Statistica

Che cos'è la statistica

La statistica, in origine, si occupava di rispondere a quesiti che riguardavano il governo e la vita di uno stato, ed è proprio dalla parola **stato** che deriva il termine statistica.

La statistica si occupava, inizialmente, di rispondere a quesiti come i seguenti:

- quanti sono i cittadini italiani?
- quanti sono gli agricoltori in Italia?
- quanti posti letto occorrono negli ospedali della Calabria?
- Quanti dottori saranno necessari fra cinque anni in Calabria?

Oggi la statistica è utilizzata, oltre che da organismi statali, anche da enti privati, come aziende, organi di informazione, partiti politici, ecc.

Questi enti privati hanno sempre più bisogno di conoscere in tempo reale i bisogni, i gusti, le preferenze, i pareri della popolazione, prima di aprire un centro commerciale, prima di mettere sul mercato un prodotto, prima di fare una scelta politica.

Esempi

Secondo l'Auditel, ieri sera il programma diretto da Pippo ha avuto un indice di ascolto del 40%, mentre il programma diretto da Pluto ha avuto un indice di ascolto del 30%.

Secondo l'agenzia di sondaggi elettorali "Oggi e domani" il 35% degli italiani voteranno alle prossime elezioni il partito "X", mentre il 40% degli italiani voteranno alle prossime elezioni il partito "Y".

Secondo una ricerca di mercato, il 45% degli automobilisti italiani preferisce acquistare un'automobile che abbia la doppia alimentazione benzina-gas.

Per studiare scientificamente situazioni come quelle indicate nei precedenti esempi e altre simili è necessario utilizzare gli strumenti concettuali forniti dalla statistica: una disciplina oggi in fase di grande sviluppo.

La statistica si occupa dello studio quantitativo di fenomeni collettivi (ossia che riguardano una pluralità di soggetti), osservabili nella realtà sociale, in natura o in laboratorio.

Nella società dell'informazione in cui viviamo non è difficile accedere a ogni sorta di dati, ma solo disponendo di adeguate competenze in statistica è possibile riuscire a interpretarli correttamente.

Contrariamente a quanto si potrebbe pensare, la possibilità di attingere a una grandissima quantità di dati rischia di impedirci di fatto di utilizzare anche solo una parte di questi: non basta infatti avere solo l'accesso teorico a un dato, ma occorre che esso sia effettivamente e praticamente fruibile.

Il compito principale della statistica è proprio quello di rendere utilizzabili grandi quantità di dati, difficilmente gestibili, relative agli oggetti della propria indagine. Infatti tutte le informazioni, per contribuire effettivamente ad accrescere la conoscenza di un fenomeno, hanno bisogno di essere trattate da vari punti di vista: occorrono tecniche accurate di rilevazione, occorre procedere ad accurate selezioni, occorre un lavoro di organizzazione e di sintesi.

La statistica raccoglie e restituisce in forma organizzata grandi quantità di dati. Nel fare ciò obbedisce alla duplice esigenza **descrittiva** e **predittiva**.

- Ogni comunità sente il bisogno, a fini di documentazione, di raccogliere una serie di dati sugli usi, sui costumi, sulle attività sociali ed economiche dei suoi componenti; i censimenti costituiscono uno strumento fondamentale attraverso cui la statistica esplica questa funzione.
- un'altra esigenza a cui risponde la statistica è quella predittiva: la raccolta e l'elaborazione dei dati, e quindi la "fotografia" del passato e del presente, serve per prevedere i comportamenti futuri, per operare scelte, per assumere decisioni.

Termini della statistica

Per iniziare lo studio della statistica occorre conoscere il significato di alcuni termini.

Si chiama **popolazione** (o universo o collettivo) l'insieme degli individui oggetto di un'indagine statistica; ciascun elemento della popolazione viene detto **unità statistica**.

Esempi: una popolazione umana, una colonia di batteri, un insieme di automobili.

In alcune indagini statistiche è possibile interpellare tutti i membri della popolazione. In altri casi, per ovvi motivi di economicità, di tempestività, o di impossibilità reale, l'indagine viene effettuata su un sottoinsieme della popolazione, che viene chiamata **campione**. Naturalmente il campione deve essere il più rappresentativo possibile della popolazione. La parte della statistica che stabilisce i criteri di rappresentatività si chiama **inferenza statistica** o statistica induttiva.

Il **carattere** di un'indagine statistica è la proprietà che si intende studiare in una popolazione

Esempi: l'altezza delle persone, l'età delle persone, la marca delle automobili, ecc.

Si chiama **modalità** ciascuna delle varianti con cui un carattere può presentarsi. Le modalità osservate si chiamano **dati**.

Esempi: il carattere "altezza" può assumere, in corrispondenza di un dato individuo, la modalità "172 cm", in corrispondenza di un altro la modalità "178 cm"; il carattere "età" può assumere per un dato individuo la modalità "84 anni" oppure "78 anni"; il carattere marca di un'automobile può assumere per un dato elemento la modalità "Fiat", "Mercedes", ecc.

Un carattere può essere di tipo quantitativo o di tipo qualitativo.

Un carattere le cui modalità sono espresse da numeri è detto **quantitativo** (o variabile).

Un carattere le cui modalità non sono espresse da numeri è detto **qualitativo** (o mutabile).

Esempi:

Caratteri quantitativi	Caratteri qualitativi
L'altezza di una persona	Il colore delle automobili vendute
L'età di una persona	Il tipo di alimentazione delle automobili
La quantità di pane consumata in un giorno	Il tipo di vacanza preferito dalle persone

I caratteri quantitativi o variabili si classificano ulteriormente in variabili discrete e variabili continue.

Un carattere quantitativo o variabile è detto **discreto** quando può assumere soltanto un numero finito di valori o al più un insieme di valori che può essere posto in corrispondenza biunivoca con i numeri naturali.

Un carattere quantitativo o variabile è detto **continuo** quando può assumere tutti i valori reali di un determinato intervallo.

Le variabili discrete sono quelle che si rilevano contando, mentre le variabili continue sono quelle che si rilevano mediante misurazioni.

Esempio: il numero degli studenti di una scuola è una variabile discreta; la temperatura massima giornaliera registrata a Trebisacce in un dato giorno è una variabile continua.

Le fasi dell'indagine statistica

Un'indagine statistica è un processo complesso suddiviso in più fasi.

Le fasi principali di un'indagine statistica sono:

1. Pianificazione dell'indagine statistica

   Si individuano:

   - il carattere che interessa studiare
   - la popolazione (o eventualmente il campione) su cui si intende studiare tale carattere

2. Rilevazione dei dati

   La rilevazione dei dati può avvenire in diversi modi, a seconda della popolazione presa in esame. Se la popolazione è costituita da esseri umani il metodo più usato è l'intervista: essa consiste nel rivolgere alcune domande alle unità che compongono la popolazione presa in esame e nel registrare le risposte in un apposito modello, detto questionario. L'intervista può avvenire secondo diverse tecniche: può essere diretta, cioè avvalersi della presenza fisica di un intervistatore che pone direttamente le domande a un individuo, oppure può avvenire per autocompilazione del questionario, o telefonicamente.

3. Elaborazione dei dati

   L'obiettivo di questa fase è di fare emergere da una gran mole di dati le informazioni che interessano. Come primo approccio può essere utile riordinare i dati o raggrupparli in modo conveniente. Successivamente si cerca di sintetizzarli attraverso il calcolo di pochi numeri significativi. La parte della statistica che si occupa delle tecniche volte a questo scopo si chiama statistica descrittiva.

   Se i dati non sono stati rilevati sull'intera popolazione ma solo su un campione, la loro elaborazione è più complessa, perché occorre anche porsi l'obiettivo di estendere i risultati ottenuti dal campione all'intera popolazione e dunque anche alla parte non osservata. La parte della statistica che si occupa delle tecniche adatte a questi scopi si chiama **statistica inferenziale**.

4. Presentazione dei risultati

   In questa fase si costruiscono tabelle, diagrammi o grafici che rappresentano i risultati ottenuti dalle elaborazioni dei dati, allo scopo di rendere tali risultati più evidenti e di facile lettura. Quindi si rendono pubblici gli esiti dell'indagine.

5. Interpretazione dei risultati

   L'interpretazione dei risultati non è sempre immediata e richiede un attento esame del contesto, soprattutto per evitare estrapolazioni indebite. È buona norma fare chiare distinzioni tra i dati oggettivi rilevati nel corso dell'indagine e le interpretazioni soggettive di tali dati.

Elaborazione dei dati

Esaminiamo in questo capitolo la fase dell'elaborazione dei dati.

Distribuzione di frequenze

Per spiegare il significato di distribuzione di frequenze è utile il seguente esempio:

Consideriamo i paesi di provenienza degli studenti della classe 1 A, formata da 18 studenti.

I risultati della rilevazione sono raccolti nella seguente tabella:

Studente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Paese di provenienza	T	P	A	T	T	P	A	V	V	P	P	T	T	V	V	T	P	T

(T = Trebisacce P = Plataci A = Albidona V = Villapiana)

Una prima forma di elaborazione dei dati, volta a ottenere una maggiore sintesi, consiste nel costruire una tabella in cui riportare, per ciascuna delle modalità osservate (T, P, A, V), il numero di individui su cui è stata rilevata.

Paese di provenienza	**Studenti** *(Numero)*
Trebisacce	7
Plataci	5
Albidona	2
Villapiana	4

I numeri scritti a fianco di ciascuna modalità osservata sono detti frequenze assolute; esse indicano il numero degli studenti che provengono da quel paese.

La **frequenza assoluta** è il numero di volte in cui la modalità è stata osservata.

La funzione che associa a ogni modalità di un carattere la rispettiva frequenza è detta **distribuzione delle frequenze**.

Consideriamo i paesi di provenienza degli studenti della classe 1 B, formata da 24 studenti.

Studente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Paese di provenienza	T	P	A	T	T	P	A	V	V	P	P	T	T	V	V	T	P	T	T	P	A	V	T	P

Paese di provenienza	**Studenti** *(Numero)*
Trebisacce	9
Plataci	7
Albidona	3
Villapiana	5

Se si vuole confrontare le distribuzioni delle frequenze delle due classi, occorre tenere conto del fatto che queste sono composte da un numero diverso di alunni.

Per poter confrontare correttamente le due classi è necessario depurare le frequenze assolute dall'influenza dovuta al numero di alunni della classe. A tale scopo si introduce la frequenza relativa.

La **frequenza relativa** di una modalità è il rapporto fra la sua frequenza assoluta e il numero complessivo del collettivo.

La frequenza relativa può essere espressa anche sotto forma di percentuale: in tal caso è detta **frequenza percentuale**.

Classe 1 A			
Paese di provenienza	**Frequenza assoluta**	**Frequenza relativa**	**Frequenza percentuale**
Trebisacce	7	$\frac{7}{18} = 0{,}389$	$38{,}9\%$
Plataci	5	$\frac{5}{18} = 0{,}278$	$27{,}8\%$
Albidona	2	$\frac{2}{18} = 0{,}111$	$11{,}1\%$
Villapiana	4	$\frac{4}{18} = 0{,}222$	$22{,}2\%$
Totale	*18*	*1*	*100%*

Classe 1 B			
Paese di provenienza	**Frequenza assoluta**	**Frequenza relativa**	**Frequenza percentuale**
Trebisacce	9	$\frac{9}{24} = 0{,}375$	$37{,}5\%$
Plataci	7	$\frac{7}{24} = 0{,}292$	$29{,}2\%$
Albidona	3	$\frac{3}{24} = 0{,}125$	$12{,}5\%$
Villapiana	5	$\frac{5}{24} = 0{,}208$	$20{,}8\%$
Totale	*24*	*1*	*100%*

Distribuzione per classi

Per spiegare il significato di intervallo o classe è utile il seguente esempio:

Cinquanta studenti sono stati sottoposti a un test di matematica. Oggetto dell'indagine è il tempo impiegato da uno studente per svolgere il test.

Studente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Tempo *(min)*	26	30	35	27	37	31	36	28	44	34	45	29	30	34	35	45	25	36	38	37	30	43	42	33	39

Studente	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Tempo *(min)*	32	38	35	41	33	38	25	35	37	40	28	35	34	27	37	33	38	26	36	32	31	36	29	36	39

Se si costruisse la distribuzione delle frequenze non si otterrebbe una **sintesi significativa**, perché si discosterebbe di poco dalla tabella precedente poiché l'indagine presenta molte più modalità.

In casi come questo occorre accorpare le modalità in intervalli disgiunti di uguale ampiezza.

Nell'esempio conviene ripartire i tempi rilevati in intervalli di ampiezza di 5 minuti, come di seguito indicato:

Intervallo *(minuti)*	**Studenti** *(Numero)*	**Frequenza relativa**	**Frequenza percentuale**
$25 \leq m < 30$	10	$10/50 = 0{,}20$	20%
$30 \leq m < 35$	13	$13/50 = 0{,}26$	26%
$35 \leq m < 40$	20	$20/50 = 0{,}40$	40%
$40 \leq m \leq 45$	7	$7/50 = 0{,}14$	14%

Gli intervalli $[25, 30[$ $[30, 35[$ $[35, 40[$ $[40, 45]$ rappresentano le **classi**.

In una classe $[a, b[$ il numero a è detto estremo sinistro e il numero b è detto estremo destro.

Il numero $b - a$ è detta ***ampiezza*** della classe *(nell'esempio considerato tutte le classi hanno la stessa ampiezza 5.*

Nell'esempio, il carattere "tempo impiegato" è un carattere continuo, perché è una grandezza che può assumere qualsiasi valore in ognuno degli intervalli.

Distribuzione delle frequenze cumulate

Riconsideriamo l'esempio precedente del test di matematica.

In questa indagine si può ad esempio rispondere alla seguente domanda:

Quanti sono gli studenti che hanno completato il test in meno di 35 minuti?

Per rispondere a questa domanda si introduce un altro tipo di frequenza, detta frequenza cumulata.

La ***frequenza cumulata*** relativa a una data modalità è la somma delle frequenze di tutte le modalità minori o uguali a essa.

Intervallo *(minuti)*	**Studenti** *(Numero)*	**Frequenza cumulata**
$25 \leq m < 30$	10	10
$30 \leq m < 35$	13	10 + 13 = 23
$35 \leq m < 40$	20	23 + 20 = 43
$40 \leq m \leq 45$	7	43 + 7 = 50

Dalla colonna delle frequenze cumulate è possibile ottenere la risposta alla domanda che ci siamo posto: *gli studenti che hanno completato il test in meno di 35 minuti sono 23.*

Rappresentazioni grafiche

Dopo la raccolta dei dati, un metodo spesso utilizzato per sintetizzare i risultati dell'indagine è la rappresentazione grafica

Le principali rappresentazioni grafiche sono le seguenti:

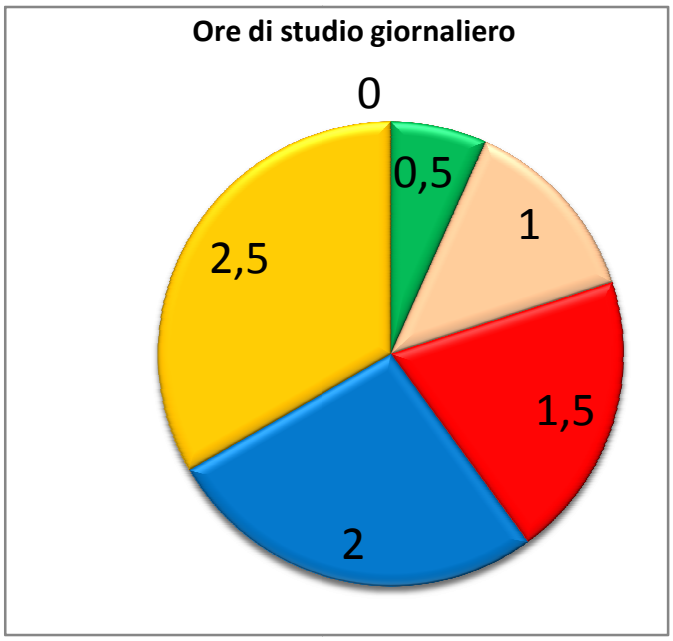
Diagrammi circolari

Il **diagramma circolare** è un cerchio suddiviso in settori circolari, con angoli al centro di ampiezza proporzionale alla frequenza di ciascuna modalità.

INDAGINE STATISTICA

Quante ore al giorno studi ?

N° Ore	N° Alunni
0	1
0,5	2
1	4
1,5	5
2	9
2,5	12



Per calcolare l'ampiezza α del settore circolare relativo alla frequenza f di ciascuna modalità occorre applicare la seguente proporzione: $\alpha : 360° = f_{\%} : 100$ oppure occorre moltiplicare la frequenza relativa per 360.

In simboli: $\alpha = f_R \cdot 360$

INDAGINE STATISTICA			
Quante ore al giorno studi ?			
N° Ore	**Frequenza assoluta**	**Frequenza relativa**	**Ampiezza angolo**
0	1	$1/33 = 0{,}03$	$\alpha_1 = 0{,}03 \cdot 360 = 11°$
0,5	2	$2/33 = 0{,}06$	$\alpha_2 = 0{,}06 \cdot 360 = 22°$
1	4	$4/33 = 0{,}12$	$\alpha_3 = 0{,}12 \cdot 360 = 44°$
1,5	5	$5/33 = 0{,}15$	$\alpha_4 = 0{,}15 \cdot 360 = 55°$
2	9	$9/33 = 0{,}27$	$\alpha_5 = 0{,}27 \cdot 360 = 98°$
2,5	12	$12/33 = 0{,}36$	$\alpha_6 = 0{,}36 \cdot 360 = 130°$
Totale	*33*	*1*	*360°*

Il diagramma circolare è utile per rappresentare i caratteri che non possono essere ordinati. Esso permette un'agevole lettura e un confronto immediato dei dati.

Questo tipo di grafico si utilizza per visualizzare le diverse parti in cui "un tutto" è suddiviso. Per esempio si presta bene a rappresentare la composizione del parlamento da parte dei diversi partiti politici. Mentre non è adatto per rappresentare le altezze degli alunni di una classe.

Istogrammi

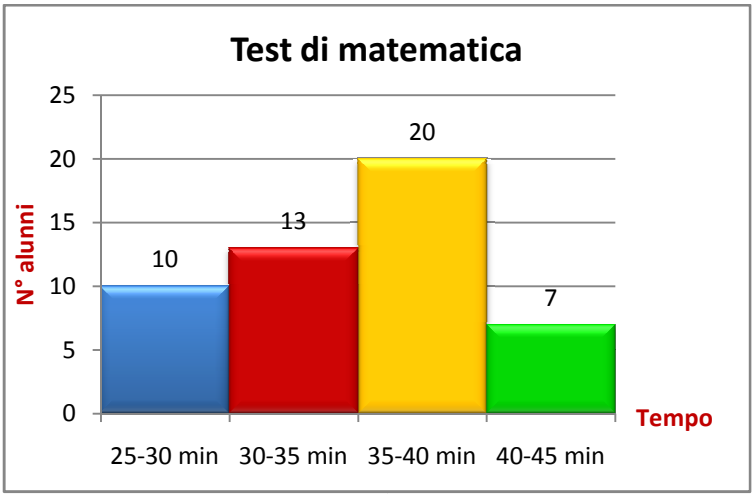
Un **istogramma** è un grafico costituito da rettangoli non distanziati, ciascuno dei quali ha un'area proporzionale alla frequenza della classe che rappresenta.

Gli **istogrammi** sono utilizzati per rappresentare distribuzioni di caratteri suddivisi in classi.

Riprendendo l'esempio del test di matematica

TEST DI MATEMATICA

Tempo	N° Alunni
25-30 min	10
30-35 min	13
35-40 min	20
40-45 min	7



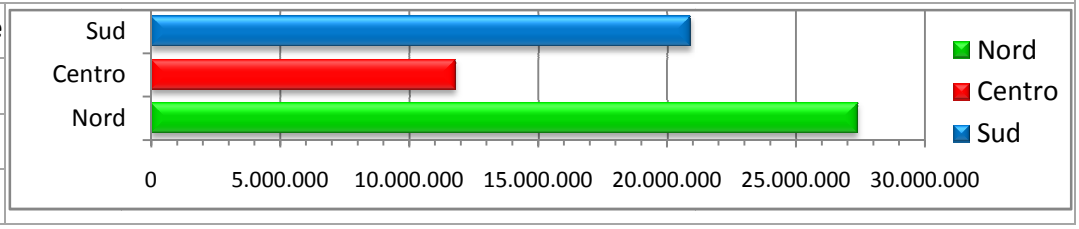
Un istogramma si ottiene riportando sull'asse orizzontale segmenti adiacenti di lunghezza uguale all'ampiezza della classe. L'altezza di ciascun rettangolo si determina facendo il rapporto tra la frequenza e l'ampiezza della relativa classe.

Ortogrammi

Un **ortogramma** o **diagramma a barre** è un grafico costituito da rettangoli verticali o orizzontali aventi tutti le basi di eguale misura (arbitraria) che poggiano sull'asse orizzontale o verticale e altezze proporzionali alle frequenze delle classi che rappresentano.

POPOLAZIONE RESIDENTE AL 1° GENNAIO 2009

Area geografica	Popolazione
Nord	27.400.000
Centro	11.800.000
Sud	20.900.000



Nello stesso ortogramma è possibile rappresentare contemporaneamente due o più caratteri.

Per esempio è possibile confrontare la popolazione del 2009 con quella del 2003.

POPOLAZIONE ITALIANA

Area geografica	2003	2009
Nord	25.500.000	27.400.000
Centro	12.000.000	11.800.000
Sud	22.000.000	20.900.000

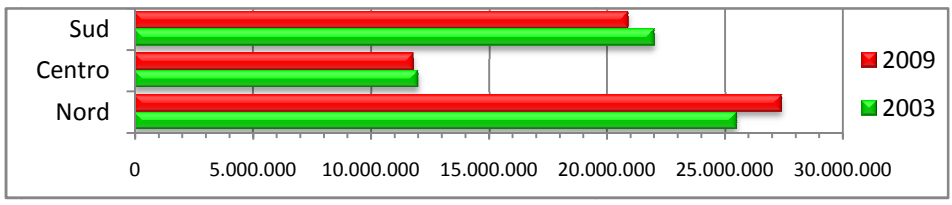


Diagramma cartesiano

Il **diagramma cartesiano** è il grafico ottenuto congiungendo con dei segmenti i punti che hanno come ascisse i valori osservati e come ordinate le corrispondenti frequenze.

In statistica questo tipo di rappresentazione è particolarmente indicato per rappresentare le cosiddette **serie temporali**, cioè quei fenomeni che vengono osservati in determinati periodi di tempo.

Popolazione Italiana

Anno	N° Abitanti
2001	56.980.000
2002	57.320.000
2003	57.890.000
2004	58.460.000
2005	58.750.000
2006	59.130.000
2007	59.620.000
2008	60.050.000
2009	60.340.000
2010	60.630.000

Indici di posizione

Per analizzare e comprendere l'andamento di un fenomeno è utile ricavare dai dati raccolti alcuni valori particolarmente significativi, detti indici di posizione.

Media aritmetica semplice

La media aritmetica semplice di n valori $x_1, x_2, \dots, x_n$ è il numero

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

Esempio

Con tale formula è possibile determinare la media dei voti ottenuti dallo studente Mario Rossi.

Italiano	*Inglese*	*Matematica*	*Fisica*	*Informatica*	*Scienze*
9	*8*	*6*	*8*	*6*	*8*

La media dei voti dello studente Mario Rossi è: $M = \frac{9+8+6+8+6+7}{6} = 7{,}5$

La media aritmetica ha la proprietà di mantenere inalterata la somma dei valori, quando è sostituita a ciascuno di essi.

Media aritmetica ponderata

La media aritmetica ponderata di n valori $x_1, x_2, \dots, x_n$ aventi rispettivamente frequenze $f_1, f_2, \dots, f_n$ è il numero:

$$M_P = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \dots + x_n \cdot f_n}{f_1 + f_2 + \dots + f_n}$$

Osservazione

Nel caso in cui siano note le frequenze relative $f_{R1}, f_{R2}, \dots, f_{Rn}$ *anziché le frequenze assolute, la formula della media aritmetica ponderata risulta essere:* $M_P = x_1 \cdot f_{R1} + x_2 \cdot f_{R2} + \dots + x_n \cdot f_{Rn}$.

Esempio

Da un'indagine effettuata su un campione di famiglie della regione Calabria si è ottenuto la distribuzione di frequenze sotto riportata. Esaminando tale distribuzione, calcolare il numero medio di figli per famiglia.

Numero figli per famiglia	**Frequenza**
0	812
1	1223
2	2126
3	1535
4	934
5	231

Soluzione

$$M_P = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \dots + x_n \cdot f_n}{f_1 + f_2 + \dots + f_n} = \frac{0 \cdot 812 + 1 \cdot 1223 + 2 \cdot 2126 + 3 \cdot 1535 + 4 \cdot 934 + 5 \cdot 231}{812 + 1223 + 2126 + 1535 + 934 + 231} =$$

$$= \frac{0 + 1223 + 4252 + 4605 + 3736 + 1155}{6861} = \frac{14971}{6861} \simeq 2{,}182\ figli\ .$$

Media aritmetica ponderata di un carattere suddiviso per classi

Se un carattere è suddiviso per classi, non è possibile calcolare il valore esatto della media aritmetica, perché non si conoscono esattamente i valori osservati all'interno di ciascuna classe. In casi come questo si conviene di assumere come media aritmetica il valore approssimato che si ottiene sostituendo ciascuna classe con il suo valore centrale.

Esempio

Da un'indagine effettuata su un campione di persone si è ottenuto la distribuzione di frequenze sotto riportata. Esaminando tale distribuzione, calcolare il peso medio delle persone della popolazione considerata.

Peso *(kg)*	**Frequenza**
$40 \leq m < 50$	8
$50 \leq m < 60$	24
$60 \leq m < 70$	22
$70 \leq m < 80$	18
$80 \leq m < 90$	4
$90 \leq m < 100$	2

Soluzione

Sostituendo ogni classe con il suo valore centrale si la seguente distribuzione di frequenze

Peso *(kg)*	**Frequenza**
45	8
55	24
65	22
75	18
85	4
95	2

Da questa distribuzione di frequenze si ricava il peso medio:

$$M_P = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \ldots + x_n \cdot f_n}{f_1 + f_2 + \cdots + f_n} = \frac{45 \cdot 8 + 55 \cdot 24 + 65 \cdot 22 + 75 \cdot 18 + 85 \cdot 4 + 95 \cdot 2}{8 + 24 + 22 + 18 + 4 + 2} =$$

$$= \frac{360 + 1320 + 1430 + 1350 + 340 + 190}{78} = \frac{4990}{78} \simeq 63{,}974\ kg\ .$$

Media armonica

La **media armonica** è il reciproco della media aritmetica dei loro reciproci.

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Esempio 1

Due studenti, Davide e Luana devono preparare un esame su un libro di 1200 pagine.

Davide decide di studiare il libro due volte: la prima al ritmo di 20 pagine al giorno e la seconda al ritmo di 30 pagine al giorno. Luana invece, intende studiare il libro sempre due volte, ma sempre con lo stesso ritmo. Quante pagine al giorno deve studiare Luana per studiare il libro due volte nello stesso tempo che impiega Davide?

Soluzione

La risposta di un lettore disattento sarebbe 25 pagine al giorno.

Cioè la media dei due valori delle velocità di lettura di Davide $V_D = \frac{20+30}{2} = 25$. *Ma la risposta è errata.*

Infatti, il tempo impiegato da Davide per studiare il libro due volte è:

$t_D = \frac{1200}{20} + \frac{1200}{30} = 60 + 40 = 100\ giorni$.

Luana, per poter studiare il libro due volte in questo tempo, deve avere una velocità di lettura: $V_L = \frac{2400}{100} = 24$.

Questo risultato si può ottenere calcolando la media armonica delle due velocità di lettura di Davide.

$$V_L = \frac{2}{\frac{1}{20} + \frac{1}{30}} = \frac{2}{\frac{3+2}{60}} = \frac{2}{\frac{5}{60}} = 2 \cdot \frac{60}{5} = 24$$

In generale, indicando con p *il numero delle pagine del libro e con* v_1 *e* v_2 *le velocità di lettura, si ha:*

$\frac{p}{v_1} + \frac{p}{v_2} = \frac{p}{v} + \frac{p}{v}$; $\frac{1}{v_1} + \frac{1}{v_2} = \frac{1}{v} + \frac{1}{v}$; $\frac{1}{v_1} + \frac{1}{v_2} = \frac{2}{v}$; $passando\ ai\ reciproci$:

$\frac{v}{2} = \frac{1}{\frac{1}{v_1} + \frac{1}{v_2}}$; $v = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}}$

che rappresenta la formula della media armonica nel caso di $n = 2$.

Esempio 2

Una gara ciclistica è stata organizzata su un circuito da percorrere tre volte. Il ciclista Gianbattista vince la gara percorrendo i tre giri rispettivamente alle velocità medie di $25\ km/h$, $30\ km/h$ *e* $20\ km/h$ *. Se Gianbattista avesse percorso i tre giri sempre alla stessa velocità media di* $25\ km/h$ *avrebbe vinto ugualmente? Calcola poi, la velocità media, uguale in tutte e tre i giri, che gli avrebbe consentito di vincere la gara nello stesso tempo.*

Soluzione

Ragionando similmente al problema precedente, ponendo uguale a s *la lunghezza del circuito, si ha:*

Il tempo impiegato da Gianbattista per effettuare i 3 giri del circuito è:

$t = \frac{s}{25} + \frac{s}{30} + \frac{s}{20} = \frac{12s+10s+15s}{300} = \frac{37s}{300}\ h$.

Alla velocità media di $25\ km/h$ *Gianbattista avrebbe percorso uno spazio:*

$s = v \cdot t = 25\ \frac{km}{h} \cdot \frac{37s}{300} h \cong 3{,}08\ s$ *valore superiore alla lunghezza della gara pari a* $3s$*. Pertanto Gianbattista avrebbe vinto la gara in un tempo minore.*

Gianbattista avrebbe vinto la gara nello stesso tempo da lui realizzato se avesse percorso i 3 giri alla velocità media di:

$v = \frac{3}{\frac{1}{25}+\frac{1}{30}+\frac{1}{20}} = \frac{3}{\frac{12+10+15}{300}} = \frac{3}{\frac{37}{300}} = 3 \cdot \frac{300}{37} \cong 24{,}32\ km/h.$

Media geometrica

La **media geometrica** è la radice n-esima del loro prodotto.

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ \ldots \ \cdot x_n}$$

Si ricorre alla media geometrica quando si vuole conservare il prodotto di più valori. Tale formula può essere applicata per risolvere problemi come quello seguente.

Esempio

Un risparmiatore ha sottoscritto un investimento di 10.000 euro con durata 3 anni: alla fine del primo anno gli verrà corrisposto un tasso di interesse percentuale del 2% sul capitale investito, alla fine del secondo anno un tasso di interesse percentuale del 3% e alla fine del terzo anno un tasso di interesse percentuale del 5%. Il secondo anno il tasso di interesse viene calcolato sul montante ottenuto come somma del capitale iniziale e dell'interesse maturato l'anno precedente; analogamente, il terzo anno il tasso di interesse viene calcolato sul montante ottenuto come somma del capitale iniziale e degli interessi maturati nei due anni precedenti. Qual è il tasso di interesse che, applicato per tutti e tre gli anni, lascia invariato il montante finale ottenuto allo scadere dell'investimento?

Per calcolare il montante al termine del terzo anno occorre moltiplicare il capitale investito per i coefficienti di incremento $(1 + tasso\ di\ interesse)$ *corrispondenti a ciascuno dei tre anni.*

Il montante finale sarà quindi dato da:

$M = 10.000 \cdot \ (1 + 0{,}02) \cdot (1 + 0{,}03) \cdot (1 + 0{,}05).$

Il coefficiente di incremento (non il tasso di interesse) che, applicato per tutti e tre gli anni, lascia invariato il capitale finale è quello che conserva il prodotto dei tre coefficienti (1 + 0,02), (1 + 0,03) e (1 + 0,05), quindi è la loro media geometrica:

$$M = \sqrt[3]{1{,}02 \cdot 1{,}03 \cdot 1{,}05} \backsimeq 1{,}03326$$

Poiché $1{,}03326 = 1 + 0{,}03326 = 1 + \frac{3{,}326}{100}$*, deduciamo che il valore del tasso di interesse che, applicato tutti e tre gli anni, lascia invariato il montante finale è approssimativamente uguale a 3,326%.*

Media armonica ponderata

La media geometrica ponderata di n valori $x_1, x_2, \ldots, x_n$ aventi rispettivamente frequenze $f_1, f_2, \ldots, f_n$ è il numero:

$$H = \frac{f_1 + f_2 + \ \ldots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots + \frac{f_n}{x_n}}$$

Media geometrica ponderata

La media geometrica ponderata di n valori $x_1, x_2, \ldots, x_n$ aventi rispettivamente frequenze $f_1, f_2, \ldots, f_n$ è il numero:

$$G = \sqrt[n]{{x_1}^{f_1} \cdot {x_2}^{f_2} \cdot \ \ldots \cdot {x_n}^{f_n}}$$

Le tre medie, armonica, geometrica e aritmetica hanno il seguente ordine di grandezza: $H \leq G \leq M$.

Mediana

La mediana di un insieme di n numeri ordinati in senso crescente o decrescente è:

- il numero che occupa la posizione centrale, se n è dispari
- la media aritmetica dei due numeri che occupano le posizioni centrali $\frac{n}{2}$ e $\frac{n}{2}+1$, se n è pari.

Esempi

La mediana dell'insieme di valori $A = \{5,\ 8,\ 11,\ 23,\ 56\}$ *è* 11.

La mediana dell'insieme di valori $A = \{5,\ 8,\ 11,\ 23,\ 56,\ 58\}$ *è* $\frac{11+23}{2} = 17$.

Mediana di una distribuzione di frequenze

Per calcolare la mediana di una distribuzione di frequenze occorre:

1. calcolare le frequenze cumulate
2. calcolare la mediana utilizzando l'ordinamento delle frequenze cumulate

Esempio

In una festa di beneficenza sono stati raccolti 400 offerte. 90 persone hanno offerto 50 euro a testa, 130 hanno offerto 200 euro e 180 hanno offerto 100 euro. Qual è l'offerta mediana?

Soluzione

Costruiamo la tabella delle frequenze cumulate:

Offerta	Frequenza	Frequenza cumulata
50 €	90	90
100 €	180	90 + 180 = 270
200 €	130	270 + 130 = 400

Il numero delle offerte è un numero pari $n = 400$.

Pertanto le posizioni centrali $\frac{n}{2}$ *e* $\frac{n}{2}+1$ *sono: 200 e 201.*

Dalla colonna delle frequenze cumulate si deduce che le offerte di posto n° 200 e n°201, ammontano a 100 €.

La mediana pertanto è 100 € (media aritmetica fra 100 € e 100 €).

Il significato della mediana, in questo esempio, è:

"almeno il 50% delle offerte è minore uguale a 100 € e almeno il 50% delle offerte è maggiore uguale a 100 €.

Mediana di un carattere suddiviso per classi

Per calcolare la mediana di una distribuzione di frequenze di un carattere suddiviso per classi occorre:

1. determinare la classe che contiene la mediana detta **classe mediana**
2. calcolare il valore approssimato della mediana

Esempio

Qual è il peso mediano della seguente distribuzione di frequenze:

Peso *(kg)*	**Frequenza**
$40 \leq m < 50$	8
$50 \leq m < 60$	24
$60 \leq m < 70$	22
$70 \leq m < 80$	18
$80 \leq m < 90$	4
$90 \leq m < 100$	2

Soluzione

Costruiamo la tabella delle frequenze cumulate:

Peso *(kg)*	**Frequenza**	**Frequenza cumulata**
$40 \leq m < 50$	8	8
$50 \leq m < 60$	24	8 + 24 = 32
$60 \leq m < 70$	22	32 + 22 = 54
$70 \leq m < 80$	18	54 + 18 = 72
$80 \leq m < 90$	4	72 + 4 = 76
$90 \leq m < 100$	2	76 + 2 = 78

Il campione è costituito da 78 individui (78 è pari).

La mediana è data pertanto dalla media fra il 39° e il 40° peso registrato.

Questi due elementi appartengono alla classe $60 \leq m < 70$.

Pertanto la mediana appartiene a tale classe, detta classe mediana.

In definitiva il valore approssimato della mediana è il valore centrale di tale classe, cioè: $\frac{60+70}{2} = 65$.

La moda

La **moda** è la modalità che si presenta con la massima frequenza.

La moda, a differenza dei primi due indici statistici media e mediana, si può determinare anche nel caso di caratteri qualitativi.

Esempio

La moda dell'insieme dei valori $A = \{5,\ 8,\ 8,\ 6,\ 2,\ 8,\ 5,\ 6\}$ *è* 8 .

L'insieme dei valori $B = \{5,\ 8,\ 11,\ 8,\ 23,\ 56,\ 11\}$ *ha due mode: 8 e 11* .

Nell'insieme dei valori $C = \{5,\ 5,\ 5,\ 5,\ 5,\ 5,\ 5\}$ *la moda non esiste.*

La moda dell'insieme dei valori sotto riportati è Trebisacce.

Studente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Paese di provenienza	T	P	A	T	T	P	A	V	V	P	P	T	T	V	V	T	P	T

(T = Trebisacce P = Plataci A = Albidona V = Villapiana)

Classe modale per un carattere suddiviso per classi

Nel caso di caratteri quantitativi suddivisi in classi, la moda viene sostituita dalla classe modale.
La **classe modale** :

- se le classi hanno la stessa ampiezza, è la classe che ha la frequenza più alta
- se le classi hanno ampiezze diverse, è la classe che ha la densità di frequenza più alta

La **densità di frequenza** è il rapporto fra la frequenza e l'ampiezza della classe.

Esempio 1

Da un'indagine effettuata su un campione di persone si è ottenuto la distribuzione di frequenze sotto riportata.

Peso (kg)	$40 \leq m < 50$	$50 \leq m < 60$	$60 \leq m < 70$	$70 \leq m < 80$	$80 \leq m < 90$	$90 \leq m < 100$
Frequenza	*8*	*24*	*22*	*18*	*4*	*2*

Le classi hanno la stessa ampiezza. Pertanto la classe modale è quella che ha maggior frequenza, cioè la classe $[50, 60[$.

Esempio 2

Da un'indagine effettuata su un campione di persone si è ottenuto la distribuzione di frequenze sotto riportata.

Peso (kg)	$40 \leq m < 50$	$50 \leq m < 60$	$60 \leq m < 75$	$75 \leq m < 80$	$80 \leq m < 90$	$90 \leq m < 100$
Frequenza	*8*	*24*	*22*	*18*	*4*	*2*

Le classi non hanno la stessa ampiezza. Pertanto occorre calcolare le densità di frequenza.

Peso (kg)	$40 \leq m < 50$	$50 \leq m < 60$	$60 \leq m < 75$	$75 \leq m < 80$	$80 \leq m < 90$	$90 \leq m < 100$
Frequenza	*8*	*24*	*22*	*18*	*4*	*2*
Densità di Frequenza	$\frac{8}{10} = 0{,}8$	$\frac{24}{10} = 2{,}4$	$\frac{22}{15} \cong 1{,}5$	$\frac{18}{5} = 3{,}6$	$\frac{4}{10} = 0{,}4$	$\frac{2}{10} = 0{,}2$

Pertanto la classe modale è $[75, 80[$.

La variabilità

Gli indici di posizione, visti nel capitolo precedente, non sempre sono sufficienti a dare una corretta visione d'insieme di un fenomeno. Tali valori non dicono quanto ciascun dato si discosta dal valore di sintesi considerato. Quanto detto è evidenziato dal seguente semplice esempio.

Esempio

Consideriamo le pressioni arteriose massime registrate in due pazienti A e B .

Paziente	Valori										Media
A	160	170	120	195	165	170	110	95	100	115	140
B	135	140	140	145	130	140	145	140	140	145	140

La media aritmetica dei valori delle pressioni arteriose massime dei due pazienti è, per entrambi, 140.

Ma, come si evince dalla tabella, il paziente B non ha sbalzi di pressione, mentre il paziente A ha dei valori molto alti (195) e dei valori molto bassi (95).

Pertanto la media aritmetica non riesce a cogliere questa variabilità.

La **variabilità** è l'attitudine di un fenomeno a manifestarsi sulle varie unità statistiche con modalità diverse e distanti tra loro. Per misurare la variabilità di un fenomeno esistono i cosiddetti **indici di variabilità**.

I più importanti indici di variabilità sono:

- il campo si variazione
- la varianza
- lo scarto quadratico medio.

Campo di variazione

Il campo di variazione è la differenza fra la più piccola e la più grande fra le modalità osservate.

Esempio

Nell'esempio precedente, il campo di variazione delle pressioni arteriose del paziente A è: $195 - 90 = 105$,
mentre il campo di variazione delle pressioni arteriose del paziente B è: $145 - 130 = 15$.

Anche questo indice però, fornisce una misura molto grossolana della variabilità delle modalità osservate. Esso dipende solo dalle due modalità estreme e non è influenzato dal variare di tutte le altre.

Varianza

Una misura più raffinata del campo di variazione è dato dalla varianza.

Indicando con $\bar{x}$ la media aritmetica delle n modalità $x_1,\ x_2,\ \dots, x_n$ del fenomeno osservato, sono chiamati **scarti** le differenze fra le n modalità $x_1,\ x_2,\ \dots, x_n$ e il valore medio, cioè: $x_1 - \bar{x}$, $x_2 - \bar{x}$, $\dots$ $x_n - \bar{x}$.

La **varianza** V è la media aritmetica dei quadrati degli scarti:

$$V = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ \dots \ + (x_n - \bar{x})^2}{n}$$

Per calcolare la varianza, in particolare modo quando non si utilizza un computer, è possibile utilizzare anche la seguente formula:

$$V = \frac{{x_1}^2 + {x_2}^2 + \ \dots \ + {x_n}^2}{n} - \bar{x}$$

Ma anche la varianza ha un **difetto**: a causa dell'elevamento al quadrato degli scarti, non presenta la stessa unità di misura delle modalità del carattere. Per tal motivo viene definito un nuovo indice che ristabilisce la stessa unità di misura con le modalità del carattere: lo scarto quadratico medio.

Scarto quadratico medio

Lo scarto **quadratico medio** o **deviazione standard**, è la radice quadrata della varianza.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ \ldots \ + (x_n - \bar{x})^2}{n}}$$

Proprietà

1. Se a tutte le modalità di un carattere quantitativo $x_1,\ x_2,\ \ldots, x_n$ si aggiunge (o si toglie) uno stesso numero reale k, lo scarto quadratico medio resta invariato.
2. Se tutte le modalità di un carattere quantitativo $x_1,\ x_2,\ \ldots, x_n$ vengono moltiplicati per uno stesso numero reale k, lo scarto quadratico medio della nuova serie di valori risulta moltiplicata per $|k|$.

Nota

Se invece sono note le frequenze con le quali si presentano i dati si utilizza la formula:

$$s = \sqrt{\frac{f_1 \cdot (x_1 - \bar{x})^2 + f_2 \cdot (x_2 - \bar{x})^2 + \ \ldots \ + f_n \cdot (x_n - \bar{x})^2}{n}}$$

Esempio 1

Riconsideriamo l'esempio precedente.

Paziente	Valori										Media
A	160	170	120	195	165	170	110	95	100	115	140
B	135	140	140	145	130	140	145	140	140	145	140

Pressioni arteriose massime registrate in due pazienti A e B

Paziente A	Modalità	Scarto	$(\text{Scarto})^2$
1	160	20	400
2	170	30	900
3	120	-20	400
4	195	55	3025
5	165	25	625
6	170	30	900
7	110	-30	900
8	95	-45	2025
9	100	-40	1600
10	115	-25	625

Media	140		Varianza	1140
Moda	170		S. q. m.	33,8
Mediana	140			

Paziente B	Modalità	Scarto	$(\text{Scarto})^2$
1	135	-5	25
2	140	0	0
3	140	0	0
4	145	5	25
5	130	-10	100
6	140	0	0
7	145	5	25
8	140	0	0
9	140	0	0
10	145	5	25

Media	140		Varianza	20
Moda	140		S. q. m.	4,5
Mediana	140			

Il valore basso dello $\boldsymbol{S.q.m. = 4,5}$ *del paziente B è rassicurante per la sua salute.*

Più preoccupante è il valore alto dello $\boldsymbol{S.q.m. = 33,8}$ *del paziente A.*

Esempio 2

Supponiamo che quattro studenti, che indicheremo con A, B, C, D, abbiano conseguito i seguenti punteggi in una serie di 4 test di ammissione ad un corso di specializzazione.

A	*26*	*16*	*24*	*30*
B	*10*	*26*	*30*	*30*
C	*25*	*26*	*23*	*22*
D	*26*	*24*	*24*	*22*

Se solo due di essi potranno essere ammessi al corso, come stendere una graduatoria di ammissione?

La prima cosa che viene in mente di fare è calcolare la media aritmetica dei punteggi conseguiti da ognuno di essi: tale media è però 24 in tutti e quattro i casi; quindi non ci possiamo basare su di essa per il confronto fra gli studenti.

Se però confrontiamo le distribuzioni dei punteggi nei quattro casi, ci accorgiamo che essi si distribuiscono in modo molto diverso uno dall'altro rispetto alla media. Questo fatto ci suggerisce di studiare la variabilità come studio della dispersione intorno ad un valore fissato, detto polo, che di solito coincide con una delle misure di posizione, nel nostro caso la media aritmetica.

Cominciamo allora a calcolare la distanza di ciascuno dei dati dalla media. Si ha che:

per lo studente A gli scarti sono: $26-24=2$ $16-24=-8$ $24-24=0$ $30-24=6$

per lo studente B gli scarti sono: $10-24=-14$ $26-24=2$ $30-24=6$ $30-24=6$

per lo studente C gli scarti sono: $25-24=1$ $26-24=2$ $23-24=-1$ $22-24=-2$

per lo studente D gli scarti sono: $26-24=2$ $24-24=0$ $24-24=0$ $22-24=-2$

Per sintetizzare questi scarti potremmo calcolare la loro media; tuttavia, poiché sappiamo che la somma degli scarti dalla media aritmetica è nulla, questo calcolo non ci darebbe informazioni aggiuntive sulla dispersione.

Allora, riflettendo sul fatto che la somma degli scarti è nulla perché gli scarti negativi compensano quelli positivi, possiamo pensare di eliminare l'influenza del segno considerando i quadrati degli scarti e facendone poi la media che chiameremo media quadratica.

Nel caso dei nostri studenti avremo dunque che la media quadratica degli scarti è:

Studente A	*Studente B*	*Studente C*	*Studente D*
$\sqrt{\frac{2^2+(-8)^2+0^2+6^2}{4}}=5{,}1$	$\sqrt{\frac{14^2+2^2+6^2+6^2}{4}}=8{,}2$	$\sqrt{\frac{1^2+2^2+1^2+2^2}{4}}=1{,}6$	$\sqrt{\frac{2^2+0^2+0^2+2^2}{4}}=1{,}4$

Si può allora concludere che lo studente D presenta una minor variabilità, seguito nell'ordine dagli studenti C, A, B.

I due studenti ammessi al corso saranno quindi D e C, in quanto il loro rendimento è più costante.

Coefficiente di variazione o indice di variabilità relativo

Per la proprietà 2, lo S.q.m. è influenzato dall'unità di misura prescelta. Se la stessa indagine viene effettuata con unità di misura differenti i valori degli scarti sono differenti (proporzionali all'unità di misura scelta rispetto all'altra). Infatti se ad esempio, si calcola lo S.q.m. delle altezze degli alunni di una classe, una prima volta riportando i dati in metri e un'altra riportando i dati in centimetri si ottiene che lo S.q.m. del primo calcolo è 100 volte lo S.q.m. del secondo.

Il coefficiente di variazione invece, consente di confrontare distribuzioni di dati che si riferiscono a fenomeni diversi e/o con unità di misura diverse. Esso è un numero puro non influenzato dall'unità di misura scelta.

$$C_V = \frac{Scarto\ quadratico\ medio}{media\ aritmetica}$$

Per spiegare il significato del coefficiente di variazione è utile il seguente esempio.

Esempio

Di un campione di cento ragazzi vengono registrati i pesi in chilogrammi e le altezze in centimetri.

Lo scarto quadratico medio delle altezze fornisce una misura della variabilità, in centimetri, delle altezze misurate.

Mentre lo scarto quadratico medio dei pesi fornisce una misura della variabilità, in chilogrammi, dei pesi dei ragazzi.

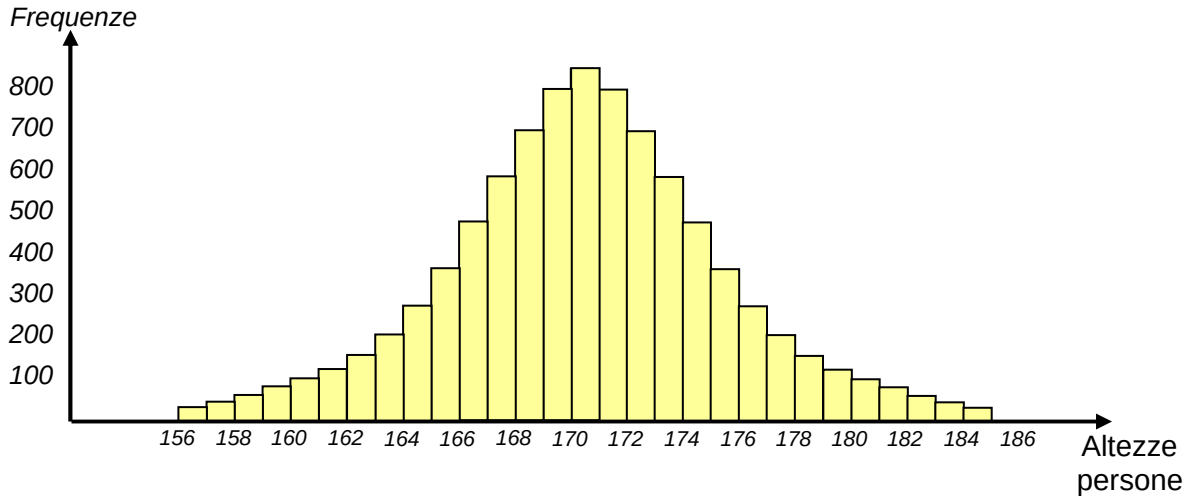
Se, a questo punto, si volesse analizzare se sono più omogenei tra loro le altezze oppure i pesi, i due s.q.m. non sarebbero confrontabili, perché si riferiscono a unità di misure differenti.

In casi come questo è utile utilizzare il coefficiente di variazione, o indice di variabilità relativo.

Grafico della distribuzione di frequenza

Una distribuzione di frequenza può essere rappresentata mediante istogrammi.

L'istogramma consiste in una serie di rettangoli affiancati (la cui base inferiore poggia sull'asse orizzontale del grafico, è centrata sul valor centrale ed è larga quanto l'ampiezza della classe) la cui altezza è proporzionale al valore rappresentato.



Per approfondire scaricare la seguente dispensa: http://www.mimmocorrado.it/mat/pro/statistica.pdf