Università di Cagliari
Corso di Laurea in Farmacia

Matematica Statistica

Sonia Cannas

A.A. 2019/2020

Introduzione alla statistica descrittiva

Statistica

La **statistica** è una disciplina il cui fine è lo studio quantitativo di fenomeni collettivi (ossia riguardanti più individui) osservabili nella realtà sociale, in natura o in laboratorio per fini descrittivi o per prendere delle decisioni.

Statistica descrittiva

La **statistica descrittiva** è quella parte della statistica che si occupa di illustrare e sintetizzare i dati raccolti in un esperimento/studio attraverso i suo strumenti grafici e i suoi indici.

Introduzione alla statistica descrittiva

La statistica si interessa di ottenere informazioni su un insieme completo di oggetti detto *popolazione*.

Popolazione

Per **popolazione** si intende l'insieme degli elementi oggetto di un'indagine statistica. Ciascun elemento facente parte della popolazione è detto anche **unità statistica**. Le popolazioni possono essere umane, animali, collezioni di batteri, abbigliamento, ecc...

In alcune indagini statistiche non è possibile esaminare un'intera popolazione, quindi se ne estrae un campione.

Campione

Per **campione** si intende un insieme finito di n unità che si può ritenere rappresentativo di un'intera popolazione.

Introduzione alla statistica descrittiva

Indagine statistica

Un'indagine statistica si articola nelle seguenti fasi:

1. pianificazione dell'indagine;
2. rilevazione dei dati;
3. elaborazione dei dati;
4. presentazione e interpretazione dei risultati.

Ci occuperemo prevalentemente della fase di elaborazione dei dati.

Introduzione alla statistica descrittiva

Carattere

La proprietà oggetto di studio in un'indagine statistica è detta **carattere**.

Esempi di caratteri in un'indagine statistica

Colore degli occhi, misura dell'altezza, misura del peso, l'età degli individui di una popolazione.

Modalità

Ciascuna delle varianti con cui un carattere può presentarsi è detta **modalità**.

Esempi di modalità di caratteri in un'indagine statistica

Il carattere "colore degli occhi" può assumere le modalità: verdi, azzurri, marroni, ecc...
Il carattere "misura dell'altezza" può assumere le modalità: 170 cm, 157 cm, 186 cm, ecc...

Le modalità osservate sono dette **dati**.

Introduzione alla statistica descrittiva

Caratteri quantitativi e qualitativi

Un carattere le cui modalità sono espresse da numeri è detto **quantitativo**, in caso contrario è detto **qualitativo**.

Esempi di caratteri quantitativi

Misura dell'altezza, misura del peso, l'età degli individui di una popolazione, presenza di malattie cardiovascolari, ecc...

Esempi di caratteri qualitativi

Colore degli occhi, tipo di alimentazione, genere, credo religioso, ecc...

Caratteri quantitativi discreti e continui

Un carattere quantitativo è detto **discreto** se può assumere una quantità di valori finita oppure numerabile (cioè può essere posto in corrispondenza biunivoca con $\mathbb{N}$); si dice **continuo** se può assumere una quantità continua di valori, cioè se può assumere tutti i valori reali di un determinato intervallo.

Introduzione alla statistica descrittiva

Esempio

Indagine	Popolazione	Carattere	Modalità	Tipo carattere
Colore degli occhi dei sardi	Tutti i sardi	Colore degli occhi	Marroni, verdi, ecc...	Qualitativo
Peso degli studenti universitari	Tutti gli studenti universitari	Misura del peso	50 kg, 51 kg, ...	Quantitativo continuo
Anno di nascita delle matricole di unica	Matricole di unica	Anno di nascita	2000, 1999, 1998, ...	Quantitativo discreto

Distribuzioni di frequenze

Definizione (Frequenza assoluta)

Si definisce **frequenza assoluta** *(o semplicemente* **frequenza***) di una modalità il numero di volte in cui una modalità è stata osservata.*

Esempio

Supponiamo di voler studiare l'evoluzione genetica del colore degli occhi di una determinata popolazione. Nell'indagine sono stati rilevati i colori degli occhi di una generazione il cui campione è formato da 180 individui, le cui frequenze dei colori sono sintetizzate nella seguente tabella.

Colore occhi	Numero individui
Nero	70
Marrone	50
Verde	40
Azzurro	20
Totale	180

Distribuzioni di frequenze

Funzione distribuzione delle frequenze

Ad ogni modalità di un carattere possiamo associare la rispettiva frequenza. Tale relazione è detta **funzione di distribuzione delle frequenze**.
Solitamente si rappresenta tramite una tabella con due colonne, dove la prima rappresenta le modalità e la seconda le relative frequenze (vedi tabella dell'esempio precedente). Nell'ultima riga si è soliti riportare la somma delle frequenze, il cui numero deve essere pari al numero di individui analizzati. Se il carattere è quantitativo solitamente le modalità sono ordinate in senso crescente o decrescente.

Distribuzioni di frequenze

Osservazione

Se il carattere oggetto di studio presenta modalità con frequenza molto bassa la funzione di distribuzione delle frequenze non determinerebbe una sintesi significativa. Il problema può essere superato accorpando alcune modalità in intervalli disgiunti detti **classi** e costruendo la funzione di distribuzione delle frequenze delle classi.

Esempio

Supponiamo di voler analizzare la statura dei 18 atleti di una squadra di calcio e che i dati siano i seguenti.

Atleta	1	2	3	4	5	6	7	8	9
Altezza	173	164	174	180	182	176	184	185	170
Atleta	10	11	12	13	14	15	16	17	18
Altezza	172	186	167	188	183	168	176	184	178

Le modalità hanno frequenza 1 o al massimo 2. Conviene analizzare la distribuzione di frequenze delle classi.

Distribuzioni di frequenze

Nel nostro caso possiamo suddividere le possibili altezze negli intervalli:

$$(160, 165] \quad (165, 170] \quad (170, 175] \quad (175, 180] \quad (180, 185] \quad (185, 190]$$

e costruire la seguente tabella di distribuzione di frequenze delle classi

Altezza atleti	Numero atleti
(160, 165]	1
(165, 170]	3
(170, 175]	3
(175, 180]	4
(180, 185]	5
(185, 190]	2

Distribuzioni di frequenze

Definizione (Frequenza relativa)

Si definisce **frequenza relativa** *di una modalità il rapporto tra la sua frequenza assoluta e il numero complessivo di individui oggetto di studio.*

Osservazione

Le frequenze relative possono essere espresse sotto forma di percentuali: in tal caso si parla di **frequenza percentuale**.

Distribuzioni di frequenze

Esempio

Consideriamo nuovamente lo studio sull'evoluzione genetica del colore degli occhi di una determinata popolazione e riportiamo in tabella anche le frequenze relative e quelle percentuali.

Colore occhi	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
Nero	70	$\frac{70}{180} \simeq 0,389$	$38,9\%$
Marrone	50	$\frac{50}{180} \simeq 0,278$	$27,8\%$
Verde	40	$\frac{40}{180} \simeq 0,222$	$22,2\%$
Azzurro	20	$\frac{20}{180} \simeq 0,111$	$11,1\%$
Totale	180	$\frac{180}{180} = 1$	100%

Osservazione

La somma delle frequenze assolute è uguale al numero complessivo di individui. La somma delle frequenze relative è uguale a 1. La somma delle frequenze percentuali è uguale al 100%.

Rappresentazioni grafiche dei dati

La funzione di distribuzione delle frequenze, come ogni funzione, può essere rappresentata graficamente. In generale in statistica vengono utilizzati vari grafici. I più importanti possono essere ricondotti a quattro categorie.

Principali rappresentazioni grafiche

- **Diagramma a barre**: viene utilizzato prevalentemente per caratteri qualitativi o quantitativi discreti.
- **Diagramma circolare**: viene utilizzato prevalentemente per caratteri qualitativi, quando si vuole visualizzare la distribuzione del carattere.
- **Istogramma**: viene utilizzato prevalentemente per caratteri suddivisi in classi, in particolare caratteri quantitativi continui.
- **Diagramma cartesiano**: viene utilizzato prevalentemente per rappresentare fenomeni osservati in determinati periodi di tempo (serie temporali).

Rappresentazioni grafiche dei dati

Diagramma a barre

Uso principale: rappresentare caratteri qualitativi o quantitativi discreti. Rappresentano la funzione di distribuzione delle frequenze tramite rettangoli distanziati tra loro, aventi la stessa base e altezza proporzionale alla frequenza della modalità osservata.

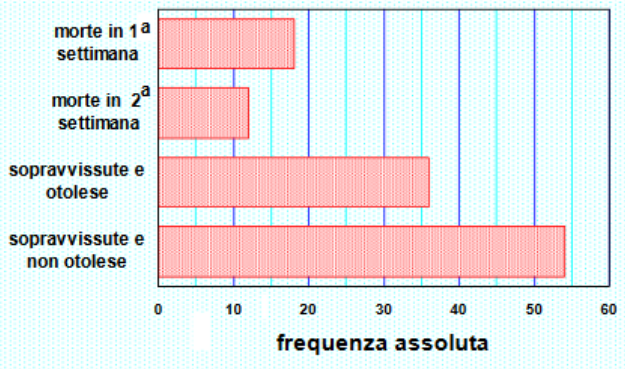


Figura: Diagramma a barre orizzontale per carattere qualitativo

Rappresentazioni grafiche dei dati

Diagramma circolare (o a torta)

Uso principale: rappresentare caratteri qualitativi, quando si vuole visualizzare la distribuzione del carattere.
Rappresentano la funzione di distribuzione delle frequenze attraverso settori circolari di ampiezza proporzionale alla frequenza.

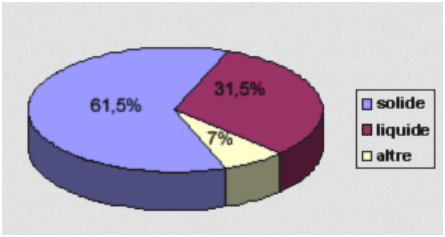


Figura: Diagramma a torta delle forme farmaceutiche commercializzate nel 2000

Rappresentazioni grafiche dei dati

Istogramma

Uso principale: rappresentare caratteri suddivisi in classi, in particolare caratteri quantitativi continui.
È un grafico rappresentato da rettangoli non distanziati, ciascuno dei quali ha un'area proporzionale alla frequenza della classe rappresentata. Se le classi hanno la stessa ampiezza i rettangoli hanno tutti la stessa base e altezza uguale alla misura della frequenza.

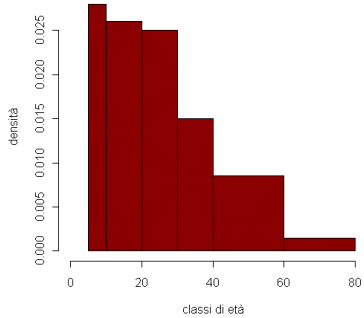


Figura: Istogramma consumatori di nutella per classi di età

Rappresentazioni grafiche dei dati

Gli istogrammi possono risultare utili anche nel confronto tra due insiemi di dati.

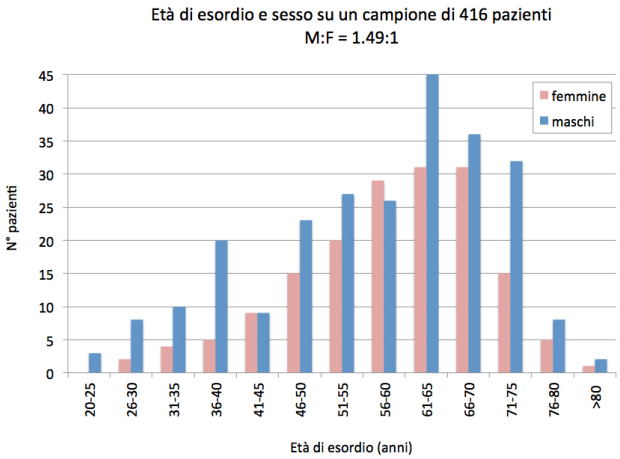


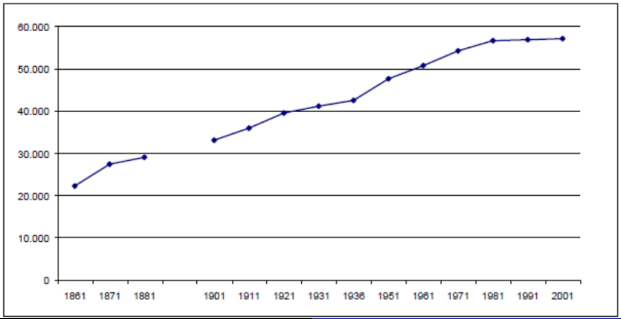
Figura: Istogramma SLA

Rappresentazioni grafiche dei dati

Diagramma cartesiano

Uso principale: rappresentare fenomeni osservati in determinati periodi di tempo (serie temporali).
L'asse delle ascisse rappresenta i tempi, quello delle ordinate i valori osservati corrispondenti a ciascun evento temporale. I punti ottenuti vengono uniti da segmenti in modo da formare una poligonale che rappresenta, con buona approssimazione, l'andamento del fenomeno nel tempo.

Evoluzione della popolazione italiana. Anni 1861-2001 *(migliaia di unità)*



Indici di posizione

Indici di posizione

Gli **indici di posizione**, detti anche **misure di tendenza centrale**, consentono di sintetizzare una distribuzione di dati attorno ad un loro valore ritenuto centrale. Ci sono diversi modi per esprimere un valore centrale dei dati a disposizione. Analizzeremo i seguenti indici di posizione:

- la media;
- la mediana;
- la moda.

Indici di posizione: media

Definizione (Media aritmetica)

Supponiamo di considerare il carattere quantitativo di una popolazione e siano $x_1, x_2, \ldots, x_n$ i dati oservati. La loro media aritmetica $\bar{x}$ (o μ) è data dal rapporto fra la loro somma e il numero totale n di misurazioni:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (1)$$

Indici di posizione: media

Esempio

Supponiamo che uno studente del corso di Laurea in Farmacia abbia sostenuto i seguenti esami con le seguenti valutazioni

Esame	Voto
Matematica	28
Chimica generale ed inorganica	21
Biologia animale	26
Biologia Vegetale e Botanica Farmaceutica	29

La sua media aritmetica è

$$\bar{x} = \sum_{i=1}^{4} \frac{x_i}{4} = \frac{28 + 21 + 26 + 29}{4} = \frac{104}{4} = 26$$

Ai fini del calcolo del voto di laurea si considera la media ponderata dei voti ottenuti negli esami.

Indici di posizione: media

Definizione (Media aritmetica ponderata)

Consideriamo n dati numerici $x_1, x_2, \ldots, x_n$ e i corrispondenti pesi $p_1, p_2, \ldots, p_n$. La media ponderata è data da

$$\bar{x}_p = \frac{x_1 p_1 + x_2 p_2 + \cdots + x_n p_n}{p_1 + p_2 + \cdots + p_n} = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i} \tag{2}$$

Esempio

Calcoliamo la media ponderata dei voti dello studente di farmacia

Esame	Voto	CFU
Matematica	28	6
Chimica generale ed inorganica	21	10
Biologia animale	26	6
Biologia Vegetale e Botanica Farmaceutica	29	10

La sua media ponderata è

$$\bar{x}_p = \frac{\sum_{i=1}^{4} x_i p_i}{\sum_{i=1}^{4} p_i} = \frac{28 \cdot 6 + 21 \cdot 10 + 26 \cdot 6 + 29 \cdot 10}{6 + 10 + 6 + 10} = \frac{824}{32} = 25,75$$

Indici di posizione: media

Definizione (Media geometrica)

Consideriamo n dati numerici $x_1, x_2, \ldots, x_n$. La loro **media geometrica** *è data da:*

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} \tag{3}$$

Esempio

Supponiamo che una popolazione, composta inizialmente da P_0 elementi, subisca il primo anno un aumento del 14%, il secondo anno del 12%, il terzo anno del 10% e il quarto anno del 9%, sempre rispetto alla popolazione dell'anno precedente. Indicando con r_i il tasso di crescita annuale, la popolazione nei vari anni è data da:

$$P_1 = P_0 + r_1 P_0 = P_0(1 + r_1) = P_0\left(1 + \frac{14}{100}\right) = \frac{114}{100} P_0$$

$$P_2 = P_1 + r_2 P_1 = P_1(1 + r_2) = \frac{114}{100} P_0 \left(1 + \frac{12}{100}\right) = P_0 \frac{114}{100} \cdot \frac{112}{100}$$

Indici di posizione: media

$$P_1 = P_0 + r_1 P_0 = P_0(1 + r_1) = P_0\left(1 + \frac{14}{100}\right) = \frac{114}{100}P_0$$

$$P_2 = P_1 + r_2 P_1 = P_1(1 + r_2) = \frac{114}{100}P_0\left(1 + \frac{12}{100}\right) = \frac{114}{100} \cdot \frac{112}{100}P_0$$

$$P_3 = P_2 + r_3 P_2 = P_2(1 + r_3) = \frac{114}{100} \cdot \frac{112}{100}P_0\left(1 + \frac{10}{100}\right) = \frac{114}{100} \cdot \frac{112}{100} \cdot \frac{110}{100}P_0$$

$$P_4 = P_3 + r_4 P_3 = P_3(1 + r_4) = \frac{114}{100} \cdot \frac{112}{100} \cdot \frac{110}{100} \cdot \frac{109}{100}P_0 \simeq 1,53P_0$$

Quindi dopo 4 anni la popolazione è circa $1,53P_0$.
L'aumento percentuale medio è quindi dato da

$$\sqrt[4]{\frac{114}{100} \cdot \frac{112}{100} \cdot \frac{110}{100} \cdot \frac{109}{100}}$$

Indici di posizione: mediana

Definizione (Mediana)

Supponiamo di avere n dati $x_1, x_2, \ldots, x_n$ ordinati in senso crescente:

$$x_1 \leq x_2 \leq \cdots \leq x_n$$

Si definisce loro **mediana**:

- *il dato $x_{\frac{n+1}{2}}$ che occupa la posizione centrale, se n è dispari;*
- *la media aritmetica dei due dati nelle posizioni centrali $x_{\frac{n}{2}}$ e $x_{\frac{n}{2}+1}$, se n è pari.*

Indici di posizione: mediana

Esempio

Le stature (in cm) di un campione di 9 persone sono

$$173 \quad 182 \quad 177 \quad 174 \quad 175 \quad 179 \quad 164 \quad 181 \quad 182$$

Per determinare la mediana ordiniamo i dati in senso crescente:

$$x_1 = 164, \quad x_2 = 173, \quad x_3 = 174, \quad x_4 = 175, \quad x_5 = 177,$$
$$x_6 = 179, \quad x_7 = 181, \quad x_8 = 182, \quad x_9 = 182$$

Poiché i dati sono in numero dispari la mediana è il valore centrale $x_5 = 177$.
Se il campione fosse stato di 8 persone e il dato escluso fosse stato l'ultima misura (182 cm), la mediana sarebbe stata la media aritmetica tra i due valori centrali, quindi

$$\frac{x_4 + x_5}{2} = \frac{175 + 177}{2} = 176$$

Indici di posizione: mediana

Osservazione

La mediana divide i dati in due parti di ugual numerosità. Sia nel caso che n sia pari, sia nel caso che n sia dispari, risulta che almeno la metà dei dati sono minori o uguali alla mediana, e almeno la metà sono maggiori o uguali a essa. È proprio questa l'informazione che la mediana ci dà sul fenomeno oggetto dell'indagine statistica: almeno il 50% della popolazione presenta modalità minori o uguali alla mediana e almeno il 50% della popolazione presenta modalità maggiori o uguali alla mediana.

Quartili

Se si dividono i dati in 4 parti di uguale numerosità i 3 elementi che dividono i dati assegnati sono detti **quartili**. Il **primo quartile** divide il 25% dei dati che assumono il valore più basso dal 75% che assumono il valore più alto; il **secondo quartile** è la mediana; il **terzo quartile** separa il 75% dei dati più bassi dal 25% di quelli più alti.

Indici di posizione: moda

Definizione (Moda)

In un'indagine statistica si definisce **moda** *la modalità che si presenta con la massima frequenza.*

Esempio

Nella scelta della scuola effettuata dagli studenti italiani in possesso della licenza di scuola media inferiore, nel 1948 la moda era rappresentata dai Licei Classici.

Esempio

Nell'ultima sessione l'esame di matematica è stato superato da 7 persone con i seguenti voti: 30, 19, 22, 27, 25, 22, 21. Tutte le modalità hanno frequenza pari a 1, eccetto 22 che ha frequenza 2, quindi la moda è 22.

Indici di dispersione

La sintesi attraverso gli indici di posizione spesso non è sufficiente, in quanto non tiene conto di come i dati sono distribuiti attorno al valore centrale. Ad esempio entrambe le distribuzioni di dati

8	9	10	11	12
1	2	10	18	19

ammettono come media e mediana il valore 10, ma nel primo caso tutti i dati sono raccolti attorno alla media, mentre nel secondo sono molto più dispersi. Per questo motivo può essere utile utilizzare gli **indici di dispersione**, i quali danno informazioni sulla variabilità dei dati.

Indici di dispersione: intervallo di variazione

Definizione (Intervallo di variazione)

Supponiamo di avere n dati $x_1, \ldots, x_n$. Indichiamo con x_{max} e x_{min} rispettivamente il più grande e il più piccolo degli n dati. Si definisce **intervallo di variazione** *I la differenza tra il più grande e il più piccolo dei dati, cioè*

$$I = x_{max} - x_{min}$$

Esempio

Nella distribuzione

$$8 \quad 9 \quad 10 \quad 11 \quad 12$$

l'intervallo di variazione $I = 12 - 8 = 4$.
Nella distribuzione

$$1 \quad 2 \quad 10 \quad 18 \quad 19$$

l'intervallo di variazione $I = 19 - 1 = 18$.

Indici di dispersione: intervallo di variazione

Osservazione

Il valore I dipende solo dal dato più alto e da quello più basso, non tiene conto degli altri dati. Perciò due dati estremi molto diversi dagli altri possono cambiare in modo significativo l'intervallo di variazione I, nonostante siano statisticamente irrilevanti.

Scarti dalla media

Per cominciare ad avere una misura della dispersione dei dati si può calcolare la loro deviazione dalla media aritmetica, ossia la differenza tra ognuno di essi e la media $\bar{x}$. Tale differenza è detta **scarto** dalla media.

Indici di dispersione: varianza

Definizione (Varianza)

Si definisce **varianza** *di n dati* $x_1, x_2, \ldots, x_n$*, e si indica con* $Var(x_i)$ *o* σ^2*, la media aritmetica dei quadrati degli scarti dalla media* $\bar{x}$*:*

$$Var(x_i) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (4)$$

Quando non si utilizza il computer, per calcolare la varianza spesso bisogna eseguire molti calcoli. In aletrnativa si può utilizzare un'altra formula, equivalente alla (4), ma che comporta meno calcoli.

Indici di dispersione: varianza

Proposizione

La varianza $Var(x_i) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ *può essere equivalentemente espressa come*

$$Var(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \tag{5}$$

Dimostrazione.

$$\begin{aligned} Var(x_i) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n}\frac{x_i}{n} + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + \frac{1}{n}\cdot n\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \end{aligned}$$

Indici di dispersione: scarto quadratico medio

Osservazione

A causa dell'elevamento al quadrato degli scarti, la varianza porta a delle distorsioni dimensionali: non presenta la stessa unità di misura delle modalità del carattere.

Per superare tale problema si è definito un altro indice di dispersione.

Definizione (Scarto quadratico medio)

Si definisce **scarto quadratico medio** *o* **deviazione standard** *dei dati* $x_1, x_2, \ldots, x_n$*, e si indica con* σ*, la radice quadrata della loro varianza:*

$$\sigma = \sqrt{Var(x_i)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \tag{6}$$

Indici di dispersione: varianza e scarto quadratico medio

Esempio

Nella classe 5A di una scuola tutti i 20 studenti hanno preso 6 nel compito in classe. Nella classe 5B, anch'essa di 20 studenti, la metà ha preso 4, l'altra metà 8. Calcoliamo la varianza e la deviazione standard.
Nella 5A gli studenti hanno preso tutti lo stesso voto, quindi non c'è variabilità. Infatti sia varianza che deviazione standard sono uguali a 0:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{20}\left(6^2 + \cdots + 6^2\right) - \left(\frac{6 + \cdots + 6}{20}\right)^2 = 0$$

$$\sigma = \sqrt{Var(x)} = \sqrt{0} = 0$$

Per la 5B osserviamo che la media dei voti $\bar{x} = \frac{4 \cdot 10 + 8 \cdot 10}{20} = \frac{120}{20} = 6$.

$$Var(x) = \frac{1}{20}\left(\underbrace{4^2 + \cdots + 4^2}_{10 volte} + \underbrace{8^2 + \cdots + 8^2}_{10 volte}\right) - 6^2 = 4$$

$$\sigma = \sqrt{4} = 2$$

Indici di dispersione: varianza e scarto quadratico medio

Osservazione

Varianza e deviazione standard

- assume valore minimo, uguale a 0, se tutti i dati osservati sono uguali;
- assumono valori positivi, via via crescenti all'aumentare della variabilità.

Correlazione

Talvolta non si ha a che fare con singoli dati, ma con coppie di dati.

Esempio

I livelli di emoglobina prima della terapia sono in relazione con i livelli di emoglobina dopo la terapia?

La dipendenza tra due caratteri quantitativi X e Y è detta **correlazione**. Un indice statistico molto utilizzato per valutare la correlazione tra due caratteri quantitativi è la *covarianza*.

Definizione (Covarianza)

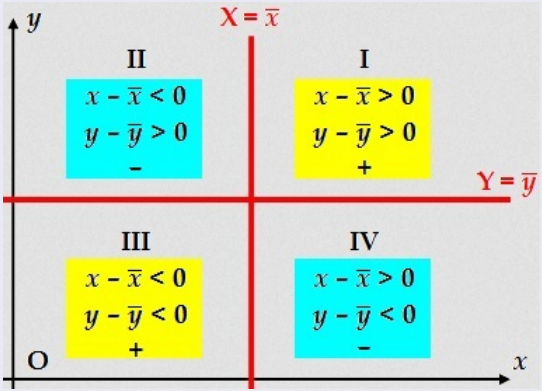
Siano X e Y due caratteri quantitativi di medie $\bar{x}$ e $\bar{y}$, rilevati da una popolazione di n unità. Siano $x_1, \ldots, x_n$ i valori osservati di X e $y_1, \ldots, y_n$ i valori osservati di Y. Si definisce **covarianza** *di X e Y il numero*

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \tag{7}$$

Covarianza

Significato covarianza

Rappresentando su un piano cartesiano i punti (x_i, y_i) si ottiene una "nuvola" di punti.
Consideriamo il punto $(\bar{x}, \bar{y})$ il baricentro di tale "nuvola" e tracciamo passanti per esso e parallele agli assi. Le rette dividono il piano in 4 semipiani che numeriamo in senso antiorario partendo da quello in alto a destra (I), come in figura.



Covarianza

Significato covarianza

A seconda del semipiano in cui si trova il punto (x_i, y_i) gli scarti $(x_i - \bar{x})$ e $(y_i - \bar{y})$ assumono diversi segni, quindi la covarianza σ_{XY} assume diversi segni. In particolare

- se $\sigma_{XY} > 0$ la maggioranza dei prodotti $(x_i - \bar{x})(y_i - \bar{y})$ sono positivi, quindi la maggior parte dei punti (x_i, y_i) si trova nei semipiani *I* e *III*. Quindi la nuvola di punti si concentra in tali semipiani: quindi vi è una relazione di tipo lineare crescente tra X e Y;
- se $\sigma_{XY} < 0$ la maggioranza dei prodotti $(x_i - \bar{x})(y_i - \bar{y})$ sono negativi, quindi la maggior parte dei punti (x_i, y_i) si trova nei semipiani *II* e *IV*. Quindi la nuvola di punti si concentra in tali semipiani: quindi vi è una relazione di tipo lineare decrescente tra X e Y;
- se $\sigma_{XY} = 0$ i punti (x_i, y_i) sono sparpagliati senza alcuna regolarità, oppure sono disposti secondo relazioni diverse da quella lineare.

Coefficiente di correlazione lineare

Una volta appurata una correlazione tra due caratteri X e Y si pone il problema di stabilire se essa è forte o debole. Per stabilire ciò esiste un indice apposito.

Definizione (Coefficiente di correlazione lineare)

Dati due caratteri X e Y, si definisce **coefficiente di correlazione lineare** *il numero*

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \tag{8}$$

dove σ_X e σ_Y sono le deviazioni standard rispettivamente di X e Y.

Coefficiente di correlazione lineare

Osservazioni

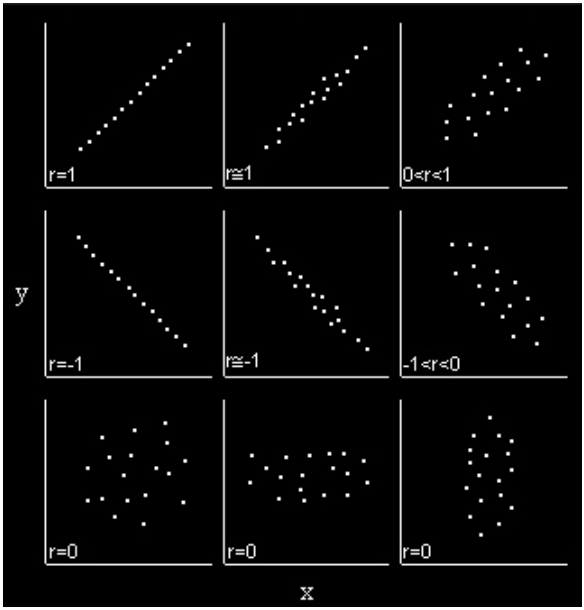
Poiché

$$-\sigma_X \sigma_Y \leq \sigma_{XY} \leq \sigma_X \sigma_Y$$

dalla formula della correlazione lineare $r = \dfrac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$ possiamo dedurre che:

- $-1 \leq r \leq 1$;
- il coefficiente di correlazione lineare r ha lo stesso segno della covarianza σ_{XY} e dà informazioni analoghe: $r > 0$ indica una relazione lineare crescente, $r < 0$ indica una relazione lineare decrescente;
- si può dimostrare che $r = \pm 1$ equivale ad una relazione lineare perfetta. Tanto più r è vicino a 0 quanto più il legame tra X e Y (se esiste) è distante da quello lineare.

Coefficiente di correlazione lineare



Regressione lineare

Se esiste una relazione lineare tra due caratteri X e Y possiamo determinare la funzione lineare che rappresenta tale legame.

Definizione (Retta di regressione lineare)

Dati due caratteri X e Y, si definisce **retta di regressione** *che esprime Y in funzione di X la retta che passa per il punto di coordinate $(\bar{x}, \bar{y})$ e che ha come coefficiente angolare m il coefficiente di regressione così definito:*

$$m = \frac{\sigma_{XY}}{\sigma_X^2}$$

Ricordando la formula della retta passante per un punto e di direzione assegnata, l'equazione della retta di regressione è

$$y - \bar{y} = m(x - \bar{x}) \qquad m = \frac{\sigma_{XY}}{\sigma_X^2} \tag{9}$$

Esempio

La relazione tra il peso X alla nascita di un neonato e la stima del peso Y alla nascita di un neonato (ecografia) può essere rappresentata con la seguente retta di regressione lineare, costruita a partire dai vari pesi x_i, dai pesi stimati y_i.