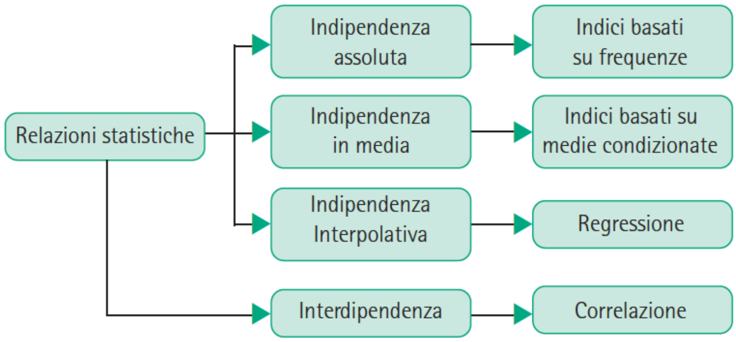
6. RELAZIONI STATISTICHE



1) DISTRIBUZIONI DOPPIE

La **distribuzione doppia** è la distribuzione congiunta di due caratteri X e Y, i quali si suppone possano essere legati da una relazione, ed è esaminata rispetto al contemporaneo verificarsi di una modalità x_i per X e di una modalità y_j per Y.

TABELLE A DOPPIA ENTRATA

Una **tabella a doppia entrata** è una tabella statistica in cui figurano le frequenze assolute o relative riguardanti le diverse combinazioni di modalità o classi di modalità di due caratteri X e Y, desumibili da una distribuzione doppia. Si consideri la tabella seguente:

Tabella 1

$X \backslash Y$	y_1	y_2	....	y_j	....	y_c	Totale
x_1	n_{11}	n_{12}	....	n_{1j}	....	n_{1c}	$n_{1.}$
x_2	n_{21}	n_{22}	....	n_{2j}	....	n_{2c}	$n_{2.}$
:	:	:		:		:	:
x_i	n_{i1}	n_{i2}	....	n_{ij}	....	n_{ic}	$n_{i.}$
:	:	:		:			
x_r	n_{r1}	n_{r2}	....	n_{rj}	....	n_{rc}	$n_{r.}$
Totale	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.c}$	n

La prima riga della tabella è detta *riga madre*, in essa figurano le modalità y_1, y_2, …, y_c del carattere Y; la prima colonna, invece, è detta *colonna madre*, in essa figurano le modalità x_1, x_2, …, x_r del carattere X. Il corpo della tabella è una matrice $r \times c$ (con r righe e c colonne), in essa figurano frequenze del tipo n_{ij} ($i = 1, 2, …, r$; $j = 1, 2, …, c$) in cui il primo indice rappresenta la riga e il secondo la colonna, la frequenza n_{ij} indica il numero di elementi della popolazione n che possiedono le modalità x_i di X e y_j di Y, simultaneamente.
Nell'ultima riga, detta *riga marginale*, figurano le **frequenze marginali** $n_{.j}$, che rappresentano i totali delle c colonne.
Per esempio, in una tabella che rileva le frequenze doppie di peso (X) e altezza (Y), la frequenza marginale $n_{.j}$ indica quanti individui sono alti un certo numero di centimetri, a prescindere dal loro peso. Nell'ultima colonna, detta *colonna marginale*, figurano le **frequenze marginali** $n_{i.}$ rappresentanti i totali delle r righe.
Pertanto, nella medesima tabella peso (X) e altezza (Y), la frequenza $n_{i.}$ indica quanti individui pesano un certo numero di chilogrammi, a prescindere dalla loro altezza.

2) DISTRIBUZIONI CONDIZIONATE E MARGINALI

Da una tabella a doppia entrata si desumono distribuzioni che consentono di evidenziare caratteristiche diverse di una distribuzione doppia di frequenza, esse sono denominate **distribuzioni parziali** e sono: le *distribuzioni condizionate* e le *distribuzioni marginali*.

DISTRIBUZIONE CONDIZIONATA

Una **distribuzione condizionata** è una distribuzione semplice ottenuta associando, in una tabella a doppia entrata, la **riga madre** con **una qualsiasi delle r righe successive**, oppure associando la **colonna madre** con **una qualsiasi delle c colonne successive**.

IL MATERIALE È STATO TRATTO DA "**STATISTICA...IN TASCA**", EDITO DA
che ne ha concesso la pubblicazione su **www.maturansia.it**
EDIZIONI SIMONE ®
Gruppo Editoriale **Simone**
Copyright © 2013 Simone S.r.l.

DISTRIBUZIONE MARGINALE

Una **distribuzione marginale** è una distribuzione semplice ottenuta associando, in una tabella a doppia entrata, la **riga madre** con la **riga marginale**, oppure la **colonna madre** con la **colonna marginale**.

3) INDIPENDENZA ASSOLUTA

È noto, dalla matematica, che una variabile *Y* si dice **indipendente** da una variabile *X* se la prima rimane costante al variare dei valori assunti dalla seconda. In caso contrario si dice che *Y* è **funzione** di *X*.

L'assenza di una qualsiasi relazione tra due caratteri *X* e *Y* desumibili da una distribuzione doppia di frequenza è detta **indipendenza assoluta**. La stessa si evince esaminando le **distribuzioni condizionate** che derivano dalla distribuzione doppia. Esattamente, il carattere *Y* è indipendente dal carattere *X*, se al variare di *X* le distribuzioni condizionate $Y|(X = x_i)$ sono costanti per $i = 1, 2, ..., r$. Analogamente, il carattere *X* è indipendente dal carattere *Y*, se al variare di *Y*, le distribuzioni condizionate $X|(Y = y_j)$ sono costanti per $j = 1, 2, ..., c$. Il concetto di indipendenza è simmetrico per cui, se *Y* è indipendente da *X*, allora anche *X* è indipendente da *Y*.

L'indipendenza tra due caratteri si verifica esaminando le frequenze con cui si presentano le modalità di entrambi i caratteri. Infatti, affinché esista indipendenza tra i due caratteri è necessario che le *frequenze relative delle distribuzioni condizionate siano uguali tra loro e uguali alle frequenze marginali relative*; in simboli:

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$$

ossia:

$$n'_{ij} = \frac{n_{i.}n_{.j}}{n}$$

4) INDICI DI CONNESSIONE

Per accertare l'esistenza di una relazione tra due caratteri *X* e *Y* si confronta la distribuzione doppia osservata con la distribuzione teorica sotto l'ipotesi di indipendenza.

Gli **indici statistici** in grado di evidenziare l'indipendenza di un carattere statistico da un altro sono **basati sulle frequenze** osservate e teoriche, e sono denominati **indici di connessione**. Essi assumono valori tanto più piccoli quanto più esiste indipendenza tra i caratteri investigati.

L'INDICE χ^2

L'indice di **Pearson**, o χ^2, costituisce un criterio di valutazione della differenza esistente tra frequenze teoriche e frequenze osservate; la sua espressione analitica è:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}}$$

È un indice assoluto, ammette valore minimo 0 se $n_{ij} = n'_{ij}$, ossia se esiste indipendenza tra i caratteri, ma non ammette valore massimo in senso matematico, ovvero ammette il massimo relativo che dipende dalla dimensione *n*.

IL MATERIALE È STATO TRATTO DA **"STATISTICA...IN TASCA"**, EDITO DA
che ne ha concesso la pubblicazione su **www.maturansia.it**
EDIZIONI SIMONE®
Gruppo Editoriale **Simone**
Copyright © 2013 Simone S.r.l.

Esempio

Sia data la seguente tabella:

Tabella 2

Carattere X	Carattere Y			Totale
	y_1	y_2	y_3	
x_1	2	5	15	22
x_2	4	14	10	28
x_3	7	6	12	25
Totale	13	25	37	75

a) Determiniamo le frequenze teoriche in caso di indipendenza;
b) determiniamo la tabella di contingenza rappresentata dalle differenze tra frequenze osservate e frequenze teoriche;
c) verifichiamo che la somma dei valori assoluti di tutte le contingenze è minore di $2n$;
d) determiniamo l'indice χ^2.

a) Le frequenze teoriche sono tali che:

$$n'_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

dove:
$n_{i.}$ è la somma delle frequenze della riga i-esima;
$n_{.j}$ è la somma delle frequenze della colonna j-esima.

Per cui, applicando la formula si ottiene la seguente tabella:

Tabella 3 - Frequenze teoriche

Carattere X	Carattere Y			Totale
	y_1	y_2	y_3	
x_1	3,813	7,333	10,853	22
x_2	4,853	9,333	13,813	28
x_3	4,333	8,333	12,333	25
Totale	13	25	37	75

b) Le contingenze sono le differenze tra frequenze osservate e frequenze teoriche; in simboli:

$$c_{ij} = n_{ij} - n'_{ij}$$

IL MATERIALE È STATO TRATTO DA "**STATISTICA...IN TASCA**", EDITO DA
che ne ha concesso la pubblicazione su **www.maturansia.it**
EDIZIONI SIMONE®
Gruppo Editoriale **Simone**
Copyright © 2013 Simone S.r.l.

applicando la formula si ottiene la seguente tabella:

Tabella 4 - Contingenze

Carattere X	**Carattere Y**			**Totale**
	y_1	y_2	y_3	
x_1	–1,813	–2,333	4,147	0
x_2	–0,853	4,667	–3,813	0
x_3	2,667	–2,333	–0,333	0
Totale	0	0	0	0

Dalla tabella di contingenza costruita si evince che la somma algebrica delle contingenze di una riga o di una colonna è nulla.

c) La somma dei valori assoluti di tutte le contingenze è:

$$\sum_{i,j}|c_{ij}| = 1{,}813 + 0{,}853 + 2{,}667 + 2{,}333 + 4{,}667 + 2{,}333 + 4{,}147 + 3{,}813 + 0{,}333 = 22{,}959$$

che è minore di $2 \cdot 75 = 150$.

d) L'indice χ^2 è:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}}$$

per cui, applicando la formula ai dati si ha:

$$\chi^2 = \frac{(-1{,}813)^2}{3{,}813} + \frac{(-2{,}333)^2}{7{,}333} + \frac{(4{,}147)^2}{10{,}853} + \frac{(-0{,}853)^2}{4{,}853} + \frac{(4{,}667)^2}{9{,}333} + \frac{(-3{,}813)^2}{13{,}813} + \frac{(2{,}667)^2}{4{,}333} +$$
$$+ \frac{(-2{,}333)^2}{8{,}333} + \frac{(-0{,}333)^2}{12{,}333} = 9{,}029$$

5) INDIPENDENZA IN MEDIA E RAPPORTO DI CORRELAZIONE

Per distribuzioni di frequenza di due caratteri, di cui almeno uno quantitativo, se invece di considerare le frequenze con cui si presentano la modalità x^1 di un carattere *X* e la modalità y_j di un carattere *Y*, se ne considerano le **medie condizionate**, è possibile misurare l'**indipendenza in media**, detta anche **connessione in media**. Un carattere ***Y*** è **indipendente in media** da un carattere ***X*** se, al variare delle modalità di *X*, le medie condizionate di *Y* risultano costanti e uguali alla media generale del carattere *Y*. Generalmente, le medie delle distribuzioni condizionate $Y \mid (X = x_i)$ sono indicate con la simbologia $\bar{y}_i$, e la loro espressione è la seguente:

$$\bar{y}_i = \sum_{j=1}^{c}\frac{y_j n_{ij}}{n_{i.}} \qquad i = 1, 2, \ldots, r$$

IL MATERIALE È STATO TRATTO DA "**STATISTICA...IN TASCA**", EDITO DA EDIZIONI SIMONE Gruppo Editoriale **Simone**
che ne ha concesso la pubblicazione su **www.maturansia.it**
Copyright © 2013 Simone S.r.l.

Analogamente, un carattere ***X*** è **indipendente in media** da un carattere ***Y*** se, al variare delle modalità di *Y*, le medie condizionate di *X*, indicate, generalmente, con la simbologia $\overline{x}_j$, restano costanti. Se entrambi i caratteri investigati sono quantitativi, ha senso considerare sia l'indipendenza in media di *Y* da *X* sia l'indipendenza in media di *X* da *Y*.
Al contrario dell'indipendenza in distribuzione, l'indipendenza in media non è simmetrica, nel senso che l'indipendenza in media del carattere *X* dal carattere *Y* non implica quella di *Y* da *X*.
Inoltre, mentre l'indipendenza in distribuzione di un carattere *X* da un carattere *Y* ne implica l'indipendenza in media, non è vero il contrario, ossia l'indipendenza in media di un carattere *X* da un carattere *Y* non ne implica l'indipendenza in distribuzione.

Per verificare l'eventuale indipendenza in media di un carattere *Y* da un carattere *X*, si ricorre ad un indice basato sulla **decomposizione della devianza** del carattere *Y*, fornita dalla seguente espressione:

$$D(Y)=\sum_{i=1}^{r}\sum_{j=1}^{c}\left(y_j-\overline{y}\right)^2 n_{ij}$$

in cui $\overline{y}$ è la media generale del carattere *Y*, data da:

$$\overline{y}=\mu_y=\frac{\sum_{j=1}^{c} y_j n_{.j}}{n}$$

DECOMPOSIZIONE DELLA DEVIANZA

La devianza del carattere Y può essere scritta in forma equivalente, infatti, aggiungendo e sottraendo la quantità $\overline{y}_i$, si ha:

$$D(Y)=\sum_{i=1}^{r}\sum_{j=1}^{c}\left(y_j-\overline{y}\right)^2 n_{ij}=\sum_{i=1}^{r}\sum_{j=1}^{c}\left(y_j-\overline{y}_i+\overline{y}_i-\overline{y}\right)^2 n_{ij}$$

Sviluppando il quadrato:

$$D(Y)=\sum_{i=1}^{r}\sum_{j=1}^{c}\left(y_j-\overline{y}_i\right)^2 n_{ij}+2\sum_{i=1}^{r}\sum_{j=1}^{c}\left(y_j-\overline{y}_i\right)\left(\overline{y}_i-\overline{y}\right)n_{ij}+\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\overline{y}_i-\overline{y}\right)^2 n_{ij}$$

Nell'espressione ottenuta, il doppio prodotto può essere scritto nella forma $2\sum_{j=1}^{c}\left(y_j-\overline{y}_i\right)\sum_{i=1}^{r}\left(\overline{y}_i-\overline{y}\right)n_{ij}$, che, comparando la somma degli scarti dalla media aritmetica, si annulla. Inoltre, l'ultimo termine può essere scritto nella forma $\sum_{i=1}^{r}\left(\overline{y}_i-\overline{y}\right)^2 n_{i.}$

Pertanto, la devianza del carattere Y è:

$$D(Y)=\sum_{i=1}^{r}\left(\sum_{j=1}^{c}\left(y_j-\overline{y}_i\right)^2 n_{ij}\right)+\sum_{i=1}^{r}\left(\overline{y}_i-\overline{y}\right)^2 n_{i.}$$

ossia:

$$D(Y)=\sum_{i=1}^{r}\left[D\left(Y|X=x_i\right)\right]+\sum_{i=1}^{r}\left(\overline{y}_i-\overline{y}\right)^2 n_{i.}$$

IL MATERIALE È STATO TRATTO DA "**STATISTICA...IN TASCA**", EDITO DA
che ne ha concesso la pubblicazione su **www.maturansia.it**
EDIZIONI SIMONE
Gruppo Editoriale **Simone**
Copyright © 2013 Simone S.r.l.

L'espressione appena ottenuta è detta **devianza totale** del carattere Y, in essa il primo termine a secondo membro è la **devianza entro i gruppi**, mentre il secondo termine è la **devianza tra i gruppi**; in simboli:

$$D(Y) = D(W) + D(B)$$

dove W deriva dal termine inglese *within* (entro) e B da *between* (tra).

Con il crescere della variabilità tra i gruppi cresce la variabilità delle medie tra i gruppi, misurata da $D(B)$, dunque, cresce la dipendenza in media di Y da X. Un indicatore della relazione tra le variabili che tenga conto di questo aspetto è il **rapporto di correlazione di Pearson**, la cui espressione analitica è la seguente:

$$\eta^2_{Y|x} = \frac{D(B)}{D(Y)} = \frac{\sum_{i=1}^{r} \left(\bar{y}_i - \bar{y}\right)^2 n_{i.}}{\sum_{j=1}^{c} \left(y_j - \bar{y}\right)^2 n_{.j}}$$

Esso assume:

— valore minimo pari a 0, nel caso in cui la differenza tra le medie di Y al variare delle modalità di X siano tutte costanti, e dunque la devianza entro i gruppi al numeratore del rapporto sia nulla; in tal caso si ha indipendenza in media di Y da X;
— valore massimo pari a 1, nel caso in cui l'intera variabilità di Y sia attribuibile alla variabilità tra i gruppi, ed è tale che ad ogni valore di X corrisponde un solo valore di Y; in tal caso si ha massima dipendenza di Y da X.

6) ANALISI DELLA DIPENDENZA E REGRESSIONE

Nell'analisi statistica, la **regressione** è volta alla ricerca di un modello atto a descrivere la relazione esistente tra una **variabile dipendente** e una o più **variabili indipendenti** o **esplicative**; si sceglie come indipendente la variabile che sia *logicamente antecedente* rispetto all'altra.
In un modello di regressione, le variabili esplicative (dette anche **regressori**) spiegano, prevedono, simulano, controllano la variabile dipendente.
Il termine *regressione* fu coniato da Galton che, nel misurare la relazione tra statura dei padri e quella dei figli, osservò una regressione dei valori delle altezze dei figli verso la media.
Per effettuare una regressione si fa riferimento a modelli teorici di vario tipo: lineare, parabolico, esponenziale, logaritmico etc.

Dall'osservazione del diagramma a scatter di una distribuzione doppia si può evincere la linearità dell'associazione tra i caratteri se i punti empirici sono distribuiti attorno ad una retta. Quest'ultima è denominata **retta di regressione** e la sua espressione analitica, considerando come variabile dipendente Y e come variabile indipendente X, è la seguente:

$$\hat{Y} = \beta_0 + \beta_{yx} X$$

in cui i coefficienti β_0 e β_{yx}, detti **coefficienti di regressione**, sono incogniti.

A seconda del valore assunto dal coefficiente β_{yx} si desume l'associazione tra X e Y; infatti:

— se $\beta_{yx} > 0$ al crescere di X anche Y cresce;

— se $\beta_{yx} < 0$ al crescere di X, la variabile Y decresce;

— se $\beta_{yx} = 0$ non esiste associazione lineare tra X e Y.

Se si assume, invece, Y come variabile indipendente e X come variabile dipendente, la retta di regressione è:

$$\hat{X} = \beta'_0 + \beta_{xy} Y$$

in cui i coefficienti β'_0 e β_{xy} sono incogniti.

IL MATERIALE È STATO TRATTO DA **"STATISTICA...IN TASCA"**, EDITO DA
che ne ha concesso la pubblicazione su **www.maturansia.it**
EDIZIONI SIMONE®
Gruppo Editoriale **Simone**
Copyright © 2013 Simone S.r.l.

I coefficienti incogniti di entrambe le rette di regressione sono determinabili ricorrendo al metodo dei minimi quadrati. Le loro espressioni sono riportate di seguito.

REGRESSIONE DI Y SU X

$$\beta_{yx} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad \beta_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

REGRESSIONE DI X SU Y

$$\beta_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} \qquad \beta_0' = \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}$$

7) INDICE DI DETERMINAZIONE LINEARE

Una volta individuata la retta di regressione si dispone di diversi indici per valutare il grado di affidabilità del modello. Tra questi assume particolare rilievo l'**indice di determinazione lineare**. Trattasi di un indice della **bontà di accostamento** della retta di regressione alla nuvola di punti osservati; misura, infatti, la parte di variabilità totale spiegata dalla retta di regressione.

DECOMPOSIZIONE DELLA DEVIANZA

Per introdurre l'indice si considera che la devianza del carattere Y, avente media $\bar{y}$, è:

$$D(Y) = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Il primo termine dell'uguaglianza rappresenta la **devianza totale** $D(Y)$ delle y_i intorno alla loro media.

La quantità $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ costituisce quella parte della devianza totale che risulta spiegata dalla regressione lineare tra le variabili X e Y e rappresenta la **devianza di regressione** $D(\hat{Y})$.

Infine, il terzo termine rappresenta la parte di devianza totale dei valori di Y non spiegati dalla regressione lineare ed è chiamato **devianza residua** $D(E)$.

L'indice di determinazione lineare è dato dal rapporto tra devianza di regressione e devianza totale:

$$R^2 = \frac{D(\hat{Y})}{D(Y)} = 1 - \frac{D(E)}{D(Y)}$$

Esso è in grado di fornire la forza della relazione rappresentata dalla retta di regressione. Infatti, quando le differenze $(y_i - \hat{y}_i)$ sono piccole, la devianza residua, *D*(*E*), è piccola per cui è grande la devianza spiegata, di conseguenza R^2 è grande.

IL MATERIALE È STATO TRATTO DA "**STATISTICA...IN TASCA**", EDITO DA che ne ha concesso la pubblicazione su **www.maturansia.it**

EDIZIONI SIMONE® Gruppo Editoriale **Simone** Copyright © 2013 Simone S.r.l.

Il coefficiente di determinazione può assumere valori compresi tra 0 e 1. Quando vale 0 la retta di regressione e la retta $y = \bar{y}$ coincidono, di conseguenza, la variabilità dei valori di *Y* non risulta affatto spiegata dalla regressione. Quando vale 1 tutti i punti sperimentali giacciono sulla retta di regressione, per cui la regressione spiega una gran parte della variabilità dei valori di *Y* e quindi il modello di regressione è appropriato.

8) CORRELAZIONE TRA CARATTERI

Nell'analisi statistica di una distribuzione doppia di caratteri entrambi quantitativi, una trattazione a parte è dedicata allo studio di una particolare relazione: l'**interdipendenza**.

Per misurare la correlazione tra due variabili è necessario fare riferimento alla **covarianza**, la cui espressione analitica è la seguente:

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

dove $\bar{x}$ e $\bar{y}$ sono il valore medio, rispettivamente, della variabile *X* e della variabile *Y*.

Si parla di:

— **concordanza** quando scarti positivi o negativi della variabile *X* tendono ad associarsi, rispettivamente, a scarti positivi o negativi della variabile *Y*, allora i loro prodotti saranno positivi, dunque la covarianza risulterà positiva;

— **discordanza**, invece, quando scarti positivi della variabile *X* tendono ad associarsi a scarti negativi della variabile *Y* o viceversa, allora i loro prodotti saranno negativi e la covarianza risulterà negativa.

COEFFICIENTE DI CORRELAZIONE LINEARE

La covarianza costituisce il numeratore di un'importante misura del grado di dipendenza lineare tra due variabili: il **coefficiente di correlazione lineare di Bravais - Pearson**:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

in cui, con i simboli σ_x e σ_y, si sono indicati lo scarto quadratico medio, rispettivamente, della variabile *X* e della variabile *Y*.

Esso può essere espresso, in maniera equivalente, in questo modo:

$$r = \frac{n\sum_{i=1}^{n}x_i y_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2}}$$

Copyright © 2013 Simone S.r.l.

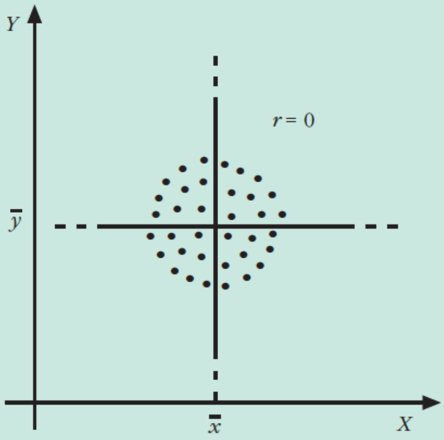
Il coefficiente di correlazione presenta le seguenti caratteristiche:

— è un numero puro;

— assume valori compresi tra –1 e +1, in simboli:

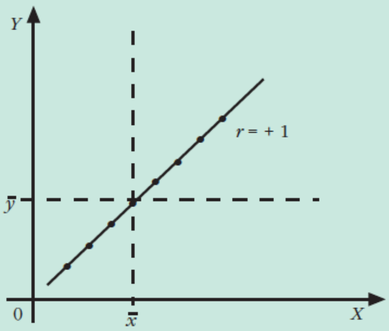
$$-1 \leq r \leq 1$$

— se $r = 0$ non vi è relazione di tipo lineare tra i due caratteri:

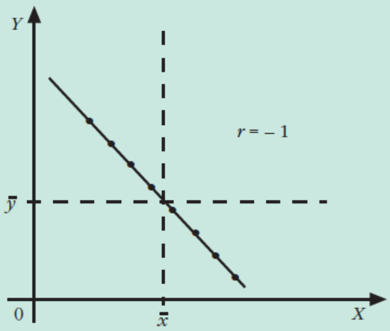


(Fig. 1)

— se $r = \pm 1$ esiste, tra i due caratteri, un legame lineare perfetto di tipo concorde ($r = +1$), o discorde ($r = -1$);



(Fig. 2)



(Fig. 3)

— se le due variabili *X* e *Y* sono indipendenti tra loro allora $r = 0$, anche se è possibile che esso sia nullo anche quando le due variabili *X* e *Y* sono legate da una relazione diversa da quella lineare. Ciò conferma che il coefficiente di correlazione non è, in generale, un indice di dipendenza, ma di concordanza; è un indice di dipendenza solo nel caso in cui quest'ultima sia di tipo lineare;

— talvolta *r* può assumere un valore elevato pur non sussistendo alcuna relazione tra le variabili, ma per l'influenza esercitata sulle stesse da uno o più fattori comuni, in tal caso si dice che esiste **correlazione spuria**.

IL MATERIALE È STATO TRATTO DA "**STATISTICA...IN TASCA**", EDITO DA
che ne ha concesso la pubblicazione su **www.maturansia.it**
EDIZIONI SIMONE®
Gruppo Editoriale **Simone**
Copyright © 2013 Simone S.r.l.

Si dimostra che il coefficiente di correlazione lineare di Bravais-Pearson è pari alla radice quadrata dell'indice di determinazione lineare; in simboli:

$$r = \pm\sqrt{R^2}$$

Esempio

La tabella seguente riporta il peso (in kg) e la corrispondente altezza (in cm) di 9 lanciatori di giavellotto:

Tabella 5

Peso	93	79	86	94	84	83	80	70	75
Altezza	184	168	180	184	185	188	180	177	178

Misuriamo la relazione esistente tra peso e altezza.

La variabile *Peso* sarà indicata con la lettera *X*, mentre la variabile *Altezza* sarà indicata con la lettera *Y*.

Per misurare la relazione tra peso e statura si fa ricorso al coefficiente di correlazione lineare di Bravais-Pearson che è un numero puro. Trattasi di una misura simmetrica di concordanza o discordanza, particolarmente adatta. Il calcolo del coefficiente di correlazione lineare viene effettuato ricorrendo al seguente schema:

Schema 3

	x_i	y_i	x^2_i	y^2_i	x_iy_i
	93	184	8.649	33.856	17.112
	79	168	6.241	28.224	13.272
	86	180	7.396	32.400	15.480
	94	184	8.836	33.856	17.296
	84	185	7.056	34.225	15.540
	83	188	6.889	35.344	15.604
	80	180	6.400	32.400	14.400
	70	177	4.900	31.329	12.390
	75	178	5.625	31.684	13.350
Totale	744	1.624	61.992	293.318	134.444

Pertanto:

$$r = \frac{9 \cdot 134.444 - 744 \cdot 1.624}{\sqrt{9 \cdot 61.992 - (744)^2}\sqrt{9 \cdot 293.318 - (1.624)^2}} = 0,5265838$$

Il valore assunto dal coefficiente fa rilevare una correlazione diretta tra le due variabili.

IL MATERIALE È STATO TRATTO DA "**STATISTICA...IN TASCA**", EDITO DA
che ne ha concesso la pubblicazione su **www.maturansia.it**

EDIZIONI SIMONE®
Gruppo Editoriale **Simone**
Copyright © 2013 Simone S.r.l.