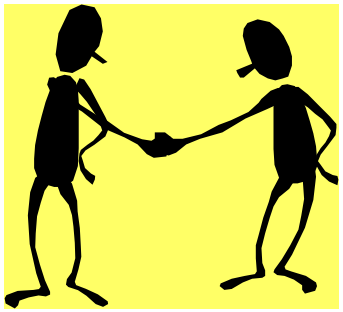
IL LEGAME TRA DUE VARIABILI

I METODI
DELLA CORRELAZIONE

CORRELAZIONE

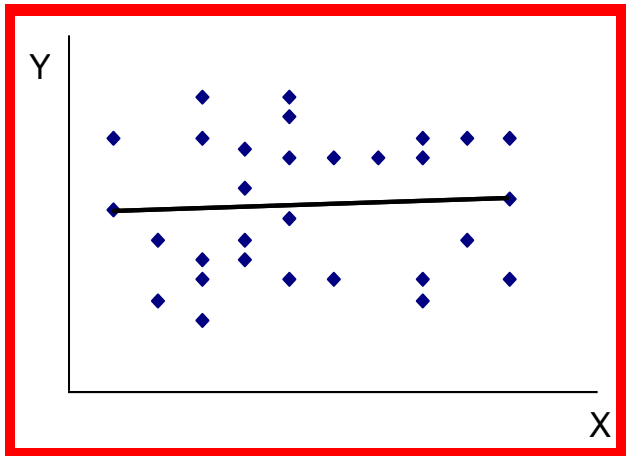
Legame - Associazione - Accordo – Relazione tra variabili

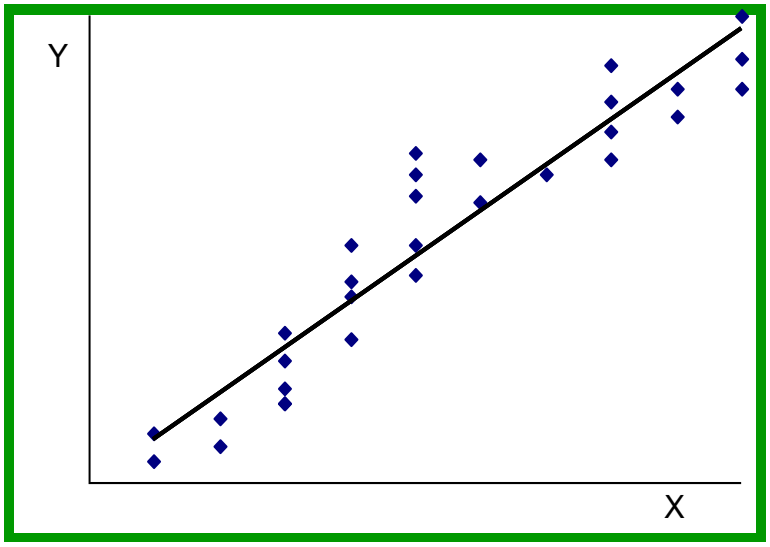


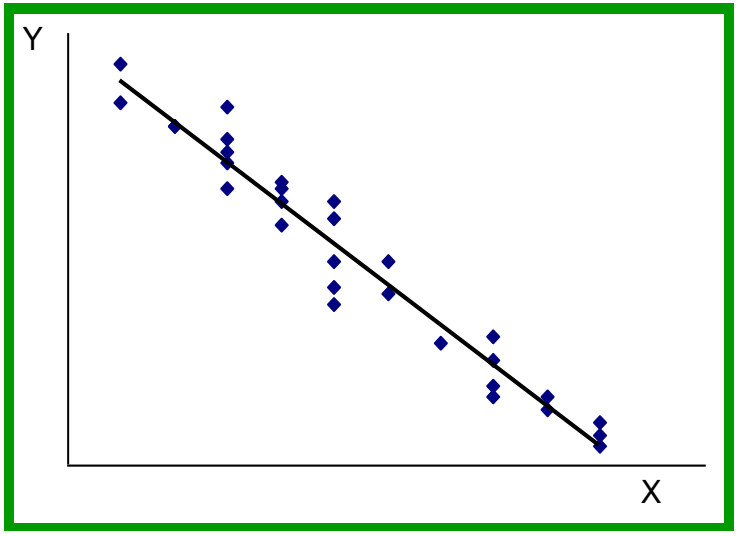
☞ valutare il grado di reciproca influenza tra due variabili;

☞ valutare il grado di associazione di due variabili che sono influenzate entrambe da una causa esterna.

La relazione esistente tra due variabili può essere analizzata graficamente ponendo i dati osservati in un diagramma a dispersione :







IL COEFFICIENTE DI CORRELAZIONE

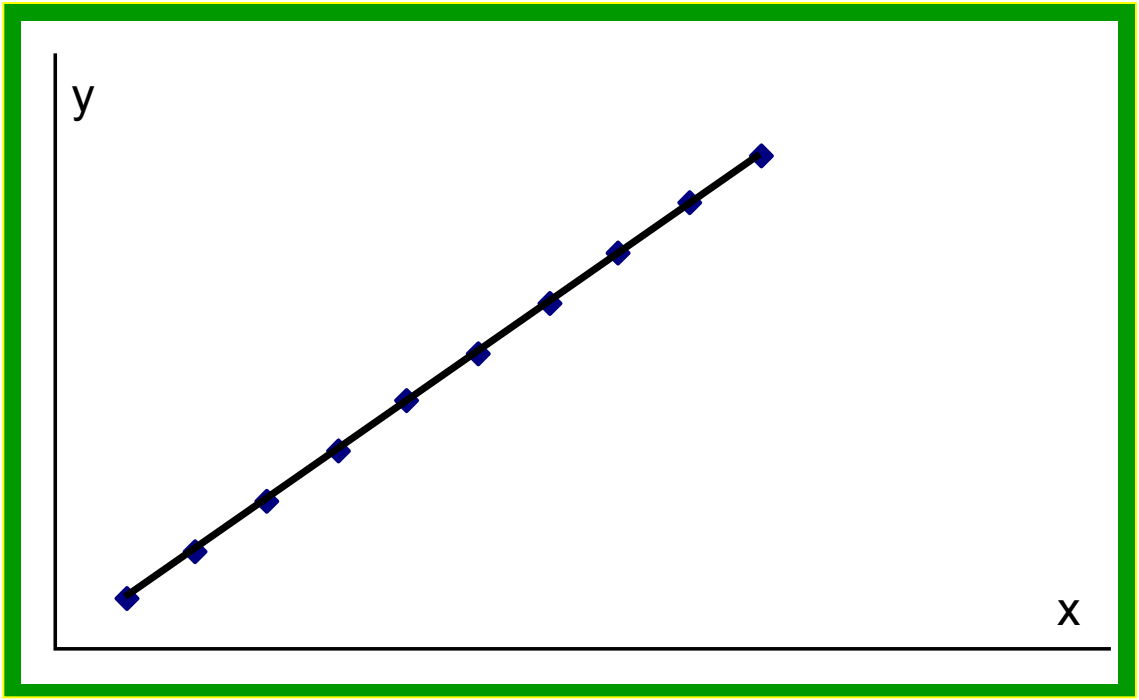
La **misura** della forza della **associazione** tra le due variabili è data dal **coefficiente di correlazione di Pearson**:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

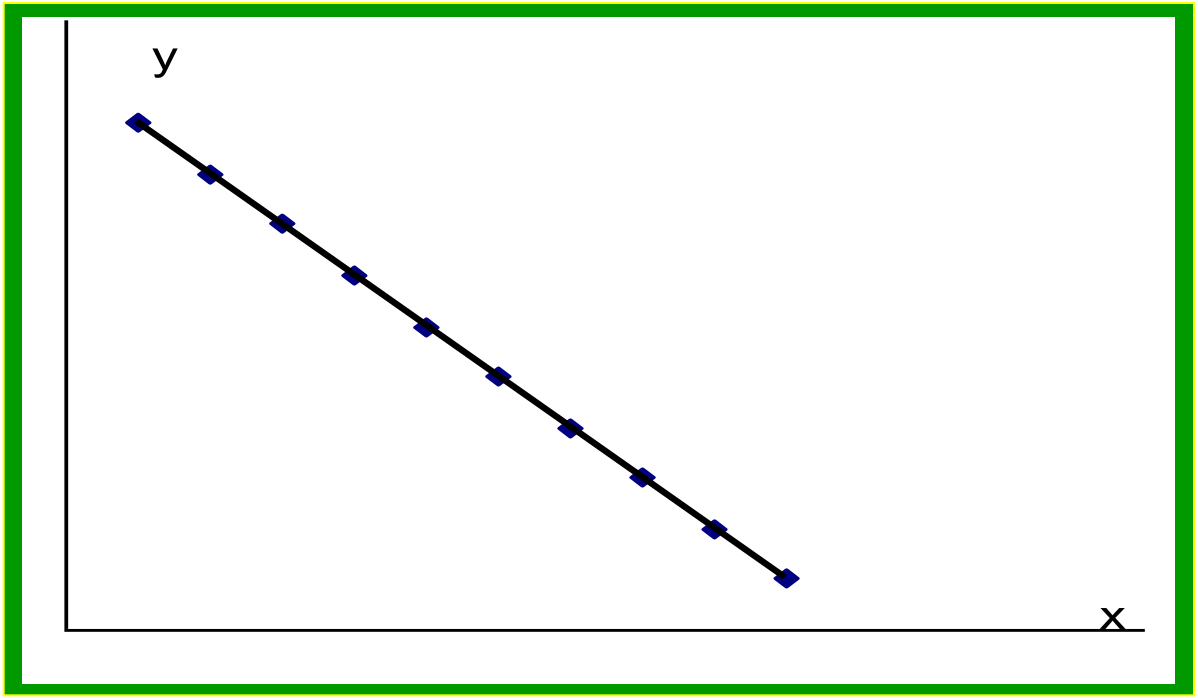
Con **$-1 \leq r \leq +1$**

La correlazione studia **l'associazione lineare** esistente tra due variabili.

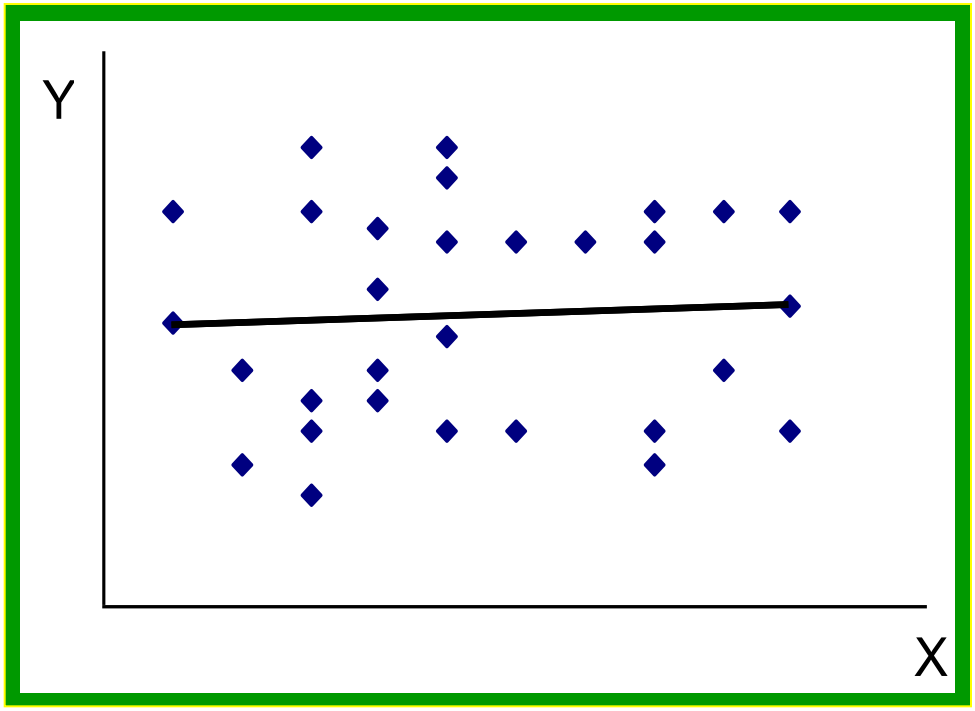
☞ **r = +1** : massima correlazione con **<u>proporzionalità diretta</u>** tra le due variabili, al crescere della X cresce anche la Y



☞ **r = -1** : massima correlazione con **proporzionalità inversa** tra le due variabili, al crescere della X decresce la Y (e viceversa).



☞ **r = 0** : vuol dire che **non esiste correlazione** tra le due variabili.



☞ Se si può assumere che le due variabili seguano **una distribuzione normale bivariata** allora la **non correlazione significa anche indipendenza**

☞ Se **non** si può assumere la **distribuzione normale bivariata allora si deve pensare ad un'altra forma di legame** (parabola, esponenziale, sigmoide, …).

IL TEST DI VERIFICA DI IPOTESI

Il valore di r è comunque una stima campionaria del coefficiente di correlazione ρ della popolazione.
E' possibile eseguire un test di verifica relativa alla significatività del nostro r campionario.
Tale test verifica anche l'indipendenza delle due variabili se si assume che queste seguano una distribuzione normale bivariata.

ASSUNZIONI

- La distribuzione di X e Y congiunte è una distribuzione normale bivariata.

LA DISTRIBUZIONE NORMALE BIVARIATA

La funzione che descrive la distribuzione normale bivariata è caratterizzata da 5 parametri:

1. la media di X
2. la deviazione standard di X
3. la media di Y
4. la deviazione standard di Y
5. il coefficiente ρ

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right\}$$

Se $\rho = 0$ allora si ha:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

Applicando la proprietà degli esponenziali secondo la quale l'esponenziale di una somma è uguale al prodotto degli esponenziali: **exp (a+b) = exp (a) × exp (b)** posso riscrivere la formula:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]$$

Ricordando che $\mathbf{2\pi = \sqrt{2\pi} \times \sqrt{2\pi}}$
e raggruppando opportunamente avrò: $f(x,y) = f(x) \times f(y)$

Conclusione: **solo se si può assumere la distribuzione normale bivariata il risultato r = 0** significa **indipendenza** delle variabili.

IPOTESI

$$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$$

STATISTICA TEST

$$T = r\sqrt{\frac{n-2}{1-r^2}}$$

DISTRIBUZIONE DELLA STATISTICA TEST

La statistica test ha una distribuzione t-Student con n-2 gradi di libertà.

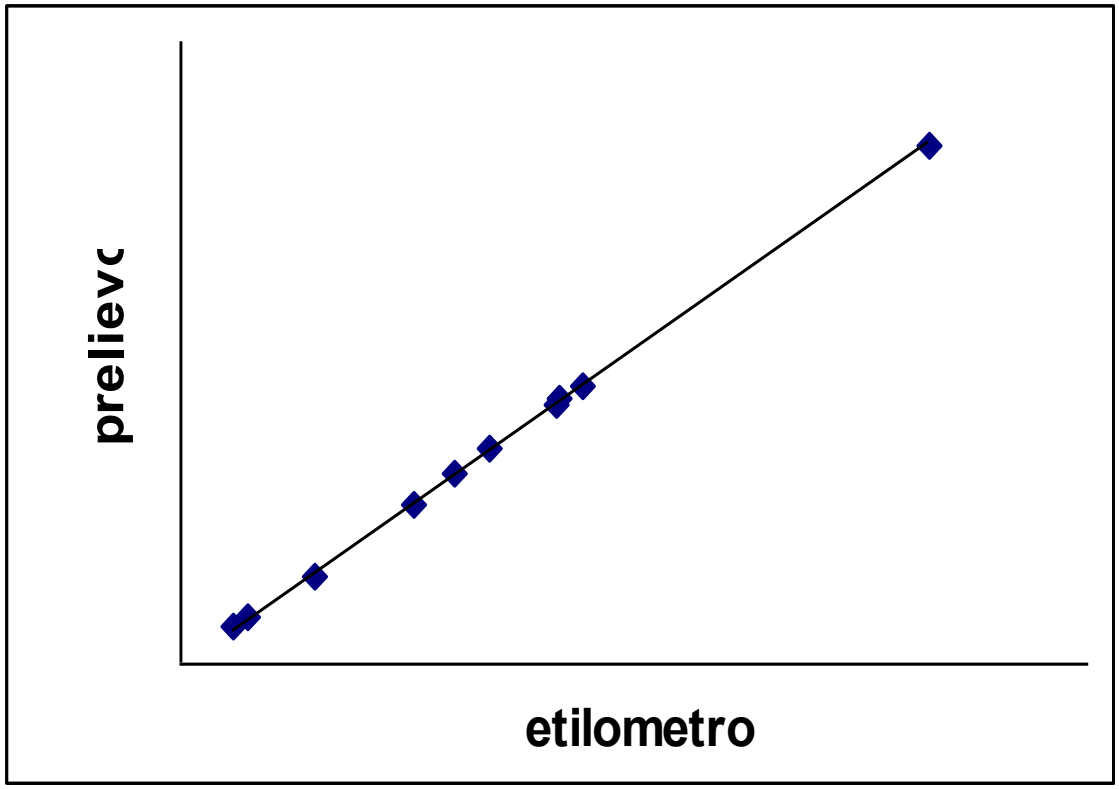
REGOLA DI DECISIONE

Conoscendo la distribuzione della statistica test, i suoi gradi di libertà e il livello di significatività ($\alpha = 0,05$), individuerò il valore tabulato con cui confrontare il valore calcolato.
Se $|t_{calc}| > |t_{tab}|$ allora rifiuto H_0.

Si voglia studiare il legame esistente tra i livelli di alcoolemia in mg % ml stimata con l'etilometro e con prelievo di sangue venoso.

Etilometro (X)	Prelievo (Y)
44	44
265	269
250	256
153	154
88	83
180	185
35	36
494	502
249	249
204	208

Proviamo a porre i dati del nostro esempio in un diagramma a dispersione :



Per effettuare più facilmente i calcoli conviene modificare la formula come segue:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} =$$

$$= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]\left[\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right]}} =$$

Etilometro (X)	**Prelievo (Y)**	**XY**	**X^2**	**Y^2**
44	44	1936	1936	1936
265	269	71285	70225	72361
250	256	64000	62500	65536
153	154	23562	23409	23716
88	83	7304	7744	6889
180	185	33300	32400	34225
35	36	1260	1225	1296
494	502	247988	244036	252004
249	249	62001	62001	62001
204	208	42432	41616	43264
1962	**1986**	**555068**	**547092**	**563228**

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} =$$

$$= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]\left[\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right]}} =$$

$$= \frac{555068 - \left(\frac{1986 \times 1962}{10}\right)}{\sqrt{\left(547092 - \frac{1962^2}{10}\right)\left(563228 - \frac{1986^2}{10}\right)}} =$$

$$= \frac{1654148}{165444,48} = 0,99$$

$$T = r\sqrt{\frac{n-2}{1-r^2}} = 0{,}99\sqrt{\frac{8}{1-0{,}99^2}} = 19{,}84$$

$t_{tab\ \alpha=0,05;gl=8} = 2{,}306$

$t_{calc} > t_{tab}$ ➡ rifiuto H_0

Decisione del ricercatore:

i valori di alcoolemia determinati con il prelievo e con l'etilometro sono correlati, quindi misurano lo stesso indicatore pur con metodi e su substrati diversi.

IL COEFFICIENTE DI CORRELAZIONE DI SPEARMAN

Nel caso in cui non sia possibile fare assunzioni sulla distribuzione delle variabili il coefficiente di correlazione da usare è :

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

Con $-1 \leq r_s \leq +1$

dove d_i sono le differenze dei ranghi attribuiti ai valori delle due variabili.

L'ipotesi nulla è di non correlazione delle due variabili.

La decisione verrà presa confrontando il valore di r_s calcolato con il valore di r_s tabulato.

Il valore tabulato si cerca sulle tavole di Spearman in corrispondenza del livello di significatività del test ($\alpha = 0,05$) e del numero di coppie di osservazioni delle due variabili

Se $|r_s \text{ calc}| > |r_s \text{ tab}|$ rifiuterò l'ipotesi nulla.

Si ordinano in maniera crescente i valori della variabile Y

Si ordinano in maniera crescente i valori della variabile X

Si assegnano i ranghi ai valori della variabile Y

Si assegnano i ranghi ai valori della variabile X

A valori uguali si assegneranno ranghi pari alla media dei ranghi che i valori avrebbero avuto se fossero stati diversi

Si determinano le differenze d_i tra i ranghi assegnati alla variabile X e i ranghi assegnati alla variabile Y e si calcola il coefficiente di correlazione di Spearman r_s

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

Si individua il valore tabulato per α fissato (0,05) e il numero di coppie di osservazioni

Si confronta r_s calcolato con il valore tabulato: se risulta maggiore si rifiuta l'ipotesi nulla di indipendenza

I dati del problema con i calcoli da effettuare sono riportati nella seguente tabella

N sig. fumate (X)	Peso neonato (Y)	Ranghi X	Ranghi Y	d_i	d_i^2
1	3864	1	10	9	81
2	3318	2	5	3	9
3	3727	3	9	6	36
4	3636	4	8	4	16
5	2955	5	4	-1	1
6	3364	6	6	0	0
7	3591	7	7	0	0
8	2818	8	3	-5	25
9	2545	9	1	-8	64
10	2773	10	2	-8	64
					296

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} =$$

$$= 1 - \frac{6(296)}{10(10^2 - 1)} = 1 - \frac{1776}{990} = -0{,}794$$

Nel nostro caso r_s tab = 0,648 < 0,794

Rifiuto l'ipotesi nulla, c'è correlazione tra le due variabili.

VERIFICA DI IPOTESI SUL LEGAME TRA VARIABILI QUALITATIVE

DATI

Si vuole verificare l'esistenza di un legame tra il gruppo sanguigno e la gravità di una certa patologia. Si dispone del numero di individui che presentano contemporaneamente la patologia ad certo grado di gravità e un dato gruppo sanguigno.

	Gruppo sanguigno				
Patologia	A	B	AB	0	Totale
Assente	543	211	90	476	1320
Media	44	22	8	31	105
Grave	28	9	7	31	75
Totale	615	242	105	538	1500

La generalizzazione della tabella precedente è:

	1° criterio						
2° criterio	1	2	…	j	…	c	Tot.
1	O_{11}	O_{12}	…	O_{1j}	…	O_{1c}	$n_{1.}$
2	O_{21}	O_{22}	…	O_{2j}	…	O_{2c}	$n_{2.}$
…	…	…	…	…	…	…	…
i	O_{i1}	O_{i2}	…	O_{ij}	…	O_{ic}	$n_{i.}$
…	…	…	…	…	…	…	…
r	O_{r1}	O_{r2}	…	O_{rj}	…	O_{rc}	$n_{r.}$
Tot.	$n._1$	$n._2$	…	$n._j$	…	$n._c$	N

ASSUNZIONI

Le variabili di cui disponiamo sono qualitative.

Se consideriamo **una sola cella** la presenza contemporanea delle due caratteristiche è il "successo", sugli N casi possibili: si può assumere una **distribuzione binomiale**.

I dati in tabella nel loro insieme seguono una distribuzione multinomiale.

IPOTESI

$$\begin{cases} H_0: & p_{ij} = p_i \times p_j \\ H_1: & p_{ij} \neq p_i \times p_j \end{cases}$$

$p_{ij} = O_{ij} / N$

$p_i = n_i. / N$

$p_j = n._j / N$

Se le due variabili sono **indipendenti** la probabilità di avere la caratteristica 1 **e** la caratteristica 2 sarà data dal **prodotto** delle probabilità (legge del prodotto).

I VALORI ATTESI

Vera l'ipotesi nulla e posta l'assunzione di distribuzione binomiale in ciascuna cella allora posso calcolare il valore atteso E_{ij} ("media") per ciascuna cella:

$$\mathbf{E_{ij} = N\, p_{ij} = N\, p_i\, p_j = N\, (n._j / N)\, (n_i. / N) = (n._j\, n_i.) / N}$$

Si può quindi costruire una tabella di valori attesi:

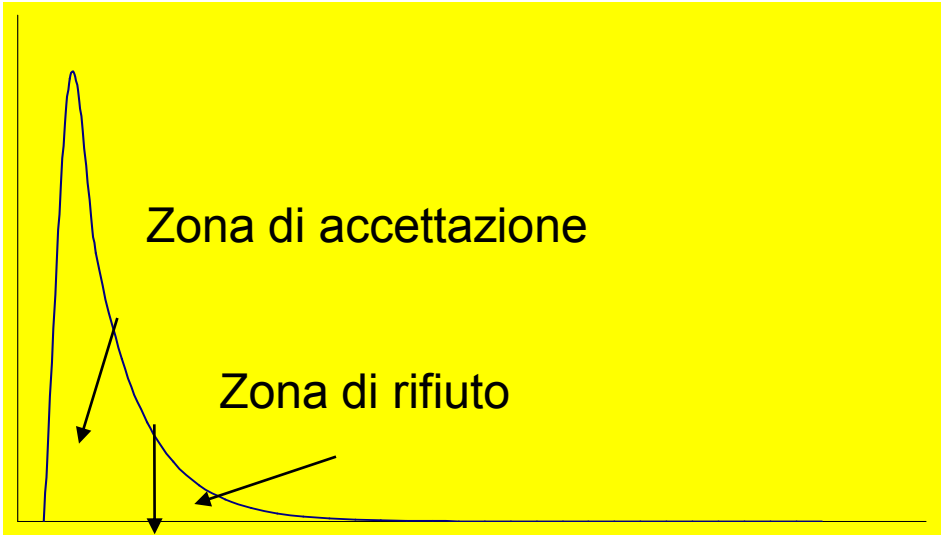
		1° criterio					
2° criterio	1	2	…	j	…	c	Tot.
1	E_{11}	E_{12}	…	E_{1j}	…	E_{1c}	$n_1.$
2	E_{21}	E_{22}	…	E_{2j}	…	E_{2c}	$n_2.$
…	…	…	…	…	…	…	…
i	E_{i1}	E_{i2}	…	E_{ij}	…	E_{ic}	$n_i.$
…	…	…	…	…	…	…	…
r	E_{r1}	E_{r1}	…	E_{rj}	…	E_{rc}	$n_r.$
Tot.	$n._1$	$n._2$	…	$n._j$	…	$n._c$	N

STATISTICA TEST

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

DISTRIBUZIONE DELLA STATISTICA TEST

La distribuzione della statistica test è una X^2 ed è caratterizzata dai gradi di libertà.



REGOLA DI DECISIONE

Fissato α accettabilmente piccolo (0,05), troverò sulle tavole X^2 un valore in corrispondenza di α prescelto e dei gradi di libertà della statistica.
Se il valore calcolato è maggiore del valore tabulato rifiuterò l'ipotesi nulla, se invece il valore calcolato è minore del tabulato accetterò l'ipotesi nulla.

I GRADI DI LIBERTA'

In questo caso i gradi di libertà sono:

g.l. = (r-1) (c-1)

dove r = numero delle righe

c = numero delle colonne

$\Sigma p_{.j} = \Sigma n_{.j} / N = 1$

$\Sigma p_{i.} = \Sigma n_{i.} / N = 1$

fissato N potrò cambiare "liberamente" $n_{i.}$, totali di riga, meno 1 che mi deve garantire la somma delle probabilità di riga ($\Sigma p_i = 1$).

fissato N potrò cambiare "liberamente" $n_{.j}$, totali di colonna, meno 1 che mi deve garantire la somma delle probabilità di colonna ($\Sigma p_j = 1$).

Tabella valori osservati

	Gruppo sanguigno				
Patologia	A	B	AB	0	Totale
Assente	543	211	90	476	1320
Media	44	22	8	31	105
Grave	28	9	7	31	75
Totale	615	242	105	538	1500

Tabella valori attesi

	Gruppo sanguigno				
Patol.	A	B	AB	0	Totale
Assente	541,2	212,96	92,40	473,44	1320
Media	43,05	16,94	7,35	37,66	105
Grave	30,75	12,10	5,25	26,90	75
Totale	615	242	105	538	1500

CALCOLO DELLA STATISTICA TEST

$$X^2 = \frac{(543-541{,}2)^2}{541{,}2} + \frac{(211-212{,}96)^2}{212{,}96} + \frac{(90-92{,}40)^2}{92{,}40} + \\ + ... + \frac{(31-26{,}90)^2}{26{,}90} = 5{,}12$$

$X^2_{\alpha=0{,}05,\ gl\ 6} = 12{,}592$

DECISIONE STATISTICA

5,12<12,592 accetto l'ipotesi nulla, le due variabili sono indipendenti

DECISIONE DEL RICERCATORE

Non c'è una evidenza di associazione tra un gruppo sanguigno e l'essere affetto dalla malattia in esame.

IPOTESI

$$\begin{cases} H_0: & p_{ij}=p_1 x\ p_2 \\ H_1: & p_{ij}=p_1 x\ p_2 \end{cases}$$

STATISTICA TEST

$$X^2 = \frac{N(ad-bc)^2}{(a+b)\times(c+d)\times(a+c)\times(b+d)}$$

$$X^2 = \frac{N(|ad-bc|-0{,}5N)^2}{(a+b)\times(c+d)\times(a+c)\times(b+d)}$$

Nella seconda formula c'è la correzione per la continuità di Yates

TEST PER IL CONFRONTO DI PIU' PROPORZIONI

Nel caso di Tabelle di contingenza 2xk dove k rappresentano i gruppi da porre a confronto e si hanno due possibili risposte, il precedente test del $\chi 2$ può essere usato per verificare:

$$H_0: \quad p_1 = p_2 = \ldots\ldots p_k = p$$
$$H_1: \quad p_r \neq p_s$$

La statistica test

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

ha una distribuzione $\chi 2$ con k-1 gradi di libertà