Introduzione alla statistica

Scheda introduttiva

Nell'esperienza quotidiana siamo posti di fronte a molteplici decisioni da prendere: per esempio decidere come investire i nostri risparmi, se acquistare un'automobile in contanti o a leasing, ...

Mentre alcune scelte vengono basate su un semplice ragionamento logico, altre richiedono la disponibilità di precise informazioni e la capacità di interpretarle correttamente. Per prendere decisioni corrette è necessario disporre dei dati relativi alla scelta da compiere: ma i dati grezzi spesso non ci rivelano immediatamente il loro vero significato.

La statistica è uno strumento fondamentale per il supporto alle decisioni, in ogni settore applicativo. Non basta infatti disporre di semplici dati per fare le scelte giuste: i dati vanno raccolti, analizzati ed elaborati con strumenti adatti (per esempio tabelle e grafici). Essi vanno poi interpretati e valutati con gli opportuni metodi statistici

$$\text{Dati} \quad + \quad \text{Metodi statistici} \quad = \quad \text{Informazioni}$$

Come si svolge un'indagine statistica

1. Definire un tema (situazione, problema,...). Individuare con precisione l'obiettivo che l'indagine si propone di raggiungere, definendo con accuratezza i termini del problema a cui bisogna dare risposta (per esempio: analizzare l'afflusso di clienti in un negozio secondo gli orari per determinarne in seguito l'orario di apertura/chiusura e la presenza di personale in certe fasce orarie; il legame tra produzione industriale e consumo di energia elettrica).

2. Definire le variabili che ci interessano in maniera da poter individuare, senza possibilità di equivoco, i valori che esse assumono nelle singole unità.

3. Fissare metodi (su tutta la popolazione / su un campione), mezzi (interviste, questionari, misurazioni, osservazioni, ...) e tempi entro i quali effettuare il rilevamento e l'elaborazione dei dati.

4. Rilevazione dei dati secondo il piano di lavoro deciso.

5. Spoglio dei dati e loro sistemazione in forme di facile lettura quali tabelle e grafici.

6. Elaborazione dei dati: mediante osservazione grafica ed operazioni matematiche si sintetizzano i risultati e si dà un'idea concreta della ripartizione dei caratteri rilevati.Analisi dei dati tramite:

   - indici di centralità: media (media ponderata), mediana, moda
   - indici di dispersione e distribuzione: campo di variazione, scarto quadratico medio, quartili, centili.

7. Interpretare i dati dando un giudizio di merito sul significato dei risultati utile per sviluppare nuovi approfondimenti ed ipotesi.

Dati statistici

La statistica induttiva e la statistica descrittiva

Immagina di parlare con uno sconosciuto e di raccogliere informazioni sulle sue abitudini, sui suoi gusti, sul suo stato di salute. Potresti dedurre un ritratto significativo di questa persona. Se raccogliessi le stesse informazioni per un gruppo di persone, diciamo mille, e ti accorgessi che alcune risposte si assomigliano e altre differiscono completamente le une dalle altre, cosa potresti dedurne? Potresti fare, in qualche modo, un ritratto di gruppo?

A volte anche molte informazioni possono essere inutili, se non sono ben organizzate. In tal caso può essere utile raggruppare e sintetizzare i dati: in questo modo si rinuncia a parte dell'informazione che essi contengono, ma si guadagna in leggibilità e facilità di interpretazione. In particolare si possono elaborare tanti dati relativi a individui singoli per trarne informazioni sulla popolazione nel suo complesso. A seconda poi di come questi dati vengono raggruppati è possibile studiare aspetti diversi del problema in esame.

La statistica si occupa proprio dei modi di raccogliere e analizzare dati relativi a un certo gruppo di persone (gli studenti di una scuola, gli abitanti di un quartiere, gli elettori di una regione, ...) o di oggetti (le automobili, i dischi, i libri, ...), per trarne conclusioni e fare previsioni. Il gruppo preso in considerazione viene anche detto popolazione. Spesso viene presa in esame soltanto una parte della popolazione, detta campione, scelta in modo che rappresenti l'intero gruppo. Per esempio, per conoscere il parere dei telespettatori su un certo programma, si può intervistare soltanto un piccolo numero di essi, che si ritenga però un campione rappresentativo. Dalle osservazioni relative al campione possono essere tratte conclusioni valide per tutta la popolazione. I metodi per ottenere risultati soddisfacenti in questo delicato procedimento di passaggio dal campione alla popolazione sono studiati da quella parte della statistica detta statistica induttiva (o inferenza statistica).

Noi non ci occuperemo di statistica induttiva ma ci limiteremo invece a studiare alcuni degli strumenti matematici utilizzati per descrivere i dati relativi a un certo gruppo. In questo caso si parla di statistica descrittiva.

I caratteri qualitativi e i caratteri quantitativi

Gli elementi di una popolazione si chiamano anche unità statistiche. È possibile studiare diverse caratteristiche di tali unità e ogni caratteristica rappresenta un carattere della popolazione.

I caratteri possono essere di due tipi:

- qualitativi se vangono descritti con parole
- quantitativi se invece vengono descritti mediante numeri.

Per esempio, se scegliamo come unità statistiche gli studenti di una scuola, alcuni caratteri qualitativi sono il sesso, il paese di provenienza, il mezzo di trasporto usato per raggiungere la scuola; sono invece caratteri quantitativi l'etäààà, il peso, la statura.

Ogni carattere viene descritto mediante le modalità con cui esso si può manifestare.

Esempio:

1. Il carattere sesso ha due modalità: maschile e femminile.
2. Il carattere mezzo di trasporto ha più modalità: treno, autobus, motorino, ...
3. Anche il carattere età ha più modalità: 14, 15, 16, ...(se espresso in anni).

Dai censimenti ai sondaggi d'opinione

L'utilizzazione di dati statistici per ottenere informazioni utili per il governo degli stati, quali il numero di abitanti, di soldati, di addetti ai vari mestieri,... risale ai popoli antichi, in particolare ai Cinesi e agli Egizi.

Nella Bibbia sono descritti più censimenti fra gli Ebrei, tra i quali il più noto è quello di Mosé nel deserto del Sinai. Anche i Romani fecero diversi censimenti: famoso quello durante il quale nacque Gesù.

Un passo avanti nella elaborazione statistica si ebbe in Inghilterra, intorno alla metà del 1600, con l' "aritmetica politica", principalmente a opera del matematico John Graunt. A causa delle pestilenze, a Londra venivano pubblicate settimanalmente le liste delle morti e quelle delle nascite. Graunt utilizzò quel materiale osservando, attraverso il calcolo di percentuali, regolarità quali il maggior numero di nascite maschili rispetto a quelle femminili, il legame fra suicidi e professioni, la diminuzione delle nascite nei periodi di carestia. Era la prima volta che venivano cercate delle relazioni tra i dati raccolti.

Un ulteriore momento importante nella storia della statistica si ebbe quando, nell'Ottocento, si trovò un collegamento con la probabilità.

Infine è dell' ultimo secolo uno sviluppo sempre più ampio della statistica come scienza matematica a sé stante. L'applicazione di tale scienza, mediante indagini a campione, investe i campi più svariati, dai fenomeni sociali a quelli meteorologici.

Le tabelle di frequenza

Esempio:

In un questionario abbiamo chiesto ai 28 studenti di una classe di indicare con le seguenti lettere i mezzi di trasporto con cui vengono di solito a scuola:

A: automobile;

P: a piedi;

B: autobus o pullman;

M: motorino o scooter;

C: bicicletta.

Abbiamo ottenuto i seguenti risultati:

A, B, M, M, P, A, A, B, P, B, C, A, B, B, B, C, P, B, A, C, C, A, M, B, M, B, A, C.

Dalla lettura di questa sequenza, è difficile trarre informazioni, perché i risultati si susseguono in modo disordinato.

Costruiamo allora una tabella, dove nella prima colonna mettiamo le diverse modalità. percorriamo poi la sequenza dei risultati facendo un segno, per esempio una barra /, di fianco alle diverse modalità ogni volta che esse si verificano. Alla fine contiamo il numero di segni per ogni modalità e lo scriviamo nella terza colonna. Tale numero rappresenta la frequenza della modalità considerata. L'automobile ha una frequenza 7, la bicicletta una frequenza 5 e così via.

Osservazione: spesso la frequenza prende il nome di **frequenza assoluta**.

Tabella: distribuzione di frequenza delle modalità

Modalità	Occorrenze	Frequenza
Automobile	///// //	7
A piedi	///	3
Autobus/pullman	///// ////	9
Motorino/scooter	////	4
Bicicletta	/////	5
Totale		28

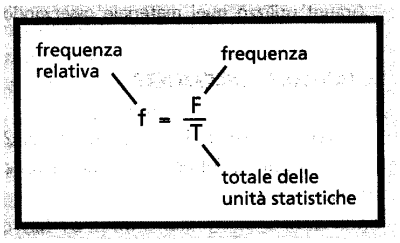
Più spesso interessa il valore della frequenza confrontato con il numero totale delle unità statistiche.

Infatti siamo in situazioni diverse se, per esempio, la frequenza di una modalità è 7 rispetto a un totale di 28 o se, invece, è 7 rispetto a un totale di 280.

Per questo motivo viene calcolata la frequenza relativa, di cui diamo la definizione.

Definizione:

La **frequenza relativa** di una modalità è il quoziente fra la frequenza della modalità e il numero totale delle unità statistiche.



Nell'esempio precedente la frequenza della modalità automobile è 7, ossia 7 studenti su 28 sono accompagnati in automobile; pertanto la frequenza relativa è:

$$f_A = \frac{7}{28} = \frac{1}{4} = 0,25$$

La frequenza relativa può essere espressa anche in **percentuale**, moltiplicandola per 100: la frequenza percentuale della modalità automobile è 25%. Questo significa che, in una distribuzione con le stesse caratteristiche di quella data, su un campione di 100 studenti 25 verrebbero in automobile.

La tabella seguente riassume le frequenze relative delle diverse modalità del precedente esempio:

Modalità	Frequenza	Frequenza relativa	Frequenza relativa percentuale
Automobile	7	$\frac{1}{4}$	25%
A piedi	3	$\frac{3}{28}$	11%
Autobus/pullman	9	$\frac{9}{28}$	32%
Motorino/scooter	4	$\frac{1}{7}$	14%
Bicicletta	5	$\frac{5}{28}$	18%
Totale	28	1	100%

La somma delle frequenze relative alle diverse modalità è 1, e in percentuale è 100%.

Definizione:

La **frequenza cumulata** è la somma della frequenza del singolo dato e delle frequenze dei dati che lo precedono nell'ordine.

Esempio:

Sono state intervistate 30 famiglie di un certo quartiere e i risultati sono stati risportati nella tabella seguente:

Nr di figli per famiglia	frequenza	fr. relativa percentuale	fr. cumulata percentuale
1	12		
2	9		
3	6		
>3	3		

Quale percentuale rappresentano le famiglie con al massimo 2 figli?

Le classi di frequenze

Studiamo i risultati ottenuti da un gruppo di studentesse che, nell'ora di Educazione Fisica, hanno eseguito una prova di salto in lungo da ferme (tabella di sotto).

Gruppo A

Nr	Salto	Nr	Salto	Nr	Salto
1	1,36	9	1,61	17	1,50
2	1,46	10	1,90	18	1,67
3	1,62	11	1,65	19	1,65
4	1,54	12	1,53	20	1,78
5	1,94	13	1,36	21	2,12
6	1,85	14	1,67	22	1,86
7	1,75	15	1,46		
8	1,88	16	1,60		

Gruppo B

Nr	Salto	Nr	Salto
1	1,95	9	1,30
2	2,16	10	1,62
3	1,95	11	1,72
4	1,84	12	1,58
5	1,62	13	1,75
6	1,74	14	1,45
7	1,78	15	1,73
8	1,64	16	1,48

In casi come questo, in cui è raro che le modalità si manifestino più volte, è utile raggrupparle in classi, determinando la frequenza di ogni classe. Nella tabella seguente consideriamo cinque classi.

Classe	Valore centrale	Fr. assoluta	Fr. relativa %	Fr. cumulata	Fr cumulata %
1,20-1,40	1,30	2	9 %	2	9 %
1,40-1,60	1,50	5	23 %	7	32 %
1,60-1,80	1,70	9	40 %	16	72 %
1,80-2,00	1,90	5	23 %	21	95 %
2,00-2,20	2,10	1	5 %	22	100 %

Il raggruppamento in classi fornisce meno informazioni (per esempio, non sappiamo quanto valgono esattamente i 6 salti compresi fra 1,40 e 1,60 m), però fornisce una sintesi più leggibile della prova. Di ogni classe è spesso utile calcolare il valore centrale, che si ottiene dividendo per 2 la somma degli estremi della classe. Per esempio, il valore centrale della classe 1,60-1,80 è (1,60+1,80)/2, ossia 1,70.

Osservazioni:

Le frequenze relative percentuali delle tabelle sono approssimate alle unità.

Di solito l'estremo inferiore di ciascuna classe viene considerato incluso dalla classe, mentre quello superiore è escluso. Per esempio, nella tabella Classi di frequenza, il valore 1,60 è relativo alla classe 1,60-1,80 e non alla classe 1,40-1,60.

L'ampiezza della classe è la differenza dei suoi estremi. Nell'esempio 1,40 − 1,20 = 0,20. Solitamente le classi hanno tutte la stessa ampiezza (possono fare eccezione la prima e l'ultima classe).

Come suddividere i dati in classi?

Innanzitutto bisogna decidere il numero e l'ampiezza delle classi.

Esistono delle leggi matematiche che danno delle indicazioni a partire dal numero totale dei dati della popolazione o del campione.

Nei problemi che tratteremo suddivideremo i dati in classi utilizzando il buon senso (es: pochi dati meno classi, più dati più classi). Di solito utilizzeremo un numero da 5 a 10/12 classi.

Frequenze cumulate = somma della frequenza del singolo dato e delle frequenze dei dati che lo precedono nell'ordine.

Dalle frequenze relative alle frequenze

Se vengono fornite le frequenze relative f e il numero totale T delle unità statistiche, è possibile calcolare le frequenze F di ogni modalità. Infatti, essendo:

$f = \frac{F}{T}$ conoscendo f e T possiamo ricavare F: $F = f \cdot T$

La frequenza di una modalità è il prodotto tra la frequenza relativa e il numero totale delle unità statistiche.

Esempio:

Se sappiamo che, in un campione di 3500 persone, il 27% ha guardato una certa trasmissione televisiva, il numero delle persone del campione che ha guardato la trasmissione è:

$$0,27 \cdot 3500 = 945$$

La rappresentazione grafica dei dati

I dati statistici e le loro frequenze si possono rappresentare graficamente.

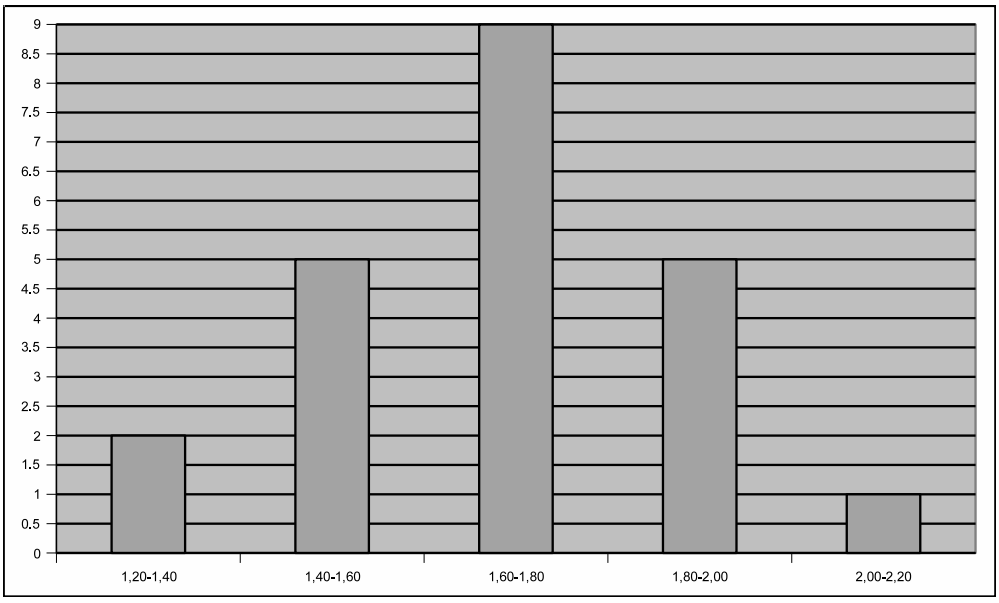
Esaminiamo due tipi di rappresentazione grafica riprendendo gli esempi del paragrafo precedente.

L'istogramma

Per rappresentare la distribuzione della tabella precedente delle classi di frequenza, riportiamo sull'asse orizzontale i valori degli estremi delle classi, ottenendo così dei segmenti le cui lunghezze rappresentano le ampiezze degli intervalli (figura 1). Disegniamo poi dei rettangoli che hanno per basi tali segmenti e la cui area è proporzionale alla frequenza della classe. Otteniamo così una rappresentazione grafica detta istogramma.

Se le classi, come nel nostro esempio, hanno tutte la stessa ampiezza, è sufficiente prendere rettangoli con le altezze proporzionali alle frequenze.

Figura 1. Un istogramma è costituito da rettangoli che hanno le basi proporzionali alle ampiezze delle classi e le aree proporzionali alle frequenze.



Osservazione:

Istogramma deriva dai termini greci histos, che significa trama , tela , e gramma, che significa segno.

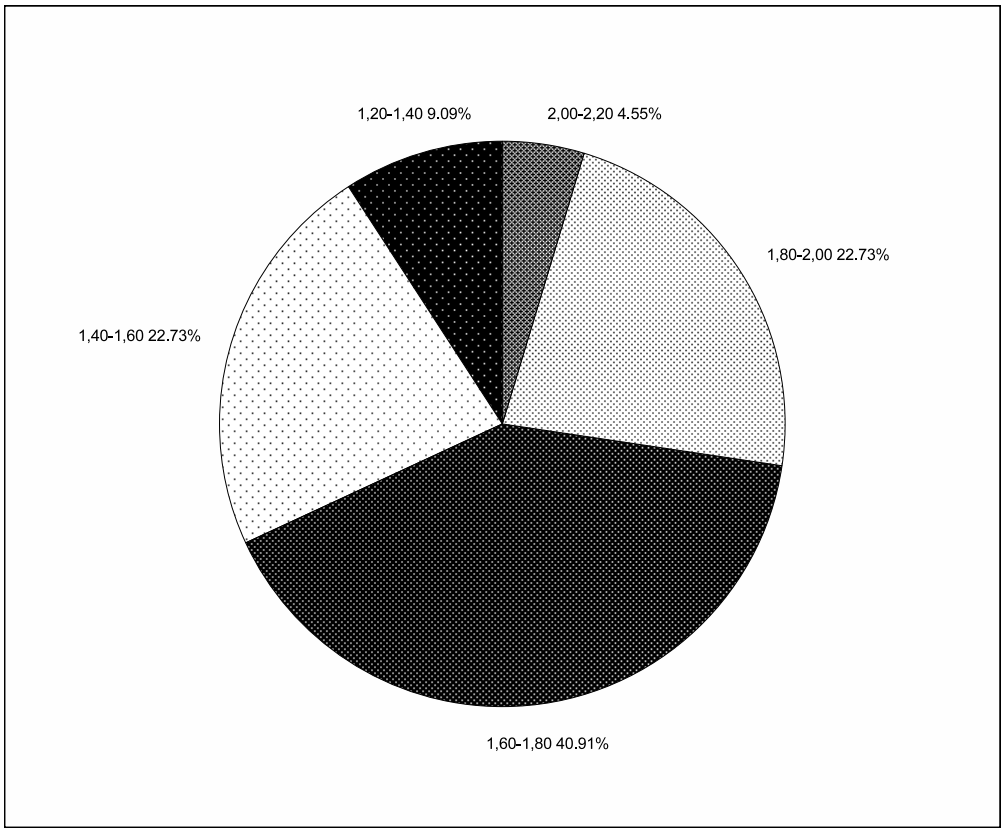
Se in un istogramma si congiungono i punti medi dei lati superiori dei rettangoli si ottiene una linea spezzata chiamata anche poligono delle frequenze. (Generalmente si "arrotonda" a mano libera la linea spezzata fino ad ottenerne una curva)

L'aerogramma

Questo tipo grafico, detto anche diagramma circolare o diagramma a torta, è particolarmente utile per rappresentare le frequenze relative percentuali.

Un cerchio viene suddiviso in tanti settori circolari, ognuno dei quali corrisponde a una classe.

L'angolo al centro del settore ha ampiezza proporzionale alla frequenza percentuale.



In un areogramma gli angoli al centro dei settori sono proporzionali alle frequenze relative percentuali.

Esempio:

Consideriamo le frequenze relative percentuali della tabella già utilizzata del salto in lungo. Per determinare l'ampiezza x del settore corrispondente alla frequenza 23% scriviamo la seguente proporzione:

$\frac{x}{360°} = \frac{23}{100} \rightarrow x = \frac{360° \cdot 23}{100} = 82,8°$

Allo stesso modo si ricavano le ampiezze degli altri settori.

Gli indici di posizione centrale

Esistono dei valori che riassumono e rappresentano un insieme di dati. Essi ci permettono di dedurre le caratteristiche di una situazione statistica e di confrontare diverse situazioni. Tali valori rappresentativi si trovano in corrispondenza delle posizioni centrali, cioè cadono in mezzo, all'interno dell'insieme di dati.

La media aritmetica

Supponiamo di voler confrontare i risultati delle prove di salto del gruppo A di studentesse del paragrafo 1 con quelli delle studentesse di un secondo gruppo (chiamiamolo gruppo B), di cui riportiamo i risultati in tabella.

Affiancando le tabelle delle frequenze dei due gruppi (tabella seguente), scopriamo che non è facile stabilire se la prova è andata meglio per il gruppo A o per il gruppo B.

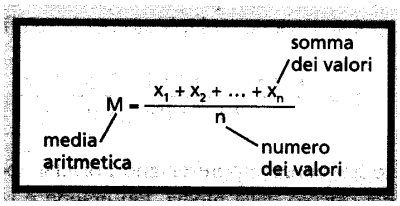
Tabella 4: confronto delle frequenze

Classe	Fr gruppo B	Fr. gruppo A
1,20-1,40	1	2
1,40-1,60	3	5
1,60-1,80	8	9
1,80-2,00	3	5
2,00-2,20	1	1

Calcolando invece la media aritmetica relativa ai due gruppi di dati, otteniamo un'informazione sintetica della distribuzione dei dati.

Definizione:

La media aritmetica $M = \bar{x}$ di n numeri x_1, x_2, ..., x_n è il quoziente tra la loro somma e il numero n.



$$M = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_n x_n}{n}$$

La media del gruppo A è

$$M_A = \frac{1,36 + 1,46 + 1,62 + ... + 1,78 + 2,12 + 1,86}{22} \cong 1,671$$

Quella del gruppo B

$$M_B = \frac{1,95 + 2,16 + 1,95 + ... + 1,45 + 1,73 + 1,48}{16} \cong 1,706$$

Poiché $M_B > M_A$ possiamo dire che le studentesse del grupo B hanno mediamente saltato meglio di quelle del gruppo A.

Nell'esempio precedente abbiamo utilizzato la media come valore di sintesi, ossia come un valore che riassume una caratteristica di un insieme di dati. Inoltre possiamo notare che, in questi esempi, la media si trova proprio nella zona della distribuzione dove si addensano maggiormente i risultati. Quando un valore di sintesi ha questa proprietà diciamo che è un buon indice di posizione centrale. Come vedremo, non sempre la media è un buon indice di posizione centrale.

La media ponderata

Consideriamo la seguente tabella, relativa ai voti di una classe ottenuti in un compito e calcoliamo la media.

Voti x	Frequenza f	$f \times x$	Frequenza relativa
4	2		9 %
5	7		32 %
6	8		36 %
7	3		14 %
8	2		9 %

$$M = \frac{4+4+5+5+5+5+5+5+5+6+6+6+6+6+6+6+6+7+7+7+8+8}{22} \cong 5,82$$

Al numeratore possiamo scrivere in modo più compatto: $4 \times 2 + 5 \times 7 + 6 \times 8 + 7 \times 3 + 8 \times 2$. Ogni voto viene moltiplicato per la sua frequenza. La media è allora:

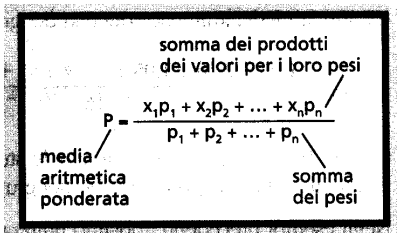
$$M = \frac{4 \times 2 + 5 \times 7 + 6 \times 8 + 7 \times 3 + 8 \times 2}{2+7+8+3+2} \cong 5,82$$

Le frequenze rappresentano i diversi "pesi" che devono avere i singoli voti nel calcolo della media. Più grande è la frequenza di un voto, maggiore è l'influenza che esso ha sul valore medio. La media calcolata in questo modo può essere considerata come caso particolare di un più generale tipo di media.

Definizione:

Media aritmetica ponderata

Dati i numeri $x_1, x_2, ..., x_n$ e associati ad essi i numeri $p_1, p_2, ..., p_n$ detti pesi chiamiamo media aritmetica ponderata p il quoziente fra la somma dei prodotti dei numeri per i loro pesi e la somma dei pesi stessi.



$$P = \frac{x_1 p_1 + x_2 p_2 + ... + x_n p_n}{p_1 + p_2 + ... + p_n}$$

Osservazione:

La media aritmetica può essere considerata un caso particolare di media ponderata in cui tutti i pesi sono uguali a 1.

Calcolo della media nel caso in cui i dati siano raccolti in classi.

Se calcoliamo la media aritmetica ponderata nel caso di classi, possiamo assumere come valori $x_1, x_2, ..., x_n$ i valori centrali di ogni classe e come pesi le frequenze. Il valore ottenuto può essere diverso dalla media aritmetica.

Esempio:

Calcoliamo la media aritmetica ponderata relativa alla tabella 3:

$$M = \bar{x} = \frac{x_1 f_1 + x_2 f_2 + ... + x_n f_n}{f_1 + f_2 + ... f_n}$$

$$M = \bar{x} = \frac{1,30 \times 2 + 1,50 \times 6 + 1,70 \times 8 + 1,90 \times 5 + 2,10 \times 1}{2+6+8+5+1} \cong 1,673$$

Il valore ottenuto è diverso, anche se di poco, dalla media aritmetica 1,671, in quanto in ogni classe abbiamo sostituito ai valori della classe il valore centrale.

Per come abbiamo usato la media ponderata nei precedenti esempi, cioè facendo coincidere i pesi con le frequenze, essa non è altro che la media ordinaria scritta in modo leggermente diverso. La media ponderata tuttavia è particolarmente significativa quando i pesi servono per indicare l'importanza dei diversi valori.

Esempio:

In un quadrimestre vengono svolte prove alle quali viene attribuita una diversa importanza (compiti in classe, relazioni, interrogazioni, test). Per un certo studente i voti riportati e i pesi da attribuire ai voti sono quelli della seguente tabella.

Tabella 6: voti pesati

Voto	5	6	5	5	7	6
Peso	1	2,5	1	1	2,5	3

Calcoliamo la media ponderata:

$$P = \frac{5 \times 1 + 6 \times 2,5 + 5 \times 1 + 5 \times 1 + 7 \times 2,5 + 6 \times 3}{1 + 2,5 + 1 + 1 + 2,5 + 3} = 5,95$$

Il valore che otteniamo è maggiore di quello della media aritmetica semplice (circa 5,67), perché i voti positivi sono stati ottenuti nelle prove alle quali è stata data maggiore importanza.

La mediana

Abbiamo già detto che la media non è sempre un buon indice di posizione centrale.

Consideriamo i seguenti sette valori: 8, 12, 7, 9, 4, 10, 55.

Calcoliamo la media aritmetica:

$$M = \frac{8 + 12 + 7 + 9 + 4 + 10 + 55}{7} = 15$$

15 non è un buon indice di posizione centrale in quanto tutti i numeri, tranne 55, sono minori di 15. È proprio la presenza del numero 55, molto maggiore degli altri, che "sposta" il valore medio rispetto alla posizione centrale. Preferiamo allora scegliere l'indice di posizione centrale nel seguente modo:

disponiamo i numeri in ordine crescente (o decrescente): 4, 7, 8, 9, 10, 12, 55;

scegliamo il valore 9 che sta al centro. Tale valore è detto mediana.

Si può determinare la mediana anche nel caso in cui il numero dei dati è pari.

Cerchiamo, per esempio, la mediana dei seguenti otto valori: 36, 22, 41, 8, 33, 46, 38, 44.

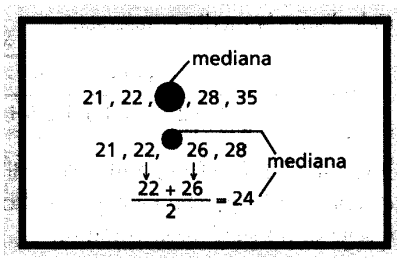
Dopo averli disposti in ordine crescente: 8, 22, 33, 36, 38, 41, 44, 46 prendiamo come mediana la media dei due valori centrali, 36 e 38.

$$\frac{36 + 38}{2} = 37$$

Definizione:

Mediana

Data la sequenza ordinata di n numeri x_1, x_2, ..., x_n la mediana è: il valore centrale, se n è dispari; la media aritmetica dei due valori centrali, se n è pari.



Osservazioni:

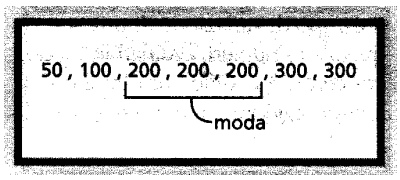
La mediana di una sequenza di numeri suddivide la sequenza in due gruppi contenenti lo stesso numero di elementi.

Il calcolo della mediana nel caso in cui i dati siano raccolti in classi è più complicato ma comunque possibile! Noi purtroppo non lo tratteremo.

La moda

Immaginiamo di dover rilevare, su una popolazione di bambini in un asilo, qual è il colore dei capelli dominante. Trattandosi di una variabile statistica qualitativa e non quantitativa non è possibile né calcolare la media aritmetica, né individuare una mediana (le modalità non sono ordinabili con un criterio oggettivo). Bisogna quindi utilizzare un altro indice di posizione chiamato moda.

Definizione: Moda Dati i numeri x_1, x_2, ..., x_n si chiama moda il valore a cui corrisponde la frequenza massima.



Consideriamo i seguenti valori:

3, 8, 2, 3, 5, 1, 7, 3, 5, 3, 15, 2, 10, 3, 12, 4

e ordiniamoli in senso crescente:

1, 2, 2, 3, 3, 3, 3, 3, 4, 5, 5, 7, 8, 10, 12, 15.

Osserviamo che il 3 ha una frequenza molto maggiore rispetto agli altri e vicino al 3 si trovano molti degli altri valori presenti. In questo caso si preferisce assumere come indice di posizione centrale tale numero, che viene chiamato moda.

La moda indica il valore più ńpresenteż nella distribuzione. Ci sono serie di dati che hanno più di una moda. Consideriamo, per esempio, i seguenti risultati di un compito in classe.

voti	4	5	6	7	8
frequenza	2	9	3	9	1

La distribuzione risulta bimodale, avendo per moda sia 5 sia 7. Ciò significa che nella classe si possono distinguere due gruppi di studenti: uno ha ben compreso gli argomenti del compito, l'altro ha bisogno di studiarli ancora! Questo tipo di informazione sarebbe andato perso se avessimo riassunto i risultati del compito con la media o la mediana, che, come puoi verificare, valgono entrambe 6.

Quando e quale indicatore di posizione centrale usare?

In conclusione, quale delle tre grandezze conviene considerare? E' opportuno usare:

la media aritmetica quando si stanno studiando delle quantità che si modificano in modo lineare (quando non ci sono valori "anomali" cioè o troppo grandi o troppo piccoli);

la moda quando si vuole evidenziare la caratteristica più diffusa;

la mediana quando è necessario conoscere il valore centrale, quello che divide a metà i dati raccolti, oppure quando ci sono dei valori "anomali" e non ci si vuol fare influenzare da questi.

Si tratterà di volta in volta di scegliere la grandezza più significativa. Ma vediamo subito un esempio:

I salari mensili di una fabbrica sono rappresentati mediante la seguente tabella:

Paga mensile	N° di persone che la ricevono
23'000	1 (il proprietario)
9'400	1
6'500	2
2'600	3
2'200	19
1'500	22
1'300	2

Calcoliamo ora i vari indici di posizione centrale studiati:

Media aritmetica =

Mediana =

Moda =

Cosa possiamo dedurre da queste informazioni? La media aritmetica ci dice che se il denaro fosse distribuito equamente (cioè in modo che ognuno ricevesse la stessa somma) ciascun dipendente avrebbe diritto a 2'612 franchi al mese. In questo caso, però, la media non è un buon indice di posizione centrale perché il salario del proprietario è un valore anomalo.

La mediana ci indica che circa la metà degli impiegati ricevono un salario di 2'200 franchi e l'altra metà di più. Non ci indica però quanto di più o quanto di meno rispetto ai 2'200 franchi.

La moda ci dice che la paga mensile più comune è di 1'500 franchi.

L'esempio ora dato ci mostra che media, mediana e moda rappresentano cose diverse.

Quindi se siete il proprietario della fabbrica e volete fare buona pubblicità alla vostra azienda utilizzerete la media aritmetica e direte: "Lo stipendio medio dei miei dipendenti è di ben 2'612 franchi mensili". Se invece rappresentate i lavoratori all'interno di un sindacato utilizzerete la moda e potrete dire: "Lo stipendio medio all'interno di questa fabbrica è di soli 1'500 franchi mensili!". Se invece siete un dipendente dell'ufficio tasse e volete tassare equamente i dipendenti utilizzerete la mediana, ma di certo non farete contenti i 22 impiegati che ricevono 1'500 franchi al mese! Ecco un piccolo esempio che vi mostra come la statistica può "mentire" se usata impropriamente! (riprenderemo questo discorso più avanti)

Gli indici di variabilità (o di dispersione)

Oltre che i valori centrali, la statistica studia come i diversi dati si situano intorno ai valori medi, quanto sono distanti, cioè quanto si disperdono o al contrario quanto sono vicini, cioè quanto si raccolgono attorno ad essi.

Consideriamo le due sequenze di valori:

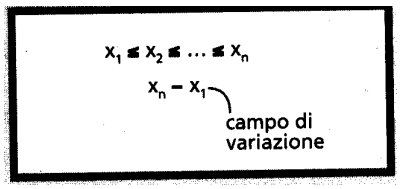
a) 12, 24, 32, 43, 56, 74, 88;

b) 42, 43, 44, 46, 49, 52, 53.

Esse sono costituite dallo stesso numero di valori e, per entrambe, la media è 47. Tuttavia la distribuzione dei valori intorno al valore medio 47 è diversa per le due sequenze: i valori della seconda sequenza sono più vicini al valore medio, mentre quelli della prima sono più sparsi. In statistica, per indicare questo fatto, si dice che le due sequenze hanno diversa dispersione o variabilità. Per misurare la variabilità si usano degli indici di variabilità quali il campo di variazione, lo scarto semplice medio e lo scarto quadratico medio. Noi tralasceremo lo scarto semplice medio e lo scarto interquartile.

Il campo di variazione

Definizione: Campo di variazione Il campo di variazione di una sequenza di numeri, ordinati in modo crescente, è la differenza fra il numero maggiore e il minore.



Nella sequenza a il campo di variazione è 88 -12 = 76, nella sequenza b è 53- 42 = 11.

Una misura della dispersione che elimini l'inconveniente dato dal campo di variazione che non riesce a descrivere come si distribuiscono i dati che si trovano fra il minimo ed il massimo si può ottenere adottando il seguente punto di vista: misurare la dispersione dei dati intorno ad uno dei valori di sintesi.

I quartili

Si può cominciare col valutare la dispersione intorno alla mediana M grazie allo scarto interquartile.

Il calcolo dei quartili in realtà è abbastanza complicato ma noi ci restringeremo a dei semplici esempi.

Come la mediana divide la serie statistica in due parti di uguale importanza, i quartili sono valori della variabile statistica che dividono la serie in quattro gruppi di uguale importanza.

Si indica con

Q_1 il primo quartile o quartile inferiore

Q_2 il secondo quartile che coincide con la mediana

Q_3 il terzo quartile o quartile superiore.

$Q_3 - Q_1$ è detto scarto interquartile

Cerchiamo di capire meglio con l'aiuto di un problema che conduce ad osservare la dispersione dei dati.

Riportiamo i voti del compito di matematica in una classe di 25 alunni:

ragazze: $4, 4\frac{1}{2}, 5, 5, 6, 6, 6\frac{1}{2}, 6\frac{1}{2}, 7, 7$

ragazzi: $3, 3, 4, 4, 4, 4, 4, 6, 6, 6, 8, 8, 8, 9, 9$

I voti delle ragazze e dei ragazzi hanno lo stesso andamento? Questi dati possono essere esaminati con i procedimenti mostrati in precedenza. Si può considerare:

la rappresentazione grafica con due istogrammi

la media, che in entrambi i casi vale circa 5,7

la mediana, che in entrambi i casi è 6

Noi vogliamo valutare la dispersione dei dati intorno alla mediana.

Consideriamo i voti delle ragazze, in questo caso abbiamo un numero pari di dati e quindi la mediana risulta essere il valor medio fra i due dati centrali

$$4, 4\frac{1}{2}, 5, 5, 6, 6, 6\frac{1}{2}, 6\frac{1}{2}, 7, 7$$

La mediana, che è 6, divide i dati in due parti ugualmente numerose, che sono le seguenti:

a) $4, 4\frac{1}{2}, 5, 5, 6$

b) $6, 6\frac{1}{2}, 6\frac{1}{2}, 7, 7$

Di ciascuna di queste parti si può di nuovo calcolare la mediana, individuando:

- nel primo gruppo il dato 5;

- nel secondo gruppo il dato $6\frac{1}{2}$.

In questo modo i dati vengono suddivisi in quattro parti ugualmente numerose per questo i valori prima individuati prendono i seguenti nomi:

$Q_1 = 5$ $\qquad$ $Q_2 = M = 6$ $\qquad$ $Q_3 = 6\frac{1}{2}$

Calcoliamo ora lo scarto interquartile $Q_3 - Q_1 = 6\frac{1}{2} - 5 = 1\frac{1}{2}$

Valutiamo ora i quartili e la differenza interquartile relativi ai voti dei ragazzi, in questo caso abbiamo un numero dispari di dati e la mediana risulta quindi essere il dato centrale.

$$3, 3, 4, 4, 4, 4, 4, \mathbf{6}, 6, 6, 8, 8, 9, 9$$

Abbiamo quindi i dati suddivisi in due parti ugualmente numerose che sono:

a) $3, 3, 4, 4, 4, 4, 4$

b) $6, 6, 8, 8, 9, 9$

Di ciascuna di queste parti si può di nuovo calcolare la mediana, individuando:

- nel primo gruppo il dato 4;

- nel secondo gruppo il dato 8 .

Si trova allora:

$Q_1 = 4$ $\qquad$ $Q_2 = M = 6$ $\qquad$ $Q_3 = 8$

Calcoliamo ora lo scarto interquartile: $Q_3 - Q_1 = 8 - 4 = 4$

Lo scarto interquartile dei voti delle ragazze ($1\frac{1}{2}$) è minore di quello dei ragazzi ($= 4$), si può così concludere che i voti delle ragazze sono meno dispersi attorno alla mediana rispetto a quelli dei ragazzi.

Osservazione: Nello stesso modo si può dividere una distribuzione statistica in decili, centili, ecc., soprattutto se le osservazioni statistiche sono molto numerose. Decili: suddivisione in 10 gruppi della stessa importanza; si hanno 9 decili ed il 5° coincide con la mediana.

Lo scarto quadratico medio

Gli scarti dalla media.

Oltre alla mediana, c'è un altro importante indice di posizione centrale: la media. Anche la media, da sola, ignora la dispersione dei dati e dunque ha bisogno di essere accompagnata da un indice di variabilità. Per esempio, consideriamo ancora una volta i voti delle ragazze:

$$4, 4\frac{1}{2}, 5, 5, 6, 6, 6\frac{1}{2}, 6\frac{1}{2}, 7, 7$$

La media di questi dati si calcola rapidamente:

$$M = \frac{4 + 4,5 + 5 + 5 + 6 + 6 + 6,5 + 6,5 + 7 + 7}{10} = 5,75$$

Come valutare la dispersione dei voti intorno alla media?

Un modo semplice di risolvere il problema potrebbe essere organizzato nel modo seguente: valutare la differenza di ogni dato dalla media, differenza che prende il nome di scarto dalla media.

Per facilitare il calcolo organizziamoci con una tabella e completiamo la prima colonna:

x_i nota	$x_i - M$	$(x_i - M)^2$
4		
$4\frac{1}{2}$		
5		
5		
6		
6		
$6\frac{1}{2}$		
$6\frac{1}{2}$		
7		
7		
Totale		

Si è arrivati dunque ad un risultato molto particolare: la somma degli scarti dalla media vale 0. Questo risultato è un caso legato ai dati esaminati o ha un valore più generale?

La somma degli scarti dalla media vale sempre zero.

Per valutare la dispersione intorno alla media si dovrà dunque eliminare l'inconveniente degli scarti positivi che compensano quelli negativi. Un metodo che la statistica utilizza molto spesso è il seguente: calcolare la media non più degli scarti, ma dei quadrati degli scarti, quadrati che sono tutti certamente positivi.

Si ottiene, nel caso esaminato, l'espressione:

$$\sigma^2 = \frac{(-1,75)^2 + (-1,25)^2 + 2 \cdot (-0,75)^2 + 2 \cdot (0,25)^2 + 2 \cdot (0,75)^2 + 2 \cdot (1,25)^2}{10} \cong 1,01$$

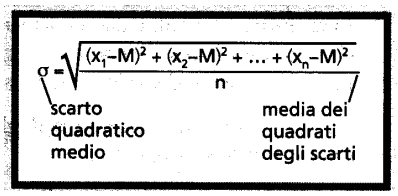
Oppure, più semplicemente riempendo la seconda colonna della tabella è sufficiente prenderne l'ultimo elemento e dividerlo per il numero dei dati, in questo caso 10.

Il risultato prende anche il nome di varianza; si ha dunque che la varianza di più dati si ottiene calcolando la media dei quadrati degli scarti dalla media.

Per sottolineare la presenza dei quadrati degli scarti, la varianza si indica spesso con il simbolo adottato prima, e cioè: varianza= σ^2 La lettera greca σ (si legge "sigma") indica lo scarto quadratico medio. Quindi per ottenere lo scarto quadratico medio si fa la radice quadrata della varianza.

Definizione:

Scarto quadratico medio Lo scarto quadratico medio di una sequenza di numeri $x_1, x_2, ..., x_n$ è la radice quadrata della media aritmetica dei quadrati degli scarti dei numeri stessi dalla loro media aritmetica.



$$\sigma = \sqrt{\frac{(x_1 - M)^2 + (x_2 - M)^2 + ... + (x_n - M)^2}{n}}$$

Osservazione: Lo scarto quadratico medio viene anche detto deviazione standard.

Varianza e scarto quadratico medio sono i più noti e diffusi indici di variabilità intorno alla media.

Così, confrontando ancora una volta i voti dei ragazzi e delle ragazze, si trova:

-voti dei ragazzi: media $M = 5,73$ $\sigma^2 = 4,46$ $\sigma = 2,11$

-voti delle ragazze: media $M = 5,75$ $\sigma^2 = 1,01$ $\sigma = 1,01$

e quindi, anche se la media è circa la stessa, si nota subito che i voti delle ragazze sono dispersi intorno alla media meno di quelli dei ragazzi.

Per sintetizzare più dati occorre il valore di sintesi accompagnato da un indice di variabilità

Le considerazioni svolte in questi ultimi due paragrafi suggeriscono di osservare sempre attentamente i dati statistici che tanto spesso sono presentati dai mezzi di informazione. Per sintetizzare più dati in modo corretto ed esauriente, occorre fornire un indice di posizione centrale, accompagnato da un indice di variabilità; così si ha che:

- la mediana senza la differenza interquartile dà un 'informazione incompleta;

- la media può fornire una sintesi scorretta se non è accompagnata dalla varianza o dallo scarto quadratico medio.

Osservazione: Nel caso in cui disponiamo di dati raccolti in classi è possibile ugualmente calcolare lo scarto quadratico medio. Si assume come valore rappresentativo il valore centrale x_i di ogni classe e la relativa frequenza f_i. Lo scarto quadratico medio allora:

$$\sigma = \sqrt{\frac{f_1 \cdot (x_1 - M)^2 + f_2 \cdot (x_2 - M)^2 + ... + f_n \cdot (x_n - M)^2}{f_1 + f_2 + ... + f_n}}$$

Esempio: Consideriamo la tabella seguente che indica le altezza s.l.m di alcuni comuni

X: altitudini	Frequenze: comuni
$0 - 50$	8
$50 - 100$	70
$100 - 150$	71
$150 - 200$	62
$200 - 250$	27
$250 - 300$	7
$300 - 350$	3

Costruiamo la tabella seguente che ci permetterà di calcolare lo scarto quadratico medio.

altitudini	Valore centrale x_l	Frequenza	$x_l \cdot f$	$(x_l - M)^2$	$(x_l - M)^2 \cdot f$
$0-50$	25	8	200	$12792,29$	$101610,32$
$50-100$	75	70	5250	$3931,29$	$275190,30$
$100-150$	125	71	8875	$161,29$	$11451,59$
$150-200$	175	62	10850	$1391,29$	$86259,98$
$200-250$	225	27	6075	$7621,29$	$205774,83$
$250-300$	275	7	1925	$18851,29$	$131959,03$
$300-350$	325	3	975	$35081,29$	$105243,87$
Totale		248	34150	$79739,03$	$917489,92$

Dalle prime tre colonne si ricava che la media è:

$$M = \frac{34150}{248} \cong 137,7$$

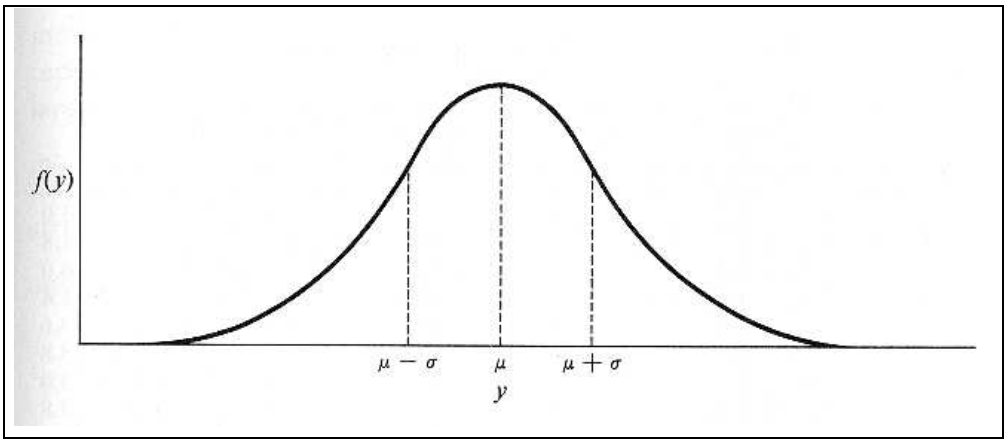
Lo scarto quadratico medio è allora

$$\sigma = \sqrt{\frac{917489,92}{248}} \cong 60,824$$

Significa quindi che l'altitudine media dei comuni è di $137,7\,[m]$, ma ci si deve preparare a superare un dislivello medio sopra e sotto di essa pari a $\sigma = 60,824\,[m]$.

La distribuzione gaussiana

Consideriamo ancora la distribuzione relativa ai risultati del salto di un gruppo di studentesse. Il suo poligono delle frequenze1 ha una forma particolare, detta anche ńa campanaż . Se aumentassimo il numero dei risultati, prendendo in considerazione, per esempio, tutte le studentesse di una stessa scuola o quelle di più scuole, il poligono delle frequenze molto probabilmente si avvicinerebbe sempre di più a una particolare curva teorica detta curva normale o gaussiana (o di Gauss).



Il calcolo dello scarto quadratico medio σ assume particolare importanza nelle distribuzioni gaussiane, perché è collegato al modo in cui le frequenze si distribuiscono attorno al valore medio M.

Da un'analisi del grafico si possono fare alcune osservazioni:

- la simmetria della curva rispetto alla retta x = M significa che intorno al valore medio tutti gli altri si distribuiscono con la stessa frequenza per valori equidistanti da M;

- nei punti $M - \sigma$ e $M + \sigma$ la curva presenta due flessi. Pertanto se σ ha un valore piccolo (e quindi c'è poca dispersione attorno al valor medio) la curva è stretta; se invece σ è grande, la curva è larga e c'è molta dispersione attorno al valor medio.

Questo significa che la forma della curva dipende da σ.

Si può dimostrare che:

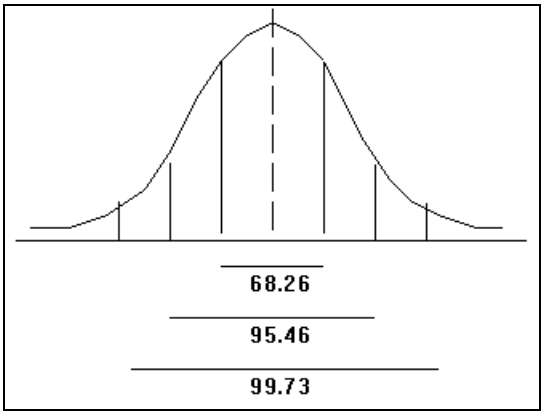
- il 68,27% dei casi osservati è compreso tra $M-\sigma$ e $M+\sigma$

- il 95,45% dei casi osservati è compreso tra $M-2\cdot\sigma$ e $M+2\cdot\sigma$

- il 99,73% dei casi osservati è compreso tra $M-3\cdot\sigma$ e $M+3\cdot\sigma$

Tali percentuali sono valide anche per distribuzioni moderatamente asimmetriche.

Possiamo quindi rappresentare le seguenti percentuali con il grafico



Da queste informazioni, essendo la distribuzione simmetrica rispetto alla media M, se ne possono ricavare altre. Per esempio, è vero che il 15,87% dei valori è maggiore di $M+\sigma$.

Infatti i valori maggiori di $M+\sigma$ o minori di $M-\sigma$ sono il 100% - 68,27% = 31,73%.

Quindi quelli maggiori di $M+\sigma$ sono:

$$\frac{31,73\%}{2}=15,87\%$$

In modo analogo si ricava che il 2,28% dei valori è maggiore di $M+2\cdot\sigma$ (o minore di $M-2\cdot\sigma$).

Esempio:

La statura in una popolazione adulta composta da 24'000'000 di persone ha una distribuzione gaussiana. Sapendo che nella popolazione studiata la media è hM = $1,75m$ e lo scarto quadratico medio $\sigma=0,05m$, quante persone hanno un'altezza compresa tra 1,70 m e 1,80 m? Quante maggiore di 1,85? E quante minore di 1,70?

Poiché $1,70=1,75-0,05$ e $1,80=1,75+0,05$, la prima domanda chiede quante sono le persone con altezza compresa tra hM $-\sigma$ e hM $+\sigma$; sappiamo che sono il 68,27%:

$24'000'000\cdot\frac{68,27}{100}=16'384'800$

La seconda domanda chiede quante persone hanno un'altezza maggiore di hM + $2\cdot\sigma$, poiché 1, $85=1,75+2\cdot0,05$. Esse sono il 2,28%.

$24'000'000\cdot\frac{2,28}{100}=547'200$

La terza domanda chiede quante persone hanno un'altezza minore di hM $-\sigma$, $(1,70=1,75-0,05)$; esse sono il 15,87%, cioè:

$24'000'000\cdot\frac{15,87}{100}=3'808'800$

È ovvio che queste cifre vadano considerate in modo approssimato, perché la popolazione non segue rigorosamente una distribuzione gaussiana, ma sono di solito abbastanza attendibili.

Scoprire come si può mentire con la statistica

Mentire con i grafici

Attività 1 Leggere e discutere la seguente storia di "fantapolitica", che avviene in un paese immaginario.

Il ministro del commercio con l'estero riflette sull'aumento delle esportazioni negli ultimi tre mesi e, valutando le esportazioni in milioni di dollari, trova i seguenti valori: in gennaio 151; in febbraio 159; in marzo 165. Il giorno seguente il ministro afferma alla televisione: "La nostra economia va molto bene; le nostre esportazioni sono aumentate in febbraio ed hanno avuto un ulteriore aumento in marzo". Per illustrare i dati, la televisione mostra il grafico di fig. 1. Lo stesso giorno, il ministro del lavoro commenta nel modo seguente il crescente numero di operai in cassa integrazione: "Per quanto riguarda gli operai messi in cassa integrazione in questi ultimi mesi si hanno i seguenti dati in migliaia: in gennaio 151; in febbraio 159; in marzo 165. Il numero di operai in cassa integrazione è dunque aumentato, ma il grafico dimostra che l' aumento del fenomeno non è poi così alto come qualcuno vorrebbe far credere". Per illustrare i dati, la televisione mostra il grafico di fig. 2.

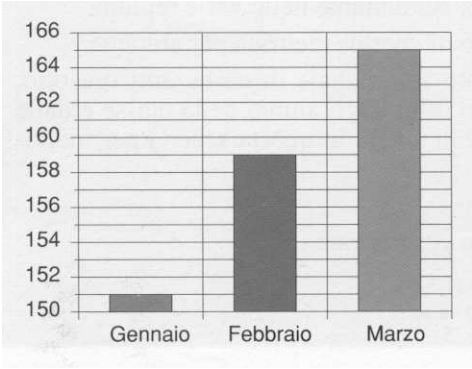


Figura 1

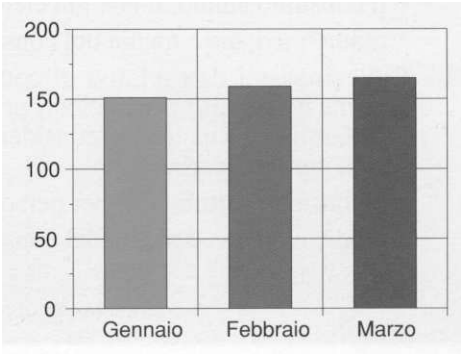


Figura 2

Mentire ignorando la frequenza relativa

Attività 2 Dalle statistiche delle compagnie di assicurazione risulta che in Italia sono più numerosi gli incidenti provocati da vetture che vanno a velocità moderata rispetto agli incidenti provocati da vetture che viaggiano ad una velocità superiore a 150 km/h. A partire da queste informazioni, si può concludere che gli italiani guidano meglio a forte velocità?

Attività 3 Dalle stesse statistiche risulta che in Italia sono molto più numerosi gli incidenti causati da guidatori che hanno più di venti anni, rispetto agli incidenti provocati da giovani di età da diciotto a venti anni. A partire da queste informazioni, si può concludere che in Italia i giovani guidano meglio degli adulti con più di venti anni?

Mentire con i valori di sintesi

Attività 4

In fig. 3 (sotto) sono rappresentati i 25 impiegati di una piccola industria ed il loro stipendio mensile. Esaminare le seguenti frasi che descrivono la situazione economica degli impiegati:

a)"In quell'industria si guadagna bene: lo stipendio medio è di 3,4 milioni al mese";

b)"Non ti conviene andare a lavorare in quell'industria: i giovani guadagnano appena 1 milione al mese";

c)"Quell'industria offre buoni stipendi: la metà degli impiegati arriva a guadagnare almeno 3 milioni al mese".

Scoprire in ogni frase qual è il valore di sintesi che descrive la situazione; spiegare perché un solo valore di sintesi descrive la situazione in modo inadeguato.

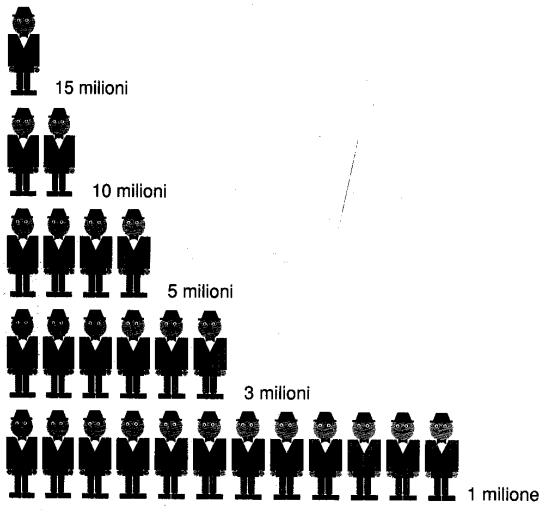


Figura 3: gli stipendi di una piccola industria

Ricerca di altre menzogne statistiche

Attività 5

Scoprire qualche menzogna statistica in un quotidiano, su una rivista o in un programma televisivo recente.

Attività 6

Inventare una storia o dei dati che presentino delle menzogne statistiche.