Dati e stime:

variabilità e insicurezza nella vita quotidiana

(M.S. Bernabei)

Introduzione

Spesso noi pensiamo in modo "statistico" nella nostra vita quotidiana senza che ce ne accorgiamo: ad esempio quando decidiamo di prendere l'ombrello prima di uscire, quando facciamo commenti sul numero degli incidenti del sabato sera, o notato che rendiamo più durante il corso che durante gli esami.
La parola statistica deriva dal latino "status" e significa stato. Per lungo tempo la statistica è stata identificata soltanto con la rappresentazione di dati e relativi grafici relativi al mondo dell'economia, politico, ecc.. Solo durante il ventesimo secolo la statistica è cresciuta notevolmente come disciplina scientifica.
La Statistica va ben oltre la semplice rappresentazione dei dati, infatti essa si occupa

- della raccolta dei dati;
- della descrizione e sintesi dei dati,
- dell' interpretazione dei dati in modo da trarre conclusioni sul fenomeno in esame.

La Statistica si applica a tutti i fenomeni che coinvolgono la raccolta e l'analisi dei dati: sondaggi di opinione (indagini di tipo sociale, economico o sulla salute), esperimenti nel campo dell'agricoltura (su nuovi semi, pesticidi), studi clinici di vaccini, inseminazione di una nuvola per la produzione di pioggia artificiale, economico (disoccupazione, costo della vita, soddisfazione per un certo prodotto), politico (elezioni, miglioramento dei servizi dello stato), industria (qualità e miglioramento della produzione), ecc..
I principi e le metodologie della statistica sono utili per rispondere a domande del tipo:

- Che tipo e quanti dati è necessario raccogliere?
- Come dovremmo organizzare e interpretare i dati?
- Come possiamo analizzare i dati e dedurre le conclusioni?
- Come stimare la potenza delle conclusioni e giudicare la loro incertezza?

Dai dati e tabelle distribuite dai mass media, dai reports, la statistica inferenziale fornisce criteri per determinare quali conclusioni sono supportate dai dati e quelle che non lo sono. La credibilità delle conclusioni dipende fortemente dai metodi statistici utilizzati.

Esempi

Programma di formazione. I programmi di formazione sono specifici per un certo tipo di utenti (studenti, operai, persone con handicap fisico, bambini ritardati, ecc.) e vengono continuamente monitorati, valutati e modificati per migliorare la loro utilità verso la società. Per confrontare l' efficacia fra i diversi tipi di programmi, è essenziale raccogliere dati sul raggiungimento o miglioramento delle abilità dei tirocinanti alla fine di ogni programma.

Richieste di controllo pubblicitario. Il pubblico è costantemente bombardato dalla pubblicità che dichiara la superiorità di un prodotto di una certa marca rispetto ad altri. Quando tali confronti sono basati su una valida dimostrazione sperimentale, servono ad educare ed orientare il consumatore. Non di rado, comunque, dei messaggi pubblicitari fuorvianti sono basati su una insufficiente sperimentazione, su analisi dei dati errate, o anche su vistose manipolazioni dei risultati sperimentali. Le agenzie di governo e i gruppi di consumatori devono essere preparati a verificare la relativa qualità dei prodotti usando una adeguata procedura dei dati raccolti e appropriati metodi di analisi statistica.

Riproduzione delle piante. Per far crescere la produzione di cibo, gli scienziati nel campo dell'agricoltura sviluppano nuovi ibridi dall' incrocio di diverse specie di piante. Le nuove specie di piante devono essere confrontate con quelle già esistenti. La loro relativa produttività è stimata piantando in certo numero di posti ciascuna di ogni specie. I prodotti sono conservati e quindi

analizzati per comprendere se ci sono sostanziali differenze. Le specie potrebbero essere anche confrontate sulla base di resistenza alle malattie o capacità di riprodursi.

Per ogni analisi statistica è molto importante avere informazioni reali. La branca della statistica chiamata **disegno sperimentale** potrebbe guidare colui che investiga nel progettare in che modo e in che quantità raccogliere i dati. Dopo che i dati sono stati raccolti, ci sono dei metodi statistici che sintetizzano e descrivono le caratteristiche principali dei dati, essi sono comunemente conosciuti come **statistica descrittiva**. Oggi, fornisce una maggiore conoscenza del fenomeno in esame la stima dell' informazione presente nei dati. Tale area è chiamata **statistica inferenziale**.

Ad esempio se chiedessimo a 1000 persone per chi voteranno il giorno prima delle elezioni locali, avremmo la possibilità di prevedere quale partito potrebbe vincere o la proporzione verosimile. Non saremo però in grado di dire con sicurezza chi vincerà le elezioni. La statistica potrebbe valutare la "chance" di un qualcosa che sta avvenendo e che potrebbe rivelarsi vero.
Nella vita reale la statistica è usata di frequente per porre in evidenza le differenze tra gruppi individui o luoghi

Bisogna capire che spesso la ricerca scientifica è tipicamente un processo di prove che potrebbero contenere errori. Raramente un fenomeno potrebbe essere compreso completamente o una teoria perfezionata per mezzo di un singolo e unico esperimento, E' troppo aspettarsi di ottenere tutto in una sola prova. I dati ottenuti da un esperimento fornisce nuova conoscenza. Tale conoscenza spesso suggerisce una revisione di una esistente teoria, e ciò potrebbe richiedere ulteriori ricerche attraverso altri esperimenti e analisi dei dati.

Anche se gli esempi precedenti provengono da campi molto lontani tuttavia ci sono delle caratteristiche che li accomuna:

- per acquisire nuova conoscenza si ha bisogno di raccogliere dati rilevanti per il fenomeno in esame.
- anche se spesso le osservazioni siano fatte nelle "stesse condizioni" tuttavia è impossibile eliminare la variabilità dei dati. Ad esempio il trattamento di una allergia potrebbe creare un lungo sollievo per qualche individuo, mentre potrebbe portare solo un momentaneo sollievo o nessuno per tutti gli altri pazienti.
- L'accesso all'intero insieme di dati è praticamente impossibile o da un punto di vista pratico non facile per le limitazioni di tempo, delle risorse e delle facilitazioni, così che dobbiamo lavorare con una informazione incompleta, cioè con i dati che abbiamo a disposizione nel corso di una studio sperimentale.

Dobbiamo distinguere l'insieme dei dati ottenuti attraverso le osservazioni sperimentali (**campione**) e tutte le potenziali osservazioni che possono essere fatte in un certo contesto (**popolazione**). Ogni misura in un insieme di dati è originata da una distinta sorgente (**unità**) che può essere un paziente, una famiglia, un albero, una fattoria.

Unità: una singola unità è di solito una persona o un oggetto le cui caratteristiche o variabilità sono di interesse per l'indagine.
Popolazione: è l'insieme di tutte le unità su cui si svolge l'indagine statistica. Essa può essere finita o infinita.
Campione: sottoinsieme finito di unità della popolazione che sono state raccolte nel corso dell'indagine. Il campione dovrebbe essere il più possibile "rappresentativo" per la popolazione.

Per esempio se vogliamo conoscere le preferenze musicali di una certa città e consideriamo come campione l'insieme degli ascoltatori che chiamano alla radio per esprimere il loro gusti non è rappresentativo perché dipende dal tipo di musica trasmessa da quella radio. Inoltre gli ascoltatori che chiamano ed aspettano per prendere la linea sono delle persone determinate nelle loro opinioni, per cui il risultato dell'indagine sarà falsato dal tipo di campione scelto. Un campione rappresentativo potrebbe essere scelto formando dei numeri di telefono scegliendo a caso tra le cifre 0,1,2,…, 9. Il computer potrebbe simulare tale esperimento (generazione di numeri casuali). E' importante l'ampiezza del campione?
Certamente le abitudini di acquisto di una persona potrebbe non rappresentare quella dell'intera popolazione. Comunque, nonostante le diverse abitudini di acquisto degli individui, potremmo ottenere un'informazione accurata sulle abitudini dell'intera popolazione se prendiamo un campione "sufficientemente" ampio.
Raccogliere i dati per rispondere ad una particolare domanda, in modo da fornire le basi per decidere una azione o migliorare un processo. Bisognerebbe a tal fine stabilire di un obiettivo che sia specifico e non ambiguo.

Esempio. Ogni giorno una città deve testare l'acqua di un lago per determinare se l'acqua è sicura per nuotare. La difficoltà più grande è la crescita delle alghe e il limite di sicurezza è stato stabilito in termini di limpidezza dell' acqua.
L'obiettivo in questo caso è determinare se la limpidezza dell'acqua sulla spiaggia è o no al di sotto del limite di sicurezza.

Obiettivi della Statistica

- fare inferenza sulla popolazione dall'analisi delle informazioni che sono contenute nel campione dei dati. Questo include una stima dell'incertezza contenuta in tale inferenza,
- delineare il processo e l'ampiezza del campione in modo che le osservazioni formino una base per fare valide inferenze.

Statistica descrittiva

La Statistica Descrittiva si occupa di illustrare e sintetizzare i dati osservati.
Fasi:

- Raccolta dei dati
- Organizzazione dei dati in tabelle e grafici.
- Sintesi dei dati mediante gli indici sintetici in modo da descrivere le caratteristiche essenziali.

Variabile: X grandezza che varia all'interno di una popolazione. Essa può essere **numerica** se i valori che essa può assumere sono numeri, in particolare essa si dice **discreta** se l'insieme dei valori è finito o numerabile, p.e. numero di chiamate ad un centralino, e **continua** se è continuo, p.e. altezze di una popolazione, peso, distanze percorse per andare al lavoro. Se la variabile non è numerica si dice **categorica** o **qualitativa**, per esempio gruppi sanguinei, sesso, materia studiata.

Se la variabile è categorica o numerica discreta con un numero di valori non molto alto allora si potrebbe costruire una tabella delle frequenze in cui a nella prima colonna ci sono i valori assunti dalla variabile nella seconda colonna le corrispondenti frequenze assolute e nella terza le frequenze relative. Nella quarta la frequenza cumulata.

Esempio 1
(Variabile categorica): gruppi sanguigni di un campione estratto da una certa popolazione:

gruppo	**Frequenza assoluta**
A	60
B	16
AB	7
O	66
TOTALE	149

Frequenza assoluta: f_i è il numero di volte in cui la variabile X assume il valore quantitativo o qualitativo x_i Se la variabile X assume i valori $x_1, x_2, \ldots, x_k$, con frequenze rispettivamente $f_1, f_2, \ldots, f_k$ allora

$$\mathbf{f_1 + f_2 + \ldots + f_k = n}$$

dove n è il numero totale degli oggetti del campione.

Frequenza relativa: è il rapporto tra la frequenza assoluta e il numero degli elementi del campione, cioè

$p_i = f_i / n$

$p_1 + p_2 + \ldots + p_n = 1.$

Frequenza cumulata: p_{ci} è la somma delle frequenze relative fino alla p_i, cioè

$p_{ci} = p_1 + p_2 + \ldots + p_i.$

N.B.: le distribuzioni delle frequenze relative o percentuali sono indispensabili quando si confrontano due o più gruppi di misure, che presentano un diverso numero di osservazioni.

Esempio 2
(variabile numerica discreta): Numeri e percentuali di donne inglesi di 40 anni e più intervistate sul numero di figli avuti.

Numero di figli	Numero delle donne Freq. Ass.	% delle donne Freq. rel.	Percentuale cumulata
0	354	12,5	12,5
1	414	14,6	27,1
2	1130	39,9	67,0
3	567	20,0	87,0
4	246	8,7	95,7
5	66	2,3	98,0
6	33	1,2	99,2
7	11	0,4	99,6
8	6	0,2	99,8
9	1	0,0	99,8
10	1	0,0	99,8
TOTALE	2829	99,8	

Fonte: General Household Survey 1995-96

Distribuzione in classi per dati continui

Nel caso in cui si avesse un grande numero di dati o che contengono misure su una virtuale scala continua potrebbe risultare difficile includerli tutti senza rendere la tabella complicata e difficile da leggere si potrebbe allora raggruppare i dati in classi.
Osservazione: Raggruppando i dati si perde informazioni riguardo la distribuzione dei dati all'interno di ogni intervallo. Scegliendo un numero di classi troppo basso si rischia di sintetizzare troppo i dati perdendo informazioni sui dati. D'altra parte con un numero di classi troppo elevato rispetto al numero dei dati, le frequenze potrebbero variare in modo caotico e non sarebbe possibile riconoscere un certo andamento della distribuzione, a causa della eccessiva dispersione dei dati.
Osservazione: Quando il numero di dati è molto alto conviene calcolare la distribuzione di frequenza utilizzando il computer.

Esempio: La seguente tabella rappresenta il consumo di alcool in unità di alcool data da un bicchierino di cognac o da un boccale di birra forniti dall'Ufficio Nazionale di Statistica, 1998.

Classi di età	freq. assol. f_i	freq. rel. p_i	freq. cumulata
16-24	1850	0,12	0,12
25-44	5800	0,37	0,49
45-64	4724	0,30	0,79
65-84	3281	0,21	1,00
TOTALE	156555	1,00	

Grafici di distribuzioni di frequenze

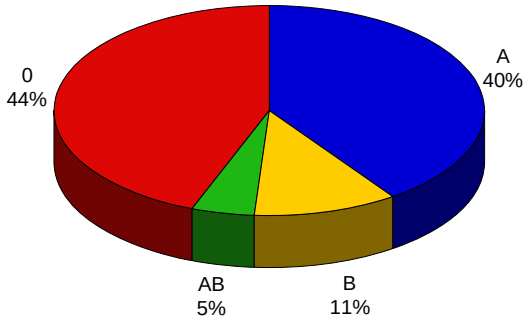
Le distribuzioni contenute nella distribuzione di frequenza possono essere rappresentate graficamente. Ciò permette di sintetizzare i dati.

Diagrammi per dati qualitativi

Aerogrammi a torta

Serve per rappresentare soprattutto le variabili categoriche in cui ogni "fetta" o settore di cerchio rappresenta la frequenza relativa della categoria. L'area di ciascun settore è proporzionale alla frequenza:

a_i:360°= f_i : n

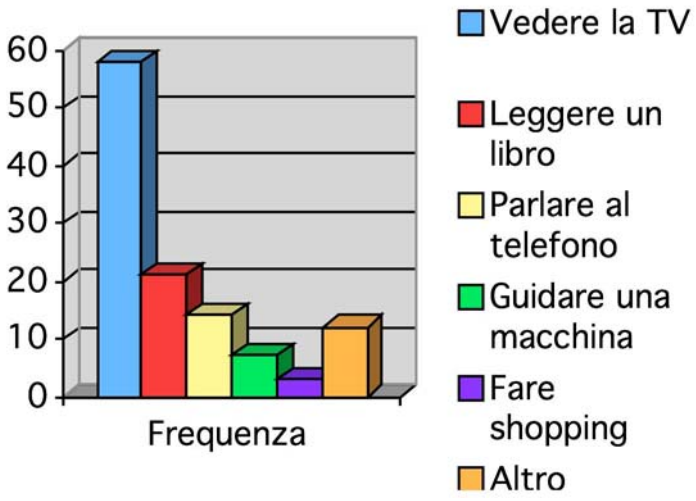


Diagrammi di Pareto

Secondo la legge empirica di Pareto fu formulata da Joseph M. Juran, ma che è nota anche con il nome di principio di Pareto, e che è sintetizzabile nell'affermazione: la maggior parte degli effetti è dovuta ad un numero ristretto di cause. Nei diagrammi di Pareto si ordinano le categorie da quella con frequenza più alta e via via fino a quella più bassa. Sull'asse delle ascisse ci sono le caratteristiche della variabile qualitativa mentre sull'asse delle ordinate ci sono le frequenze assolute delle varie categorie.

Esempio: Un gruppo di studenti che frequentano un corso di psichiatria sono stati interrogati sulla abitudine che dovrebbe essere migliorata. Per ridurre l' effetto di tale abitudine dovrebbero raccogliere dati sulla frequenza e la circostanza in cui si manifesta. Uno studente ha raccolto le seguenti frequenze relativamente al vizio di mangiarsi le unghie in un periodo di due settimane:

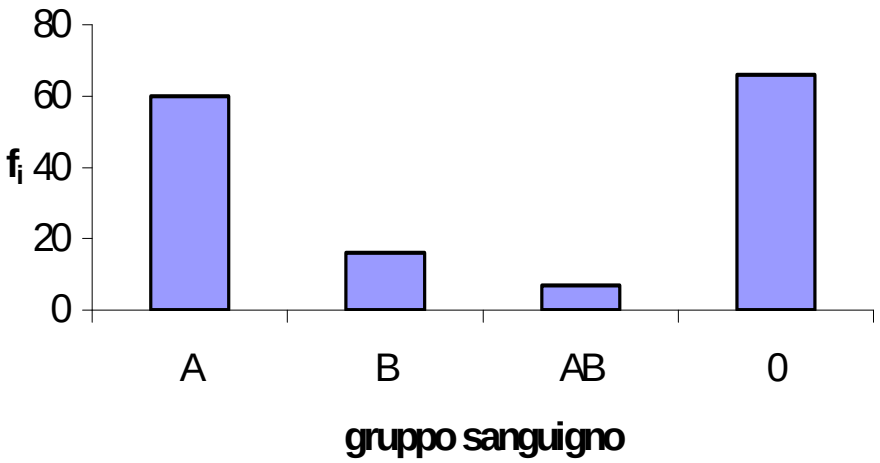
Attività	Frequenza
Vedere la TV	58
Leggere un libro	21
Parlare al telefono	14
Guidare una macchina	7
Fare shopping	3
Altro	12



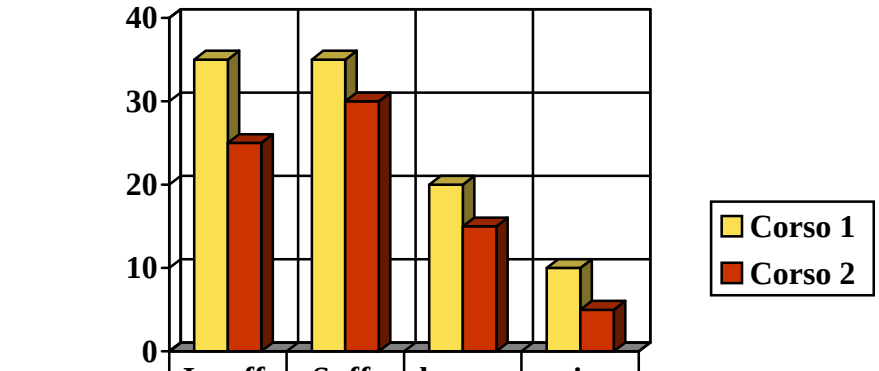
Il diagramma di Pareto mostra la relazione tra le varie attività e l mangiarsi le unghie. Guardare la TV rappresenta il 50,4% del totale.

Diagrammi a barre e istogrammi

I diagrammi a barre (o a linee) vengono utilizzati sia per rappresentare distribuzioni di variabili categoriche, sia variabili numeriche discrete o numeriche con pochi valori. Ad ogni classe corrisponde una barra (o linea) la cui ampiezza della base (per tutte uguali) non ha significato, mentre l'altezza rappresenta la frequenza (assoluta o relativa) della classe. Se le barre sono adiacenti allora si ha un' istogramma, e l'ordine delle barre ha significato nel caso in cui i valori della variabile si possono ordinare. Consideriamo l'esempio 1. Il relativo diagramma a barre è il seguente:



Il seguente diagramma rappresenta i risultati di un esame per due corsi distinti:



	Insuff.	Suff.	buono	ottimo
Corso 1	35	35	20	10
Corso 2	25	30	15	5

Oppure si potrebbe rappresentare nel seguente modo:

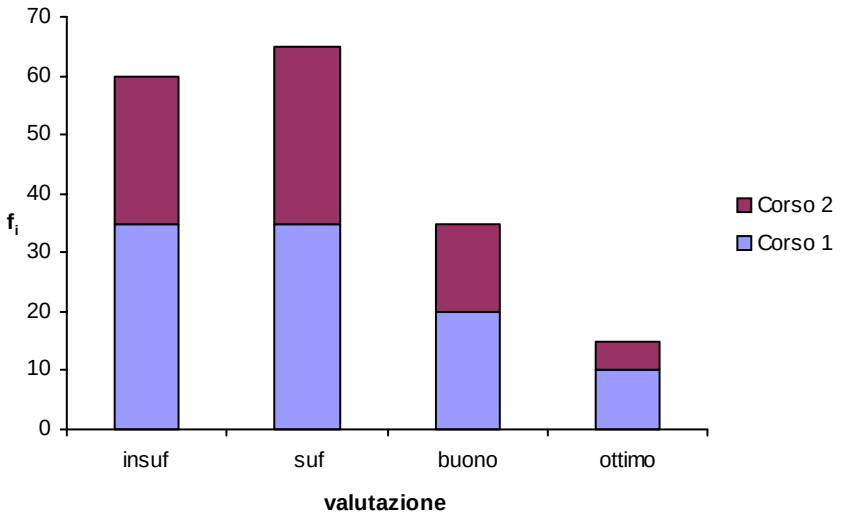
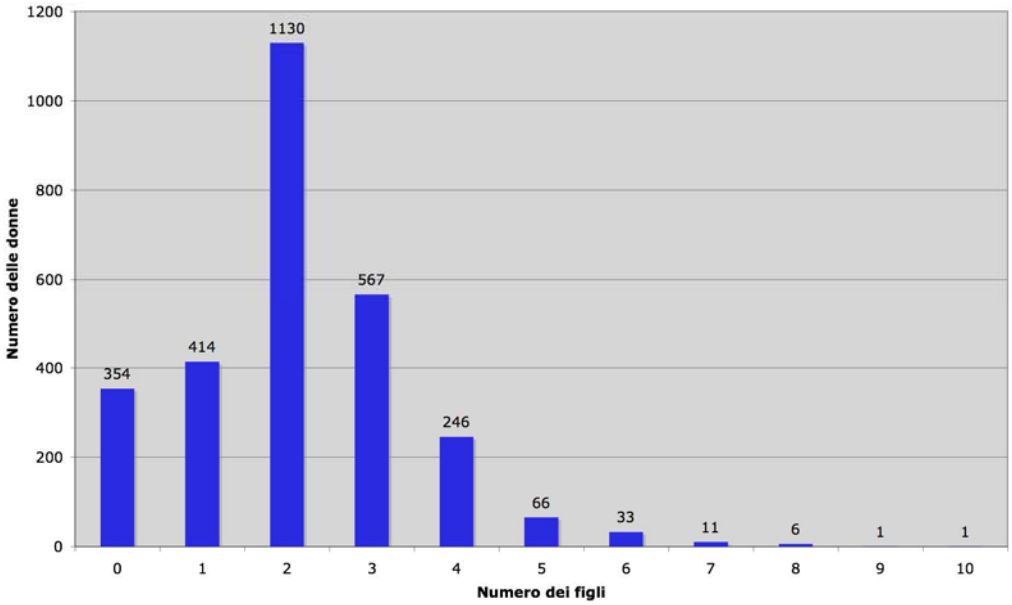


Diagramma relativo all'esempio sul numero dei figli

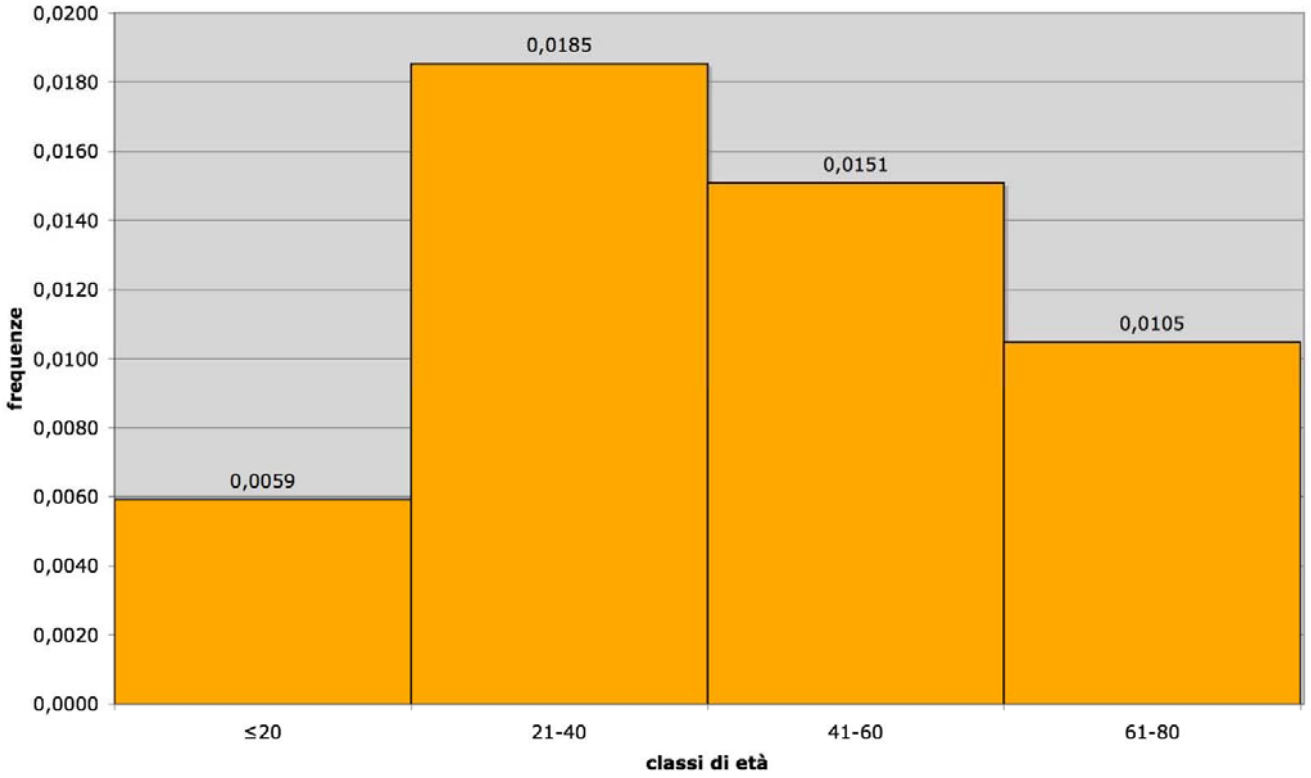


Istogrammi

L'istogramma è costituito mediante rettangoli adiacenti, le cui basi sono gli intervalli che definiscono le classi, e le altezze rappresentano le densità di frequenza, cioè

- F_i = frequenza relativa della classe i / ampiezza classe = p_i / ampiezza
- Le classi sono supposte di uguale ampiezza.

L'istogramma relativo all'esempio precedente



L'area totale dell'istogramma è 1. Infatti l'area di ogni rettangolo è proprio la frequenza relativa della classe e l'area dell'istogramma è la somma delle aree di ciascun rettangolo che lo costituisce, quindi la somma delle frequenze relative, che è uguale a 1.

Nel caso considerato tutte le classi hanno la stessa ampiezza per cui la densità di frequenza risulta proporzionale alla frequenza. In tal caso si potrebbe considerare la frequenza al posto della densità di frequenza.

Indici Sintetici dei dati

Essi danno delle informazioni quantitative sull'ordine di grandezza delle osservazioni (misure di posizione), sulla variabilità delle osservazioni (misure di dispersione, misure di forma). Indichiamo i dati osservati con i seguenti simboli: x_1, x_2, …, x_n.
La misura di posizione localizza il valore centrale di una distribuzione di frequenza. Le più comuni sono la media, la mediana e la moda.

Nel caso di dati continui si prende come valore x_i, i=1,2,…,n, il valore centrale della classe i-esima.

Misure di posizione

Media

- per dati semplici:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + ... + x_n)$$

- per dati ponderati:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r} x_i = \frac{1}{n}(x_1 f_1 + x_2 f_2 + ... + x_r f_r)$$

Osservazione: la media aritmetica sarà indicata con μ quando si riferisce alla popolazione, quando si riferisce al campione sarà indicata con $\bar{x}$

Nella Tabella 2 la media è:

$$\bar{x} = \frac{1}{2829}(0 \cdot 354 + 1 \cdot 414 + 2 \cdot 1130 + 3 \cdot 567 + 4 \cdot 246 + 5 \cdot 66 + 6 \cdot 33 + 7 \cdot 11 + 8 \cdot 6 + 9 \cdot 1 + 10 \cdot 1) = 2{,}13$$

Mediana

- E' il valore che occupa la posizione centrale in un insieme ordinato di dati.
- Non è influenzata dai valori estremi.
- Si usa per attenuare l'effetto dei valori estremi molto alti o molto bassi
- Per calcolare la mediana bisogna ordinare i valori.
- Se il campione ha un numero dispari di valori la mediana è il valore che occupa la posizione centrale.

Per esempio la mediana di 1, 4, 6, 7, 8 è 6. Se il campione ha un numero pari di valori, la mediana è il valor medio dei due valori che occupano la posizione centrale. Per esempio la mediana di 1, 4, 6, 7, 8, 9 è (6+7)/2=6.5.

Confronto tra media e mediana

Esempio: Il numero di giorni di sopravvivenza per i primi sei pazienti che hanno avuto un trapianto di cuore a Stanford sono stati: 15, 3, 46, 623, 126, 64.

Il valor medio è

$$\bar{x} = \frac{15+3+46+623+126+64}{6} = \frac{877}{6} = 146{,}2$$

La mediana è invece (46+64)/2 =55. Infatti i valori centrali della distribuzione ordinata sono 46 e 64 il cui valore medio è 55:

3, 15, 46, 64, 123, 623

Notiamo che il valore molto alto 623 influenza la media, che risulta significativamente più grande della mediana. In tal caso la mediana è un indicatore di posizione migliore della media.

L'esempio sopra dimostra che la mediana non è affetta da poche osservazioni molto basse o molto alte, mentre la presenza ti tali estremi potrebbe avere un effetto considerevole sulla media. Per distribuzioni molto asimmetriche la mediana risulta essere una misura di posizione più sensibile misura del centro della distribuzione rispetto che la media.

Moda

- E' il valore più frequente di una distribuzione
- Nelle distribuzioni di frequenza per dati raggruppati essa è molto sensibile alla modalità di costruzione delle classi.

Misure di dispersione

Due insiemi di dati che hanno dei valori centrali confrontabili potrebbero avere una variabilità molto diversa tra loro. Per studiare una distribuzione di dati le misure di posizione non bastano, ma è necessario anche analizzare come i dati variano rispetto al valore centrale.

Per definire la varianza introduciamo prima la deviazione dalla media:

$$D_i = x_i - \bar{x}, \qquad i = 1, 2, ..., n$$

Naturalmente la somma delle deviazioni dalla media è nulla: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Varianza

Per ottenere una misura di variabilità dei dati dovremmo considerare la deviazione senza segno perchè le parti negative potrebbero compensare quelle negative, e facciamo il quadrato. Otteniamo in tal modo la varianza:

Per i dati semplici

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

Per i dati ponderati

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot f_i$$

Altro modo per calcolare la varianza

Per i dati semplici:

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

Per i dati ponderati:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}f_i \cdot x_i^2 - \bar{x}^2$$

Esempio: Calcolare la varianza nei due modi descritti dei valori 5, 6, 7, 7, 8, 10

$$\bar{x} = \frac{5+6+7+7+8+10}{6} = \frac{43}{6} = 7.17$$

Allora si ha:

I metodo: $s^2 = \frac{1}{6}\left[(5-7.17)^2 + (6-7.17)^2 + (7-7.17)^2 + (7-7.17)^2 + (8-7.17)^2 + (10-7.17)^2\right] = 2.47$

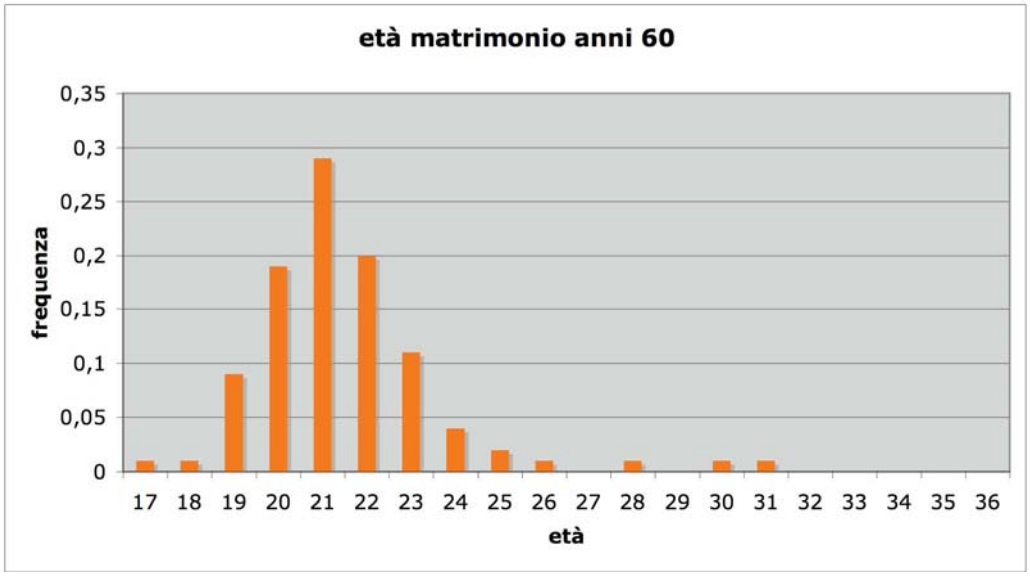
II metodo: $s^2 = \frac{1}{6}\left[5^2 + 6^2 + 7^2 + 7^2 + 8^2 + 10^2\right] - 7.17^2 = 2.42$

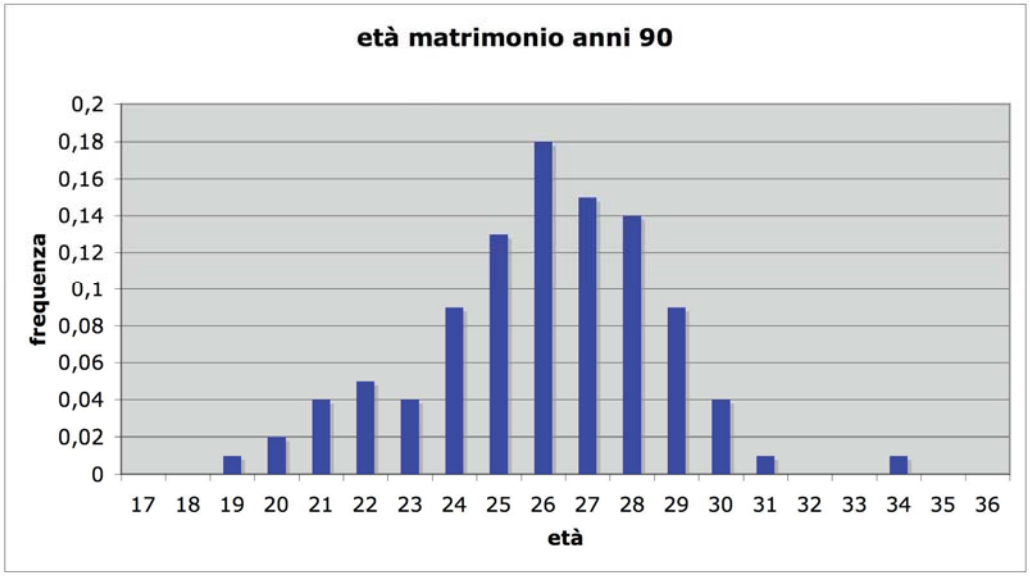
Deviazione standard (o Scarto quadratico medio) del campione

E' la radice quadrata della varianza

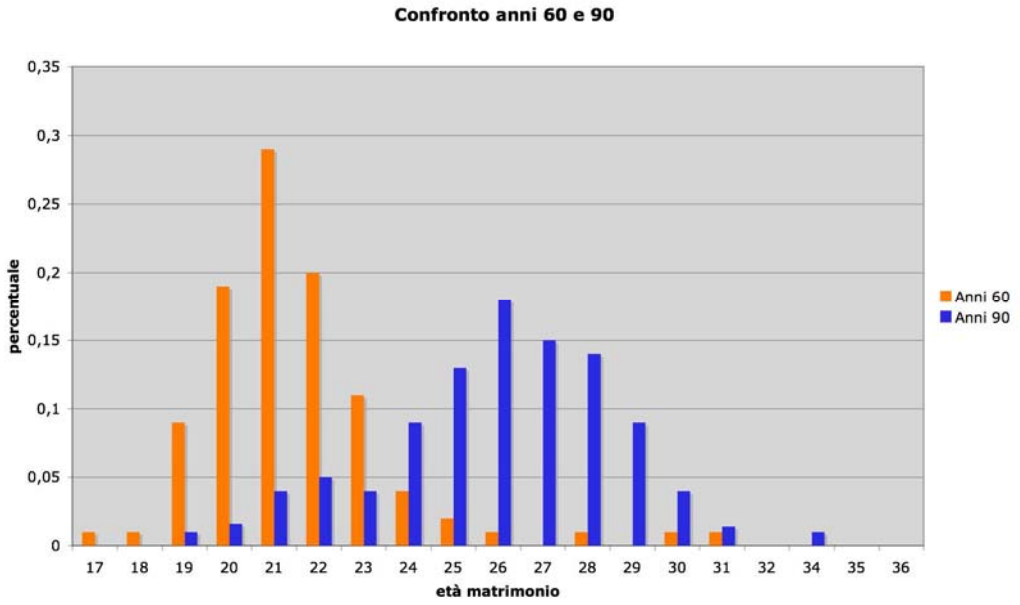
$$s = \sqrt{s^2}$$

Esercizio: Si confronti le due seguenti distribuzioni che rappresentano l'età media del matrimonio negli anni 60 e 90 su un campione. Le frequenze sono espresse in percentuali:





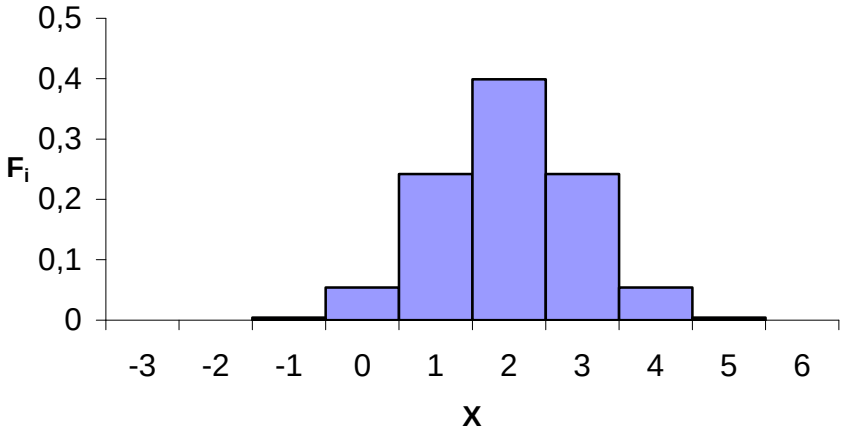
Quale distribuzione ha varianza maggiore? I valori medi sono rispettivamente 21,5 e 26.



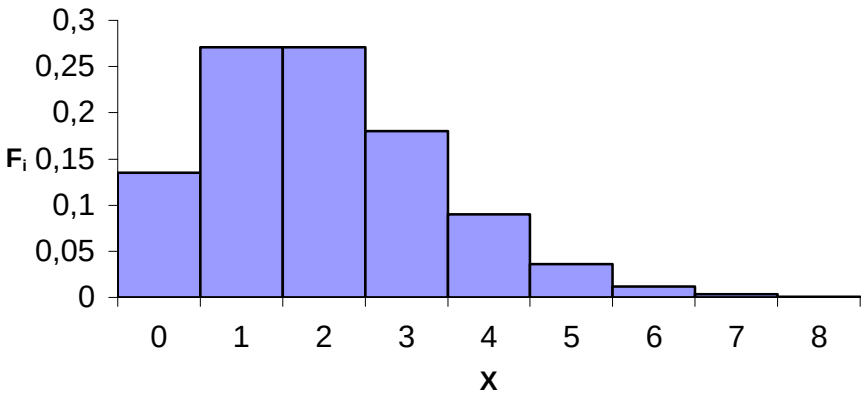
Misure di Forma

Servono per quantizzare due caratteristiche di una distribuzione di frequenza: **Simmetria**.

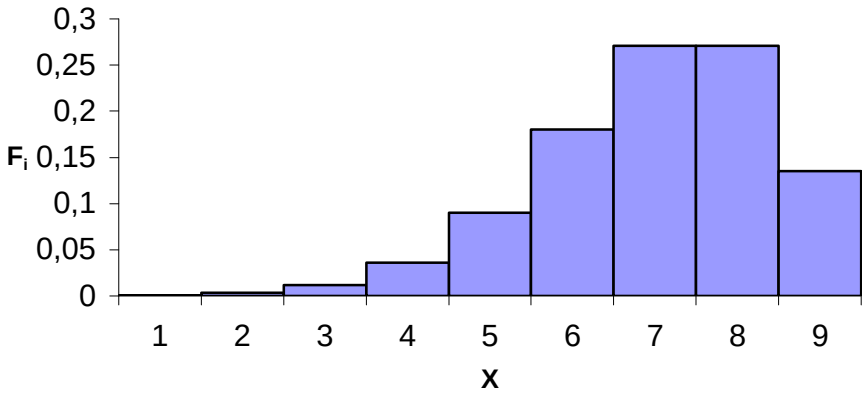
Si ha **Simmetria** se in una distribuzione di frequenza i valori equidistanti dalla media (o mediana) presentano la stessa frequenza. Si ha **Asimmetria Positiva** quando la distribuzione ha una coda verso destra e **Asimmetria Negativa** quando ha una coda verso sinistra.



Esempio di distribuzione simmetrica



Esempio di distribuzione con asimmetria positiva



Esempio di distribuzione con asimmetria negativa

Trasformazione dei dati

Supponiamo di voler confrontare due alunni in base ai voti ottenuti alla fine dell'esame sostenuto in 4 materie da due studenti A e B.
In generale chi dei due ha avuto una resa migliore? Non basta calcolare il voto medio di ciascun alunno. Per un confronto non basta vedere i voti dei singoli studenti, ma abbiamo bisogno di conoscere il voto medio e lo scarto quadratico ottenuto in ciascuna materia.
Supponiamo che:

Materia	Voto A	Voto B
Italiano	7	
Storia	7	7
Filosofia	6,5	6
Matematica		6,5
Fisica		6,6
Inglese	7,5	
	6,9	6,6

Calcolando il voto medio per ogni studente, il risultato risulterebbe sbilanciato a favore di chi ha sostenuto l'esame di inglese. Ora vogliamo trovare un indice, per ciascuna materia, che ci indichi la posizione relativa di ogni studente di quella materia. Per fare ciò dobbiamo standardizzare i dati.

$$Z_i = \frac{X_i - \overline{X}}{s}$$

dove

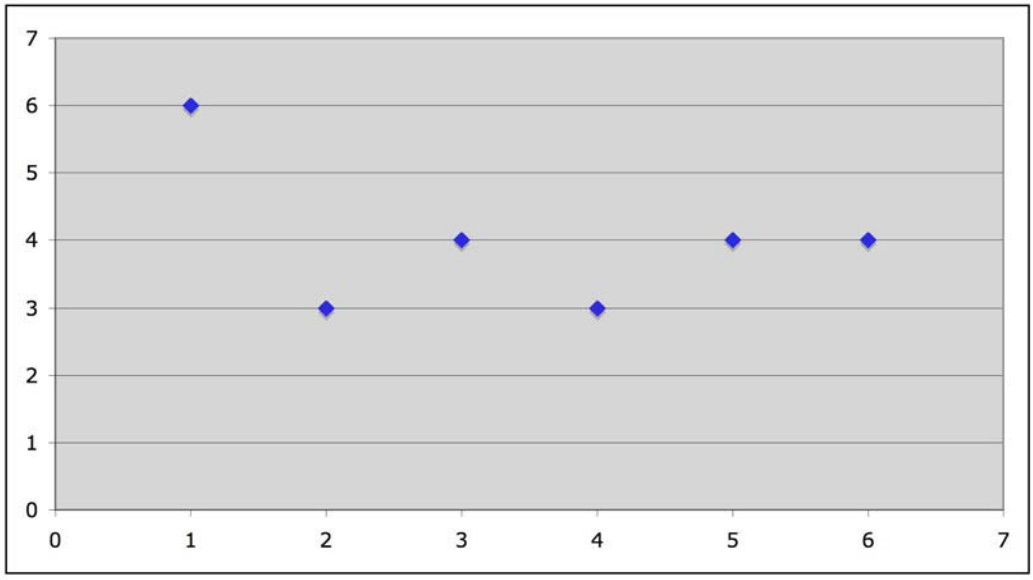
- Xi è la i-esima osservazione, i=1,2,…,n
- $\overline{X}$ è il valor medio di esse
- s è lo scarto quadratico medio di esse

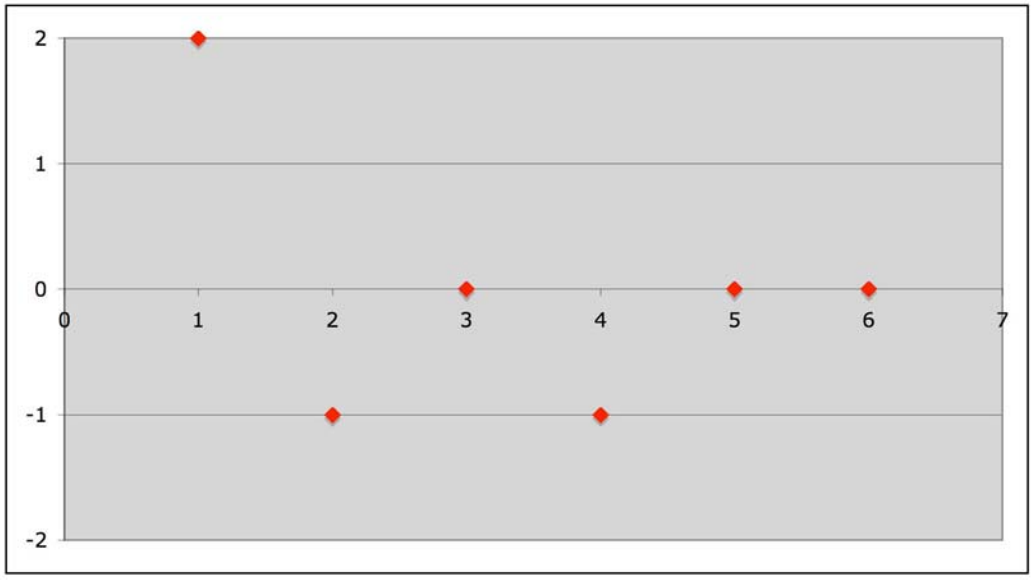
Osserviamo che l'operazione di sottrarre una stessa costante trasforma I dati in nuovi dati con media pari alla media originale meno la costante, lasciando inalterata lo scarto quadratico medio. In particolare se si sottrae la media si ottiene una serie di dati con media nulla.

Esempio: Supponiamo che 5 atlete dopo un allenamento abbiano perso ciascuna 5 chili:

Peso prima	Peso dopo
50	45
54	49
58	53
61	56
62	57
Media 57	52
Dev. Stand. 5	5

Esempio: Dati I numeri 6,3,4,3,4,4, il loro valor medio è 4. Sottraendo la media otteniamo 2,-1,0,-1,0,0 che ha media nulla





Supponiamo ora di moltiplicare ciascun dato per una stessa costante. In tal caso la media sarà moltiplicata per la stessa costante e lo stesso accade per il valor medio.

Consideriamo le 5 atlete che dopo l'allenamento hanno perso un decimo del proprio peso:

Peso prima	**Peso dopo**
50	45
54	48,6
58	52,2
61	54,9
62	55,8
Media 57	51,3
Dev. Stand. 5	4,5

dove

$5,9 = 0,9 \cdot 57$

$4,5 = 0,9 \cdot 5$

Tornando all'esempio degli studenti, supponiamo di avere

Materia	**media**	**scarto quadratico**
Italiano	7	0,1
Storia	7	0,2
Filosofia	6	0,5
Matematica	6	0,9
Fisica	5,8	1,1
Inglese	7,8	0,84

se ora standardizziamo i dati di partenza si ottiene

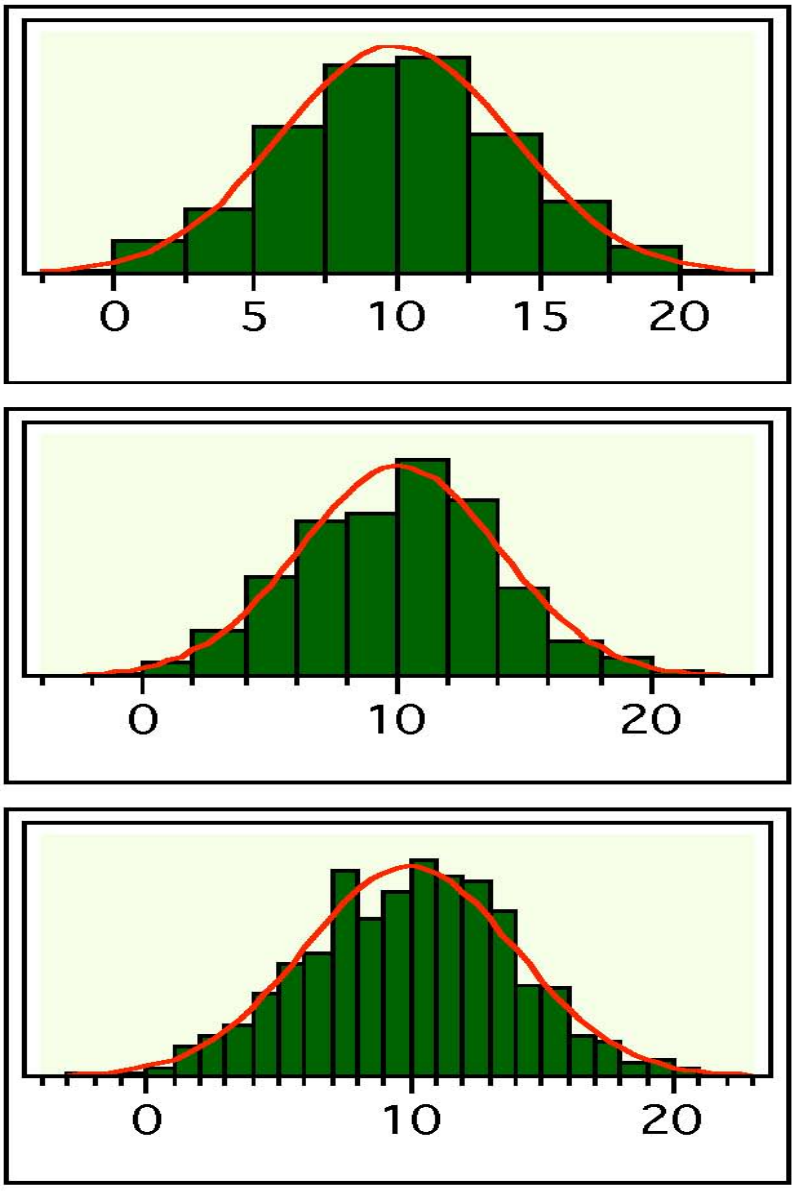
Materia	Voto norm A	Voto norm B
Italiano	0	
Storia	0	0
Filosofia	1	0
Matematica		0,556
Fisica		0,727
Inglese	-0,357	
	0,16	0,32

e quindi la media dei voti standardizzati risulta essere maggiore per lo studente B, contrariamente alla media di ogni studente dei voti non standardizzati.

Quando il voto standardizzato viene 0, ciò vuol dire che lo studente sta sulla media, se viene 1 si trova ad una scarto sopra la media e così via.

Distribuzione continue

Nel caso discreto abbiamo costruito per ogni campione e per una distribuzione discreta il relativo istogramma. In ogni caso l'area sottesa dall'istogramma è uguale a 1. Se infittissimo le classi dell'istogramma aumentando i dati raccolti otterremmo degli istogrammi con più classi, e andando avanti con questa procedura otterremmo che l'istogramma si avvicina sempre di più ad una curva continua



Data una successione di misurazioni fatte nelle stesse condizioni (o indipendenti) $\{X_n\}$ e consideriamo la media aritmetica o media campionaria

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

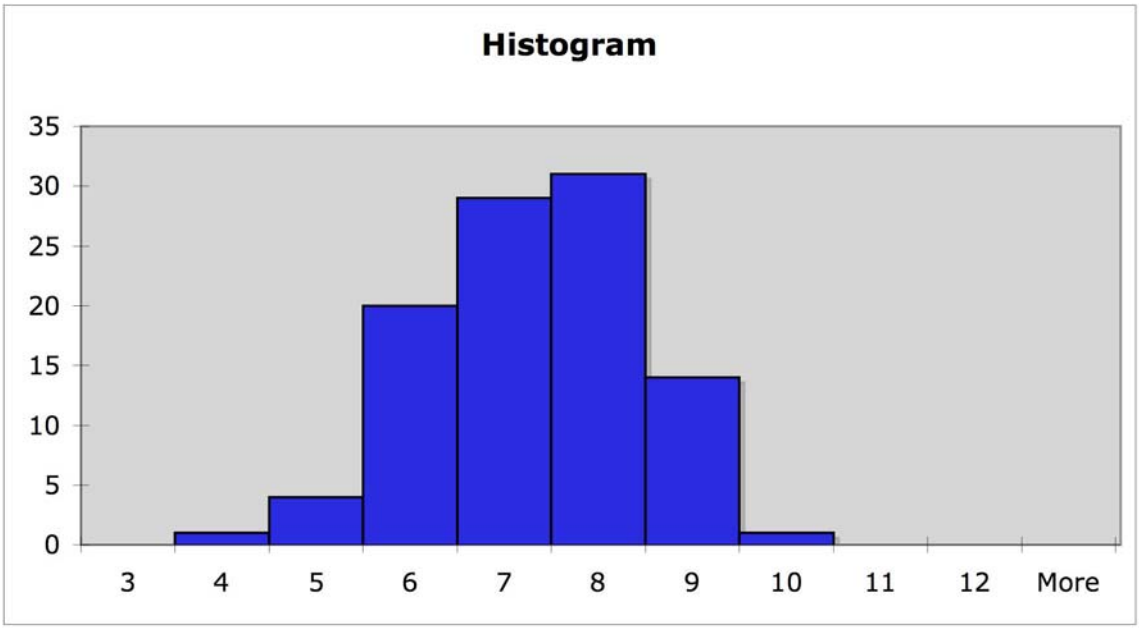
Se I campioni sono effettivamente significativi, la maggior parte di essi avrà un valore medio molto vicino a quello della popolazione. Tuttavia ci sarà sempre un certo numero di campioni che avrà una media distante dal valor medio della popolazione.
La distribuzione delle medie campionarie segue, grosso modo, la distribuzione normale.
La media delle medie campionarie è, grosso modo, uguale al valor medio della popolazione.

Esempio: Si lanci due dadi e sia X la variabile aleatoria che misura la somma dei risultati. Supponiamo di lanciare 5 volte la coppia di dadi per 100 volte e otteniamo i seguenti risultati:

X_1	X_2	X_3	X_4	X_5	media
5	7	5	7	5	5,8
8	4	9	11	7	7,8
6	10	8	8	7	7,8
6	5	6	7	6	6
7	12	8	5	9	8,2
9	9	4	5	4	6,2
6	3	10	6	4	5,8
5	7	8	5	7	6,4
8	10	7	7	2	6,8
11	3	5	9	9	7,4
6	6	3	5	7	5,4
7	7	10	3	10	7,4
6	3	7	6	8	6
10	7	9	9	4	7,8
6	6	11	12	10	9
7	7	3	8	5	6
9	11	3	11	10	8,8
6	10	8	3	11	7,6
6	2	6	4	8	5,2
2	11	8	7	7	7
9	8	10	4	12	8,6
4	8	5	7	6	6
8	5	6	5	5	5,8
12	5	7	8	11	8,6
7	10	6	11	6	8
9	9	8	7	9	8,4
4	3	2	6	3	3,6
8	4	6	7	8	6,6
5	6	8	8	4	6,2
7	8	4	11	8	7,6
4	6	6	5	6	5,4

11	7	11	8	5	8,4
5	8	9	4	6	6,4
7	5	6	5	6	5,8
7	8	6	6	2	5,8
8	10	4	11	7	8
6	10	8	10	8	8,4
5	5	7	3	7	5,4
6	2	6	6	4	4,8
7	3	6	9	9	6,8
12	10	9	7	7	9
6	5	7	8	7	6,6
8	7	11	11	5	8,4
7	5	6	6	7	6,2
6	7	12	3	4	6,4
11	7	6	10	6	8
11	5	6	6	4	6,4
6	4	8	4	7	5,8
7	8	8	5	11	7,8
10	6	5	8	7	7,2
8	4	8	4	7	6,2
7	8	9	4	7	7
11	4	9	7	7	7,6
6	9	6	3	11	7
11	7	9	12	4	8,6
7	3	7	5	7	5,8
9	3	9	6	6	6,6
8	9	7	7	4	7
9	7	5	8	6	7
7	9	4	9	6	7
6	12	6	11	5	8
12	7	6	8	8	8,2
8	10	11	6	7	8,4
7	5	9	9	7	7,4
6	6	9	7	9	7,4
4	8	6	7	9	6,8
5	8	4	5	8	6
6	8	8	4	10	7,2
6	6	6	6	4	5,6
10	6	7	4	3	6
8	11	10	9	8	9,2
5	6	9	7	10	7,4
8	5	3	6	3	5
9	6	8	9	7	7,8
5	7	10	7	8	7,4
8	7	8	5	12	8

8	7	9	6	6	7,2
5	8	6	9	4	6,4
6	7	7	8	11	7,8
7	5	7	9	10	7,6
4	12	9	6	8	7,8
4	6	7	9	3	5,8
6	8	9	4	7	6,8
7	11	6	6	5	7
4	4	5	7	4	4,8
9	5	10	5	7	7,2
7	8	8	6	4	6,6
4	6	8	3	7	5,6
9	6	10	10	5	8
3	5	7	8	8	6,2
10	10	11	3	4	7,6
6	6	10	6	6	6,8
6	7	6	10	5	6,8
8	3	8	9	7	7
7	5	5	5	2	4,8
8	7	9	3	5	6,4
7	8	8	10	11	8,8
12	9	4	7	5	7,4
8	11	9	5	5	7,6
10	2	10	8	9	7,8



Statistica descrittiva	
Mean	6,96
Standard Error	0,11
Median	7,00
Mode	5,80
Standard Deviation	1,11
Sample Variance	1,24
Kurtosis	-0,25
Skewness	-0,25
Range	5,60
Minimum	3,60
Maximum	9,20
Sum	696,20
Count	100,00

Lo scarto quadratico medio delle medie campionarie, detto errore standard si dimostra essere uguale a $\sigma/\sqrt{n}$ dove σ è lo scarto quadratico medio della popolazione.

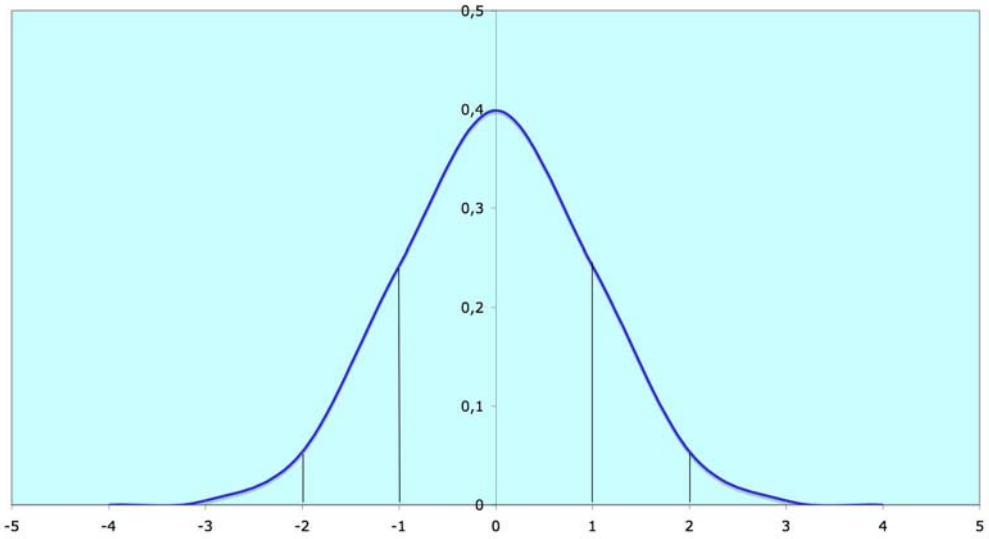
Un risultato fondamentale nella Teoria della Probabilità noto come **Teorema del Limite Centrale** stabilisce che se si estraggono da una popolazione campioni di dimensione n, la media campionaria, grosso modo, sarà distribuita normalmente con valore medio uguale alla media della popolazione e errore standard $\sigma/\sqrt{n}$.

Se l'ampiezza del campione è sufficientemente grande ($n \geq 30$) allora qualunque sia la distribuzione della successione la distribuzione della somma standardizzata campionaria può essere ben approssimata dalla distribuzione di Gauss standardizzata.

Distribuzione di Gauss

La **distribuzione di Gauss** o curva degli errori è la più importante distribuzione continua, è stata proposta da Gauss (1809) nell'ambito della teoria degli errori, è stata attribuita anche a Laplace (1812), che ne definì le proprietà principali in anticipo rispetto alla trattazione più completa di Gauss. Il nome deriva dalla convinzione che i fenomeni fisico-biologici solitamente si distribuiscono con frequenze più elevate nei valori centrali e frequenze progressivamente minori verso gli estremi, in quanto la distribuzione degli errori commessi nel misurare ripetutamente la stessa grandezza, è molto bene approssimata da questa curva.

Fig.1 Distribuzione di Gauss standard



Proprietà:

- media μ
- varianza σ^2
- è simmetrica rispetto alla media
- ha media, moda e mediana coincidenti (e pari μ)
- non raggiunge mai lo zero per ogni valore di x.

Una curva normale può essere definite in maniera univoca dal suo valore medio e dallo scarto quadratico medio.

Le curve normali hanno la stessa forma di base, sebbene possano essere alte e sottili oppure basse e larghe. Cambiando la media la forma della configurazione non cambia ma si sposta il massimo.

Figura 1b: distr. normale con stessa Dev. Stand. =1

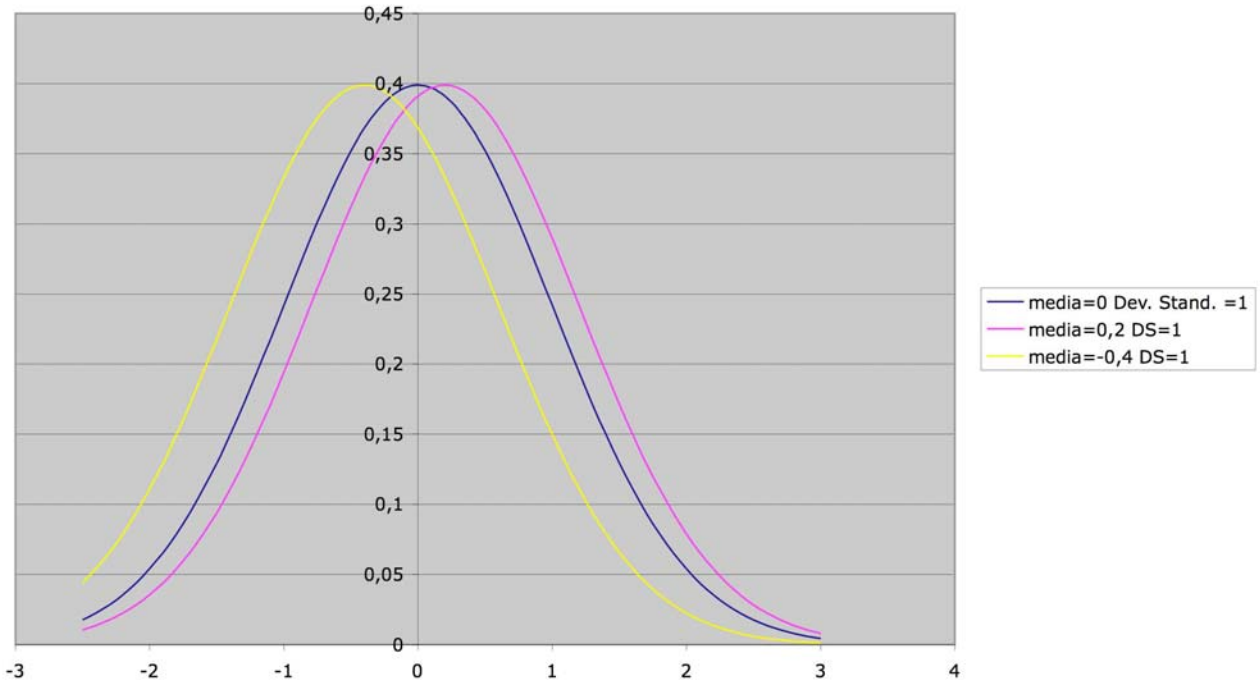
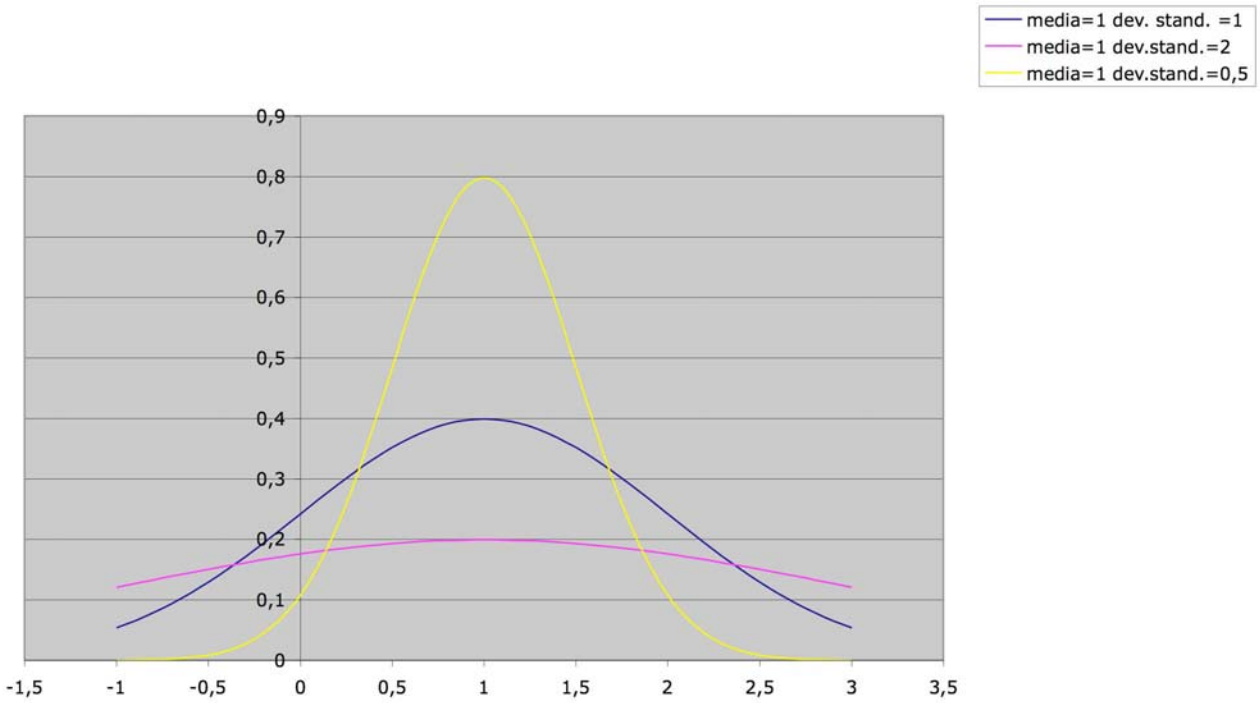


Figura 1 a: distribuzione normale con stessa media =1



A differenza del caso precedente cambiando solo la deviazione standard la forma della curva

cambia diventando più “appiattita” se essa aumenta, più “piccata” se diminuisce. La posizione del suo centro non cambia.

Ad ogni distribuzione continua si associa la funzione densità di probabilità f(x) che corrisponde alla distribuzione di probabilità nel caso discreto. Soddisfa le seguenti proprietà

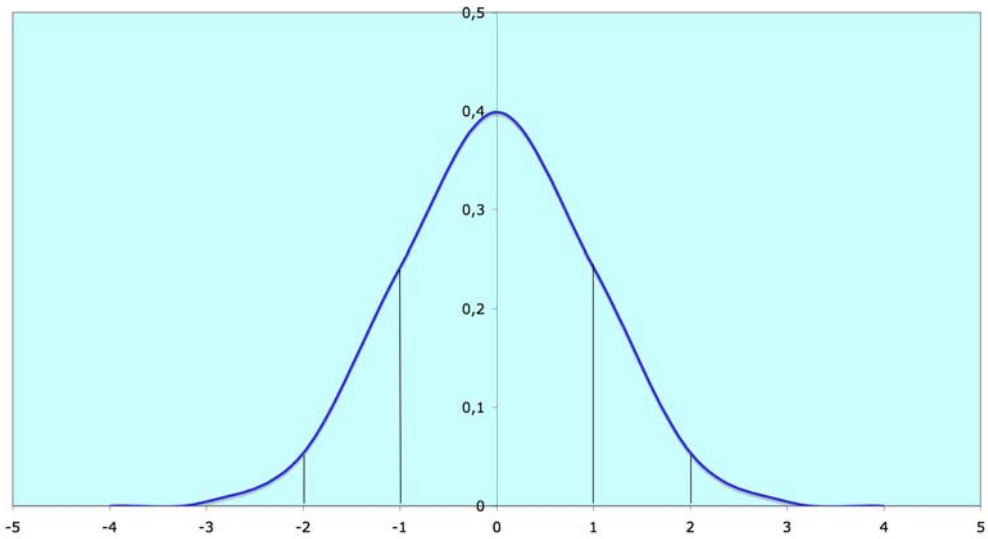
- $f(x) \geq 0$ per ogni x
- l’area totale della densità di probabilità è 1
- l’area sottesa dalla curva normale nell’intervallo (a,b) è: $P(a \leq x \leq b) = \int_a^b f(x)\,dx$

Nel caso della distribuzione di Gauss

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

In particolare:
$P(\mu-\sigma,\mu+\sigma) = 0.6827$
$P(\mu-2\sigma,\mu+2\sigma) = 0.9545$
$P(\mu-3\sigma,\mu+3\sigma) = 0.9973$

Fig.1 Distribuzione di Gauss standard



Inferenza statistica

Introduzione

L'obiettivo dell' indagine statistica è ottenere un campione il più possibile rappresentativo della popolazione, cosicché le informazioni sulle caratteristiche della popolazione che da essa si traggono siano il più possibile accurate. Si cerca dalle osservazioni particolari di un campione di comprendere il caso generale, come la variabilità dei dati nella popolazione è trasmessa nel campione attraverso la sua media.

In generale siamo interessati a conoscere qualche grandezza numerica della popolazione, ad esempio la media o la deviazione standard. Tale grandezza si chiama **parametro**. Il valore reale di un particolare parametro della popolazione è sconosciuto e può essere determinato solo dopo un'analisi su tutta la popolazione (ad esempio la media della popolazione). Se ciò è impossibile o non praticabile (per problemi di tempo, economici, ecc.) allora si utilizza l'inferenza statistica.

La **Stima dei parametri** è il procedimento con cui dal campione osservato si traggono informazioni per assegnare al parametro un valore (**stima puntuale**) o un insieme di valori (**stima per intervallo**).

Osservazioni

1. Poiché il campione è una parte della popolazione il valore della statistica non può dare l'esatto valore del parametro
2. Il valore della statistica dipende dal particolare campione selezionato.
3. Esiste una variabilità nei valori della statistica su differenti modi di campionamento.

Inferenza su una media: intervallo di fiducia

Vorremo stimare la media della popolazione sconosciuta attraverso un campione, calcolando a partire dal campione un intervallo che potrebbe contenere la media cercata con un certo grado di credibilità.

Varianza non nota Sia $X_1, \ldots, X_n$ un campione estratto da una popolazione con media incognita μ e varianza non nota σ^2. Allora in questo caso il parametro σ potrebbe essere approssimato da s', dove s' è la deviazione standard del campione con la correzione di Student, cioè

$$s' = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

Si ottiene in questo modo una distribuzione che si avvicina a Gauss purché n > 100.

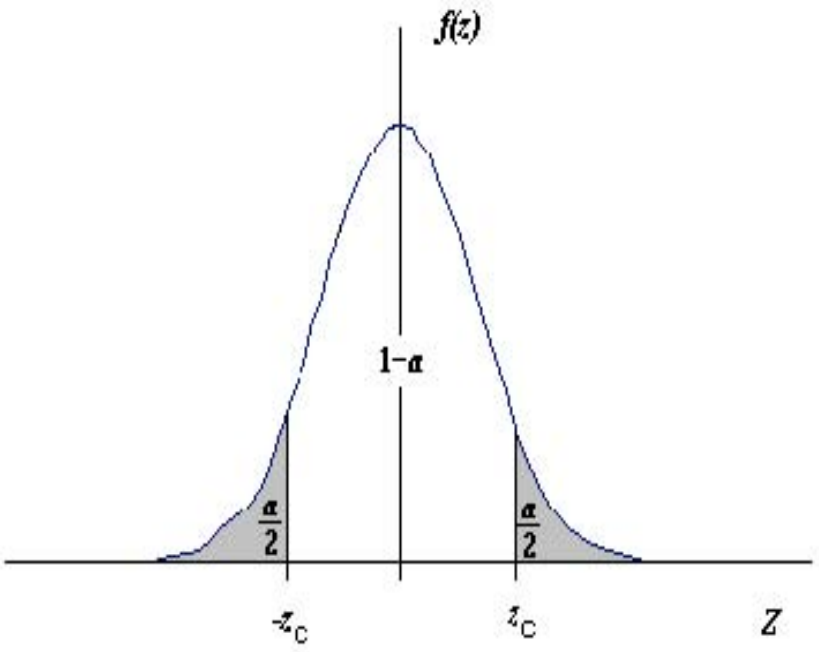
$$Z = \frac{\overline{X} - \mu}{s'/\sqrt{n}}$$

Come può essere una buona stima del valor medio di un qualunque campione, quando sappiamo che, se prendiamo diversi campioni della stessa popolazione, otterremo sempre un valore diverso?

Sia $X_1, \ldots, X_n$ un campione estratto da una popolazione con media incognita μ e varianza non nota σ^2. Vogliamo trovare dei valori $(-z_c, z_c)$ tali che Z sia compresa nell' intervallo $(-z_c, z_c)$ con una alta probabilità a (ad esempio 1-a = 0,95), cioè

$$P(-z_c \leq Z \leq z_c) = 0{,}95$$

Dalle tavole di Gauss si vede che tale valore è esattamente $z_c = 1{,}96$



Sostituendo a Z l'espressione di sopra si ottiene

$$P\left(-1{,}96 \leq \frac{\bar{X} - \mu}{s'/\sqrt{n}} \leq 1{,}96\right) = 0{,}95$$

Applicando le proprietà sulle disuguaglianze si ottiene

$$P\left(-1{,}96 \cdot s'/\sqrt{n} \leq \bar{X} - \mu \leq 1{,}96 \cdot s'/\sqrt{n}\right) = 0{,}95$$

e quindi

$$P\left(-1{,}96 \cdot s'/\sqrt{n} - \bar{X} \leq -\mu \leq 1{,}96 \cdot s'/\sqrt{n} - \bar{X}\right) = 0{,}95$$

applicando di nuovo le proprietà delle disuguaglianze si ha

$$P\left(\bar{X}-1,96\cdot {s'}/{\sqrt{n}}\le \mu \le \bar{X}+1,96\cdot {s'}/{\sqrt{n}}\right)=0,95$$

da cui si ottiene che l'intervallo di fiducia per la media con un grado di fiducia del 95% è

$$\left(\bar{X}-1,96\cdot {s'}/{\sqrt{n}},\bar{X}+1,96\cdot {s'}/{\sqrt{n}}\right)$$

Se conosciamo anche la varianza della popolazione, allora si ottiene il corrispondente intervallo di fiducia

$$\left(\bar{X}-z_c\cdot {\sigma}/{\sqrt{n}},\bar{X}+z_c\cdot {\sigma}/{\sqrt{n}}\right)$$

L'errore o livello di fiducia nel caso generale ha la seguente espressione

$$E=z_c\cdot\frac{\sigma}{\sqrt{n}}$$

Tale errore aumenta anche all'aumentare del grado di fiducia e diminuisce al crescere dell'ampiezza del campione ed è inversamente proporzionale alla variabilità dei dati.. Se volessimo fissare un errore a priori, conoscendo la varianza della popolazione, siamo in grado di trovare quanto deve essere ampio il campione in modo da avere un errore non più grande di quello fissato.

Esempio: Un urbanista é interessato alla superficie media μ delle abitazioni della propria città. Uno studio precedente indica che la deviazione standard della popolazione sia circa $8m^2$. In un campione di 50 appartamenti si osserva la media del campione è pari a $120m^2$. Sulla base di questi dati l'intervallo di confidenza per μ con un grado di fiducia del 95% è

$$\left(120-1,96\cdot {8}/{\sqrt{50}},120+1,96\cdot {8}/{\sqrt{50}}\right)$$

Se volessimo trovare quanto deve essere grande il campione in modo da avere un errore non più grande di 1, per risolvere il problema basterebbe considerare la disequazione

$$|E \le 1 \Rightarrow z_c\cdot\frac{\sigma}{\sqrt{n}}\le 1\Rightarrow$$
$$1,96\cdot\frac{8}{\sqrt{n}}\le 1\Rightarrow$$
$$\sqrt{n}\ge 8\cdot 1,96\Rightarrow$$
$$n\ge 15,68^2\Rightarrow$$
$$n\ge [245,86]$$

Quindi basterebbe scegliere n = 246 per ottenere un errore non più grande di 1.

Esempio: Il produttore di una certa marca di sigarette desidera controllare il quantitativo di catrame in esse contenuto. A questo scopo si osserva un campione di 30 sigarette in cui la media è 10.92 mg e la deviazione standard 0.50 mg . Sulla base di questi dati l' intervallo di fiducia per la media pari al 99%

$$\left(10,92-2,756\cdot 0,51/\sqrt{30},10,92+2,756\cdot 0,51/\sqrt{30}\right)\approx(10,66;11,18)$$

Il valore t_c è stato trovato considerando la 29-esima riga (gradi di libertà) e $2\gamma=0,01$ (somma delle due code)

Per ulteriori approfondimenti

- Ian Diamond, J. Jefferies, Introduzione alla statistica per le scienze sociali, McGraw-Hill.
- Bergamini Massimo, Trifone Anna, Barozzi Graziella, Moduli blu di matematica. Modulo delta: Inferenza statistica. Per le Scuole superiori, Zanichelli.
- Fraschini Marzia, Grazzi Gabriella, Probabilità e statistica. Per le Scuole superiori, AtlasQ

Per le applicazioni

- M. Middleton, Analisi statistica con Excel, ApogeoM. Middleton, Analisi statistica con Excel, Apogeo