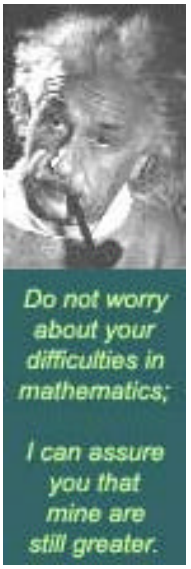
# DATISTICA (DA STATISTICA)

**a cura di Sergio Casiraghi**

**Testo ripreso dal CAPITOLO 3 del libro : ELEMENTI DI BASE DI PROBABILITÀ E STATISTICA di Eugenio Rapella, ed. Jackson, 1988. Il software che accompagna tale libro è stato realizzato da Sergio Casiraghi.**

# ELEMENTI DI STATISTICA




Problemi di calcolo delle probabilità di un certo interesse vengono affrontati con l'ausilio di strumenti matematici relativamente semplici. Un'analoga operazione in relazione alla statistica risulta più difficile in quanto i problemi di statistica si collocano a diversi e precisi livelli di difficoltà e i più interessanti (campionamenti, correlazioni ecc.) richiedono un bagaglio matematico piuttosto avanzato.

In queste pagine ci si limiterà pertanto agli argomenti di base; l lettore interessato ad approfondire questa parte troverà in altri testi il logico proseguimento e numerosi programmi.

## 1. I DATI

Siete stati nominati responsabili della produzione del calzaturificio X che inizierà l'attività nei prossimi mesi. Fra i moltissimi problemi da risolvere dovete naturalmente decidere, per ogni modello di calzatura, quanti paia di scarpe delle varie misure intendete produrre.

Due sono le situazioni che dovete assolutamente evitare: che troppi potenziali clienti non trovino le scarpe adatte ai loro piedi e che nei vostri magazzini si accumulino scarpe che nessuno vuole perché troppo grandi o troppo piccole.

Per organizzare un piano di produzione sensato (grossolani errori vi costerebbero il posto), vi servono delle informazioni e, se queste non sono già disponibili, dovete procurarvele, sintetizzarle e soprattutto interpretarle.

**Universo statistico**

Limitiamoci ad un tipo di calzatura destinata al pubblico maschile e supponiamo che siate interessati al solo mercato nazionale: la popolazione maschile residente in Italia costituirà il vostro UNIVERSO STATISTICO (o COLLETTIVO STATISTICO); le singole persone saranno le unità UNITÀ STATISTICHE.

**Carattere**

Ciò che intendete esaminare in relazione alle varie unità statistiche, in questo caso il "numero di scarpa", è il CARATTERE che, in generale, può essere espresso da un numero ( risultato di un qualche tipo di misura) o dalla presenza o meno di certi attributi.

Nel primo caso si parla di caratteri QUANTITATIVI (età, altezza, peso ecc.), nel secondo di caratteri QUALITATIVI (colore degli occhi, appartenenza ad un certo partito politico ecc.).

Fissato l'universo statistico e il carattere, o i caratteri, oggetto di studio, è necessario procurarsi le informazioni ovvero RILEVARE i DATI.

**Rilevazioni**

La rilevazione può essere TOTALE, se riferita all'intero universo; o PARZIALE, se riferita ad un CAMPIONE ovvero ad un sottoinsieme opportunamente scelto.

Nella vostra situazione dovrete per forza accontentarvi di una rilevazione parziale. Il campione dovrà essere RAPPRESENTATIVO dell'intero universo e di dimensioni tali da consentire l'estensione dei risultati a tutta la popolazione entro ragionevoli margini di sicurezza.

Poiché una rilevazione su vasta scala può essere assai costosa mentre una rilevazione troppo limitata può essere assai rischiosa, la determinazione di un campione adatto è un problema delicato; vi si accennerà nel quarto paragrafo di questo capitolo (Inferenza Statistica).

Supponiamo risolto il problema del campionamento: ora disponete di una consistente mole di dati; pagine fitte di numeri, risultati della vostra indagine statistica.

**Sintetizzare**

L'informazione che vi serve è contenuta nei dati in forma, per così dire, polverizzata:

**i dati** per ottenere qualcosa di utile è necessario un lavoro di sintesi.

Si tratta dunque di elaborare i dati in modo da ottenere un insieme limitato di valori che riflettano caratteristiche ritenute importanti.

Abbandoniamo l'esempio del calzaturificio e consideriamo una situazione più semplice:

Nella TABELLA 1 sono riportati i voti di un compito di matematica svolto nella classe 1^ sez. A. Il professore ha raccolto questi dati per cercare una risposta ai seguenti quesiti:

- il testo proposto era adeguato o eccessivamente difficile ?
- gli argomenti oggetto della prova sono stati recepiti allo stesso modo da tutti gli studenti ?

Tabella 1

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 6.8 | 8.0 | 7.1 | 8.0 | 5.8 | 5.5 | 4.5 | 5.6 | 5.1 |
| 6.6 | 6.5 | 5.5 | 8.0 | 5.0 | 3.8 | 7.5 | 5.9 | 8.0 |
| 5.0 | 7.5 | 7.0 | 6.5 | 6.8 | 4.5 | 6.1 | 6.6 | 3.9 |

(Per i voti intermedi si è posto: 6+ = 6.1; 6 1/2 =6.5; 6/7 =6.6; 7-- =6.8; 7- = 6.9 ecc.)

L'universo statistico è la classe in esame e le unità statistiche sono i singoli studenti: la rilevazione è totale e il carattere esaminato è il voto ottenuto, carattere quantitativo.

Il numero di dati è talmente esiguo che, in realtà, si potrebbe tentare una risposta semplicemente analizzando la tabella, ma è facile immaginare situazioni in cui si considerano migliaia di dati la cui elencazione non dice granché. Inoltre potrebbe risultare utile il confronto con un altro insieme di dati (ad esempio i voti ottenuti in una classe parallela in cui si è dedicato più tempo agli esercizi) e la comparazione dei dati grezzi diverrebbe difficoltosa.

Iniziamo il lavoro di sintesi che, grazie alle modeste dimensione della Tabella 1, potremo eseguire dettagliatamente senza ricorrere al calcolatore.

Una semplice riorganizzazione dei dati può già dire qualcosa; nella Tabella 2 i dati sono disposti, per righe, in ordine crescente:

Tabella 2

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3.8 | 3.9 | *4.5* | 4.5 | 5.0 | *5.0* | 5.1 | *5.5* | 5.5 |
| *5.6* | 5.8 | 5.9 | 6.1 | *6.5* | 6.5 | 6.6 | 6.6 | 6.8 |
| 6.8 | 7.0 | 7.1 | 7.5 | *7.5* | 8.0 | 8.0 | 8.0 | 8.0 |

Si nota ora che il voto più basso è 3.8, quello più alto è 8.

Se raggruppiamo i dati secondo la seguente

Tabella 3

| i voti x : | sono |
|---|---|
| 3=x<4 | 2 |
| 4=x<5 | 2 |
| 5=x<6 | 8 |
| 6=x<7 | 7 |
| 7=x<8 | 4 |
| 8=x<9 | 4 |

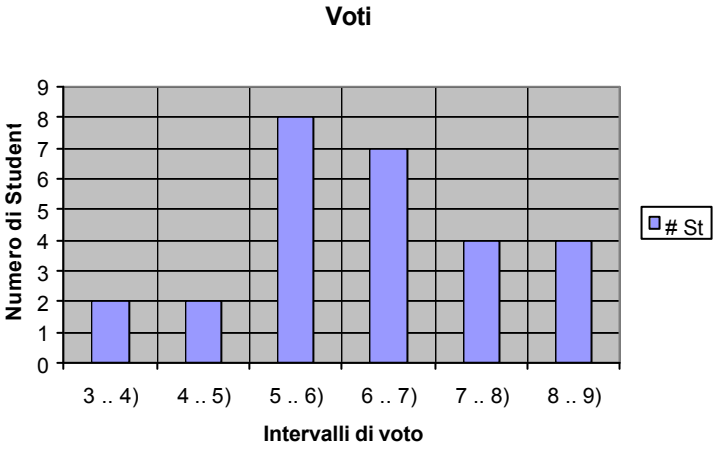



*fig.* **1** *Rappresentazione grafica dei dati della Tabella 3*

e riportiamo i valori su un grafico (fig.1), saltano all'occhio altre caratteristiche: i voti estremi sono più rari mentre la maggioranza si concentra tra il *5* e il 7.

Vi sono molti modi per rappresentare un insieme di dati ed esistono in commercio programmi come i fogli di lavoro elettronici che forniscono rappresentazioni grafiche suggestive (vedi le seguenti fig. 2 e 3).

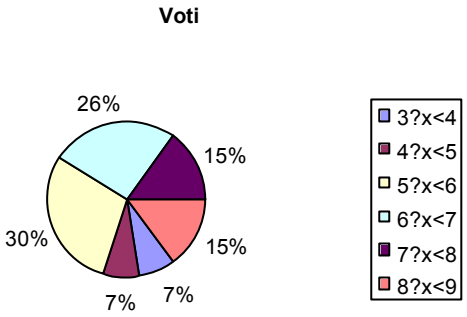



fig.**2** *Rappresentazione grafica"a torta" dei dati della Tabella 3*

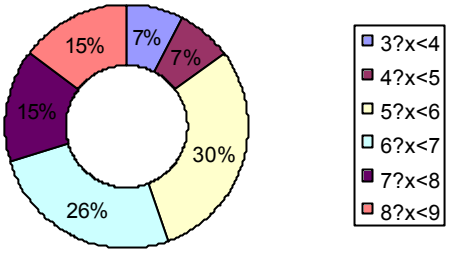



fig. **3** *Rappresentazione o torta dei dati della Tabella 3*

## 2. INDICI SINTETICI DI VARIABILI SEMPLICI: INDICI DI POSIZIONE

Consideriamo un insieme di dati relativi ad un carattere quantitativo; indichiamo con "n" il loro numero e con $x_i$ i = 1,2,...,n i valori espressi, in generale, da numeri reali (nell'esempio precedente e n=27=numero dei voti e $x_i$ =voto).

Al di la delle possibili rappresentazioni grafiche, è molto utile calcolare degli INDICI numerici che evidenzino certe caratteristiche dell'intero insieme; se un grafico fornisce una visione globale, un numero si presta ad essere confrontato con altri numeri e, come si vedrà, la cosa può essere molto utile.

Gli indici statistici più frequentemente utilizzati si distinguono in indici di POSIZIONE e indici di DISPERSIONE; i primi forniscono informazioni sulla "grandezza" dei dati, i secondi sulla loro diversità reciproca.

**Media aritmetica**

L'indice di posizione più famoso è senz'altro la MEDIA ARIT METICA: "la media aritmetica di "n" dati è quel valore che, sostituito a ciascuno dei dati, ne lascia invariata la somma".

Detto M tale valore, è, per definizione, $x_1+x_2+...+x_n = M+M+...+M = nM$ da cui

$$\mathbf{M = ( x_1+x_2+... + x_n )/n = ( \sum_i x_i )/n}$$

ESERCIZIO

Mostrare che, se x' e x" sono rispettivamente il minimo e il massimo di $x_i$, è **x' = M = x".**

Una proprietà caratteristica della media aritmetica è: $\mathbf{\sum_i (x_i - M) = 0}$

ovvero "la somma degli scarti dalla media aritmetica è zero".
In effetti: $(x_1 - M)+ (x_2 - M)+ \ldots+(x_n - M) = (x_1+x_2+...+ x_n)-(M+M+...+M)=nM-nM=0$.

Per quanto riguarda la Tab. 1 si ha M = 167.1/27 = 6.189

**Media aritmetica ponderata**

Quando i dati sono raggruppati e il dato $x_i$ compare nella statistica $f_i$ volte, l'espressione di M è data da:

$$\mathbf{M =( \sum_i f_i x_i )/ ( \sum_i f_i ) = (f_1 x_1+ f_2 x_2+... + f_k x_k )/ (x_1+ x_2+... + x_k )}$$

dove "k" è il numero di raggruppamenti (media aritmetica PONDERATA).

Volendo applicare la formula precedente ai dati della Tab. 3, potremo scegliere come $x_i$ il valore centrale di ogni intervallo della suddivisione scelta (cioè considerare uguali a 8.5 tutti i valori compresi tra 8 e 9 ecc.), ottenendo:

*M=(3.5x2+* 4.5x2 *+ 5.5x8 + 6.5x7 + 7.5x4+ 8.5x4)/(*2+2+8+7+4+4)=169.5/27=6.278

La media ottenuta in questo modo è diversa dalla precedente poichè, considerando il valore

centrale, si commette un errore di approssimazione.

Alla suddivisione in classi e alla media ponderata si ricorre soprattutto quando il carattere quantitativo è CONTINUO ovvero quando il dato può assumere, almeno teoricamente, un qualsiasi valore reale di un certo intervallo. Così, se viene misurata la statura di un gruppo di persone, converrà contare il numero di individui la cui altezza è compresa tra 175 e 180 cm, tra 180 e 185 cm ecc. evitando, già in fase di raccolta dei dati, una eccessiva frantumazione. Diverso il discorso per un carattere DISCRETO come, ad esempio, il numero di persone che costituiscono un nucleo familiare; una tabella del tipo:

| **I nuclei familiari formati da** | **sono** |
|---|---|
| 1 persona | 10 |
| 2 persone | 52 |
| 3 persone | 60 |
| 4 persone | 47 |
| ecc. | |

contiene dati raggruppati per classi, ma, poiché si tratta di un carattere discreto, la media aritmetica coincide con quella ponderata.

Nel caso del voto, il carattere andrebbe considerato come discreto, dato che esiste un numero finito di voti intermedi convenzionalmente assegnati (7+, 7 1/2 ecc.), ma la suddivisione in classi di ampiezza "1 voto" risulta comoda sia dal punto di vista grafico sia per gli scopi che si propone il professore.

**Mediana**

Un altro indice di posizione di uso molto frequente è la MEDIANA (? ) definita come "valore centrale" dei dati quando questi siano ordinati in ordine non decrescente.

Se $x_1=x_2=\ldots=x_n$

è

**? = $x_{(n+1)/2}$** se n è dispari (valore centrale)

mentre si assume

**? = $(x_{n/2} + x_{(n/2)+1})/2$** se n è pari (media aritmetica dei 2 valori centrali)

ESEMPIO 1:

DATI : 7; 2; 7; 10; 100 (n=5)
DATI ORDINATI : 2; 7; **7**; 10; 100

(n+1)/2 = 3; ? = 7

ESEMPIO 2:

DATI : 4; 1; 1; 6; 30; *5* (n=6)
DATI ORDINATI :1; 1; **4; *5;*** *6;* 30

? = (4 + 5)/2 =4.5

Nell'esempio dei voti è n=27 e la mediana sarà il 14° valore delle Tab. 2 : ? = 6.5 il che vuol dire che una metà della classe ha ottenuto voti inferiori (o uguali) a *6.5,* l'altra metà superiori (o uguali) a *6.5*.

Una grande differenza tra media e mediana sta ad indicare una distribuzione "sbilanciata": un solo valore eccezionalmente grande o eccezionalmente piccolo influenza la media aritmetica ma non la mediana (Es. DATI 1; 2; 96 : M=33; ?=2).

Media aritmetica e mediana godono di notevoli proprietà: la media aritmetica rende minima la funzione **$f(t) = ?_i (x_i - t)^2$** ; la mediana minimizza la funzione **$g(t) = ?_i |x_i - t|$**. Verifichiamole in

un esempio:

Sia n=3 : $x_1 = 1$; $x_2 = 2$; $x_3 = 12$. La media aritmetica è *5,* la mediana è 2.
La funzione
$f(t)=(1-t)^2 + (2-t)^2 + (12-t)^2 = 3t^2-30t + 149$
assume il valore minimo per $t=5$, ascissa del vertice della parabola.
La funzione
$g(t) = |1-t| + |2-t| + |12-t|$

può essere riscritta come

$$g(t) = (15-3t)*(t<1)+(13-t)*(1=t<2)+(t+9)\ (2=t<12)+(3t-15)(12=t)$$

il minimo si ottiene (vedi fig.4) per t =2.

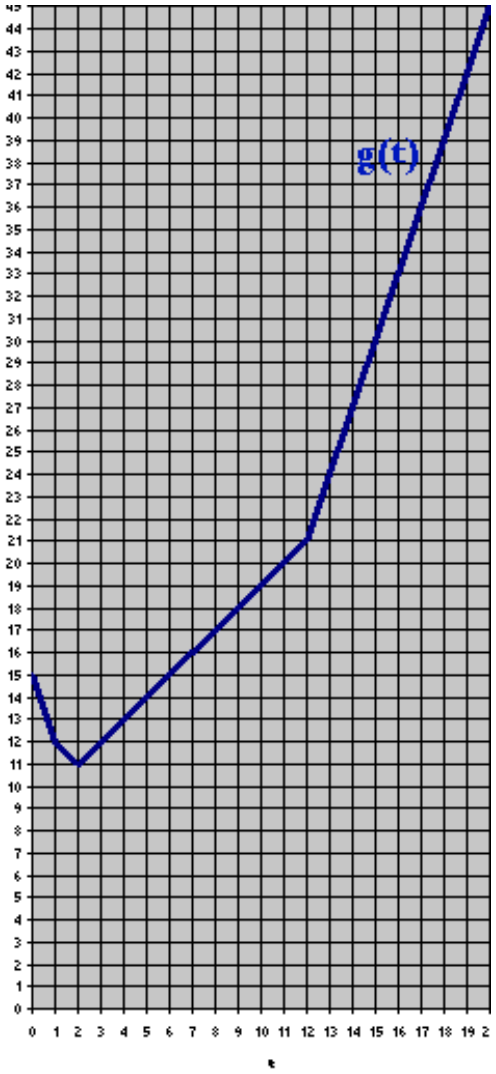



fig. **4**

**Quartili**

In analogia con la mediana, che divide i dati ordinati in due gruppi ugualmente numerosi, si possono definire altri indici di posizione. Per primo, secondo, terzo QUARTILE ($q_1$, $q_2$, $q_3$) si intendono i tre valori che dividono i dati in quattro tronconi di pari dimensione, una volta che i dati siano disposti in ordine non decrescente.

In riferimento alla Tab. 2, il 1', 2', 3' quartile corrispondono rispettivamente al 7', 14', 21' valore dell'elenco:

$q_1 = 5.1$ ; $q_2 = 6.5$ ; $q_3 = 7.1$

il secondo quartile ($q_2$) è la mediana (? ).

Generalizzando ulteriormente, si possono definire analoghi indici di posizione (quintili, decili, percentili) di un certo interesse se i dati sono molto numerosi.

**Moda**

Definiamo come MODA (o VALORE NORMALE) il dato che compare più spesso nell'insieme dei dati. Se il carattere in esame è di tipo qualitativo o quantitativo discreto, la moda sarà appunto la caratteristica o il valore cui corrisponde la massima frequenza; nel caso di un carattere continuo (o, comunque, registrato per raggruppamenti), avrà senso parlare di CLASSE MODALE in riferimento all'intervallo cui compete la maggior frequenza (supponendo di considerare intervalli della stessa ampiezza).
Nell'esempio dei voti, la moda è 8.

**Media geometrica**

Oltre la media aritmetica, sono definite altre medie utilizzate per specifici problemi:
si definisce **MEDIA GEOMETRICA** dei numeri positivi $x_1, x_2, .., x_n$ quel numero G che, sostituito ai dati, ne lascia invariato il prodotto.

Per definizione: $x_1 \cdot x_2 \cdot ... \cdot x_n = G \cdot G \cdot ... \cdot G = G^n$ da cui $G = \sqrt[n]{(x_1 \cdot x_2 \cdot ... \cdot x_n)}$

Il calcolo di G viene normalmente eseguito utilizzando i logaritmi:

$$\log G = [\log x_1 + \log x_2 + ... + \log x_n]/n$$

(il logaritmo della media geometrica è uguale alla media aritmetica dei logaritmi dei dati).

Sempre in relazione ad un insieme di dati positivi $x_1, x_2, .., x_n$ ricordiamo la definizione di altre due medie (raramente utilizzate):

- **MEDIA QUADRATICA:** $Q = \sqrt[n]{[(x_1 + x_2 + .. + x_n)/n]}$
- **MEDIA ARMONICA:** $A = n/(1/x_1 + 1/x_2 + .. + 1/x_n)$

E' possibile dimostrare che, se $x_i > 0$ ? i, è A = G = M = Q.

Esercizio. Dati i valori 3; 3; *5;* 7; 10, verificare che A ? 4.51; G? 5.0l; M=5.6; Q? 6.2.

Esercizio. Verificare che, se gli n dati sono tutti uguali ad uno stesso numero h>0, è A=G=M=Q=h.

Esercizio. Dimostrare che A = G = M = **Q** nel caso di due dati positivi.

Soluzione

1. E' $(a - b)^2 ab$ ? 0
   $[(a + b)^2 - 4ab] ab$ ? 0
   $(a + b)^2 ab$ ? $4a^2 b^2$
   (a+b) ?(ab) ? 2ab
   ?(ab) ? 2ab / (a + b ) = 2 / ( 1/a + 1/b )

   cioè G(a,b) ? A(a,b)

2. E' $(\sqrt{a} - \sqrt{b})^2$ ? 0
   a + b – 2 ?(ab) ? 0
   a+b ? 2 ?(ab)
   (a + b)/2 ? ?(ab)

   quindi M(a,b) ? G(a,b)

3. E' $(a – b)^2$ ? 0
   $(a^2 + b^2 - 2ab)/4$ ? 0
   $(2a^2 + 2b^2 - a^2 - b^2 - 2ab)/4$ ? 0
   $(a^2 + b^2)/2 – (a^2 + b^2 + 2ab)/4$ ? 0
   $(a^2 + b^2)/2$ ? $[(a + b)/2]^2$
   $\sqrt{[(a^2 + b^2)/2]}$ ? (a + b)/2

ovvero Q(a,b) ? M(a,b)

Chi si avvicina per la prima volta alla statistica sarà, a questo punto, un pò disorientato: si voleva sintetizzare un insieme di dati e ci si trova di fronte ad una miriade di indici. Il fatto è che ciascuno degli indici introdotti evidenzia certe caratteristiche dei dati; alcune informazioni potranno risultare più utili, altre meno. Se, tornando all'esempio del calzaturificio, decidete di produrre inizialmente uno stock di scarpe di un'unica misura (tanto per saggiare il gradimento del pubblico), vi interesserà la moda più della media o della mediana.

**Un esempio**

Concludiamo questo paragrafo con un'applicazione della media geometrica (l'esempio è preso da PICCINATO L. e PINTACUDA N. "Probabilità e statistica" a cura del C.N.R., 1985, pag. 16):

La popolazione di una città negli anni dal 1980 al 1984 è

| Anno | popolazione |
|---|---|
| 1980 | 40000 |
| 1981 | 42000 |
| 1982 | 50400 |
| 1983 | 57960 |
| 1984 | 63756 |

i tassi di incremento annuali sono:

| | |
|---|---|
| 80-81 | 42000/40000 = 1.05 ($a_1$) |
| 81-82 | 50400/42000 = 1.20 ($a_2$) |
| 82-83 | 57960/50400 = 1.15 ($a_3$) |
| 83-84 | 63756/57960 = 1.10 ($a_4$) |

Definiamo come "tasso medio di crescita" nel quadriennio quel tasso **"t"** che applicato ripetutamente a partire dall 1980 riproduca lo stesso ammontare della popolazione nel 1984. **"t"** è la media geometrica dei tassi di incremento annuali:

$\mathbf{t} = \sqrt[4]{}(1.05 \times 1.2 \times 1.15 \times 1.1) = \sqrt[4]{}\, 1.5939 = 1.1236091$

infatti:

$40000 \times \mathbf{t} = 44944$
$44944 \times \mathbf{t} = 50500$
$50500 \times \mathbf{t} = 56742$
$56742 \times \mathbf{t} = 63756$

In effetti: $P_{81}=a_1 P_{80}$ ; $P_{82}=a_2P_{81}$ ; $P_{83}=a_3P_{82}$ ; $P_{84}=a_4P_{83}$. Sostituendo, $P_{84}= a_4a_3a_2a_1P_{80}$. Per definizione di **"t"**, deve essere $P_{81}= \mathbf{t}\, P_{80}$ ; $P_{82}= \mathbf{t}\, P_{81}$ ecc. in modo che $P_{84}= \mathbf{t\,t\,t\,t}\, P_{80}$. Quindi $t^4=a_1a_2a_3a_4$ e $\mathbf{t} \sqrt[4]{}(a_1a_2a_3a_4)$.

## 3. INDICI SINTETICI DI VARIABILI SEMPLICI: INDICI DI DISPERSIONE

I voti dei compiti di matematica di Massimo sono 6, 8, 6, 4; quelli dei compiti di Elisabetta 6, 6, 6, 6. Per entrambi è M = 6, Mediana = 6, Moda = 6 ma, mentre Elisabetta è sempre sufficientemente preparata, il rendimento di Massimo è discontinuo.

**Variabilità**

La DISPERSIONE (o VARIABILITÀ) di un insieme di dati è senz'altro un aspetto importante e risulterà utile calcolare dei valori che ne misurino l'entità (INDICI DI DISPERSIONE).

Una prima informazione è data dal CAMPO DI VARIAZIONE (V) definito come differenza tra il maggiore e il minore dei dati; se

$x' = \min x_i$ e
$x'' = \max x_i$ è

V =x" -x'

Generalmente non si tratta di un valore molto indicativo (un solo dato eccezionalmente grande o eccezionalmente piccolo "dila ta" il campo di variazione); può rivelarsi utile se V risulta particolarmente ridotto rispetto all'ordine di grandezza dei dati.

Il campo di variazione può essere espresso attraverso la coppia (x',x") in modo da fornire anche indicazioni "di posizione": tutti i dati sono compresi nell'intervallo [x' ,x"].

**Scostamento medio**

Più interessante è lo SCOSTAMENTO SEMPLICE MEDIO definito come:

$$?' = ( ?_i |x_i - M| )/n$$

che dice di quanto, mediamente, i dati si discostano dalla loro media aritmetica.

In relazione all'ultimo esempio si ha:

- Voti di Massimo (M =6):

(x',x") =(4,8) ; V =8 - 4 = 4

$$?' = (|6-6|+|8-6|+|6-6|+|4-6|)/4=1$$

- Voti di Elisabetta (M =6):

(x',x") =(6,6) ; V =6 - 6 = 0

?' = 0.

**Scostamento quadratico medio**

Più comodo da calcolare e di uso più comune è lo SCOSTAMENTO QUADRATICO MEDIO (o DEVIAZIONE STANDARD):

$$? = ? [?_i (x_i - M^2)/n]$$

**Varianza**

il cui significato e' analogo a quello di ?' (anche se, nel computo di ?, grossi scarti hanno "maggiore peso"). Il quadrato dello scostamento quadratico medio ( $?^2$ ) è la VARIANZA.

Per quanto riguarda i voti di Massimo abbiamo:

| $x_i$ | $x_i - M$ | $(x_i - M^2)$ |
|---|---|---|
| 6 | 0 | 0 |
| 8 | 2 | 4 |
| 6 | 0 | 0 |
| 4 | -2 | 4 |

la somma dei valori dell'ultima colonna è 8 per cui ? = ?(8/4) = ?2 =1.414 ..

Ovviamente ? (Voti di Elisabetta) =0.

Esercizio. Dimostrare che $?_i (x_i - M^2) = ?_i x_i^2 - nM^2$

Soluzione:

$$?_i (x_i - M^2) = ?_i (x_i^2 - 2M x_i + M^2) = ?_i x_i^2 - 2M?_i x_i + nM^2 =$$
$$= ?_i x_i^2 - 2M(Mn) + nM^2 = ?_i x_i^2 - nM^2$$

Questa identità consente di calcolare ? come

$$? = ? [(?_i x_i^2 - nM^2)/n] = ? (?_i x_i^2/n - M^2)$$

una formula alternativa assai utile dato che non è necessario disporre preventivamente della media aritmetica (si evita di passare in rassegna i dati due volte e quindi di doverli memorizzare in un vettore).

Ricalcoliamo ? (Voti di Massimo) con la nuova formula:

$x_i$: 6 8 6 4

$x_i^2$: 36 64 36 16

$?_i x_i = 152$ ; $M^2 = 36$ ; n = 4 per cui ? = ? (152/4 – 36) = ? = ?2 .

| $x_i$ | $\|x_i - M\|$ | $(x_i - M^2)$ |
|---|---|---|
| 3.8 | 2.389 | 5.707 |
| 3.9 | 2.289 | 5.240 |
| 4.5 | 1.689 | 2.853 |
| 4.5 | 1.689 | 2.853 |
| 5.0 | 1.189 | 1.414 |
| 5.0 | 1.189 | 1.414 |
| 5.1 | 1.089 | 1.186 |
| 5.5 | 0.689 | 0.475 |
| 5.5 | 0.689 | 0.475 |
| 5.6 | 0.589 | 0.347 |
| 5.8 | 0.389 | 0.15 1 |
| 5.9 | 0.289 | 0.084 |
| 6.1 | 0.089 | 0.008 |
| 6.5 | 0.3 11 | 0.097 |
| 6.5 | 0.311 | 0.097 |
| 6.6 | 0.411 | 0.169 |
| 6.6 | 0.411 | 0.169 |
| 6.8 | 0.6 11 | 0.373 |
| 6.8 | 0.611 | 0.373 |
| 7.0 | 0.8 11 | 0.658 |
| 7.1 | 0.911 | 0.830 |
| 7.5 | 1.311 | 1.719 |
| 7.5 | 1.311 | 1.719 |
| 8.0 | 1.811 | 3.280 |
| 8.0 | 1.811 | 3.280 |
| 8.0 | 1.811 | 3.280 |
| 8.0 | 1.811 | 3.280 |
| **Totale** | **28.511** | **41.531** |

Campo di variazione = 8 - 3.8 = 4.2
Scostamento semplice medio ?' = 28.511/27 = 1.056
Deviazione standard ? = ? (41.53 1/27) =1.24
Differenza interquartile = 7.1 - 5.1 =2

Ecco un breve programma Basic per il calcolo di M e ? :

```
10 INPUT ''QUANTI DATI"; N
20 FOR I = 1 TO N
30 INPUT "DATO..."; X
40 S=S+X:Q=Q+X*X
50 NEXT I
60 M=S/N:D=SQR(Q/N-M*M)
70 PRINT M,D
```

Altro indice di dispersione è la DIFFERENZA INTER QUARTILE $q_3 - q_1$
che, come il campo di variazione, può essere espressa dalla coppia $(q_3,q_1)$. Il significato è semplice: metà dei dati cadono in questo intervallo.

Tornando all'esempio della Tab. 2 (n = 27 voti; M = 6.189), calcoliamo gli indici di dispersione:

Ricapitolando:

**CLASSE 1' SEZ. A**

| | | |
|---|---|---|
| M | = 6.189 | media aritmetica |
| ? | = 6.5 | mediana |
| MD | = 8 | moda |
| (x',x'') | = (3.8 . 8) | campo di variazione |
| V | = 4.2 | campo di variazione |
| $(q_1,q_3)$ | = (5.1 . 7.1) | differenza interquartile |
| $q_3 - q_1$ | = 2 | differenza interquartile |
| ?' | = 1.056 | scostamento semplice medio |
| ? | = 1.24 | scostamento quadratico medio |
| $?^2$ | = 1.5376 | varianza |
| | | |

La statistica mostra che il compito proposto può ritenersi adeguato (M e ? oltre "la sufficienza"); la classe è abbastanza omo genea: anche se V è piuttosto elevato, ?' e ? mostrano che, mediamente, ci si discosta di 1 voto dalla sufficienza.

Nella classe parallela, la 1' B, gli studenti hanno usufruito di un'ora di lezione in più dedicata al ripasso. I risultati, relativi allo stesso compito valutato con gli stessi criteri, sono riportati nella seguente

Tabella 4

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3.8 | 4.1 | 4.5 | 4.9 | 5.5 | 5.8 | 6.0 | 6.1 | 6.5 |
| 6.6 | 6.6 | 6.6 | 6.8 | 6.8 | 6.8 | 6.8 | 6.9 | 7.0 |
| 7.0 | 7.0 | 7.0 | 7.0 | 7.0 | 7.1 | 7.1 | 7.1 | 7.6 |

Per questa classe abbiamo:

**CLASSE 1' SEZ. B**

| | | |
|---|---|---|
| M | = 6.37 | media aritmetica |
| ? | = 6.8 | mediana |
| MD | = 7 | moda |
| (x',x'') | = (3.8 . 7.6) | campo di variazione |
| V | = 3.8 | campo di variazione |
| $(q_1,q_3)$ | = (6 . 7) | differenza interquartile |
| $q_3 - q_1$ | = 1 | differenza interquartile |
| ?' | = 0.76 | scostamento semplice medio |
| ? | = 0.97 | scostamento quadratico medio |
| $?^2$ | = 0.94 | varianza |
| | | |

Dal confronto delle due statistiche si nota come la situazione della 1' B sia decisamente migliore: voti complessivamente più alti e maggiore omogeneità. Sembra proprio che la lezione aggiuntiva sia servita.

Naturalmente, trattandosi di valutazioni scolastiche, il "6" risulta un valore particolare poiché discrimina tra prove sufficienti ed insufficienti; nella 1' A sono pienamente sufficienti ($x_i$ ? 6) 15 studenti su 27 (circa il 56%), nella 1' B 21 su 27 (circa il 78%).

**Il programma STATISTICA**

Il programma STATISTICA, che si propone scopi puramente didattici, consente di calcolare gli indici statistici descritti in relazione ad un insieme di dati positivi che possono essere registrati su disco (o nastro) e da lì richiamati.

**Esempio di Output del programma STATISTICA**

(Dati: 7, 3, 10, 10, 10, 9, 9)

| | Risultati |
|---|---|
| Media Aritmetica | 8.28571429 |
| Media Geometrica | 7.76426976 |
| Media Quadratica | 8.61891608 |
| Media Armonica | 7.01112878 |
| Moda | 10 |
| Frequenza valore modale | 3 |
| Mediana | 9 |
| Primo Quartile | 7 |
| Terzo Quartile | 10 |
| Campo di variazione | 7 |
| Minimo $x_i$ | 3 |
| Massimo $x_i$ | 10 |
| Differenza interquartile | 3 |
| Scostamento semplice medio | 1.87755102 |
| Scarto quadratico medio | 2.3733211 |
| Varianza | 5.63265305 |

# 4. INFERENZA STATISTICA

Nei paragrafi precedenti ci siamo occupati della rappresentazione e della sintesi di un insieme di dati (STATISTICA DESCRITTIVA) indipendentemente dal fatto che questi provenissero dall'intera popolazione o da una sua parte.

E' facile rendersi conto che una rilevazione totale è molto spesso da escludere: non appena l'universo statistico supera una certa dimensione, diviene praticamente impossibile rilevare i dati da tutte le unità statistiche. Inoltre una rilevazione su vasta scala può essere assai costosa sia in termini monetari (si pensi di valutare l'efficienza di un farmaco) sia in termini di tempo (nei sondaggi preelettorali, i risultati servono subito); per non parlare dei rilevamenti "distruttivi" in cui l'unità esaminata risulta poi inservibile (per saggiare la resistenza dei tondini metallici utilizzati in edilizia, li si sottopone a "prove a trazione fino a rottura").

Quando si è costretti ad un rilevamento parziale, si pone un problema: fino a che punto i risultati dedotti dal campione (normalmente assai ridotto rispetto all'universo statistico) possono estendersi alla totalità ?

Di questi tipi di problemi si occupa I'INFERENZA STATISTICA, il cui studio richiede conoscenze matematiche piuttosto avanzate; ci limiteremo a qualche considerazione.

Un primo grosso problema consiste nella determinazione del campione da utilizzare: come sceglierlo e quanto grande ?

**Campionamento casuale semplice**

Per la scelta, si ricorre spesso a metodi probabilistici; nel CAMPIONAMENTO CASUALE SEMPLICE le "k" unità che formano il campione vengono estratte casualmente dalle "n" che costituiscono l'universo.

In questo modo si evita di privilegiare, magari inconsciamente, certe categorie e di lavorare su insiemi non rappresentativi. Se la popolazione è molto variabile, ma può essere suddivisa in sottoinsiemi singolarmente più omogenei, si preferisce operare separatamente sui sottoinsiemi (con estrazioni casuali) e mediare poi i risultati (CAMPIONAMENTI PER STRATIFICAZIONE).

**ESEMPIO**

Vi sono 1000 persone di età compresa tra i 20 e i 21 anni residenti nella città X; sono interessato al loro peso medio e voglio limitare l'indagine ad un campione di 100 persone.

PROCEDIMENTO 1: scelgo a caso 100 persone, rilevo i dati e medio i pesi

PROCEDIMENTO 2: scelgo a caso 50 ragazzi e 50 ragazze, calcolo la media dei due gruppi e medio i risultati.

Che succede se la popolazione è di 600 ragazze e 400 ragazzi? Nel primo caso il campione conterrà, probabilmente, più femmine che maschi ed il valore ottenuto con il procedimento 1 sarà più attendibile di quello ricavato con il procedimento 2.

Se però so, fin dall'inizio, che l'universo è formato da 600 ragazze e 400 ragazzi, posso considerare un campione stratificato di 60+40 ottenendo un valore ancora più attendibile (si sta lavorando nell'ipotesi che i due insiemi siano singolarmente più omogenei della totalità).

Quando si utilizzano campioni casuali, l'inferenza statistica fornisce tutta una serie di strumenti che consentono di valutare con quale grado di fiducia le informazioni ricavate dal campione si possano riferire a tutta la popolazione. Viceversa, fissato il margine di errore che si è disposti a tollerare, si potrà determinare qual è la dimensione minima del campione necessario.

Concludiamo con tre esempi (il 2' e il 3' sono presi da [PINTACUDA, N. "Insegnare la probabilità" Padova: Muzio, 1981] pag. 53 e pag. 58) che diano un'idea dei problemi tipici dell'inferenza statistica e del tipo di approccio per risolverli; gli studenti degli ultimi anni della secondaria superiore troveranno in [GAMBOTTO MANZONE, A. "Matematica per ragionieri programmatori", Bresso: Tramontana, 1985] (volume 3) una trattazione molto più esauriente.

**ESEMPIO 1**

Vorrei sapere quanti, tra i 6 vicini di casa, possiedono un televisore a colori.
Decido di chiedere a tre di loro e, dopo averli numerati da 1 a 6, lancio tre dadi ed intervisto le persone corrispondenti ai numeri ottenuti. Degli intervistati, due mi dicono di avere un apparecchio a colori, il terzo no.
Come stanno le cose in relazione a tutto il vicinato?

**SOLUZIONE**

L'universo statistico è costituito da 6 unità. Il campione, di dimensione 3, viene determinato attraverso una ESTRAZIONE BERNOULLIANA (o CON RIPETIZIONE): dopo l'estrazione, l'unità statistica sorteggiata torna a far parte della popolazione e può essere nuovamente estratta (con un universo statistico di dimensioni tanto ridotte, questo modo di procedere appare ridicolo poiché la stessa persona può essere intervistata due volte o addirittura tre).
Identifichiamo i vicini con 6 palline contenute in una scatola:
i possessori di televisori a colori corrispondano a palline bianche, gli altri a palline nere.
Il problema diviene:
"Se una scatola contiene 6 palline tra bianche e nere e ne vengono estratte 3 con reimbussolamento ottenendone 2 bianche e 1 nera, che si può dire del contenuto della scatola?"
Indichiamo con x il numero di palme bianche presenti nella scatola e con Hx l'evento
Hx = ?la scatola contiene x bianche e 6-x nere?.
Prima di eseguire le estrazioni, sappiamo solo che x può essere un qualsiasi intero da 0 a 6: l'assenza di informazione equivale all'ipotesi P(Hx)=1/7 per ogni x.
Posto
A=?vengono estratte 2 bianche e 1 nera?
è

$$P(A/Hx)= C_{3,2}\ (x / 6)^2(6-x)/6$$

si tratta infatti della probabilità di avere 2 successi (successo = estrazione di una bianca) su 3 prove in uno s.p.r. di parametro incognito $p = x/6$ (vedi calcolo combinatorio). Giustamente $P(A/H_0)= P(A/H_6)=0$ dato che A non può verificarsi se le palline sono tutte bianche o tutte nere.

Il problema consiste nel valutare P(Hx/A) in quanto sappiamo che A si è verificato mentre non conosciamo x. Ci viene in aiuto la formula di Bayes; poiché Hx x=0,1,.. ,6 costituisce una partizione di ? , è

$$P(Hx/A)= P(A/Hx)\ P(Hx) / [\ ?_j\ P(A/Hj)\ P(Hj)\ ]\ ,\ j=0,\ 1,\ ..,\ 6$$

Calcoliamo il denominatore:

$$?_j\ P(A/Hj)\ P(Hj) = [?_j\ j^2\ (6-j)/36]/7 = [0+5+16+27 + 32 + 25] / 36 / 7 = 5/12$$

quindi

$P(Hx/A) = 12\ x^2(6-x)/36\ /7\ /\ 5 = x^2(6-x)\ /\ 105$

Per x=0,...,6 abbiamo:

| x | P(Hx/A) |
|---|---|
| 0 | 0 |
| 1 | 5/105 = 0.05 |
| 2 | 16/105 = 0.15 |
| 3 | 27/105 = 0.26 |
| 4 | 32/105 = 0.30 |
| 5 | 25/105 = 0.24 |
| 6 | 0 |

Il valore più alto si ottiene per x =4 (come ci si aspettava) anche se la probabilità associata non è poi elevatissima. La tabella fornisce però informazioni più utili, ad esempio:

$P[3 = x = 5] = P(3) + P(4) + P(5) = 0.8$

posso ora affermare, con un margine di sicurezza dell'80%, che il numero di vicini che possiedono un televisore a colori è compreso tra 3 e *5* mentre prima dell'esperimento statistico tale probabilità era pari a 3/7 =43%.

**ESEMPIO** 2

Una fabbrica produce componenti elettronici in confezioni da 100 pezzi che provengono da due macchine M1 e M2 e precisamente:
60 pezzi da Ml, 40 da M2.
A causa di un temporaneo malfunzionamento di M2, alcune confezioni contengono 40 circuiti difettosi. Ne controllo una verificando *5* pezzi scelti a caso: risultano perfettamente funzionanti.
Posso essere sicuro, al *95 %,* della bontà di tutti i cento pezzi della confezione in esame?

**SOLUZIONE**

Interessa calcolare la probabilità che una confezione difettosa superi il controllo ovvero la probabilità che, nelle *5* estrazioni, siano sempre usciti pezzi funzionanti quando la confezione ne contiene 40 difettosi.
La probabilità richiesta è (estrazioni SENZA reimbussolamento):

P = 60/100 ?59/99 ?58/98 ?57/97 ?56/96 = 0.07254.. ? 7.2 %

P è la "probabilità di sbagliarsi" cioè la probabilità di considerare valida una confezione difettosa. Poiché 7.2% > 5%, non posso essere sicuro al 95% (ma solo al 92.8%).
Se avessi eseguito le *5* estrazioni CON reimbussolamento, la probabilità di "errore" sarebbe stata:

$P' = (60/100)^5$ ? 0.078 =7.8 %

Supponendo di eseguire estrazioni con reimbussolamento, quanti pezzi devo controllare (come minimo) per essere sicuro della bontà della confezione al 95% ? E al 99% ?

Sia "n" il numero di circuiti da controllare; " n" e il più piccolo intero per cui:

$(0.6)^n$ ? 0.05 per il 95%
e
$(0.6)^n$ ? 0.01 per il 99%

| n | $(0.6)^n$ |
|---|---|
| 1 | 0.60000 |
| 2 | 0.36000 |
| 3 | 0.21600 |
| 4 | 0.12960 |
| 5 | 0.07776 |
| 6 | 0.04666 *** |
| 7 | 0.02799 |
| 8 | 0.01680 |
| 9 | 0.01008 |
| 10 | 0.00605 *** |

Quindi 6 pezzi per essere sicuri al 95% e 10 pezzi per il 99%. Il risultato si ottiene più rapidamente passando ai logaritmi. Deve essere:

log $(0.6)^n$ ? log 0.05 per il 95%
e
log $(0.6)^n$ ? log 0.01 per il 99%

da cui, rispettivamente:

n log (0.6) ? log 0.05 per il 95%
e
n log (0.6) ? log 0.01 per il 99%

poiché log 0.6 < 0 è

n ? log 0.05 / log 0.6 ? 5.864 per il 95% : il minimo intero è 6;

mentre

n ? log 0.01 / log 0.6 ? 9.0151 per il 99% : il minimo intero è 10.

**ESEMPIO 3**

Si vuol sapere qual è stato il numero medio di giorni di degenza dei pazienti ricoverati presso un certo ospedale nel 1986.

Marinella e Sergio, incaricati dell'indagine, procedono in modo diverso: Marinella, procuratasi un elenco delle persone ricoverate, ne estrae un campione su cui lavorare; Sergio sceglie a caso un certo numero di date del 1986 e prende in considerazione le persone che risultano ricoverate in quei giorni. Chi dei due otterrà risultati più attendibili ?

**SOLUZIONE**

Il campione utilizzato da Sergio risulta essere un CAMPIONE DISTORTO; a parità di dimensioni, il valore da lui calcolato sarà, probabilmente, più elevato di quello valutato da Marinella.

Con il metodo di Sergio infatti, le persone costrette a lunghe degenze hanno maggiore probabilità di entrare a far parte del campione contribuendo ad aumentare il valore medio.

Consideriamo un "caso limite" in cui, di 100 pazienti, 50 siano stati ricoverati dal 1' gennaio al 10 gennaio e i rimanenti 50 dall'11 gennaio al 31 dicembre (una situazione davvero poco invidiabile). Mentre Marinella estrarrà un campione che, in linea di massima, conterrà metà elementi del primo gruppo e metà del secondo, quello di Sergio sarà costituito quasi esclusivamente da persone del secondo gruppo, poiché le date scelte casualmente cadranno quasi tutte oltre il 10 gennaio.

Nella realtà, il carattere in esame risulta molto variabile e converrebbe far uso di campioni stratificati, studiando separatamente i vari reparti in diversi periodi dell'anno.

continua nella rubrica Statistiche