Lezione 18/10/2022

Statistica descrittiva

Per *misurare* la variabilità di un carattere si definiscono degli indici opportuni, detti indici di variabilità. Un indice di variabilità è la *varianza* definita nella lezione precedente.

Definizione (Una formula alternativa della varianza)**.** *Date n modalità x_1, x_2, ..., x_n, di media aritmetica $\bar{x}$, la loro varianza è espressa dalla formula:*

$$\sigma^2 = \frac{x_1^2 + \ldots + x_n^2}{n} - \bar{x}^2$$

Un altro indice utilizzato per misurare la variabilità dei dati è lo *scarto semplice medio.*

Definizione (Scarto semplice medio)**.** *Date n modalità x_1, x_2, ..., x_n, di media aritmetica $\bar{x}$, il loro scarto semplice medio è espresso dalla formula:*

$$\frac{|x_1 - \bar{x}| + \ldots + |x_n - \bar{x}|}{n}.$$

Si preferisce usare la varianza invece dello scarto quadratico medio perchè non contiene valori assoluti che complicano i calcoli.

Come nel caso della media si può calcolare la varianza a partire dalle frequenze assolute e relative.

Definizione. *Date le modalità $x_1, \ldots, x_n$ con frequenze assolute $f_1, \ldots, f_n$ e frequenze relative $f_{r_1}, \ldots, f_{r_n}$, la varianza è espressa dalle seguenti formule.*

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 \cdot f_1 + \ldots + (x_n - \bar{x})^2 \cdot f_n}{f_1 + \ldots + f_n};$$

$$\sigma^2 = \frac{x_1^2 \cdot f_1 + \ldots + x_n^2 \cdot f_n}{f_1 + \ldots + f_n} - \bar{x}$$

$$\sigma^2 = (x_1 - \bar{x})^2 \cdot f_{r_1} + \ldots + (x_n - \bar{x})^2 \cdot f_{r_n};$$

$$\sigma^2 = x_1^2 \cdot f_{r_1} + \ldots + x_n^2 \cdot f_{r_n} \; - \; \bar{x}.$$

Osservazione: Più è piccola la varianza e più le modalità sono vicine alla loro media, più la varianza è grande e più le modalità sono lontane dalla loro media.

Anche la varianza ha dei difetti: a causa dell'elevamento al quadrato degli scarti, non presenta la stessa unità di misura delle modalità del carattere ed è molto più grande degli scarti. Per questo motivo si definisce un ulteriore indice, che ristabilisce l'unità di misura e l'ordine di grandezza delle modalità, e che risulta perciò solitamente preferibile alla varianza.

Definizione (deviazione standard)**.** *La deviazione standard (o scarto quadratico medio) è la radice quadrata della varianza e si indica con* σ*.*

Esempio. *Consideraiamo i voti di due studenti A e B:*
A: 22 24 24 26
B: 30 29 18 19
La media dei voti di A e B è 24, anche se essi presentano una variabilità diversa. I primi sono meno variabili dei secondi, infatti si avvicinano maggiormente a 24. Calcolando la varianza infatti otteniamo $\sigma_A^2 = 4$ *e* $\sigma_B^2 = 28$*. La varianza però non ci dice quanto i voti sono lontani dalla media; basti pensare, ad esempio, che 28 è più grande degli scarti 28-24, 30-24, 19-24 e 18-24. Calcoliamo dunque la deviazione standard:* $\sigma_A = \sqrt{4} = 2$ *e* $\sigma_B = \sqrt{28} = 5.2$*.*

Esempio. *Considero i pesi di 3 oggetti: 100 g, 200 g e 300 g. La loro varianza è* $6700\ g^2$*. Osservo che*

- *l'ordine di grandezza di* 6700 *è diverso da quello delle modalità (la varianza è un numero molto più grande delle modalità).*
- *l'unità di misura della varianza è* g^2 *che è diversa da quella delle modalità (g).*

Con la deviazione standard supero questi problemi, infatti è uguale a $\sqrt{6700g^2} = 81,4\ g$*.*

Proprietà della varianza e della deviazione standard

Consideriamo le modalità $x_1, \ldots, x_n$ di varianza σ^2 e deviazione standard σ e il numero reale k, allora

- $x_1 + k, \ldots, x_n + k$ hanno varianza σ^2 e deviazione standard σ^2;
- $k \cdot x_1, \ldots, k \cdot x_n$ hanno varianza $k^2 \cdot \sigma^2$ e deviazione standard $|k| \cdot \sigma$.

Esempio. *Le modalità 5 1 2 3 e 105 101 102 103 hanno la stessa variaza e la stessa deviazione standard, infatti 105=**100**+5, 101=**100**+1, 102=**100**+2 e 103=**100**+3.*

Esempio. *4 e 8 hanno varianza 4 e deviazione standard 2. Li moltiplico per tre e ottengo* $\boldsymbol{3} \cdot 4 = 12$ *e* $\boldsymbol{3} \cdot 8 = 28$*. La varianza di 12 e 28 è* $\boldsymbol{3}^2 \cdot 4 = 36$*, mentre la loro deviazione standard è* $\boldsymbol{3} * 2 = 6$*.*

Coefficiente di variabilità

La varianza e la deviazione standard sono indici che dipendono dall'unità di misura e dall'ordine di grandezza dei dati. Pertanto non avrebbe senso, per esempio, confrontare le deviazioni standard di due fenomeni misurati con unità di misura diverse, né fenomeni misurati con la stessa unità di misura, ma tali per esempio che gli ordini di grandezza delle misure di un fenomeno siano molto maggiori degli ordini di grandezza delle misure dell'altro.Per eseguire il confronto fra la variabilità di due fenomeni, occorre utilizzare una misura della variabilità *depurata* dall'influenza dell'unità di misura e dell'ordine di grandezza dei dati. Questo obiettivo si può raggiungere costruendo il rapporto tra la deviazione standard e un valore che sintetizzi l'ordine di grandezza delle modalità del fenomeno osservato e che sia espresso nella medesima unità di misura: il valore più noto che soddisfa queste ultime proprietà è la media aritmetica. In definitiva, allora, si definisce il seguente indice.

Definizione (Coefficiente di variazione)**.** *Date n modalità* $x_1, \ldots, x_n$ *di media* $\bar{x}$ *e deviazione standard* σ*, il coefficiente di variazione è espresso dalla seguente formula:*

$$CV = \frac{\sigma}{\bar{x}}.$$

Esempio. *Consideriamo il peso e il volume di 200 pezzi tali che*

$\bar{x}_P = 9\ Kg$*;*

$\bar{x}_V = 2,7\ m^3$*;*

$\sigma_P = 1,5\ Kg$*;*

$\sigma_V = 0,6\ m^3$*.*

A partire dalle deviazioni standard non possiamo confrontare le variabilità del peso e del volume dei pezzi. Calcoliamo allora i coefficienti di variabilità:

$$CV_P = \frac{1,5\ Kg}{9\ Kg} = 0,16 \quad e \quad CV_V = \frac{0,6\ m^3}{2,7\ m^3} = 0,22.$$

Dato che $CV_V > CV_P$*, il volume dei pacchi è più variabile del loro peso.*

Quartili

- Il primo quartile Q_1 è un valore tale che il 25% dei dati ordinati è minore o uguale a Q_1.
- Il primo quartile Q_2 è un valore tale che il 50% dei dati ordinati è minore o uguale a Q_2 e coincide con la mediana.
- Il terzo quartile Q_3 è un valore tale che il 75% dei dati ordinati è minore o uguale a Q_3.

Regola per il calcolo dei quartili

1. Si ordinano gli n dati assegnati in ordine crescente;
2. si calcola il prodotto $k = np$, dove $p = 0.25$ per il primo quartile, $p = 0.5$ per il secondo e $p = 0.75$ per il terzo.
3. se k è un intero, il quartile si ottiene facendo la media del k-esimo e del (k+1)-esimo valore dei dati ordinati;
4. se k non è intero, si arrotonda k per eccesso al primo intero successivo e si sceglie come quartile il corrispondente valore dei dati ordinati.

Esempio. *Calcolare il primo e il terzo quartile dell'insieme di dati*

32.2 32.0 30.4 31.0 31.2 31.3 30.3 29.6 30.5 30.7

Dati ordinati

29.6 30.3 30.4 30.5 30.7 31.0 31.2 31.3 32.0 32.2

Primo quartile :
$n = 10$ *e* $p = 0.25$*, quindi* $k = np = 2$*.*
k non è intero, perciò si arrotonda per eccesso k = 3 : il primo quartile è il terzo dei dati ordinati

$$Q_1 = 30.4.$$

Terzo quartile:
$n = 10$ *e* $p = 0.75k = np = 7.5$
k non è intero, perciò si arrotonda per eccesso 8 k = : il terzo quartile è l'ottavo dei dati ordinati

$$Q_3 = 31.3.$$

Secondo quartile (mediana):
$n = 10$ *e* $p = 0.5$*, quindi* $k = np = 5$
k è intero, perciò si fa la media tra il quinto e il sesto dato e si ottiene

$$Q_2 = 30.85$$

(Questo valore coincide con quello che si trova con la regola della mediana).

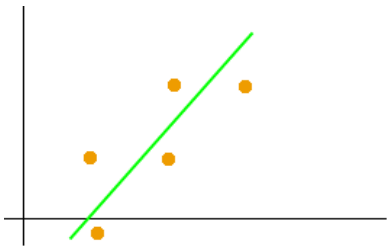
Correlazione lineare

Premettiamo che la dipendenza tra due caratteri di tipo quantitativo viene chiamata correlazione. Un indice statistico molto utilizzato per valutare la correlazione tra due caratteri quantitativi è la cosiddetta covarianza, così definita

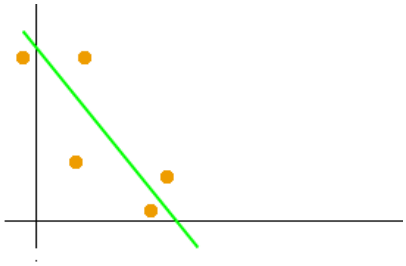
Definizione (Covarianza)**.** *Siano X e Y due variabili statistiche di medie $\bar{x}$ e $\bar{y}$, rilevate congiuntamente su un collettivo di n unità. Siano x_1, x_2, ..., x_n le modalità di X e y_1, y_2, ..., y_n le corrispondenti modalità di Y. Si chiama covarianza di X e Y, e si indica con il simbolo σ_{XY} , il numero così definito*

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

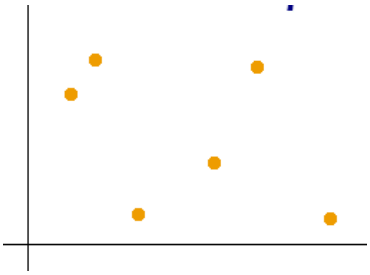
- *Se $\sigma_{xy} > 0$ allora tra X e Y c'è una correlazione lineare di tipo crescente (o diretta);*



- *Se $\sigma_{xy} < 0$ allora tra X e Y c'è una correlazione lineare di tipo decrescente (o inversa);*



- *Se $\sigma_{xy} = 0$ non c'è correlazione lineare tra X e Y (i punti sono sparpagliati o tra loro c'è una relazione che non è lineare).*



Una volta appurata se c'è una correlazione lineare tra i due caratteri dobbiamo capire se essa è forte o debole, cioè quanto i punti tendono a disporsi su una linea retta.

Massimo e minimo della covariaza

Si può dimostrare che la covarianza σ_{xy} di X e Y appartiene all'intervallo

$$[-\sigma_x\sigma_y, \sigma_x\sigma_y],$$

dove σ_x e σ_y sono la deviazioni standard di X e Y.

- Più σ_{xy} si avvicina a $\sigma_x\sigma_y$ più la correlazione lineare crescente è forte;
- Più σ_{xy} si avvicina a $-\sigma_x\sigma_y$ più la correlazione lineare decrescente è forte.

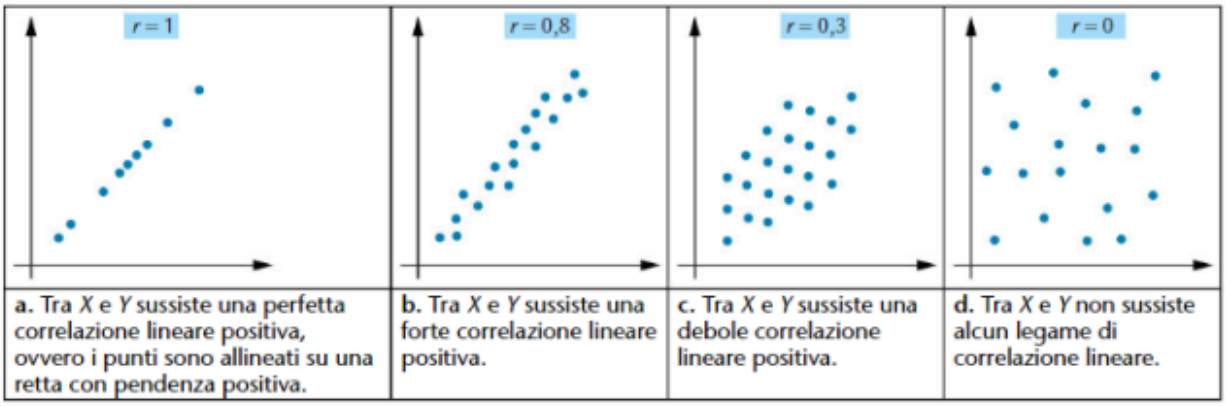
Si vuole trovare un indice che non dipenda da σ_x e σ_y e che quindi consenta di confrontare le covarianze di coppie di caratteri diversi tra di loro, ad esempio (peso, altezza) e (peso, età). Questo indice si ottine dividendo la covarianza per $\sigma_x\sigma_y$.

Definizione (Coefficiente di correlazione lineare)**.** *Si chiama coefficiente di correlazione lineare di due variabili X e Y, e si indica con il simbolo R, il numero così definito*

$$R = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

E' importante fare alcune osservazioni.

- Per come è stato definito R, risulta sempre: $-1 \leq R \leq 1$.
- Il segno del coefficiente di correlazione lineare è lo stesso della covarianza e dà informazioni analoghe: un coefficiente $R > 0$ indica una relazione lineare crescente, mentre un coefficiente $R < 0$ indica una relazione lineare decrescente.
- Si può dimostrare che l'indice di correlazione R è uguale a $+1$ o -1 se e solo se tra Y e X sussiste una perfetta relazione lineare. Tanto più R è vicino a $+1$ o -1, quanto più il modello lineare interpreta bene la relazione che sussiste tra Y e X; tanto più R è vicino a 0, quanto più il legame tra Y e X (se c'è) è distante da quello lineare, come illustrato nelle seguenti figure.

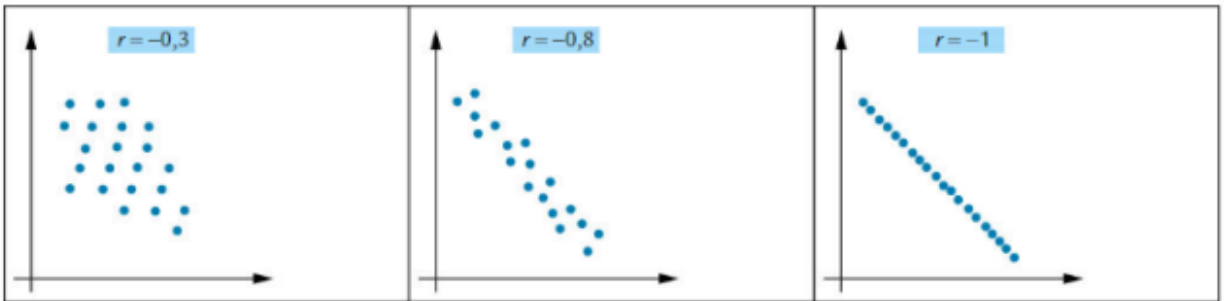


a. Tra *X* e *Y* sussiste una perfetta correlazione lineare positiva, ovvero i punti sono allineati su una retta con pendenza positiva.

b. Tra *X* e *Y* sussiste una forte correlazione lineare positiva.

c. Tra *X* e *Y* sussiste una debole correlazione lineare positiva.

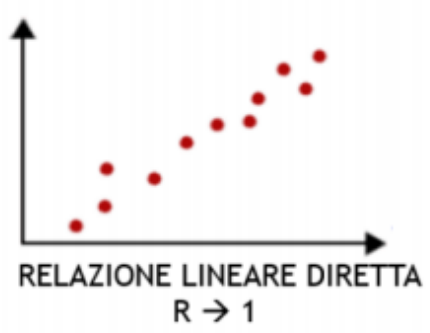
d. Tra *X* e *Y* non sussiste alcun legame di correlazione lineare.

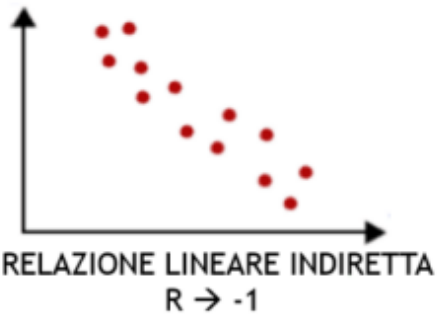


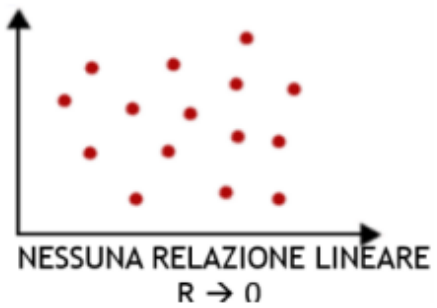
e. Tra *X* e *Y* sussiste una debole correlazione lineare negativa.

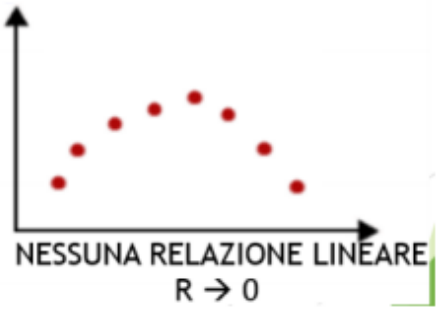
f. Tra *X* e *Y* sussiste una forte correlazione lineare negativa.

g. Tra *X* e *Y* sussiste una perfetta correlazione lineare negativa, ovvero i punti sono allineati su una retta con pendenza negativa.









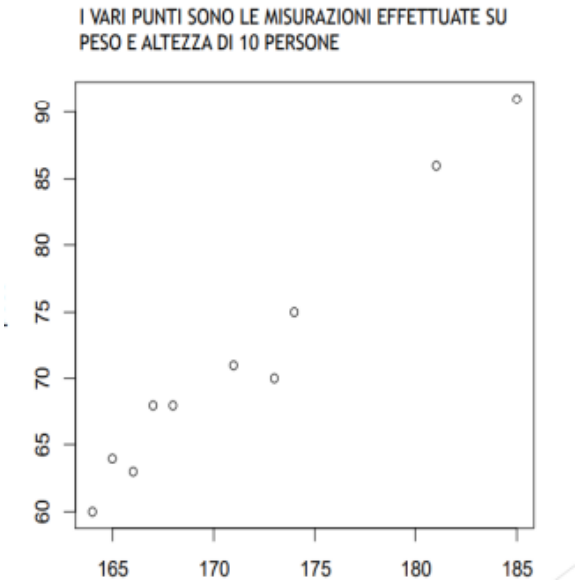
Retta di regressione

Dopo avere scoperto l'esistenza di una relazione lineare tra le due variabili X e Y, in base all'analisi di un diagramma cartesiano o al calcolo del coefficiente di correlazione (che deve essere vicino a + o - 1), ci proponiamo di determinare la funzione lineare che interpreta meglio tale legame. Vogliamo determinare, dunque, la retta che più si avvicina ai punti $(x_1, y_1), \ldots, (x_n, y_n)$, dove $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$ sono le modalità di X e Y.

Partendo da un esempio, possiamo studiare il legame tra i caratteri peso e altezza di un gruppo di 10 persone le cui modalità sono riportate nella seguente tabella:

Soggetto	Altezza	Peso
A	174	75
B	166	63
C	173	70
D	171	71
E	168	68
F	167	68
G	165	64
H	164	60
I	181	86
L	185	91

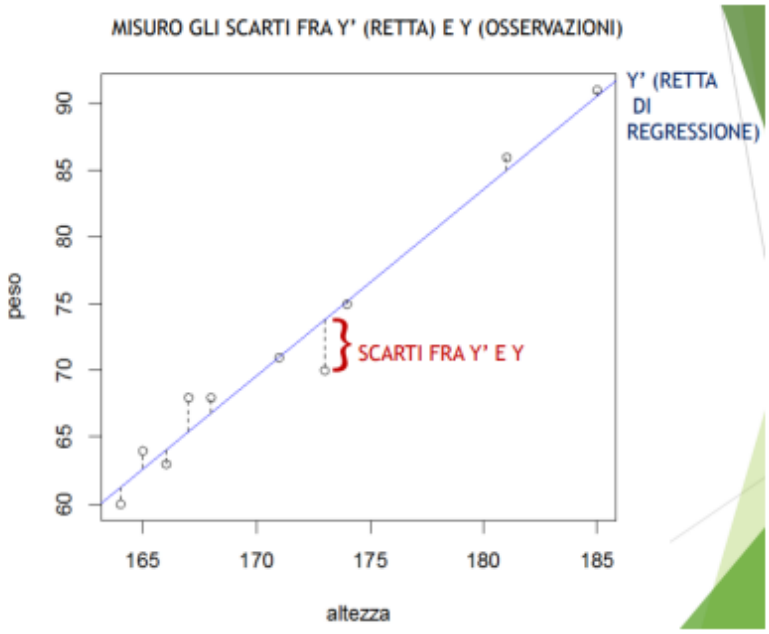
Per osservare graficamente se esiste un legame tra X e Y si può disegnare il diagramma di dispersione (o scatter-plot). Il diagramma di dispersione è un grafico dove sull'asse delle ascisse sono riportate le modalità di X e sull'asse delle ordinate le modalità di Y. Ogni coppia di modalità (x_i, y_i) è rappresentata da un punto del piano (vedi la seguente tabella).



Nell'esempio è evidente che i punti tendono a disporsi lungo una retta, dunque diciamo che tra di loro c'è una relazione lineare, ovvero che all'aumentare dell'altezza il peso tende ad aumentare nello stesso modo.

La **retta di regressione** $y = ax + b$ è una retta che meglio descrive il legame tra i due caratteri X e Y. Per trovare i parametri a e b utilizziamo il *metodo dei minimi quadrati*: la retta che meglio si avvicina ai punti $(x_1, y_1), \ldots, (x_n, y_n)$ del diagramma di dispersione è quella per cui **la somma dei quadrati degli scarti è minima**.

Lo scarto i-esimo è la quantità $y_i - (ax_i + b)$ e graficamente corrisponde alla distanza tra i punti (x_i, y_i) e $(x_i, ax_i + b)$ (vedi la seguente figura).



Dunque, cerchiamo i numeri reali a e b in maniera tale che la quantità

$$S(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))$$

sia minima.

Attraverso passaggi matematici troviamo che il valore minimo di $S(a,b)$ si ottiene in corrispondenza dei seguenti valori di a e b:

$$a = \frac{\sigma_{xy}}{\sigma_x^2} \quad \text{and} \quad b = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}.$$

Quindi, l'espressione analitica della retta di regressione è la seguente:

$$y = \frac{\sigma_{xy}}{\sigma_x^2}x + \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}.$$