NOTE DALLE LEZIONI
DI
STATISTICA MEDICA
ED ESERCIZI

IL LEGAME TRA DUE VARIABILI

I METODI
DELLA CORRELAZIONE

IL PROBLEMA

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

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.

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

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

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

Per risolvere il problema si consiglia di costruire una tabella che contenga le quantità

Etilometro (X)	**Prelievo (Y)**	**XY**	**X^2**	**Y^2**
44	44	…….	…….	…….
265	269	…….	…….	…….
…….	…….	…….	…….	…….

ADESSO EFFETTUATI I CALCOLI

E POI PROSEGUITE

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**

Quindi

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

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

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

$$= \frac{165414,8}{165444,48} = 0,99$$

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

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

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.

CALCOLATE LA STATISTICA TEST E POI VERIFICATE IL RISULTATO

Nel caso del nostro esempio

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

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

$t_{calc} > t_{tab}$ ➔ 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 PROBLEMA

Supponiamo di voler studiare se esiste un legame tra **il peso alla nascita** e **il numero di sigarette** fumate dalla mamma in gravidanza

N sig. fumate (X)	Peso neonato (Y)
1	3864
2	3318
3	3727
4	3636
5	2955
6	3364
7	3591
8	2818
9	2545
10	2773

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

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

con $-1 \leq r_s \leq +1$
dove d_i sono le differenze dei ranghi attribuiti ai valori delle due variabili.

PRIMA DI ANDARE AVANTI ATTRIBUITE I RANGHI CALCOLATE LE DIFFERENZE AL QUADRATO E QUINDI LA SOMMA DELLE DIFFERENZE

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

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.

Effettuando i calcoli per il problema in esame si ha:

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

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

Poiché r_s tabulato = 0,648 < r_s calcolato = -0,794

Rifiuto l'ipotesi nulla, e concludo che c'è correlazione tra il peso alla nascita e il numero di sigarette fumate dalla madre durante la gravidanza. Inoltre poiché il valore del coefficiente è negativo posso affermare che all'aumentare del numero di sigarette decresce il peso del bambino.