certa distribuzione di probabilità

Ciò che ci consente di compiere inferenze sulla popolazione a partire dal campione è la conoscenza della funzione densità di probabilità delle statistiche campionarie

Media Campionaria (1)

In particolare ci interessa compiere inferenze circa la media di una popolazione

X popolazione di origine con media μ e varianza σ^2

μ e σ^2 sono delle costanti ma incognite

Campione di dimensione n

Vediamo cosa la teoria della probabilità ci permette di sapere circa la v.c media campionaria M

Teorema del campionamento statistico

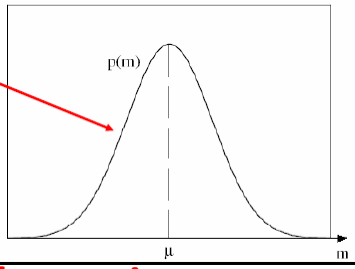
Media Campionaria (2)

La teoria della probabilità permette di affermare che:

la media campionaria M è distribuita **normalmente**, con valore medio μ e varianza σ^2/n

$$M = N\left(\mu, \frac{\sigma^2}{n}\right)$$

Distribuzione di probabilità della v.c. media campionaria M



Teorema del campionamento statistico

Inferenza

Ci limitiamo a compiere inferenza circa la media di una popolazione

stima per intervalli di confidenza

test d'ipotesi

Inferenza sulla media: Stima per intervalli di confidenza (1)

μ media della popolazione = parametro fisso ma incognito

Il nostro problema consiste nello stimare tale parametro estraendo un campione dalla popolazione e calcolandone il valore medio

Sul campione otteniamo un valore medio ben preciso (m)

Inferenza sulla media: Stima per intervalli di confidenza (2)

Possiamo porre $\mu = m$?

NO!!!

il valore m ottenuto non sarà esattamente uguale alla media della popolazione, ma sarà inficiato da un errore per eccesso o per difetto

infatti la variabile casuale M si distribuisce attorno a μ assumendo valori superiori o inferiori

Inferenza sulla media: Stima per intervalli di confidenza (3)

È più corretto pensare che μ sia compreso, con una certa probabilità, in un intervallo, noto come ***intervallo di confidenza***

$$\mu = m \pm certo\ errore = m \pm \Delta$$

L'errore Δ dipende dal grado di fiducia, cioè dalla probabilità, che vogliamo avere (più è alto il grado di fiducia più è alto l'errore)

La probabilità comunemente più utilizzata è il 95%

Si parla in questo caso di **intervallo di confidenza al 95%**

Inferenza sulla media: Stima per intervalli di confidenza (4)

Un intervallo di confidenza al 95% significa che la probabilità che l'intervallo contenga il valore vero del parametro è del 95%:

$$Pr\{M - \Delta \le \mu \le M + \Delta\} = 95\%$$

$$\Downarrow$$

$$Pr\{\mu - \Delta \le M \le \mu + \Delta\} = 95\%$$

Un intervallo di confidenza al 99% significa che la probabilità che l'intervallo contenga il valore vero del parametro è del 99%:

$$Pr\{\mu - \Delta \le M \le \mu + \Delta\} = 99\%$$

Inferenza sulla media: Stima per intervalli di confidenza (5)

Quando si compie un'inferenza su una media mediante intervalli di confidenza occorre distinguere due casi

- varianza della popolazione nota
- varianza della popolazione ignota

Inferenza sulla media: Stima per intervalli di confidenza, σ nota (1)

Vogliamo stimare la media μ di una popolazione tramite la media m ottenuta sul campione **nota** la deviazione standard σ della popolazione di origine

In questo caso, per calcolare l'errore Δ, trasformiamo la variabile casuale media campionaria M (che è una variabile casuale normale con valore medio μ e varianza σ^2/n) nella sua forma standardizzata Z:

$$\frac{M-\mu}{\sigma/\sqrt{n}} = Z$$

Inferenza sulla media: Stima per intervalli di confidenza, σ nota (2)

Quindi cerchiamo nella tabella della variabile normale standardizzata il valore z_0 tale che

$$Pr\{-z_0 \le Z \le +z_0\} = 95\%$$

Dalla tabella si ottiene: $z_{0.05}$ = 1.96

Inferenza sulla media: Stima per intervalli di confidenza, σ nota (3)

$$Pr\left\{-1.96 \le \frac{M-\mu}{\sigma/\sqrt{n}} \le +1.96\right\} = 95\%$$

$$Pr\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} \le M \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 95\%$$

$$Pr\left\{M - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le M + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 95\%$$

Inferenza sulla media: Stima per intervalli di confidenza, σ nota (4)

L'errore Δ è quindi dato da:

$$\Delta = z_{0.05}\frac{\sigma}{\sqrt{n}} = 1.96\frac{\sigma}{\sqrt{n}}$$

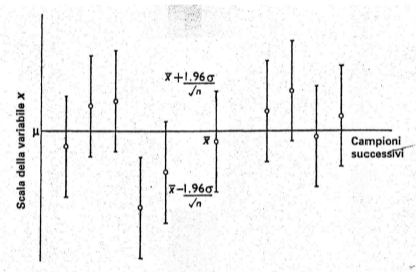
intervallo di confidenza al 95%

$$\mu = m \pm z_{0.05}\frac{\sigma}{\sqrt{n}} = m \pm 1.96\frac{\sigma}{\sqrt{n}}$$

cioè μ appartiene a questo intervallo con una probabilità del 95%

Inferenza sulla media: Stima per intervalli di confidenza, σ nota (5)

interpretazione dell'intervallo di confidenza



popolazione con valore medio μ (parametro incognito)

estraggo più campioni; ognuno fornisce un diverso valore m e quindi diversi intervalli di confidenza

il 95% di questi intervalli include μ, mentre solo nel 5% dei casi m devia da μ per più di 1.96 deviazioni standard e l'intervallo di confidenza non comprende μ

Inferenza sulla media: Stima per intervalli di confidenza, σ nota (6)

Per determinare l'intervallo di confidenza al 98% è necessario trovare il valore $z_{0.02}$

Dalla tabella si trova che $z_{0.02} \approx 2.33$

intervallo di confidenza al 98%

$$\mu = m \pm z_{0.02}\frac{\sigma}{\sqrt{n}} \approx m \pm 2.33\frac{\sigma}{\sqrt{n}}$$

Inferenza sulla media: Stima per intervalli di confidenza, σ nota (7)

Esempio

Su un campione di n = 10 individui è stato misurato il livello di colesterolo nel sangue ottenendo una media di 220 mg/dl. Sapendo che in generale la deviazione standard del livello di colesterolo è 36 mg/dl, determinare l'intervallo di confidenza per il livello di colesterolo dell'intera popolazione al 95%, 98% e 99%

I.C. 95% $\mu = m \pm z_{0.05}\frac{\sigma}{\sqrt{n}} = m \pm 1.96\frac{\sigma}{\sqrt{n}} = 220 \pm 1.96\frac{36}{\sqrt{10}} = (197.7 \quad 242.3)$

I.C. 98% $\mu = m \pm z_{0.02}\frac{\sigma}{\sqrt{n}} \approx m \pm 2.33\frac{\sigma}{\sqrt{n}} = 220 \pm 2.33\frac{36}{\sqrt{10}} = (193.5 \quad 246.5)$

I.C. 99% $\mu = m \pm z_{0.01}\frac{\sigma}{\sqrt{n}} \approx m \pm 2.58\frac{\sigma}{\sqrt{n}} = 220 \pm 2.58\frac{36}{\sqrt{10}} = (190.6 \quad 249.4)$

Inferenza sulla media: Stima per intervalli di confidenza, σ ignota (1)

Vogliamo stimare la media μ di una popolazione tramite la media m ottenuta sul campione essendo **ignota** la deviazione standard σ della popolazione di origine

In questo caso non si può seguire la procedura esposta precedentemente perché non si conosce la deviazione standard della media campionaria ($\sigma/\sqrt{n}$) ovvero non si può utilizzare la variabile normale standardizzata Z

Inferenza sulla media: Stima per intervalli di confidenza, σ ignota (2)

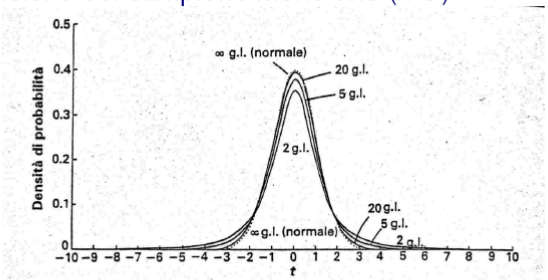
È ragionevole però sostituire σ con la deviazione standard stimata sul campione, s, ottenendo così una nuova variabile casuale

$$\frac{M-\mu}{S/\sqrt{n}} = t$$

"t di Student"

Inferenza sulla media:
Stima per intervalli di confidenza, σ ignota (3)

Le distribuzioni t formano una famiglia di distribuzioni, contraddistinte da un indice, i "gradi di libertà" (g.l.) = dimensione del campione meno uno (n-1)



All'aumentare dei gradi di libertà la distribuzione t tende alla distribuzione normale standardizzata
(perché all'aumentare delle dimensioni del campione, s approssima bene σ)

Inferenza sulla media:
Stima per intervalli di confidenza, σ ignota (4)

Anche la distribuzione t (come la distribuzione Z) è stata tabulata (per diversi gradi di libertà)

La funzione tabulata è il valore $t_{v,P}$ che lascia una probabilità P nelle due code della distribuzione t con v gradi di libertà

Ad esempio per un campione di n = 10 elementi, il numero di gradi di libertà è v = 9 e dalla tabella risulta che

$$t_{9,0.05} = 2.262$$

quindi:

$$Pr\{-2.262 \le t_9 \le +2.262\} = 95\%$$

Inferenza sulla media: Stima per intervalli di confidenza, σ ignota (5)

Il calcolo dell'intervallo di confidenza si compie come nel caso di σ nota, ma utilizzando la distribuzione t al posto della Z

intervallo di confidenza al 95%

$$\mu = m \pm t_{\nu,0.05} \frac{s}{\sqrt{n}}$$

Ad esempio, nel caso di un campione di n = 20 elementi, cioè ν = 19, l'intervallo di confidenza al 95% è:

$$\mu = m \pm t_{19,0.05} \frac{s}{\sqrt{n}} = m \pm 2.093 \frac{s}{\sqrt{n}}$$

l'intervallo al 99% è:

$$\mu = m \pm t_{19,0.01} \frac{s}{\sqrt{n}} = m \pm 2.861 \frac{s}{\sqrt{n}}$$

Inferenza sulla media: Stima per intervalli di confidenza, σ ignota (6)

Esempio

Uno psicologo sottopone un campione di 10 persone ad un test di reazione ad uno stimolo.

I tempi di reazione (sec.) sono stati:

0.21, 0.25, 0.28, 0.30, 0.43, 0.35, 0.22, 0.41, 0.33, 0.42.

Determinare l'intervallo di confidenza al 98% per la media della popolazione da cui è stato estratto il campione

m = 0.32

s = 0.0818

I.C. 98% $\mu = m \pm t_{9,0.02} \frac{s}{\sqrt{n}} = m \pm 2.821 \frac{s}{\sqrt{n}} = (0.247 \quad 0.393)$

Test di Ipotesi (1)

I test di ipotesi o test di significatività (o anche test statistici) sono tra i metodi più importanti dell'inferenza statistica

Consideriamo la seguente situazione: si vuole testare un nuovo farmaco contro l'influenza.

È stato rilevato su un grande numero di pazienti che l'influenza trattata con i farmaci tradizionali ha una durata media di 4 giorni e una deviazione standard di 2 giorni.

Un campione di n pazienti viene trattato con il nuovo farmaco e si osserva una riduzione della durata media della malattia.

Il risultato è dovuto solo ad una fluttuazione casuale?

Oppure si può affermare che il nuovo farmaco è migliore rispetto ai farmaci tradizionali?

Test di Ipotesi (2)

In situazioni di questo tipo si parla di prova delle ipotesi perché si definiscono due ipotesi in conflitto:

H_0: ipotesi nulla o del niente fuori dall'ordinario

H_1: ipotesi alternativa

L'ipotesi nulla ipotizza che il nuovo trattamento **non** differisca dai precedenti. Ovvero l'ipotesi nulla assume che il campione in esame appartenga ad una popolazione nota (di cui è nota, cioè, la media ed eventualmente anche la varianza), e che quindi non si sia scoperto nulla di nuovo rispetto a quanto già si conosce

L'ipotesi alternativa ipotizza invece l'esistenza di una diversità del nuovo trattamento rispetto a quelli esistenti

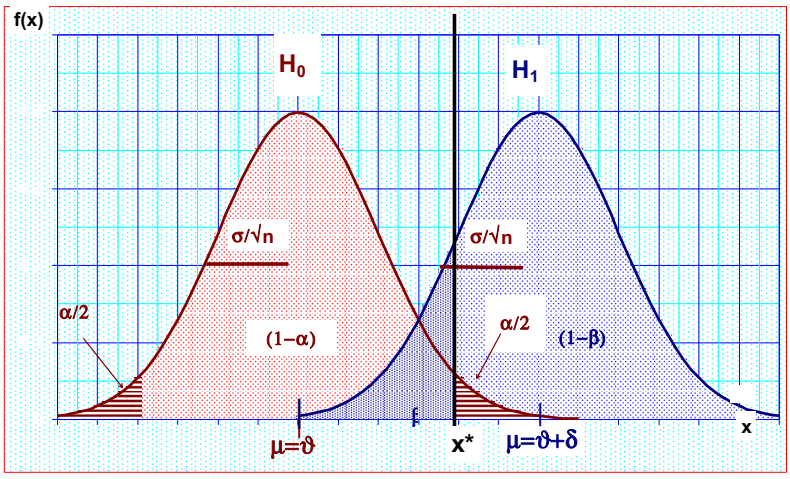
Test di Ipotesi (3)

L'ipotesi nulla è l'ipotesi che viene effettivamente sottoposta a verifica

Si assume che l'ipotesi nulla sia vera; quindi attraverso il test d'ipotesi si valuta l'***entità della discrepanza*** tra quanto osservato sul campione e quanto previsto dall'ipotesi nulla, stabilendo se la discrepanza è "significativa" o "non significativa" a ***livello α*** (***livello di significatività***)

Nel primo caso l'ipotesi nulla viene rifiutata con livello di significatività α, nel secondo caso l'ipotesi nulla non può essere rifiutata con livello di significatività α

Criterio di decisione del test



Media campionaria < oppure >x* = μ + z*σ/√n

criterio di decisione del test

Stabilire il **criterio di decisione** significa stabilire, per i valori della media campionaria, una **soglia** oltre la quale il risultato sperimentale viene ritenuto incompatibile con l'ipotesi H_0: $\mu=\theta$.

Poiché la distribuzione delle medie campionarie è nota sotto H_0, è possibile scegliere una soglia cui sia associato un **rischio d'errore di tipo I** (α) sufficientemente piccolo, e quindi una **protezione** $(1-\alpha)$ sufficientemente grande.

Il rischio di errore di tipo I (α) è detto livello di significatività.

Criterio di decisione del test

SE È VERA H_0	SE È VERA H_1	Ed in base al campione
… decisione giusta Protezione: **(1-α)**	… decisione sbagliata errore di tipo II° : **β**	… decido che è vera H_0
…decisione sbagliata errore di tipo I° : **α**	… decisione giusta Potenza: **(1-β)**	… decido che è vera H_1

Protezione (1- **α**): probabilità di accettare H_0 quando è vera H_0

Potenza del test (1-**β**): probabilità di rifiutare H_0 quando è vera H_1

Rischio di errore di tipo I (**α**): probabilità di rifiutare H_0 quando è vera H_0

Rischio di errore di tipo II (**β**): probabilità di accettare H_0 quando è vera H_1

Tabella 9.1 – RAPPRESENTAZIONE SU UNA TABELLA 2X2 DELLA DISTRIBUZIONE DI UNA POPOLAZIONE IN BASE AI RISULTATI DI UN TEST

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

M+= malati; M-= sani
T+= test positivo; T-= test negativo

Sensibilità:
capacità del test di individuare in una popolazione i soggetti malati

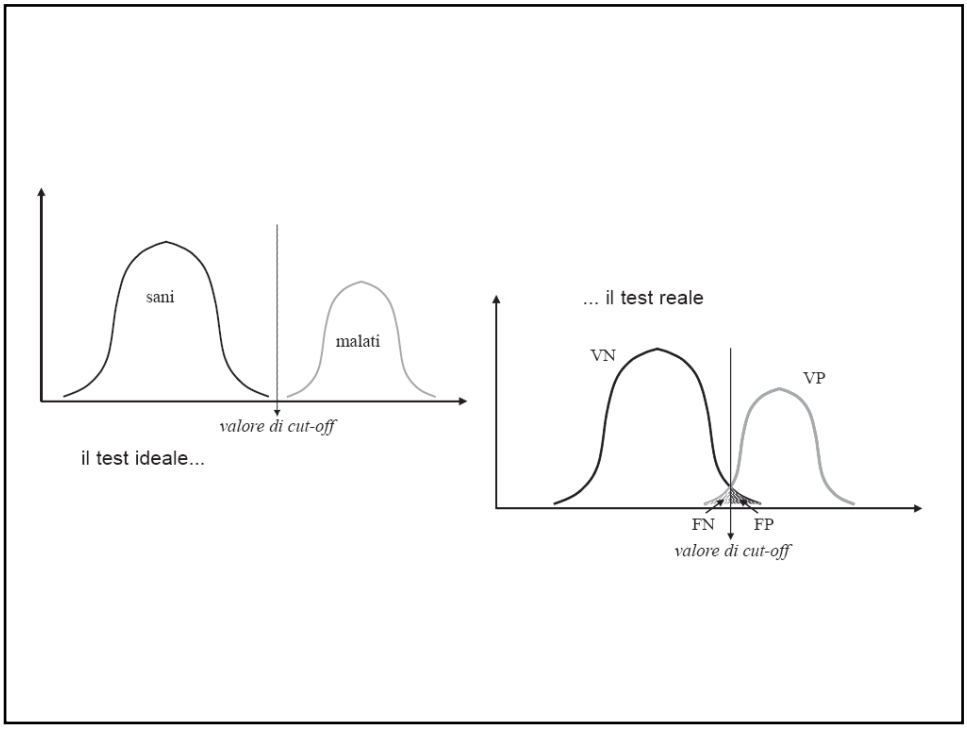
$$\frac{VP}{T_{M+}} = \frac{VP}{VP + FN}$$

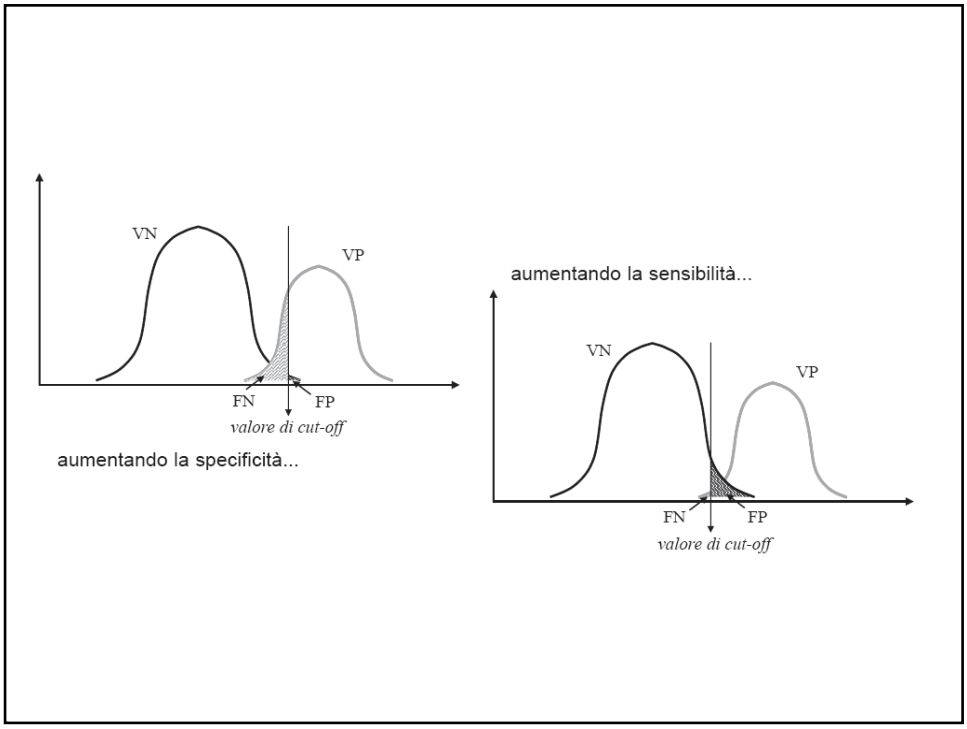
VP= veri positivi
FN= falsi negativi
T_{M+}= totale malati

Specificità:
capacità del test di individuare in una popolazione i soggetti sani

$$\frac{VN}{T_{M-}} = \frac{VN}{VN + FP}$$

VN= veri negativi
FP= falsi positivi
T_{M-}= totale sani


sani
malati
valore di cut-off
il test ideale...
... il test reale
VN
VP
FN
FP
valore di cut-off


VN
VP
FN
FP
valore di cut-off
aumentando la specificità...
aumentando la sensibilità...
VN
VP
FN
FP
valore di cut-off

Valore predittivo positivo:
esprime la probabilità che un soggetto risultato positivo ad un test sia effettivamente malato

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

VP= veri positivi
FP= falsi positivi
T_P= totale positivi

Valore predittivo negativo:
esprime la probabilità che un soggetto risultato negativo ad un test sia effettivamente sano

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

VN= veri negativi
FN= falsi negativi
T_N= totale negativi

Inferenza sulla media: Test di Ipotesi

Quando si compie un'inferenza su una media mediante test di ipotesi, occorre distinguere due casi (come per gli intervalli di confidenza)

- varianza della popolazione nota (z-test)

- varianza della popolazione ignota (t-test)

Inferenza sulla media: Test di Ipotesi, σ nota (1)

Vogliamo testare l'ipotesi nulla che il campione su cui sono stati rilevati i dati appartenga ad una popolazione nota con media μ_0 e deviazione standard σ_0

Formuliamo innanzitutto le due ipotesi:

H_0:	$\mu = \mu_0$ $\sigma = \sigma_0$	assume che il campione appartenga alla popolazione nota, (ipotesi del nulla fuori dall'ordinario)

H_1:	$\mu \neq \mu_0$	ipotizza che il campione appartenga ad una popolazione con media diversa dalla popolazione nota, ma con stessa varianza

Inferenza sulla media: Test di Ipotesi, σ nota (2)

Se è vera l'ipotesi nulla, allora la variabile casuale media campionaria (M) ha una distribuzione normale con valore medio μ_0 e deviazione standard $\sigma_0/\sqrt{n}$ (dove n è la dimensione del campione)

Il campione ha fornito un valore medio ben preciso (m)

In base al valore ottenuto sul campione si decide se accettare o rifiutare l'ipotesi nulla

In che modo si prende questa decisione?

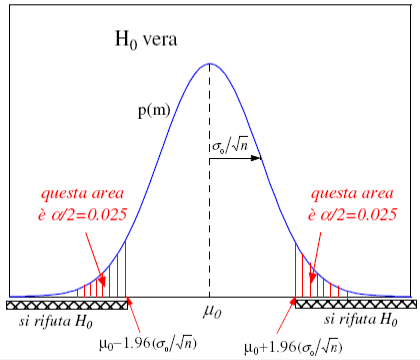
Inferenza sulla media: Test di Ipotesi, σ nota (3)

Si fissa innanzitutto il livello di significatività α

Ad esempio α = 5%.

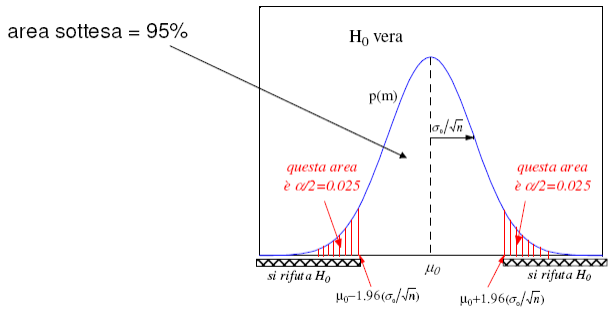
Si individuano le due code dalla funzione di probabilità di M, che sottendono ciascuna una probabilità pari ad $\alpha/2$ = 2.5%

Queste due code individuano la regione di rifiuto dell'ipotesi nulla



Inferenza sulla media: Test di Ipotesi, σ nota (4)

Infatti se fosse vera H_0, ci sarebbe una probabilità pari al 95% che la media osservata sul campione cada nell'intervallo interno e solo una probabilità del 5% che cada nelle due code esterne. Quindi se il valore osservato nel campione cade nelle due code esterne è molto probabile che l'ipotesi nulla non sia vera



Inferenza sulla media: Test di Ipotesi, σ nota (5)

Passando alla variabile normale standardizzata Z, sappiamo immediatamente qual è l'intervallo che contiene il 95% della probabilità

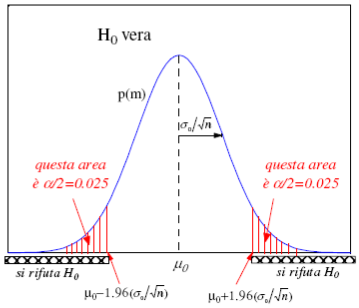
$$Pr\{-1.96 < Z < 1.96\} = 95\%$$

$$Z_{0.05} = Z_{\alpha} = 1.96$$

Quindi, se è H_0 vera c'è una probabilità del 95% che il valore di media campionaria osservato sul campione sia compreso nell'intervallo

$$\mu_0 \pm 1.96 \frac{\sigma_0}{\sqrt{n}}$$

mentre c'è solo una probabilità del 5% che cada al di fuori di tale intervallo



Inferenza sulla media: Test di Ipotesi, σ nota (6)

Avendo fissato $\alpha = 5\% \rightarrow z_{\alpha}=1.96$

Se il valore m osservato nel campione è tale che:

$$m \le \mu_0 - 1.96 \cdot \frac{\sigma_0}{\sqrt{n}} \text{ oppure } m \ge \mu_0 + 1.96 \cdot \frac{\sigma_0}{\sqrt{n}}$$

⬇

rifiuto H_0 con livello di significatività $\alpha = 5\%$

confrontare m con $\mu_0 \pm 1.96 \cdot \sigma_0/\sqrt{n}$

⬍

confrontare z con ± 1.96, essendo $z = \dfrac{m-\mu_0}{\sigma_0/\sqrt{n}}$

Inferenza sulla media: Test di Ipotesi, σ nota (7)

Quindi per testare l'ipotesi nulla con un livello di significatività α si procede in questo modo:

1. si considera la variabile normale standardizzata $Z = \frac{M - \mu_0}{\sigma_0 / \sqrt{n}}$
2. si sostituisce al posto della v.c. M il valore m osservato sul campione, ottenendo così una particolare realizzazione z della variabile casuale Z
3. si ricava dalla tabella della distribuzione Z il valore z_α

Inferenza sulla media: Test di Ipotesi, σ nota (8)

4. se $z > z_\alpha$ o $z < -z_\alpha$
   → si rifiuta H_0 e si afferma che la differenza tra il risultato ottenuto sul campione e μ_0 è significativa con livello di significatività pari a α

   se $-z_\alpha < z < z_\alpha$
   → non si può rifiutare H_0 con livello di significatività α

Poiché si utilizza la variabile normale standardizzata, questo test di ipotesi è definito anche ***z-test***

Inferenza sulla media: Test di Ipotesi, σ nota (9)

Esempio
È noto che, per una certa patologia, il tempo medio di sopravvivenza dalla diagnosi è di 38.3 mesi con deviazione standard di 43.3 mesi. Cento pazienti vengono trattati con un nuovo trattamento e il loro tempo medio di sopravvivenza è di 46.9 mesi. Si può spiegare questa differenza nella sopravvivenza come semplice fluttuazione casuale o effettivamente il nuovo trattamento differisce dai precedenti?

Testiamo l'ipotesi nulla che il campione appartenga alla popolazione nota con media μ_0 = 38.3 (mesi) e σ_0 = 43.3 (mesi)

H_0: $\mu = \mu_0 = 38.3$
$\sigma = \sigma_0 = 43.3$

H_1: $\mu \neq \mu_0$

Inferenza sulla media: Test di Ipotesi, σ nota (10)

Esempio (continua)
Calcoliamo il valore z osservato nel campione:

$$z = \frac{m - \mu_0}{\sigma_0 / \sqrt{n}} = \frac{46.9 - 38.3}{43.3 / \sqrt{100}} = \frac{8.6}{4.33} = 1.99$$

Se consideriamo un livello di significatività α = 5%, il valore di z così ottenuto va confrontato con il valore 1.96 (= $z_{0.05}$)

Il valore 1.99 (z) osservato è esterno all'intervallo -1.96÷1.96, quindi il nuovo trattamento differisce dai trattamenti esistenti con livello di significatività del 5%

Inferenza sulla media: Test di Ipotesi, σ nota (11)

Esempio (continua)

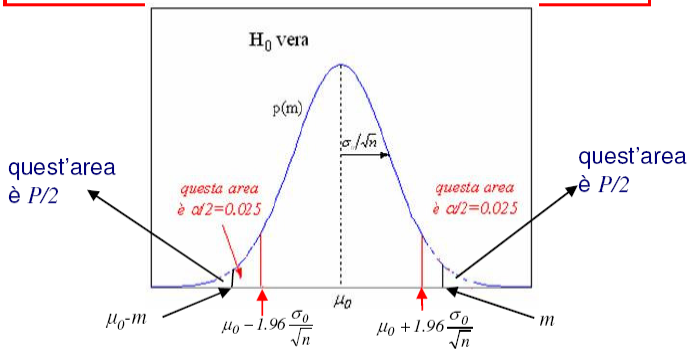
Se fissiamo il livello di significatività α all' 1% anziché al 5%, il valore ottenuto di z deve essere confrontato con il valore tabulato $z_{0.01}$ (≈2.58).

Poiché il valore z (=1.99) è compreso nell'intervallo $-z_{0.01} \div z_{0.01}$ l'ipotesi nulla non può essere rifiutata con livello di significatività dell'1%.

Il P-value (1)

I software per l'analisi statistica forniscono un valore, detto **P-value,** come risultato del test di ipotesi

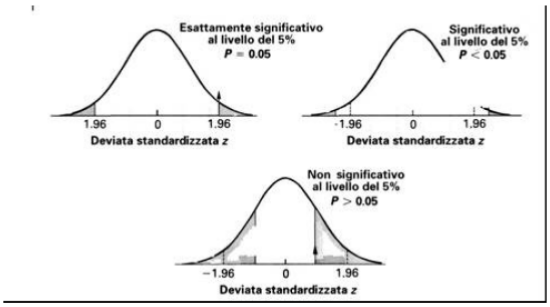
P-value = probabilità P di superare in entrambe le direzioni il valore osservato nel campione nel caso in cui l'ipotesi nulla sia vera



Il P-value (2)

Il *P-value* è una misura della credibilità dell'ipotesi nulla H_0: infatti se P è piccolo significa che il valore osservato nel campione si discosta molto da quanto previsto dell'ipotesi nulla, e quindi è improbabile che l'ipotesi nulla sia vera

In particolare: si rifiuta H_0 se e solo se $P<\alpha$



Il P-value (3)

Nel caso dell'esempio precedente, eseguendo il test al calcolatore si ottiene un valore di P pari a 0.047

P-value = 0.047

Infatti, il test risultava significativo al 5% ($P<0.05$), ma risultava non significativo al livello dell'1% ($P>0.01$)

Equivalenza tra prova delle ipotesi e intervalli di confidenza (1)

Se, a partire dal campione, volessimo fare delle inferenze sulla media μ della popolazione senza concentrarci su un singolo valore μ_0, possiamo utilizzare l'approccio mediante intervalli di confidenza

Sappiamo che c'è una probabilità del 95% che il valore medio μ della popolazione cada nell'intervallo:

$$m - 1.96\frac{\sigma}{\sqrt{n}} < \mu < m + 1.96\frac{\sigma}{\sqrt{n}}$$

Equivalenza tra prova delle ipotesi e intervalli di confidenza (2)

Nell'esempio considerato, i limiti di confidenza al 95% sono:

$$46.9 \pm 1.96 \cdot 4.33 = 38.4 \div 55.4.$$

Si osservi che tale intervallo esclude il possibile valore di 38.3, precedentemente riscontrato
Ciò corrisponde al fatto che tale valore è contraddetto da un test di significatività al livello del 5%

Al contrario, i limiti di confidenza al 99% sono:

$$m \pm 2.58\frac{\sigma}{\sqrt{n}} = 46.9 \pm 2.58 \cdot 4.33 = 35.7 \div 58.07$$

Tale intervallo include il valore 38.3, e questo corrisponde al fatto che tale valore non è contraddetto da un test di significatività al livello dell'1%