SCHEDA N. 4: VARIABILI QUANTITATIVE
(Trasformazioni lineari – Indici di covarianza e correlazione)

1) Trasformazioni lineari di variabili statistiche

A partire da una variabile se ne possono costruire altre ad essa legate; alcuni esempi ci sono familiari: operiamo una trasformazione di una variabile quando cambiamo unità di misura, ad esempio passando da dati espressi in centimetri a dati espressi in metri, oppure quando trasformando le temperature espresse in gradi Celsius in quelle in gradi Fahrenheit.
Se indichiamo con X misure espresse in centimetri e con Y le stesse espresse in metri, avremo: Y = 0.01 X
Se indichiamo con X le temperature espresse in gradi Celsius e con Y quelle in gradi Fahrenheit, avremo: Y=100/180(X-32)

In questi casi le trasformazioni sono lineari, cioè del tipo:

$$Y = a X + b$$

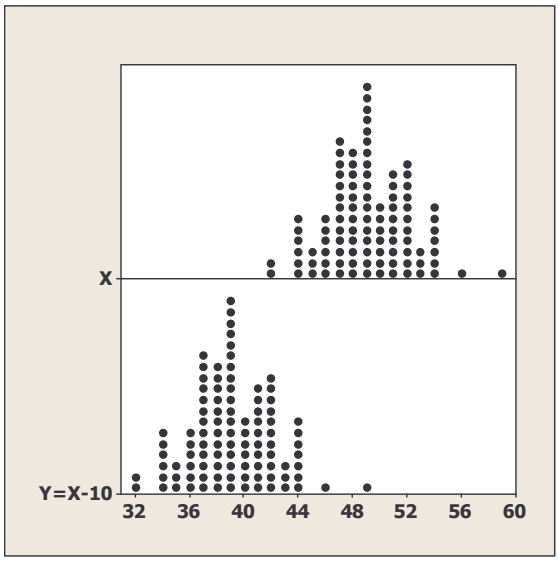
Ciascun dato viene trasformato nel seguente modo:

$$y_i = a\,x_i + b$$

Il coefficiente "**b**" rappresenta una traslazione mentre il coefficiente "**a**" è un fattore di scala che incide sulla variabile mediante una dilatazione o una contrazione (dilatazione se a > 1 e contrazione se a < 1).

Vediamo ora come si comportano media e varianza della variabile trasformata linearmente rispetto agli stessi indici della variabile originale.
Indichiamo con $\overline{x}$ e $\overline{y}$ le medie e con σ_X^2 e σ_Y^2 le varianze delle due variabili.

A) Traslazione $Y = X + b$



La media cambia: viene traslata di b, così come i singoli dati.

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i + b) = \overline{x} + b$$

La varianza resta uguale: infatti è basata sugli scarti dalla media, che restano uguali dopo la traslazione:

$$y_i - \overline{y} = x_i + b - \left(\overline{x} + b\right) = x_i - \overline{x}$$

Nell'esempio riportato a fianco si ha:

$\overline{x}$ = 49.1 e σ_X^2 = 9.07

$\overline{y}$ = 39.1 e σ_Y^2 = 9.07

B) Dilatazione/contrazione Y = a X

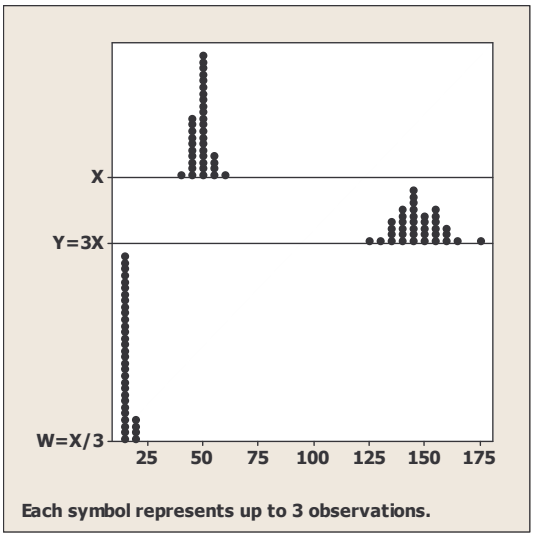
La media cambia: viene dilatata o contratta del fattore a, così come i singoli dati.

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} a\,x_i = \frac{a}{n}\sum_{i=1}^{n} x_i = a\,\bar{x}$$

La varianza cambia; gli scarti dalla media diventano:

$$y_i - \bar{y} = a\,x_i - a\,\bar{x} = a\left(x_i - \bar{x}\right)$$

e quindi $\sigma_y^2 = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} a^2\left(x_i - \bar{x}\right)^2 = a^2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = a^2\,\sigma_X^2$.



A fianco sono rappresentate, oltre alla variabile X dell'esempio precedente, una variabile Y "dilatata" 3 volte e una W "contratta" 3 volte.

$\bar{x}$ = 49.1 e σ_X^2 = 9.07

$\bar{y}$ = 147.3 e σ_Y^2 = 81.63

$\bar{w}$ = 16.37 e σ_W^2 = 1.01

Bisogna fare attenzione ai pallini: per problemi di scala nei tre grafici un pallino corrisponde a un diverso numero di osservazioni.

In presenza sia di traslazione che di dilatazione/contrazione si ha:

- la media si trasforma secondo la stessa trasformazione della variabile X, ovvero
$$\bar{y} = a\bar{x} + b\,.$$
- la varianza, invece, ha un comportamento differente
$$\sigma_y^2 = a^2\sigma_X^2\,.$$
e la deviazione standard si trasforma nel seguente modo:
$$\sigma_Y = |\,a\,|\,\sigma_X$$
infatti la deviazione standard è un indice positivo.

C) Centratura e "standardizzazione"

La trasformazione Y = X- $\bar{x}$ è detta **centratura.**

La variabile X viene traformata in una variabile Y con media **zero**.

La trasformazione $Z = \dfrac{X - \bar{x}}{\sigma_X}$ è detta **standardizzazione**.

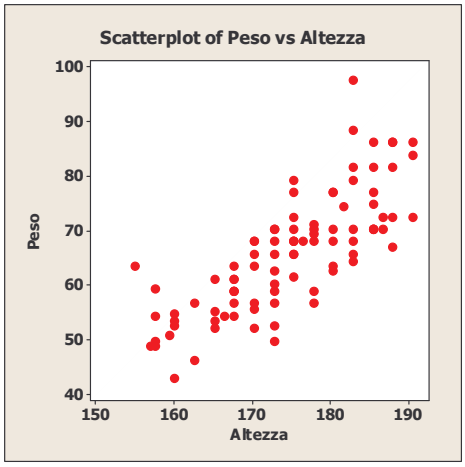
La variabile X viene traformata in una variabile Z con media **zero** e varianza **uno**.

NB: Le formule precedenti valgono **solo** per trasformazioni lineari. Ad esempio se Y= 1/X **non è vero** che $\bar{y} = 1/\bar{x}$!

2) Distribuzione congiunta di due variabili qualitative e loro rappresentazione grafica

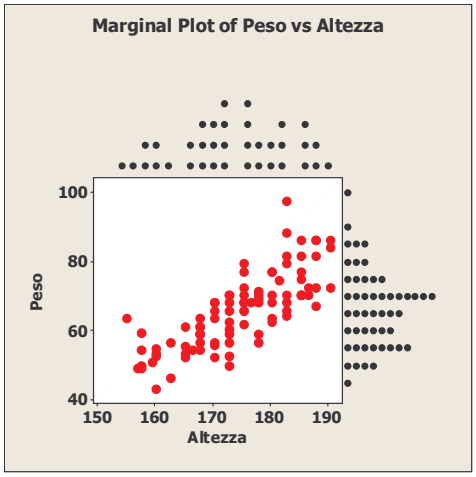
I risultati di due variabili quantitative X e Y rilevate sulla stessa popolazione possono essere rappresentati attraverso punti su un piano: a ciascuna osservazione è associato un punto le cui coordinate sono i valori di X e Y per quella osservazione, indicati con (x_i,y_i). Il grafico si chiama **diagramma di dispersione bidimensionale** o **scatterplot**.
L'insieme delle K differenti coppie di valori (x_k,y_k) e delle corrispondenti frequenze relative è detta **distribuzione congiunta** di **X** e **Y**.

ESEMPIO.
Consideriamo il grafico della distribuzione congiunta dei pesi e delle altezze dei soggetti dell'esperimento sulle pulsazioni (già visto nelle schede n. 2 e 3).



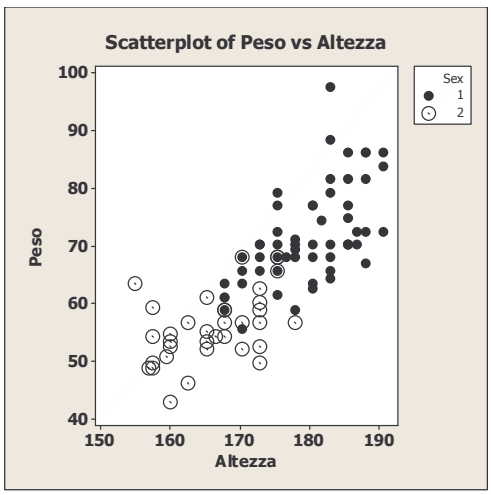
Notiamo che nel titolo dei diagrammi relativi a due variabili i software statistici scrivono:
"variabile rappresentata sulle ordinate"
rispetto (versus in inglese)
"variabile rappresentata sulle ascisse"

La rappresentazione grafica a fianco evidenzia, oltre alla distribuzione congiunta delle due variabili, anche le due distribuzioni marginali di X e Y. La situazione è del tutto analoga a quanto abbiamo visto nel caso di variabili qualitative.



Il baricentro dei dati relativi a due variabili è il punto $(\bar{x}, \bar{y})$
cioè il punto che ha coordinate i due baricentri della variabile X e della variabile Y. Anche in questo caso il baricentro è il punto di equilibrio della distribuzione.

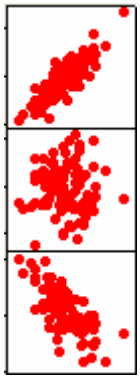
Nel grafico della distribuzione congiunta si può anche evidenziare l'appartenenza dei soggetti ai livelli di una variabile qualitativa, così come è fatto a fianco per il genere: maschi (1) e femmine (2).



3) Indici per due variabili quantitative: la covarianza e la correlazione.

Quando si hanno due variabili quantitative X e Y, definite sulla stessa popolazione di n unità, ci possiamo chiedere se esiste un legame lineare tra le due variabili e, in caso affermativo, di che tipo sia. Esamineremo come si costruiscono e che proprietà hanno due nuovi indici: la **covarianza** e la **correlazione**.

A) Gli indici di covarianza e correlazione hanno la proprietà di essere:



positivi per dati che hanno un comportamento come quello a fianco

vicini a zero per dati che hanno un comportamento come quello a fianco

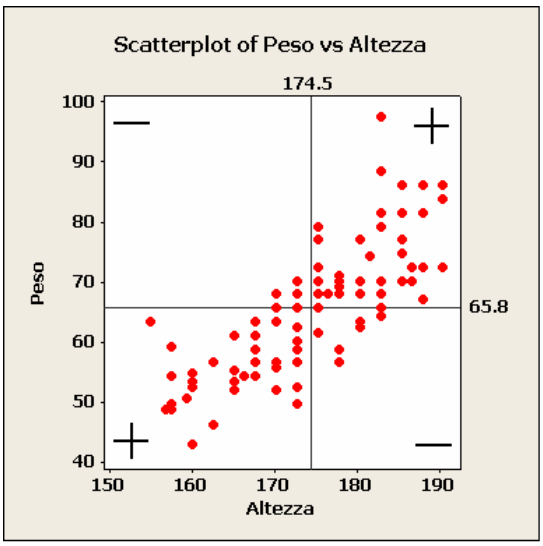
negativi per dati che hanno un comportamento come quello a fianco

B) Gli indici di covarianza e correlazione sono costruiti anzitutto centrando i dati nel baricentro. Indichiamo con $\tilde{x}$ e con $\tilde{y}$ le variabili centrate.

Osserviamo che, *una volta centrati i dati nel baricentro,* i prodotti

$$\tilde{x}_i \ \tilde{y}_i$$

sono positivi per i dati che sono rappresentati nel primo e nel terzo quadrante e negativi per i dati che sono rappresentati nel secondo e nel quarto quadrante dei nuovi assi.
Nell'esempio riportato a fianco la maggior parte dei prodotti è positiva e inoltre i prodotti negativi sono "piccoli".



La **covarianza** fra X e Y è data da

$$\text{Cov(X,Y)}= \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\,\tilde{y}_i = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \ \text{ oppure } \ \sum_{k=1}^{K} f_k(x_k-\bar{x})(y_k-\bar{y})$$

avendo indicato con (x_k, y_k) i K differenti valori assunti dalle variabili e con f_k le corrispondenti frequenze relative.
Talvolta come nel caso della varianza , l'indice di covarianza può avere (n-1) al denominatore.

Come la varianza, la covarianza può essere scritta in modo più semplice per i calcoli

$$\text{Cov(X,Y)}= \left(\frac{1}{n}\sum_{i=1}^{n}x_i\,y_i\right)-\bar{x}\,\bar{y} \ \text{ oppure } \ \left(\sum_{k=1}^{K}f_k x_k y_k\right)-\bar{x}\,\bar{y}$$

ovvero come la differenza fra la media del prodotto dei dati e il prodotto delle medie.

Una covarianza **positiva** indica che per la maggior parte dei dati:

- a valori alti della variabile X corrispondono valori alti della variabile Y
- a valori bassi della variabile X corrispondono valori bassi della variabile Y

Una covarianza **negativa** indica che per la maggior parte dei dati:

- a valori alti della variabile X corrispondono valori bassi della variabile Y
- a valori bassi della variabile X corrispondono valori alti della variabile Y

Una covarianza circa **nulla** indica che non esiste nessun legame di questo genere.

ESEMPIO: Per le variabili Altezza e Peso la covarianza vale 78,55.

Covarianza e trasformazioni lineari.

Abbiamo visto che la covarianza è ottenuta centrando le variabili e quindi non risente di eventuali traslazioni delle variabili. Quindi:

$$\mathrm{Cov}(X + b, Y + d) = \mathrm{Cov}(X,Y).$$

Invece risente, come la varianza, delle dilatazioni/contrazioni. Infatti

$$\mathrm{Cov}(a\,X, c\,Y) = \left(\frac{1}{n}\sum_{i=1}^{n} a\,x_i\ c\,y_i\right) - a\,\bar{x}\ c\,\bar{y} = a\,c\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i\ y_i\right) - \bar{x}\,\bar{y}\right) = a\,c\,\mathrm{Cov}(X,Y)$$

In generale:

$$\mathrm{Cov}(a\,X + b, c\,Y + d) = a\,c\,\mathrm{Cov}(X,Y)$$

L'unità di misura della covarianza fra X e Y (ad esempio espresse una in cm e l'altra in kg) è data dal prodotto delle unità di misura di X e di Y (quindi, cm x kg): quindi risente della scelta dell'unità di misura.

Come si potrebbe definire un indice, che dia le informazioni della covarianza ma non dipenda dalla scelta delle unità di misura di X e Y?

Bisogna trasformare le variabili X e Y operando, oltre che una centratura, anche una standardizzazione, considerando quindi variabili con varianza 1.

Indichiamo ora con $\tilde{X}$ e con $\tilde{Y}$ le variabili standardizzate: $\tilde{X} = \dfrac{X - \bar{x}}{\sigma_X}$ e $\tilde{Y} = \dfrac{Y - \bar{y}}{\sigma_Y}$.

Il **coefficiente di correlazione** $\rho(X,Y)$ è definito come $\mathrm{Cov}(\tilde{X},\tilde{Y})$:

$$\rho(X,Y) = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i\,\tilde{y}_i$$

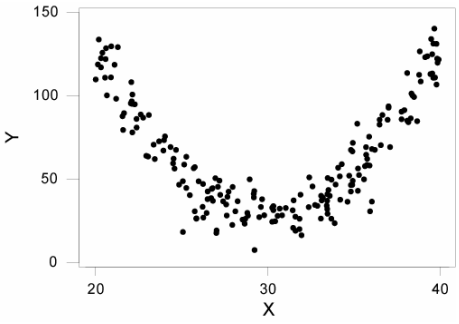
Quindi

$$\rho(X,Y) = \frac{1}{n}\sum_{i=1}^{n} \frac{X - \bar{x}}{\sigma_X}\frac{Y - \bar{y}}{\sigma_Y} = \frac{1}{n\,\sigma_X\,\sigma_Y}\sum_{i=1}^{n}(X - \bar{x})(Y - \bar{y}) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

Il segno della correlazione coincide con quello della covarianza.

L'indice di correlazione è un numero compreso fra –1 e 1. Se è vicino ai valori estremi le due variabili hanno un forte legame lineare. Se è vicino a 0 non esistono legami lineari apprezzabili fra le due variabili.

ATTENZIONE: la covarianza e la correlazione misurano solo il legame lineare fra le variabili; altri tipi di legami non sono individuati. Una covarianza o correlazione circa nulla non significa che non esista nessuna relazione fra le variabili stesse.
Il grafico a fianco mostra un caso di correlazione pressoché nulla, pur in presenza di una relazione quasi quadratica fra le variabili.



Osserviamo infine – come nel caso delle variabili qualitative – che aver individuato un legame lineare non vuol dire aver individuato una relazione di causa/effetto.
Ad esempio se da un'indagine statistica si trova che il numero di figli per famiglia e il consumo di alcool pro capite per famiglia hanno una correlazione positiva abbastanza alta, questo non vuol dire che l'avere una famiglia numerosa induce necessariamente un maggior consumo di alcolici, oppure che un alto consumo di alcolici abbia come conseguenza diretta una famiglia numerosa. In questo caso si può ipotizzare che le cause dell'alto consumo di alcolici e della numerosità dei figli siano le condizioni culturali e economiche delle famiglie, ovvero che esistono altre variabili, magari non rilevate dall'indagine, che influiscono sulle variabili studiate.

Correlazione e trasformazioni lineari.
Abbiamo visto che la correlazione è ottenuta standardizzando le variabili e quindi non risente di eventuali traslazioni e dilatazioni/contrazioni delle variabili, a parte il segno.

$$\rho\,(aX + b, cY + d) = \frac{Cov(aX+b, cY+d)}{\sigma_{aX+b}\ \sigma_{cY+d}} = \frac{ac\,Cov(X,Y)}{|a||c|} = segno(ac)\ \rho(X,Y)$$

Alcune osservazioni:

1. Si ha: $Cov(X,X) = \sigma_X^2$, $Cov(X,Y) = Cov(Y,X)$ e $\rho(X,X) = 1$, $\rho(X,-X) = -1$.

2. Date due (o più) variabili quantitative X_1 e X_2 la **matrice di varianza-covarianza** è quella matrice simmetrica contenente sulla diagonale principale $Var(X_i)$e nel posto (i,j) $Cov(X_i,X_j)$. Nel caso delle variabili Altezza e Peso si ha

	altezza	peso
altezza	86,3896	78,5528
peso	78,5528	115,950

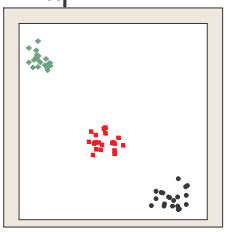
Analogamente la matrice di correlazione è quella matrice simmetrica contenente sulla diagonale principale 1 e nel posto (i,j) $\rho\,(X_i,X_j)$.
Nel caso delle variabili Altezza e Peso si ha

	altezza	peso
altezza	1	0.785
peso	0.785	1

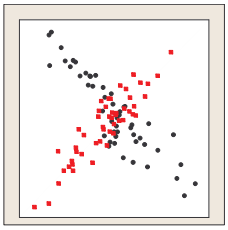
4) Covarianza e correlazione nella popolazione e in sottogruppi

Come abbiamo visto nella scheda precedente per la media e la varianza, quando la popolazione è suddivisa in sottogruppi è interessante confrontare gli indici di covarianza e di correlazione fra due variabili calcolato sull'intera popolazione con quelli calcolati nei sottogruppi.
Ad esempio, se si considerano le variabili Pesi X e Altezze Y, suddividendo la popolazione in maschi e femmine si ha che $\rho_M(X,Y)= 0.604$ e $\rho_F(X,Y)= 0.494$, mentre $\rho_{tot}(X,Y)= 0.785$.
In questo caso i tre valori sono abbastanza simili (anche se il coefficiente di correlazione totale è più alto in quanto le femmine sono mediamente più basse e più leggere dei maschi)
Talvolta, però, si presentano situazioni che possono sembrare "strane", ad esempio:

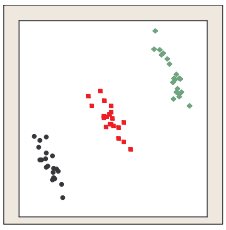
- le correlazioni nelle sottopopolazioni sono piuttosto basse (in valore assoluto) mentre quella nella popolazione totale è alta: è il caso in cui il grafico della distribuzione congiunta presenta nuvole omogenee e uniformi per le sottopopolazioni, ma la totalità dei punti presenta un andamento lineare;



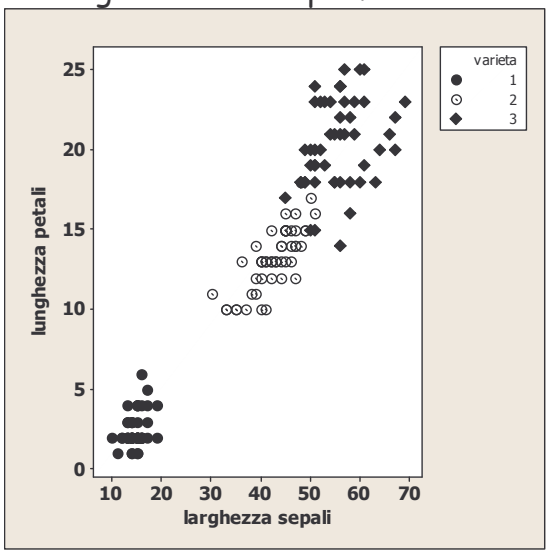
- sono piuttosto alte (in valore assoluto) mentre quello nella popolazione totale è basso: è il caso in cui il grafico della distribuzione congiunta presenta nuvole con andamento nelle sottopolazioni, ma la totalità dei punti si presenta omogenea;



- le correlazioni nelle sottopopolazioni hanno segno negativo, mentre l'indice relativo alla totalità dei dati è positivo (o viceversa); è il caso in cui il grafico della distribuzione congiunta presenta nuvole con andamento nelle sottopopolazioni di segno diverso da quello della totalità dei punti;



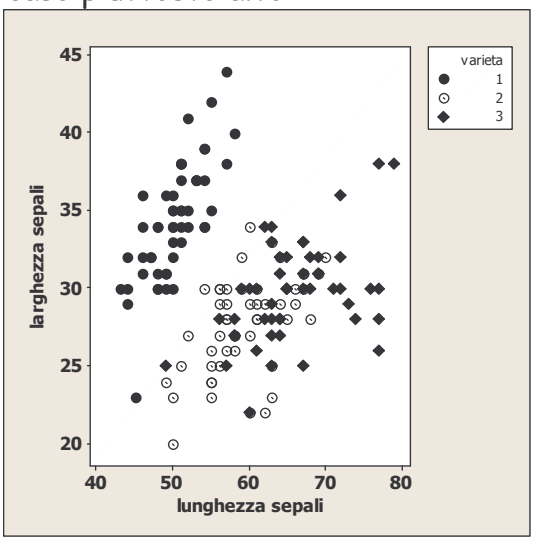
UN ESEMPIO REALE: Consideriamo alcuni dati relativi a tre varietà di Iris; sono misurate la lunghezza e la larghezza dei petali e lunghezza e la larghezza dei sepali.
Nella rappresentazione grafica a fianco sono riportate le distribuzioni congiunte della lunghezza e della larghezza dei petali di tre varietà di Iris.
Si "vede" che la correlazione complessiva fra la lunghezza e la larghezza è positiva e questo dovuto a un "fattore di scala": le tre specie sono di dimensioni diverse: la 1 è piccola, la 2 è media e la 3 è grande.
Le correlazioni fra la lunghezza e la larghezza dei petali per ciascuna varietà sono molto più basse.
Qui sotto vediamo altre due "anomalie".



ρ_{tot} =0.964 ρ_1 =0.326 ρ_2 =0.787 ρ_3 =0.322

Lunghezza e larghezza sepali:

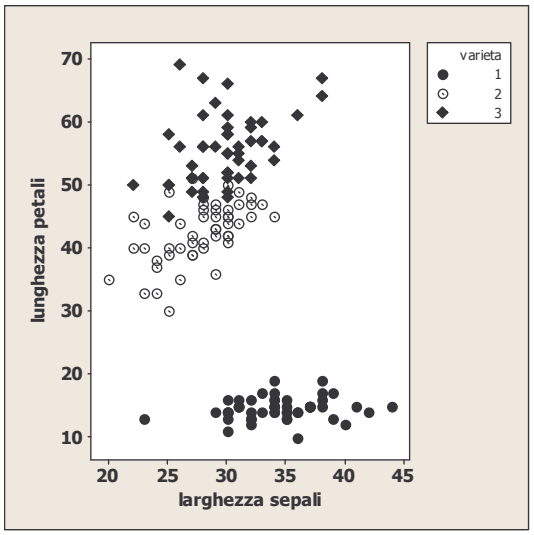
- ρ totale negativo quasi nullo;
- ρ nelle sottopopolazioni positivo e in un caso piuttosto alto



ρ_{tot} = -0.128 ρ_1 =0.748 ρ_2 =0.526 ρ_3 =0.457

Lunghezza petali e larghezza sepali:

- ρ totale negativo basso;
- ρ nelle sottopopolazioni positivo



ρ_{tot} = -0.442 ρ_1 =0.1826 ρ_2 =0.561 ρ_3 =0.401

La relazione fra la covarianza della popolazione e quelle dei sottogruppi è simile a quella che abbiamo già visto per la varianza. Supponniamo per semplicità di avere due sottogruppi A e B con frequenza relativa f_A e f_B, medie $\bar{x}_A$ e $\bar{x}_B$; la relazione è la seguente:

$Cov_{tot}(X,Y) =$

$$f_A\, Cov_A(X,Y) + f_B\, Cov_B(X,Y) + f_A(\bar{x}_A - \bar{x}_{tot})(\bar{y}_A - \bar{y}_{tot}) + f_B(\bar{x}_B - \bar{x}_{tot})(\bar{y}_B - \bar{y}_{tot})$$

Ovvero la covarianza della popolazione totale è la somma (pesata) delle covarianze delle sottopopolazioni più una quantità che può essere considerata una covarianza (pesata) fra le medie dei sottogruppi.

Come abbiamo già detto una correlazione alta non fornisce informazioni su eventuali cause/effetto fra le variabili. Talvolta però queste informazioni sono note a chi sta studiando una situazione reale: c'è una variabile (che indicheremo con X) che produce degli effetti su un'altra variabile (che indicheremo con Y).

In questi casi di correlazione alta e di dipendenza di Y da X, possiamo dire che Y è approssimabile tramite X in questo modo:

Y = aX + b + un errore

Quali coefficienti a e b si devono scegliere affinché la retta che approssima i dati osservati sia la **migliore**, ossia l'errore sia il più piccolo possibile?

Lo vedremo nella prossima scheda.

ESERCIZI

1) A fianco sono riportati i risultati di due caratteristiche quantitative effettuate sulla stessa popolazione.

X	Y
5.6	3.6
1.6	-0.3
2.4	1.8
4.1	3.7
6.9	6.4
3.2	3.7
2.1	2.0
6.4	7.4
2.5	-0.2
6.9	6.0
2.5	2.4
-0.3	-0.6

a. Costruire un diagramma di dispersione che visualizzi la distribuzione della variabile X
b. Calcolare la media di X .
c. Calcolare la varianza di X .
d. Costruire un grafico della funzione di distribuzione cumulata della variabile X .
e. Costruire un box-plot per la variabile X .
f. Sapendo che per la variabile Y si ottiene: $\sum_{i=1}^{12} y_i = 35.9$ e $\sum_{i=1}^{12} y_i^2 = 185.55$, calcolare media e varianza di Y .
g. Costruire un diagramma di dispersione bidimensionale che visualizzi la distribuzione congiunta delle variabili X e Y
h. Calcolare il coefficiente di correlazione delle variabili X e Y

2) I dati riportati nella tabella seguente sono misure di un particolare parametro di funzionalità epatica (SGOT) con il livello di colesterolo HDL nel sangue.

SGOT [x]	9.5	11	13.5	15.5	17.5	19.5	20.5
HDL (mg/dL) [y]	40	41.2	42.3	42.8	43.8	43.6	46.5

$\sum_{i=1}^{7} x_i = 107$ $\sum_{i=1}^{7} x_i^2 = 1740.5$ $\sum_{i=1}^{7} y_i = 300.2$ $\sum_{i=1}^{7} y_i^2 = 12900.2$ $\sum_{i=1}^{7} x_i\, y_i = 4637.6$

a) Calcolare media e varianza delle variabili SGOT e HDL.
b) Costruire un diagramma di dispersione bidimensionale che visualizzi la distribuzione congiunta delle variabili X e Y
c) Calcolare la covarianza fra le variabili SGOT e HDL.
d) Calcolare la correlazione fra le variabili SGOT e HDL.

3) A fianco sono riportati i 13 risultati di una rilevazione quantitativa, indicata con X.
Calcolare la media e la varianza di X.

1.537.861.977.083
1.537.861.977.080
1.537.861.977.087
1.537.861.977.087
1.537.861.977.081
1.537.861.977.125
1.537.861.977.114
1.537.861.977.082
1.537.861.977.090
1.537.861.977.090
1.537.861.977.081
1.537.861.977.080
1.537.861.977.090

4) Per alcuni, l'inizio di questo millennio è il 1 gennaio 2000, per altri è il 1 gennaio 2001. Si effettuano 150 misure di tempo riferite all'inizio del terzo millennio. Dire quale dei seguenti indici statistici riferiti alle sue 150 misure è invariante rispetto alle due scelte per l'origine:

media varianza mediana IQR