Appunti di statistica

La statistica,nata come strumento d'indagine sulla popolazione di uno Stato, è oggi una scienza che studia qualsiasi fenomeno di tipo collettivo.
Le indagini sui fenomeni collettivi vengono fatte all'interno delle popolazioni statistiche (insieme di elementi che hanno almeno una caratteristica comune).
Ciascuna caratteristica, che differenzia gli elementi di una popolazione, può essere di tipo qualitativo (come ad esempio l'attività svolta) o quantitativo (come ad esempio il peso).
L'indagine, di solito, non viene svolta sull'intera popolazione ma su un campione che ha caratteristiche analoghe all'intera popolazione.
Per effettuare un'indagine statistica si seguono, di solito, le seguenti fasi:
1) raccolta e spoglio dei dati
2) compilazione di tabelle in modo da poter osservare la distribuzione delle frequenze
3) rappresentazione grafica ed elaborazione dei dati
4) interpretazione dei risultati ottenuti.

Medie statistiche

Media aritmetica semplice

La media aritmetica semplice M di n numeri $x_1, x_2, ..., x_n$ è data da:

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

Ad esempio:
Qual è l'altezza media di cinque ragazzi che sono alti 160, 165, 170,160,164 cm?

$$M = \frac{160+165+170+160+164}{5} = 163,8\,cm$$

Media aritmetica ponderata

Se i numeri $x_1, x_2, ..., x_n$ hanno rispettivamente frequenza $f_1, f_2, ..., f_n$ la media ponderata è data da:

$$M_p = \frac{f_1 x_1 + f_2 x_2 + ... + f_n x_n}{f_1 + f_2 + ... + f_n}$$

Consideriamo ad esempio la seguente tabella di valori che si riferisce all'altezza, in cm, di 100 ragazzi:

Frequenza	5	15	10	4	20	30	6	10
Altezza	165	166	168	169	171	172	174	175

$$M_p = \frac{5\cdot 65+15\cdot 166+10\cdot 168+4\cdot 169+20\cdot 171+30\cdot 172+6\cdot 174+10\cdot 175}{100} = 170,45$$

Media geometrica

Data una distribuzione di n valori $x_1, x_2, ..., x_n$, chiamiamo media geometrica il valore m_g che, sostituito agli elementi della distribuzione, non ne cambia il prodotto, cioè:

$x_1 \cdot x_2 \cdot ..... \cdot x_n = \left(m_g\right)^n$

quindi $m_g = \sqrt[n]{x_1 \cdot x_2 \cdot ..... \cdot x_n}$

Se gli elementi x_i della distribuzione hanno frequenza f_i allora $m_g = \sqrt[n]{x_1^{f_1} \cdot x_2^{f_2} \cdot ..... \cdot x_n^{f_n}}$

E' utile ricordare che possiamo anche scrivere:

$$\ln m_g = \frac{f_1 \ln x_1 + f_2 \ln x_2 + ... + f_n \ln x_n}{n}$$

Esempio

Calcola la media geometrica della seguente serie di valori: 5; 7;8;8:

$$m_g = \sqrt[4]{\frac{5 \cdot 7 \cdot 8^2}{4}} = 4,86...$$

Media quadratica

Data una distribuzione di n valori $x_1, x_2, ..., x_n$, chiamiamo media quadratica il valore m_q che, sostituito agli elementi della distribuzione, non cambia la somma dei loro quadrati, cioè:

$x_1^{\,2} + x_2^{\,2} + ... + x_n^{\,2} = n \cdot \left(m_q\right)^2$

per cui $m_q = \sqrt{\frac{x_1^{\,2} + x_2^{\,2} + ... + x_n^{\,2}}{n}}$

Se gli elementi x_i della distribuzione hanno frequenza f_i allora

$$m_q = \sqrt{\frac{f_1 x_1^{\,2} + f_2 x_2^{\,2} + ... + f_n x_n^{\,2}}{\sum_{i=1}^{n} f_i}}$$

Esempio

Calcola la media quadratica della seguente serie di valori: 5; 7;8;8:

$$m_q = \sqrt{\frac{5^2 \cdot 7^2 \cdot 2 \cdot 8^2}{4}} = 197,98...$$

Media armonica

Data una distribuzione di n valori $x_1, x_2, ..., x_n$, chiamiamo media armonica il valore m_{ar} che, sostituito agli elementi della distribuzione, non cambia la somma dei loro reciproci, cioè:

$$\frac{1}{x_1}+\frac{1}{x_2}+.....+\frac{1}{x_n}=n\frac{1}{m_{ar}}$$

quindi: $m_{ar}=\dfrac{n}{\dfrac{1}{x_1}+\dfrac{1}{x_2}+.....+\dfrac{1}{x_n}}$

Se gli elementi x_i della distribuzione hanno frequenza f_i allora $m_{ar}=\dfrac{\sum f}{\dfrac{f_1}{x_1}+\dfrac{f_2}{x_2}+.....+\dfrac{f_n}{x_n}}$

Esempio

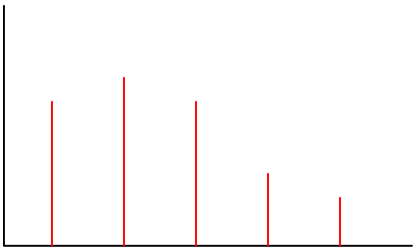
Calcola la media armonica dei numeri $\frac{1}{2};\frac{1}{3};\ \frac{1}{4};\ \frac{1}{5};\frac{1}{6}$

$$m_{ar}=\frac{5}{2+3+4+5+6}=\frac{1}{4}$$

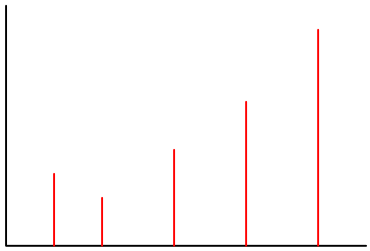
Osserviamo che la media armonica di un numero dispari di elementi che sono in progressione armonica (i loro reciproci formano una progressione aritmetica) è uguale al valore del termine centrale.

Moda e mediana

Si chiama moda o valore normale di una distribuzione di frequenze il valore al quale corrisponde la massima frequenza.
Nel caso della raccolta di dati, indicati in precedenza nella tabella delle altezze, la moda è 172 cm
Se, in particolare, la distribuzione presenta due o più frequenze massime uguali, la distribuzione è detta plurimodale. In questo caso la moda non ha un apprezzabile significato statistico.
Assegnata una distribuzione ordinata di valori, chiamiamo mediana il valore centrale della suddetta distribuzione.
Ad esempio la mediana della seguente distribuzione di valori: 5; 8; 10; 11; 25; 30; 35 è 11
Se i valori sono in numero dispari la mediana è data dalla semisomma dei due valori centrali.
La media, la moda e la mediana sono dette misure della tendenza centrale di una raccolta di dati; infatti, in situazioni normali esse occupano posizioni centrali nella distribuzione dei dati raccolti e sono di solito utili ad analizzare il fenomeno che si sta esaminando.
Oltre alle misure della tendenza centrale è necessario considerare anche lo scarto S_i che esiste tra l'elemento di indice i e la media (scarto dalla media). $S_i=x_i-M \quad (i=1,2,....,n)$
L'indice di dispersione è la differenza tra il valore massimo ed il valore minimo dei dati raccolti.



esempio di dispersione verso i valori più bassi

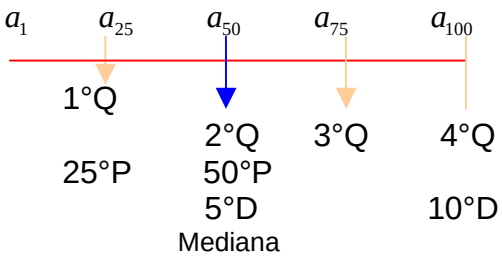


esempio di dispersione verso i valori più alti

La dispersione è misurata mediante gli indici di variabilità: i più usati sono la varianza, lo scarto quadratico medio e gli scarti dalla mediana e dalla media aritmetica.

A volte è utile considerare elementi che dividono la distribuzione in modo diverso.
Si usano i cosiddetti *quartili Q*, elementi che dividono l'insieme ordinato dei dati in *quattro* parti uguali, oppure i *decili D* (elementi che dividono la sequenza in 10 parti uguali) o i *percentili C* (elementi che dividono la sequenza in 100 parti uguali)

Ad esempio: per la distribuzione $a_0, a_1, ...a_{100}$ si ha:



Indici di variabilità

Campo di variabilità

Si chiama campo di variabilità di un insieme di valori la differenza tra il valore massimo ed il valore minimo: $C = x_{\max} - x_{\min}$ di detti valori.

Scarto semplice medio

Se $x_1, x_2, ..., x_n$ è un insieme di valori che hanno M come media aritmetica, lo scarto semplice medio è dato da:

$$s_m = \frac{|x_1 - M| + |x_2 - M| + ... + |x_n - M|}{n} = \frac{\sum_{i=1}^{n} |x_i - M|}{n}$$

Varianza

Se $x_1, x_2, ..., x_n$ è un insieme di valori che hanno M come media aritmetica, la varianza è data da:

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

Se i valori $x_1, x_2, ..., x_n$ hanno frequenze $f_1, f_2, ..., f_n$ la varianza è data da:

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

In pratica, Il calcolo della varianza può essere effettuato mediante la formula:

$\sigma^2 = \underline{M} - M^2$ dove $\underline{M}$ è la media aritmetica dei quadrati dei dati

Esempio
Consideriamo la seguente tabella:

valori	3	5	7	9	12
frequenze	8	4	3	7	2

calcoliamo le medie M e $\underline{M}$

$$M = \frac{8 \cdot 3 + 4 \cdot 5 + 3 \cdot 7 + 7 \cdot 9 + 2 \cdot 12}{8 + 4 + 3 + 7 + 2} = 6,333 \quad (M^2 = 40,106)$$

$$\underline{M} = \frac{8 \cdot 3^2 + 4 \cdot 5^2 + 3 \cdot 7^2 + 7 \cdot 9^2 + 2 \cdot 12^2}{8 + 4 + 3 + 7 + 2} = 48,916$$

e ricaviamo $\sigma^2 = \underline{M} - M^2 \simeq 8,8$

Scarto quadratico medio o deviazione standard

Si chiama scarto quadratico medio di un insieme di valori $x_1, x_2, ..., x_n$ la radice quadrata della varianza di tali valori:

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

Tra varianza e scarto quadratico medio è preferibile usare quest'ultimo perché non altera l'unità di misura dei dati statistici presi in esame.

Coefficiente di variabilità

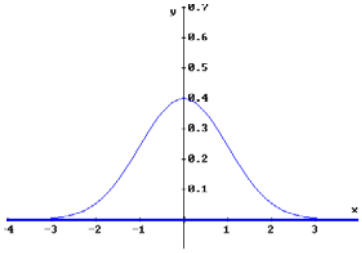
E' il rapporto tra lo scarto quadratico medio e la media aritmetica $coeff.\,\mathrm{v.} = \frac{\sigma}{M}$

E viene di solito usato in percentuale mediante: $coeff.\,\mathrm{v.} = \frac{\sigma}{M} \cdot 100$

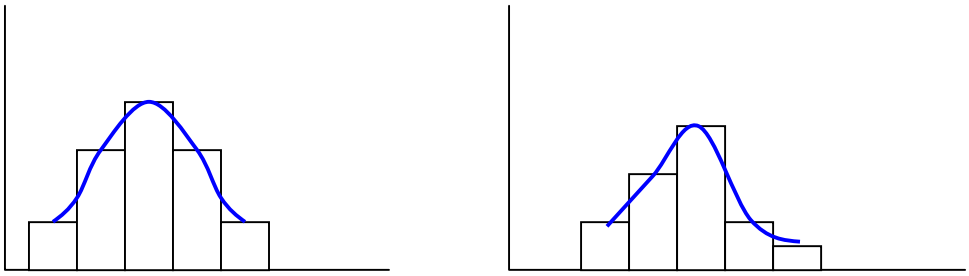
Distribuzione normale (curva di Gauss)

Una distribuzione di frequenze è detta normale se ha un andamento grafico che si avvicina alla curva di Gauss

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



Più i dati di una distribuzione normale sono concentrati, più stretta e alta è la "campana"che li rappresenta. Più i dati di una distribuzione normale sono dispersi, più larga e bassa è la "campana"che li rappresenta.



Distribuzioni statistiche doppie; distribuzioni marginali

Quando si studia una popolazione statistica è possibile che i caratteri esaminati siano più di uno. In questo caso si parla di distribuzioni statistiche multiple. Se i caratteri sono due la distribuzione è rappresentata efficacemente da una tabella a doppia entrata.

Consideriamo la seguente tabella del 1990 che si riferisce ad un'indagine campionaria sulla distribuzione delle abitazioni secondo la superficie abitata.

superficie regione	50-95 mq	96-110 mq	111-130mq	131-200 mq
Liguria	130	11	6	5
Campania	362	1805	105	122
Sicilia	1068	430	203	149

Integriamo la tabella scrivendo a destra di ogni riga e in fondo a ogni colonna la somma dei valori riportati:

superficie regione	50-95 mq	96-110 mq	111-130mq	131-200 mq	
Liguria	130	11	6	5	*152*
Campania	362	1805	105	122	*2394*
Sicilia	1068	430	203	149	*1850*
	1560	*2246*	*314*	*276*	

I valori ai margini della tabella si chiamano distribuzioni marginali della distribuzione assegnata.

Le distribuzioni marginali per regione e per superficie del nostro esempio sono:

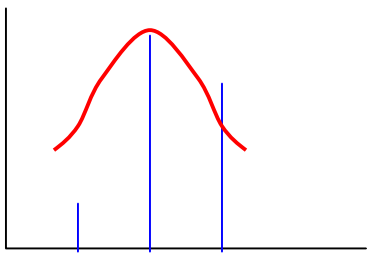
classe	*frequenza*
Liguria	152
Campania	2394
Sicilia	1850

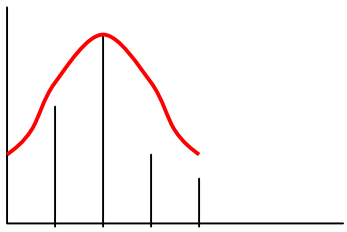
superficie	50-95 mq	96-110 mq	111-130mq	131-200 mq
frequenza	1560	2246	314	276

Osserviamo che:
se indichiamo con x_i i valori medi delle superfici e con f_i le frequenze di detti valori, il valore medio della superficie abitata è dato da: $M = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$

e la deviazione standard da: $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M)^2 \cdot f_i}{\sum_{i=1}^{n} f_i}}$

Dall'analisi degli istogrammi relativi alle distribuzioni marginali





deduciamo che la prima distribuzione non è normale perché si discosta dalla curva di Gauss.

Funzione interpolatrice

Quando si vuole ricavare la legge di un fenomeno, nel quale intervengono due grandezze variabili x e y, una indipendente dall'altra, si determinano mediante esperimenti quale valore assume la y al variare di x in un intervallo $[a;b]$. Per avere poi una visione grafica

dell'andamento del fenomeno è utile costruire sul piano xy i punti $A_0(x_0;y_0);A_1(x_1;y_1);.....;A_n(x_n;y_n)$. Dato che tra le due grandezze non esiste una ben definita legge matematica, non è possibile determinare tutti i valori che può assumere la variabile x ma solo un limitato numero di valori. Il grafico che si otterrà unendo i punti $A_0(x_0;y_0);A_1(x_1;y_1);.....;A_n(x_n;y_n)$ costituirà un diagramma approssimato del fenomeno che si sta studiando.
Affinché sia minimo lo scostamento dal diagramma reale occorre individuare una funzione $\varphi(x)$, detta *funzione interpolatrice*, che assume gli stessi valori di y nei punti di interpolazione $x_i\ (i=0;1;2;....;n-1)$.
Per capire come ottenerla consideriamo il seguente esempio:
Sia data la tabella di valori:

x	0	2	3	4
y	3	1	5	7

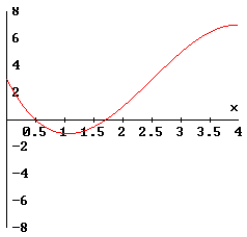
Che scaturisce da una particolare indagine sperimentale.
Poiché i punti assegnati sono 4 si dovranno determinare i coefficienti $a_0;a_1;a_2;a_3$ di un polinomio di terzo grado $(n-1)$ che ha la forma: $P_3(x)=a_0x^3+a_1x^2+a_2x+a_3$

Poiché: $\begin{cases} P_3(0)=a_0; \\ P_3(2)=8a_0+4a_1+2a_2+a_3; \\ P_3(0)=27a_0+9a_1+3a_2+a_3 \\ P_3(0)=64a_0+16a_1+4a_2+a_3 \end{cases}$ si ha: $\begin{cases} a_0=3 \\ 8a_0+4a_1+2a_2+3=1 \\ 27a_0+9a_1+3a_2+3=5 \\ 64a_0+16a_1+4a_2+3=7 \end{cases}$

Risolvendo questo sistema con il metodo di Cramer si ricava: $a_0=-\frac{2}{3};\quad a_1=5;\quad a_2=-\frac{25}{3}$

La funzione interpolatrice $\varphi(x)$ avrà quindi equazione: $y=-\frac{2}{3}x^3+5x^2-\frac{25}{3}x+3$

Questa funzione assume nei punti x_i i valori y_i elencati in tabella, ma non permette di determinare ulteriori valori di y quando si assegnano valori di x, diversi da quelli elencati.
Possiamo dire che $\varphi(x)$ approssima l'andamento di $f(x)$ nell'intervallo $[0;4]$.



L'errore che si commette in un generico punto $x\neq x_i$ è: $R(x)=f(x)-\varphi(x)\quad con\quad x\in[a;b]$.

Per determinare l'equazione della curva interpolatrice si può anche usare il metodo dei minimi quadrati.Ci si prefigge cioè di determinare una funzione in modo che la sua equazione renda minima la somma dei quadrati degli scarti dei dati teorici da quelli reali

Regressione

Se la funzione interpolatrice è una retta di equazione $y = mx + q$ e chiamiamo con m_x e m_y le medie aritmetiche di x_i e y_i, ovvero: $m_x = \frac{\sum x_i}{n}$; $m_y = \frac{\sum y_i}{n}$ avremo: $\sigma_x^{\,2} = \frac{\sum (x_i - m_x)^2}{n}$; $\sigma_y^{\,2} = \frac{\sum (y_i - m_y)^2}{n}$.

Poiché si dimostra che la retta interpolatrice ha equazione $y - m_y = r_{y,x}(x - m_x)$ dove $r_{y,x} = \frac{\sigma_{x,y}^2}{\sigma_x^{\,2}}$ con $\sigma_{x,y}^2 = \frac{\sum (x_i - m_x)(y_i - m_y)}{n}$ (detta covarianza), possiamo dire che $r_{y,x} = \frac{\sigma_{x,y}^2}{\sigma_x^{\,2}}$ è il coefficiente angolare della retta. Tale coefficiente si chiama coefficiente di regressione di y su x.

Allo stesso modo, se la retta interpolatrice ha equazione $x - m_x = r_{x,y}(y - m_y)$, il coefficiente di regressione di x su y sarà: $r_{x,y} = \frac{\sigma_{x,y}^2}{\sigma_y^{\,2}}$

Correlazione

Se tra i valori della distribuzione esiste una corrispondenza di tipo lineare si fa uso del coefficiente di correlazione lineare r di Bravais-Pearson:

$$r = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

poiché $r_{y,x} = \frac{\sigma_{xy}^2}{\sigma_x^2}$ *e* $r_{x,y} = \frac{\sigma_{xy}^2}{\sigma_y^x}$ possiamo scrivere: $|r| = \sqrt{|r_{y,x} \cdot r_{x,y}|}$ che esprime il legame esistente tra regressione e correlazione.