CAPITOLO ZERO – ELEMENTI DI STATISTICA DESCRITTIVA

§1 Introduzione

Il termine "statistica" venne introdotto nel diciassettesimo secolo col significato di "scienza dello stato", volta a raccogliere e ordinare informazioni utili all'amministrazione pubblica: entità e composizione della popolazione, movimenti migratori, mutamenti anagrafici, tavole di natalità e mortalità, dati sui commerci, sui raccolti, sulla distribuzione della ricchezza, sull'istruzione e la sanità.

Il primo passo dell'attività statistica è la raccolta di dati che, se ben organizzata, risparmia fatica nelle operazioni successive e permette la corretta impostazione del lavoro di analisi.
Si dice **unità statistica** *la minima unità della quale si raccolgono i dati*.
Si dice **popolazione** *l'insieme delle unità statistiche oggetto di studio*.
Si dicono **caratteri** *le proprietà che sono oggetto di rilevazione*.
I caratteri possono essere **qualitativi** o **quantitativi**. I caratteri *qualitativi vengono indicati mediante espressioni verbali*. Sono caratteri qualitativi lo stato civile (celibe o nubile, coniugato/a, ecc.), il sesso (maschio o femmina), il colore degli occhi (chiari, castani, neri; ma anche, se si preferisce: grigi, azzurri, verdi, castani, neri). I caratteri *quantitativi sono esprimibili numericamente* e si dividono in **discreti** e **continui**. I caratteri *discreti*, come il numero degli alunni di una classe, o di reti segnate in una partita di calcio, *possono assumere solo determinati valori*, *quasi sempre numeri interi*. I caratteri *continui*, quali i pesi, le stature e più in generale le grandezze che possono essere misurate, *possono assumere qualsiasi valore reale in un dato intervallo* (anche se usualmente si impiegano numeri decimali finiti).

Esempio. Sorge una discussione fra due amici. Uno afferma che gli abitanti della loro città vanno al cinema assai raramente, in media una volta all'anno. L'altro sostiene invece che tale forma di divertimento è tornata di moda, e che la stima dell'amico va moltiplicata almeno per venti. Decidono di dedicare qualche tempo a un'indagine statistica per risolvere la questione. Evidentemente non possono intervistare tutti i loro concittadini: si limiteranno a un campione opportunamente scelto. Si pongono subito due interrogativi:

- Quale dev'essere l'ampiezza del campione affinché la stima sia attendibile e si possa essere ragionevolmente certi di aver stimato, con un accettabile margine di errore, il dato cercato?
  Basterà intervistare trenta persone, o ne occorreranno cento, oppure mille?
- Come si può essere sicuri che il campione non sia distorto, ma sia rappresentativo dell'intera popolazione?

È evidente che sarebbe scorretto condurre l'indagine all'uscita di un cinema, o fra gli ospiti di una casa di riposo; ma è preferibile intervistare le persone per strada oppure, se non si bada alla spesa, per telefono?

La situazione proposta è un tipico problema di **statistica induttiva**: la rilevazione dei dati, anziché sull'intera popolazione, è eseguita su una parte di essa, detta **campione**, e dall'esame di quest'ultimo si desumono informazioni (quanto attendibili?) sulla prima. Si tratta di questioni piuttosto complesse, alcune delle quali saranno esaminate nel Capitolo sesto dopo che avremo trattato gli elementi di base della teoria della probabilità.

In questo capitolo introdurremo soltanto alcuni elementi di **statistica descrittiva**, il cui compito è organizzare in modo facilmente dominabile i dati raccolti sull'intera popolazione in esame, senza trarre alcuna conclusione circa gli eventuali rapporti con una popolazione più ampia. Più precisamente, ci concentreremo su alcuni parametri con i quali si riassumono i dati rilevati, ossia le **medie** e gli **indici di dispersione**. Tralasceremo totalmente, per brevità, l'importante aspetto delle rappresentazioni grafiche dei dati.

§2 Medie

Il concetto di media è del tutto familiare, in quanto l'uomo è per natura incline a riassumere dati discordanti per poter concentrare l'attenzione sull'intensità media di un carattere e poter più facilmente confrontare dati omogenei relativi a popolazioni diverse. Molte nostre valutazioni e decisioni sono assunte, talvolta inconsciamente, facendo riferimento a valori medi. Così diciamo che il clima di Napoli è più mite di quello di Torino, che gli italiani del Nord hanno un reddito maggiore di quelli del Sud, che i maschi sono più forti delle femmine, e così via.

Da una sequenza di dati si possono ottenere varie medie, che assumono nomi diversi. In sostanza, una ***media*** è un valore opportunamente scelto e compreso fra il minimo e il massimo dei dati. In tutti i casi, *la media è un numero che ne sintetizza molti, e consente di averne una visione unitaria, ovviamente nascondendo la molteplicità dei dati da cui è ottenuta*. Così, il reddito medio delle famiglie italiane è un valore unico, utile per fare confronti con altre nazioni o con periodi passati, ma non evidenzia che i redditi sono molto diversi e molte famiglie sono al di sotto della soglia della sopravvivenza, mentre altre possiedono beni in grande quantità; la statura media ci consente di dire che gli svedesi sono, in media, più alti degli italiani, ma non evidenzia che molti italiani sono più alti di parecchi svedesi. Prenderemo in esame le seguenti medie: *moda*, *mediana*, *media aritmetica*, *media quadratica*, *media geometrica* e *media armonica*.

(A) Moda

Si dice **moda** *il carattere o il valore cui corrisponde la massima frequenza*.

***Esempio* 1**.

La sequenza di numeri 5, 6, 8, 8, 8, 12, 12, 14 ha moda 8.

La sequenza di numeri 5, 6, 8, 8, 8, 12, 14, 14, 14 ha due mode: 8 e 14.

Nella sequenza di numeri: 1, 2, 3, 4, 5, 6 si potrebbe anche dire, a stretto rigore, che vi sono sei mode; ma è più ragionevole concludere che in questo caso la moda non esiste.

***Esempio* 2**. Si rileva il numero delle stanze di ciascun appartamento di un condominio:

Numero delle stanze	Frequenze
2	1
3	3
4	8
5	2
6	1
7	1

La moda è 4, e in effetti sembra appropriato dire che in media gli appartamenti di quel condominio hanno quattro stanze.

***Esempio* 3**. Il direttore di un supermercato vuole provare a includere, fra gli articoli da vendere, anche delle pantofole per donna. Decide di tenere, almeno all'inizio, un'unica misura. Per individuare quale, chiede la misura del piede a dieci abituali clienti, ottenendo i seguenti dati:

38, 39, 37, 34, 40, 39, 35, 37, 39, 36

La sua scelta cadrà, evidentemente, sulla misura del 39 cui corrisponde la massima frequenza del campione (che, detto per inciso, è troppo piccolo per essere affidabile).

***Esempio* 4**. In un cantiere lo stipendio medio mensile dei quattro apprendisti è 600 €, dei venti operai è 1.000 €, del capocantiere 2.000 €. La moda è 1.000 € e sintetizza efficacemente la paga media dei dipendenti.

(B) Mediana

La **mediana** *è il valore che occupa il posto di mezzo, quando i dati sono disposti in ordine crescente*.

In altre parole, i dati che la seguono sono tanti quanti quelli che la precedono.

***Esempio* 5**. I voti di Pierino, intelligente ma discontinuo e scansafatiche, sono, in ordine crescente: 4, 5, 5, 6, 7, 8, 9

Il voto che occupa il posto di mezzo è 6, e in effetti pare equo assumerlo per sintetizzare la situazione. Si noti che la mediana, a differenza della media aritmetica (vedi (C) più avanti), può essere usata anche quando i dati non hanno carattere numerico: è sufficiente che possano essere disposti in ordine crescente. Sostituendo i voti con dei giudizi:

gravemente insufficiente, insufficiente, insufficiente, sufficiente,
discreto, buono, ottimo

la mediana è "sufficiente".

***Esempio* 6**. Consideriamo le seguenti sequenze di numeri o giudizi:

(a) 15, 18, 18, 19 (b) mediocre, discreto, discreto, ottimo

(c) 15, 16, 18, 19 (d) mediocre, discreto, buono, ottimo.

Quando i dati sono in numero pari esistono non uno, ma due valori centrali. Se essi coincidono, è naturale assumerli come mediana, per cui in (a) la mediana è 18 e in (b) è "discreto". Se invece non coincidono, ma sono numeri, si conviene di assumere come mediana la loro media aritmetica: in (c) la mediana è $\frac{16+18}{2} = 17$. Se, infine, i due dati centrali non coincidono e non hanno carattere numerico, come in (d), *non* si può parlare di mediana.

***Esempio* 7**. La seguente tabella mostra la distribuzione delle età dei capi famiglia degli Stati Uniti nell'anno 1957:

Età del capo famiglia	Numero in milioni
fino a 25	2,22
25-29	4,05
30-34	5,08
35-44	10,45
45-54	9,47
55-64	6,63
65-74	4,16
75 o più	1,66
	43,72 (totale)

Il totale delle frequenze è, in milioni, 43,72 e la sua metà è 21,86. Poiché la somma delle frequenze delle prime quattro classi è 21,8, di poco inferiore a tale valore, l'età mediana si colloca all'inizio della quinta classe. Possiamo concludere che l'età media dei capi famiglia è (appena superiore a) 45 anni, nel senso che quelli più giovani rispetto a tale età sono tanti quanti quelli più vecchi.

Moda e mediana hanno un vasto campo di applicazione, ma può succedere che, cambiando alcuni dei dati anche in modo vistoso, restino del tutto invariate. Ciò in qualche caso toglie efficacia a tali medie e sembra andare contro il senso comune.

Consideriamo i voti di Giulio:

Primo quadrimestre: 1, 5, 5, 5, 6, 6, 6

La mediana è 5 e vi sono due mode: 5 e 6.

Secondo quadrimestre: 4, 5, 5, 5, 6, 7, 10

La mediana è 5 e l'unica moda è 5. La sostituzione dell'1 iniziale con il 4 non ha portato alcun beneficio e, paradossalmente, il 7 e il 10 hanno sortito l'effetto di far sparire, delle due mode, quella favorevole.

Esaminiamo allora le **medie ferme**, cioè quelle medie che tengono conto di tutti i dati, indipendentemente dal loro ordine. Variando, anche di poco, anche uno solo dei dati, esse variano con continuità e senza salti. Le medie ferme si possono usare solamente per dati numerici.

(C) La media aritmetica

Dati n valori $X_1, X_2, ..., X_n$, si dice **media aritmetica** (o semplicemente **media**) *il valore che si ottiene dividendo la loro somma per il loro numero n*; indicando con M_a la media aritmetica, in formula si ha:

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

***Esempio* 8**. La media aritmetica M_a dei numeri 3, 7, 8, 9, 11 e 16 è:

$$M_a = \frac{3+7+8+9+11+16}{6} = \frac{54}{6} = 9$$

***Esempio* 9**. In un cantiere lo stipendio mensile dei quattro apprendisti è 600 €, dei venti operai è 1.000 €, del capocantiere 2.000 €. La media aritmetica degli stipendi è in euro:

$$M_a = \frac{4 \cdot 600 + 20 \cdot 1.000 + 1 \cdot 2.000}{25} = \frac{24.400}{25} = 976$$

La media aritmetica è di gran lunga la più nota e usata delle medie. Uno dei motivi è il seguente. Il senso comune attribuisce al concetto di media le caratteristiche delle medie ferme alle quali si è precedentemente accennato. Tuttavia, mentre le altre medie ferme: quadratica, geometrica e armonica che esamineremo fra poco, sono abbastanza complesse per chi è digiuno di matematica, quella aritmetica presuppone nozioni facili, possedute da tutti.

Il suo uso acritico e indiscriminato va però evitato: non è vero che, se io ho due polli e tu nessuno, è come se avessimo un pollo a testa; che per due amiche sia indifferente andare a passeggio con due ragazzi alti 170 cm, o con uno alto 140 cm e l'altro alto 200 cm; e così via.

È invece indifferente se su un ascensore, di portata massima 240 Kg, salgono tre persone il cui peso è 60 Kg, 70 Kg e 110 Kg rispettivamente, o tre persone tutte del peso di 80 Kg.

In generale, ogniqualvolta ha senso sommare i dati, l'uso della media aritmetica è appropriato. In tal caso *essa esprime quale sarebbe l'intensità costante del carattere in esame, se fosse ripartita in parti uguali*.

Inoltre la media aritmetica è il valore più attendibile nei due casi seguenti:

(a) *quando si eseguono diverse misurazioni di una stessa grandezza*.

Quando si misura più volte con uno strumento una grandezza fisica, in pratica non si ottiene sempre lo stesso risultato. Ciò è dovuto a vari fattori: al fatto che, operando in tempi successivi, possono essere mutate condizioni ambientali (temperatura, umidità, pressione atmosferica,...) che influenzano la grandezza da misurare e lo strumento, alle modalità di impiego dello strumento, alle incertezze nella lettura delle scale graduate, e così via. Proprio per questo, quando si vuole conoscere con precisione la misura di una grandezza, si eseguono diverse misurazioni. Si può dimostrare che, se le differenze tra le misure ottenute sono dovute ad errori accidentali, *la media aritmetica delle misurazioni è il valore più attendibile della misura della grandezza* (che è e resta ignota).

(b) *quando si misura il valore tipico in una popolazione omogenea*.

Ad esempio, quando si producono con uno stampo dei pezzi metallici, questi dovrebbero avere tutti lo stesso peso. Ma se si pesano i pezzi prodotti, i pesi risulteranno diversi, sia per gli errori di misurazione, ai quali si è accennato nel punto precedente, sia per errori di lavorazione (il materiale metallico non è perfettamente omogeneo, i vari pezzi non hanno mai forma identica, il funzionamento dello stampo è influenzato da fattori ambientali che variano nel tempo, ecc.). Si può dimostrare che *la media aritmetica dei pesi ottenuti dà il peso tipico che dovrebbe avere ciascun pezzo* (secondo il modello ideale derivato dallo stampo).

Spesso, anziché la media aritmetica semplice, si usa la **media ponderata**: assegnati agli n valori $X_1, X_2,..., X_n$ i pesi $p_1, p_2,..., p_n$ proporzionali all'importanza che vogliamo loro attribuire, la *media aritmetica ponderata* è:

$$\frac{X_1 \cdot p_1 + X_2 \cdot p_2 + ... + X_n \cdot p_n}{p_1 + p_2 + ... + p_n}$$

***Esempio* 10**. Supponiamo che nel corso dell'anno il pane sia aumentato del 18%, il prosciutto del 42% e il burro del 30%. Se si vuole stabilire l'aumento percentuale medio del costo della vita appare naturale dare un peso maggiore all'aumento del pane che non a quello del prosciutto o del burro. Ad esempio possiamo attribuire peso 8 all'aumento del pane, peso 1 all'aumento del prosciutto e peso 3 a quello del burro. La media aritmetica ponderata dei tre aumenti percentuali risulta:

$$\frac{18\% \cdot 8 + 42\% \cdot 1 + 30\% \cdot 3}{8+1+3} = 23\%$$

***Esempio* 11**. Per superare un esame uno studente deve sostenere una prova pratica, una prova scritta e una prova orale e ottenere una media superiore a 60. La prova pratica è meno importante di quella scritta, la quale, a sua volta, è meno importante di quella orale; esse hanno pesi 1, 2 e 3. Se uno studente merita 78 nella prova pratica, 44 nella scritta e 66 nella prova orale, la sua media ponderata è: $\frac{78\cdot 1+44\cdot 2+66\cdot 3}{1+2+3}=\frac{364}{6}=60{,}67$, per cui, seppur di strettissima misura, ha superato l'esame.

(D) Media quadratica

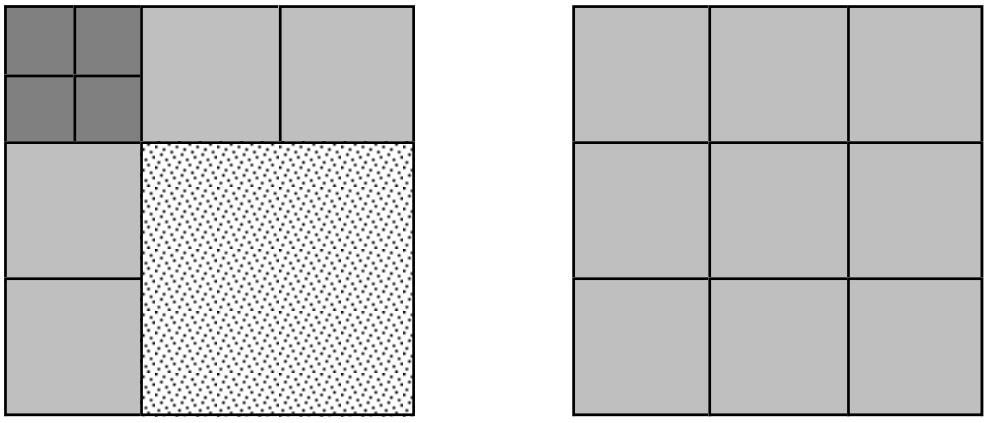
Dati n valori $X_1, X_2,..., X_n$, si dice **media quadratica** *la radice quadrata della media aritmetica dei loro quadrati*; indicando con M_q la media quadratica, in formula si ha:

$$M_q=\sqrt{\frac{X_1{}^2+X_2{}^2+...+X_n{}^2}{n}}$$

***Esempio* 12**. Calcolare la media quadratica dei seguenti numeri: 1, 1, 1, 1, 2, 2, 2, 2, 4.

$$M_q=\sqrt{\frac{1^2+1^2+1^2+1^2+2^2+2^2+2^2+2^2+4^2}{9}}=\ 2.$$

Geometricamente ciò si può interpretare dicendo che quattro quadrati di lato 1, quattro quadrati di lato 2 e un quadrato di lato 4 equivalgono a nove quadrati di lato 2:



***Esempio* 13**. Si vogliono sostituire tre tubi di raggio rispettivamente 2 cm, 3 cm e 4 cm con tre tubi di uguale raggio in modo che la portata complessiva resti inalterata. Quale deve essere il loro raggio?

Detta x la misura in cm del raggio incognito, deve essere:

$3\pi x^2=\pi 2^2+\pi 3^2+\pi 4^2$ e quindi:

$$x=\sqrt{\frac{2^2+3^2+4^2}{3}}\cong 3{,}11$$

e il raggio richiesto è la media quadratica dei raggi dei tre tubi dati.

(E) Media geometrica

Dati n valori $X_1, X_2,..., X_n$, si dice **media geometrica** *la radice n-esima del loro prodotto*; indicando con M_g la media geometrica, in formula si ha:

$$M_g=\sqrt[n]{X_1\cdot X_2\cdot...\cdot X_n}$$

***Esempio* 14**. La media geometrica dei tre numeri 10, 20, 60 è:

$$M_g = \sqrt[3]{10 \cdot 20 \cdot 60} \cong 22{,}9$$

come si trova facilmente con una calcolatrice, ed è notevolmente inferiore alla media aritmetica che è 30.

Evidentemente l'uso della media geometrica è appropriato quando il carattere in esame è moltiplicativo, cioè quando ha significato moltiplicare i dati.

***Esempio* 15**. Un trasformatore rende l'81%, un altro il 64%. Se si applicano in serie il rendimento complessivo è pari al prodotto dei due rendimenti. La loro media geometrica:

$$\sqrt{0{,}81 \cdot 0{,}64} = 0{,}72 = 72\%$$

ci dice quale rendimento uguale dovrebbero avere i due trasformatori per lasciare inalterato il loro rendimento complessivo.

Dati n valori positivi $X_1, X_2, ..., X_n$ l'inverso della loro media geometrica:

$$\frac{1}{\sqrt[n]{X_1 \cdot X_2 \cdot ... \cdot X_n}}$$

è uguale alla media geometrica dei loro inversi:

$$\sqrt[n]{\frac{1}{X_1} \cdot \frac{1}{X_2} \cdot ... \cdot \frac{1}{X_n}}$$

Questa proprietà rende la media geometrica la più opportuna quando si esamina il cambio fra due monete.

***Esempio* 16**. Il cambio fra la moneta *A* e la moneta *B* è 16/1 in una certa epoca e 25/1 un anno dopo. Come cambio medio assumiamo la media geometrica fra i due $\sqrt{16 \cdot 25} = 20$.
Ovviamente il cambio fra la moneta *B* e la moneta *A* era $\frac{1}{16}$ inizialmente e $\frac{1}{25}$ un anno dopo; il cambio medio risulta $\sqrt{\frac{1}{16} \cdot \frac{1}{25}} = \frac{1}{20}$ e i due cambi medi sono, come deve accadere, l'uno l'inverso dell'altro. È facile controllare che ciò non si verifica usando la media aritmetica.

(F) Media armonica
Dati n valori $X_1, X_2, ..., X_n$, si dice **media armonica** *l'inverso della media aritmetica dei loro inversi*; indicando con M_A la media armonica, in formula si ha:

$$M_A = \frac{1}{\frac{\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}}{n}} = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}}$$

Tale media, piuttosto complessa, può apparire astratta e lontana dalla realtà. Ha invece importanti applicazioni pratiche.

***Esempio* 17**. Percorro 21 Km alla velocità di 30 Km/h e altri 21 Km alla velocità di 70 Km/h. Qual è la velocità media?
Risolviamo il problema in generale.
Dette s la lunghezza comune dei due tratti e v_1 e v_2 le due velocità, il tempo t_1 impiegato nel primo tratto è $t_1 = \frac{s}{v_1}$. Analogamente il tempo t_2 impiegato nel secondo tratto è $t_2 = \frac{s}{v_2}$.
Il tempo complessivo è $t = t_1 + t_2 = \frac{s}{v_1} + \frac{s}{v_2}$, per cui la velocità media v_m risulta:

$$v_m = \frac{2s}{t} = \frac{2s}{t_1 + t_2} = \frac{2s}{\frac{s}{v_1} + \frac{s}{v_2}} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}}$$

cioè proprio la media armonica delle due velocità.
In questo caso la velocità media è, in Km/h:

$$\frac{2}{\frac{1}{30} + \frac{1}{70}} = 42$$

Uno dei metodi più efficaci per effettuare buoni investimenti a lunga scadenza è destinare a intervalli costanti la stessa somma all'acquisto dello stesso bene. In questo modo se ne acquista un'elevata quantità quando i prezzi sono bassi e una quantità modesta quando i prezzi sono alti, ottenendo un prezzo medio di acquisto più basso di quanto avverrebbe acquistando ogni volta una quantità costante di quel bene.

***Esempio* 18**. Un risparmiatore impiega, in ciascuno di due acquisti successivi, 2.100 € per comperare monete d'oro la cui quotazione è una volta di 70 € e l'altra volta di 30 €. Qual è il prezzo medio di acquisto?

Il risparmiatore acquista la prima volta $\frac{2.100}{70} = 30$ monete e la seconda volta $\frac{2.100}{30} = 70$ monete. Complessivamente spende 4.200 € per procurarsi 100 monete, ognuna delle quali gli è quindi costata mediamente 42 €. Tale prezzo, come si verifica facilmente, è proprio la media armonica dei due prezzi d'acquisto.

Consideriamo tre valori alquanto distanti fra loro come 1, 8, 125 e calcoliamone le quattro medie ferme:

$$M_A = \frac{3}{\frac{1}{1} + \frac{1}{8} + \frac{1}{125}} \cong 2{,}6 \qquad M_g = \sqrt[3]{1 \cdot 8 \cdot 125} = 10$$

$$M_a = \frac{1 + 8 + 125}{3} \cong 44{,}7 \qquad M_q = \sqrt{\frac{1^2 + 8^2 + 125^2}{3}} \cong 72{,}3$$

Si può dimostrare che quanto verificato in questo caso particolare vale in generale:

$$M_A \le M_g \le M_a \le M_q$$

e le uguaglianze valgono solo quando tutti i dati sono uguali fra loro.

§3 Indici di dispersione

Le medie riassumono in un unico valore il fenomeno studiato, ma non forniscono alcuna informazione sulla sua variabilità.

***Esempio* 19**. Si scopre che Marte è abitato da una specie intelligente simile alla nostra. Misurate le altezze di sette marziani adulti, si trova che moda e mediana coincidono e valgono 170 cm. Esaminiamo qualche possibile sequenza di dati che soddisfa tali condizioni.

(a) 167, 169, 170, 170, 170, 172, 172
La variabilità è piccola e pare che l'altezza dei marziani sia quasi costante.

(b) 161, 163, 170, 170, 173, 175, 178
La variabilità riscontrata è vicina a quella della statura umana.

(c) 80, 100, 120, 170, 170, 250, 300
La variabilità è notevole: su Marte vi sono nani e giganti.

Risulta quindi evidente che, ai fini di una descrizione sintetica ma significativa, è necessario definire dei parametri che indichino la dispersione dei dati o anche (è l'altra faccia di una stessa medaglia) la loro maggiore o minore concentrazione attorno a un valore medio.

(A) La più immediata misura della variabilità è il **campo di variazione**, cioè *la differenza fra il minimo e il massimo dei valori osservati*.

Nel precedente Esempio 19, il campo di variazione è 5 cm per la serie di dati (a), 17 cm per (b), 220 cm per (c) e risulta notevolmente significativo.

In genere però il campo di variazione, che tiene conto soltanto dei due valori estremi e non è influenzato in alcun modo da quelli intermedi, costituisce una misura troppo rozza della variabilità.

Assegnati n valori $X_1, X_2,..., X_n$ e indicata con M la media giudicata più opportuna nel caso in esame, appare naturale prendere in considerazione gli *scarti* da essa, ossia i valori $X_1 - M$, $X_2 - M$,..., $X_n - M$. Essendo M compresa tra il minimo e il massimo dei dati, alcuni scarti saranno positivi, altri negativi, e più o meno si compenseranno gli uni con gli altri.

Vediamo, anzi, se esiste una media tale che la somma degli scarti da essa sia nulla.

In tal caso deve essere: $(X_1 - M) + (X_2 - M) + ... + (X_n - M) = 0$

cioè: $(X_1 + X_2 + ... + X_n) - n\,M = 0$

da cui: $M = \dfrac{X_1 + X_2 + ... + X_n}{n}$

Quindi tale media esiste, ed è la media aritmetica: *la somma degli scarti dalla media aritmetica è sempre nulla*.

La somma degli scarti da qualsiasi altra media sarà diversa da zero, ma in genere piccola e poco significativa. È allora naturale considerare *non gli scarti, ma i loro valori assoluti* .

Si dice **valore assoluto** di un numero il numero stesso, se è positivo o nullo; il suo opposto, se il numero considerato è negativo.

Il valore assoluto di un numero si indica racchiudendolo fra due barre verticali: | |.

In formula: $|a| = \begin{cases} a & \text{se } a \geq 0 \\ -a & \text{se } a < 0 \end{cases}$

Ad esempio, $|+7| = +7$; $|-3| = +3$; $|-9| = +9$; $|0| = 0$

(B) Si definisce **scarto semplice medio** *da una media M la media aritmetica dei valori assoluti degli scarti da M*.

In formula:

$$\textit{scarto medio semplice} = \frac{|X_1 - M| + |X_2 - M| + ... + |X_n - M|}{n}$$

Calcoliamo, per ciascuna sequenza di dati dell'Esempio 19, lo scarto semplice medio dal valore 170.

Per (a) lo scarto semplice medio vale:

$$\frac{|167-170| + |169-170| + |170-170| + |170-170| + |170-170| + |172-170| + |172-170|}{7}$$

$$= \frac{3+1+0+0+0+2+2}{7} \cong 1{,}14\,.$$

Per (b) lo scarto semplice medio vale:

$$\frac{|161-170| + |163-170| + |170-170| + |170-170| + |173-170| + |175-170| + |178-170|}{7}$$

$$= \frac{9+7+0+0+3+5+8}{7} = \frac{32}{7} \cong 4{,}57$$

Per (c), infine, lo scarto semplice medio vale:

$$\frac{|80-170|+|100-170|+|120-170|+|170-170|+|170-170|+|250-170|+|300-170|}{7}$$

$$= \frac{90+70+50+0+0+80+130}{7} = \frac{420}{7} = 60.$$

***Esempio* 20**. Dati i valori 4, 8, 15, lo scarto semplice medio rispetto alla mediana è:

$$\frac{|4-8|+|8-8|+|15-8|}{3} = \frac{4+0+7}{3} = 3,\overline{6}.$$

e lo scarto semplice medio rispetto alla media aritmetica è

$$\frac{|4-9|+|8-9|+|15-9|}{3} = \frac{5+1+6}{3} = 4.$$

maggiore del precedente.

Quanto visto in questo esempio accade in generale. Si dimostra che:

fra tutte le medie, la mediana è quella rispetto alla quale lo scarto semplice medio assume il valore minimo.

(C) Un altro indice di dispersione è lo **scarto quadratico medio**, che si indica con *s* ed è così definito:

$$s = \sqrt{\frac{(X_1-M)^2+(X_2-M)^2+\ldots+(X_n-M)^2}{n}}$$

Si può anche dire che lo scarto quadratico medio è la media quadratica degli scarti.

Calcoliamo, per ciascuna sequenza di dati dell'Esempio 19, lo scarto quadratico medio dal valore 170.

Per (a):

$$s = \sqrt{\frac{(167-170)^2+(169-170)^2+(170-170)^2+(170-170)^2+(170-170)^2+(172-170)^2+(172-170)^2}{7}}$$

$$= \sqrt{\frac{3^2+1^2+0^2+0^2+0^2+2^2+2^2}{7}} \cong 1,60$$

Per (b): $s = \sqrt{\frac{9^2+7^2+0^2+0^2+3^2+5^2+8^2}{7}} \cong 5,71$

Per (c): $s = \sqrt{\frac{90^2+70^2+50^2+0^2+0^2+80^2+130^2}{7}} \cong 74,45$

Si vede quindi che lo scarto quadratico medio è un indice più sensibile dello scarto semplice medio.

***Esempio* 21**. Dati i valori 4, 8, 15, lo scarto quadratico medio rispetto alla mediana è:

$$s = \sqrt{\frac{(4-8)^2+(8-8)^2+(15-8)^2}{3}} \cong 4,65$$

e rispetto alla media aritmetica è:

$$s = \sqrt{\frac{(4-9)^2+(8-9)^2+(15-9)^2}{3}} \cong 4,55$$

Quanto visto in questo esempio accade in generale. Si dimostra che:

fra tutte le medie, la media aritmetica è quella rispetto alla quale lo scarto quadratico medio assume il valore minimo.

Dimostrazione. Poiché lo scarto quadratico medio è minimo quando è minima la quantità:

$$(X_1 - M)^2 + (X_2 - M)^2 + ... + (X_n - M)^2$$

basterà limitarsi a considerare quest'ultima. Indicando con M_a la media aritmetica, e posto:

$d = M_a - M$, per cui $M = M_a - d$

si ha:

$$(X_1 - M)^2 + (X_2 - M)^2 + ... + (X_n - M)^2 =$$

$$[(X_1 - M_a) + d]^2 + [(X_2 - M_a) + d]^2 + ... + (X_n - M_a) + d]^2 =$$

$$[(X_1 - M_a)^2 + (X_2 - M_a)^2 + ... + (X_n - M_a)^2] +$$
$$+ 2d[(X_1 - M_a) + (X_2 - M_a) + ... + (X_n - M_a)] + n\,d^2 =$$

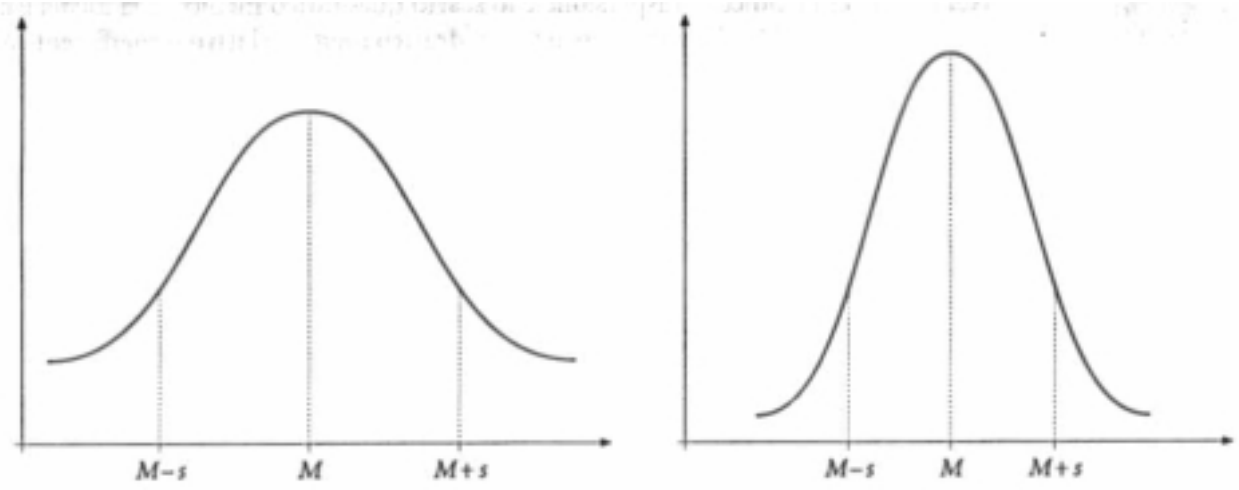
$$[(X_1 - M_a)^2 + (X_2 - M_a)^2 + ... + (X_n - M_a)^2] + n\,d^2$$

perché, come dimostrato in precedenza, la quantità:

$$(X_1 - M_a) + (X_2 - M_a) + ... + (X_n - M_a)$$

cioè la somma degli scarti dalla media aritmetica, vale 0.
Quindi: la somma dei quadrati degli scarti da una generica media M è uguale alla somma dei quadrati degli scarti dalla media aritmetica M_a, aumentata della quantità (mai negativa) $n\,d^2$.

Lo scarto semplice medio e lo scarto quadratico medio sono indici di dispersione significativi in quanto tengono conto di tutti i dati. Il secondo, nonostante sia matematicamente più complicato, è il più usato in statistica per la seguente ragione. In molte circostanze si verifica che le frequenze di un dato carattere hanno una **distribuzione normale**, ossia si distribuiscono in modo simmetrico e decrescente rispetto a un valore tipico al quale spetta la massima frequenza. L'andamento delle frequenze è allora rappresentato da una curva a campana, detta **curva di Gauss**, che ha molteplici riscontri in fenomeni reali:



Ad esempio, hanno una distribuzione normale le stature, i pesi, le misure toraciche delle persone, i valori ottenuti con misurazioni ripetute di una stessa grandezza (se esse sono soggette solo ad errori accidentali), i valori dei pezzi lavorati dalle macchine (soggetti ad errori di lavorazione e di misurazione). Nelle distribuzioni normali media aritmetica, moda e mediana coincidono nel valore M nel quale la curva raggiunge il suo valore massimo. Lo scarto quadratico medio determina la forma della curva di Gauss. Nella figura sono rappresentate due distribuzioni normali che hanno stesso valore medio M e diversa ampiezza

dovuta a differenti scarti quadratici medi: in quella a sinistra s è maggiore e la curva è meno ripida (le frequenze decrescono più dolcemente da entrambe le parti di M), in quella a destra s è minore e la curva è più ripida (le frequenze sono più addensate intorno al valore medio e decrescono più rapidamente quando ci si allontana da entrambe le parti di M).

In generale, quando un carattere ha distribuzione normale, si può dimostrare che:

(a) il 68,27% dei dati è compreso fra $M - s$ e $M + s$

(b) il 95,45% dei dati è compreso fra $M - 2s$ e $M + 2s$

(c) il 99,73% dei dati è compreso fra $M - 3s$ e $M + 3s$

Ad esempio, se tra 1000 persone si osserva un peso medio di 73 Kg con uno scarto quadratico medio di 5 Kg, si può affermare che circa 683 persone hanno un peso compreso fra 68 e 78 Kg, e circa 954 persone hanno un peso compreso tra 63 Kg e 83 Kg. Così, se le lampadine prodotte da una ditta hanno una durata media di 900 ore con uno scarto quadratico medio di 30 ore, si può affermare che il 68,27% delle lampadine avrà una durata compresa fra 870 ore e 930 ore, e la quasi totalità delle lampadine (il 99,73%) avrà una durata compresa fra 810 e 990 ore.

Nel Capitolo quinto vedremo come le varie curve gaussiane possono essere trasformate in una particolare di esse, detta **distribuzione normale standard**, la quale ha valore medio M uguale a 0 e scarto quadratico medio $s = 1$, che consente di risolvere numerosi problemi di calcolo della probabilità. Nel Capitolo sesto, infine, vedremo alcune significative applicazioni della distribuzione normale nella risoluzione di problemi di statistica induttiva.