

Correlazione e Regressione

- Analisi della dipendenza
- Analisi della correlazione
- Regressione

Analisi della Dipendenza

- Concetto di "associazione"

Si dice che tra due variabili esiste "associazione" quando la distribuzione di una è in qualche modo collegabile a quella dell'altra.

In altre parole, quando in corrispondenza della distribuzione dei valori di una, quelli dell'altra si dispongono in modo prevedibile.

Molte variabili sono in relazione tra loro, presentano cioè un'associazione:

- *peso e altezza,*
- *il peso di un bambino e la sua età,*
- *l'età di una persona e la sua pressione sanguigna,*
- *il tempo di esposizione a radiazioni e mutazione genetica,*
- *condizioni climatiche ed inquinamento,*
- *colore degli occhi e colore dei capelli, ecc.*

Dagli esempi visti possiamo quindi riscontrare che può esistere associazione sia tra 2 variabili quantitative che tra una var. quantitativa e una qualitativa, oppure tra 2 variabili qualitative.

Il concetto di associazione ci porta poi naturalmente a pensare che, se due fenomeni sono funzionalmente collegati tra loro, nel senso che i valori dell'uno tendono ad aumentare o diminuire in relazione ai valori assunti dall'altro, esiste una relazione di "causa-effetto".

Vi sono casi in cui questo è fuor di dubbio, come *p. es.* il calo ponderale in seguito ad uno scarso apporto alimentare, ma in linea di massima non è sempre vero.

E può essere pericoloso interpretare una "associazione" statistica come rapporto di causa-effetto.

Ipotizziamo, ad esempio, che venga riscontrata un'associazione tra la quantità di gelato consumato ed il numero delle persone morte per annegamento.

Una interpretazione superficiale e frettolosa ci potrebbe portare a pensare che il consumo di gelato sia una "causa" importante di annegamento.

Ma questo ovviamente va contro il più elementare buon senso: anche una elevata "associazione" non sta a significare necessariamente causalità.

Nell'esempio considerato si può ipotizzare che annegamento e consumo di gelato siano entrambi legati ad altri fattori quali temperatura dell'acqua e corporea, congestione post-prandiale ecc.

La Correlazione

Il significato di *Correlazione*

Le relazioni o le associazioni tra variabili come quelle riportate nella precedente elencazione a titolo esemplificativo (*il peso di un bambino e la sua età, l'età di una persona e la sua pressione sanguigna, ecc.*) si dicono *correlazioni*.

Le correlazioni sono misurate per variabili su scala ordinale o continua.

Correlazione positiva

Quando al crescere del valore di una variabile corrisponde l'incremento dell'altra, la correlazione viene definita ***positiva*** o ***diretta***.

L'altezza ed il peso di un bambino sono di solito positivamente correlate.

Correlazione negativa

Quando al crescere di una variabile corrisponde il decremento dell'altra, parliamo di correlazione ***negativa*** o ***inversa***.

Ad esempio il tempo di reazione è correlato inversamente con il numero di unità alcoliche assunte.

Il fatto che le variabili siano associate o correlate non significa necessariamente che una di esse determini l'altra.

L'altezza ed il peso possono essere correlate in una popolazione, ma non è certo che una variabile influenzi l'altra: entrambe sono sicuramente in relazione con fattori genetici e ambientali (*abitudini alimentari ecc.*)

Nell'uso comune la parola "correlazione" è usata per descrivere qualsiasi tipo di relazione tra oggetti ed eventi.

In statistica la correlazione ha un significato preciso; si riferisce ad una relazione quantitativa tra due variabili misurate su scala ordinale o continua.

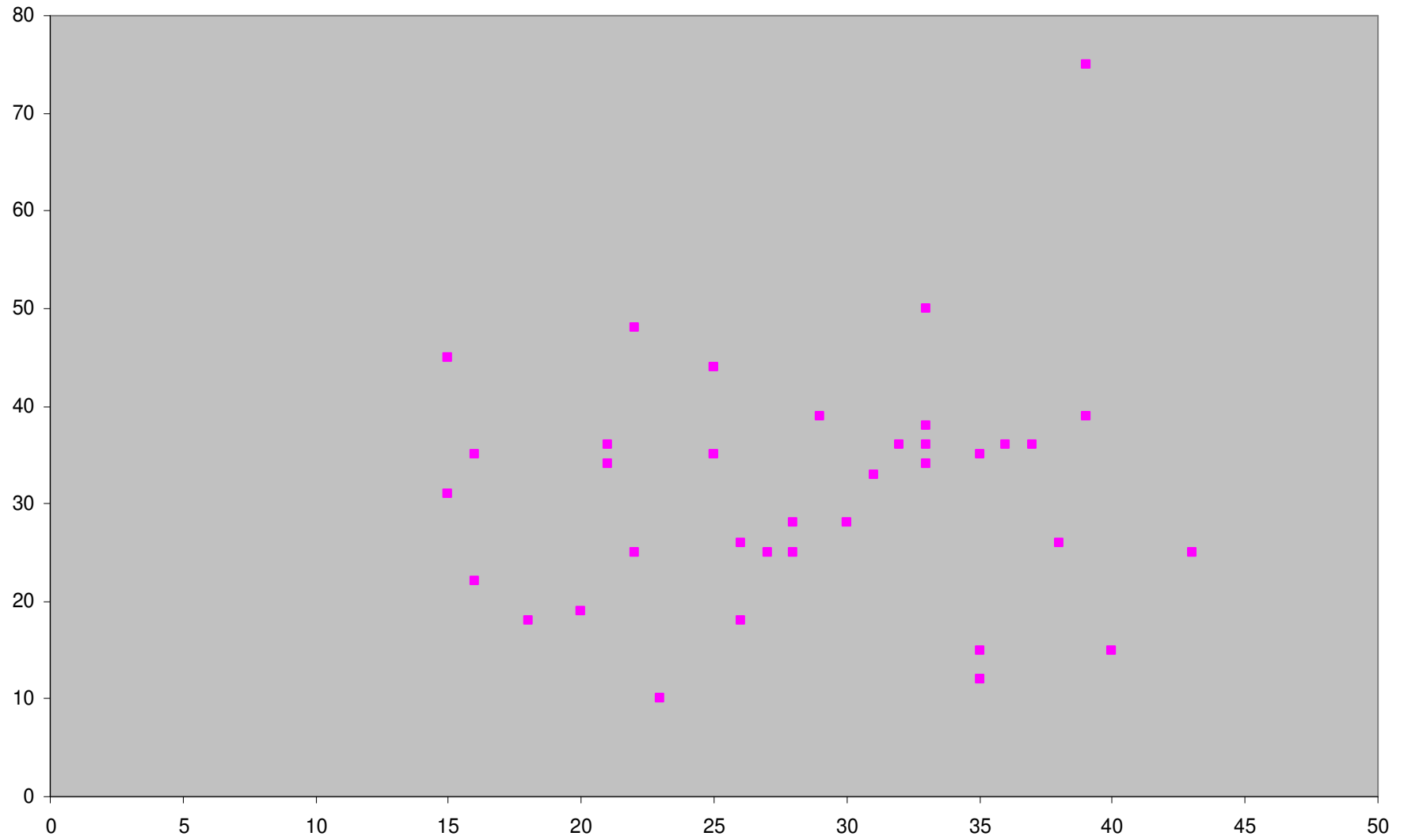
Diagramma di dispersione a due dimensioni

il Diagramma a punti

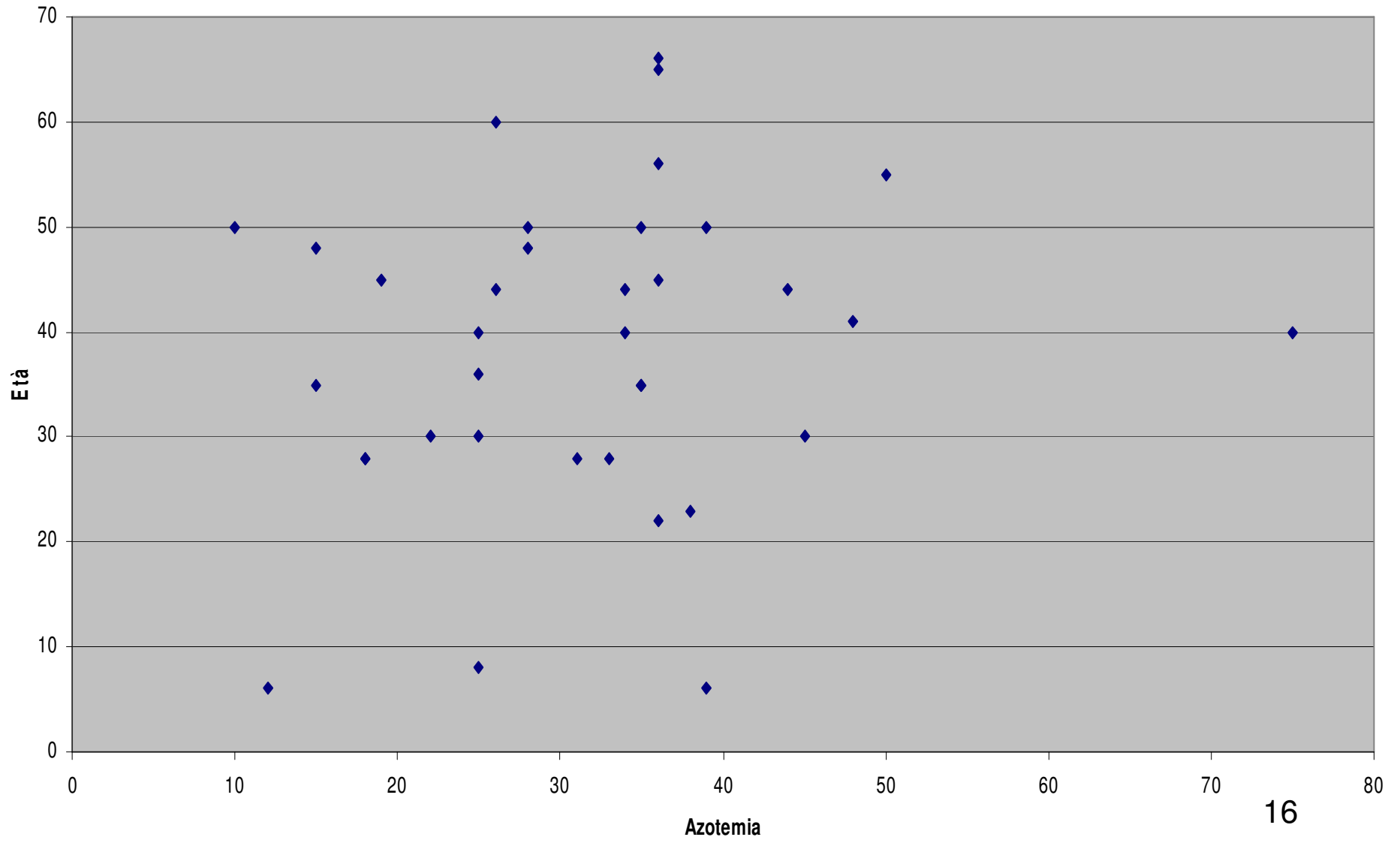
Le osservazioni bivariate – di due variabili – misurate su scale ordinali o intervallari possono essere rappresentate su un diagramma di dispersione (*scatter*).

Il *diagramma di dispersione* fornisce indicazioni sulla correlazione.

Azotemia - Got



Azotemia - Età



Spesso il diagramma è sufficiente a mostrare chiaramente la presenza di una correlazione – positiva o negativa – e la “forza” della correlazione stessa, oppure l’assenza di una qualsiasi correlazione.

In altri casi non è così facile stabilire dall’esame del diagramma se vi sia o meno correlazione.

La valutazione soggettiva di un *diagramma di dispersione* può essere sostituita allora da una tecnica statistica in grado di fornire indicazioni più precise e oggettive.

Analisi della Correlazione

Esaminiamo adesso le relazioni che possono esistere tra variabili misurate su scala intervallare o ordinale.

Una tecnica statistica spesso utilizzata per *misurare* una simile associazione è nota come *analisi di correlazione*.

Analisi della Correlazione

Considerate due variabili casuali X e Y
la correlazione tra le variabili nella
popolazione originaria è indicata dalla
lettera greca ρ (rho).

La correlazione quantizza la forza della
relazione lineare tra i risultati x e y

Analisi della Correlazione

Essa può essere intesa come la media del prodotto delle deviate normali standardizzate di X ed Y

In particolare:

$$\rho = \text{Media} \left[\frac{(X - \mu_x)}{\sigma_x} \frac{(Y - \mu_y)}{\sigma_y} \right]$$

Analisi della Correlazione

Lo *stimatore* della correlazione della popolazione (calcolato quindi sui dati campionari) è noto come

Coefficiente di correlazione di Pearson

o più semplicemente

coefficiente di correlazione

È la statistica che fornisce dunque un indice di accordo su come le due variabili siano in relazione tra loro.

Il coefficiente di correlazione è un numero adimensionale e non ha unità di misura.

Tale indice (individuato dal simbolo r) è dato dal rapporto tra la covarianza tra X e Y , e la radice quadrata del prodotto tra la varianza di X e la varianza di Y . In simboli:

$$r = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}}$$

Coefficiente di correlazione r

Il valore numerico del *coefficiente* " r " è compreso tra due valori estremi $+ 1$ e $- 1$:

- $+ 1$ perfetta correlazione *positiva*
- $- 1$ perfetta correlazione *negativa*

Nei dati intervallari la perfetta correlazione esiste quando tutti i punti di un *diagramma di dispersione* si trovano su una linea retta.

Coefficiente di correlazione r

Per i dati ordinali (punteggi), si ottiene una *relazione monotona* perfetta quando tutti i valori si trovano in ordine crescente o decrescente, non necessariamente su una linea retta.

Le correlazioni perfette - o prossime alla perfezione - (positive o negative) sono un privilegio quasi esclusivo dei fisici! Non esistono nelle scienze mediche né, di norma, in quelle sociali

Coefficiente di correlazione r

Un coefficiente di correlazione pari o vicino a 0, indica dunque mancanza di correlazione.

I coefficienti di correlazione possono essere calcolati attraverso metodi parametrici o non parametrici.

Un coefficiente parametrico è il coefficiente di correlazione lineare di Pearson.

È utilizzato per le osservazioni in scala continua (normalmente 'misure').

Coefficiente di correlazione r

Il coefficiente di correlazione lineare proposto da Bravais e Pearson misura sia l'intensità, sia il verso della dipendenza fra due caratteri che costituiscono la distribuzione statistica doppia.

Esso valuta fino a che punto una delle due distribuzioni può essere considerata una trasformata lineare dell'altra.

Coefficiente di correlazione r

la covarianza è uguale alla media dei prodotti di X e Y meno il prodotto delle medie e poiché la varianza è pari alla media dei quadrati meno il quadrato della media, possiamo riscrivere l'espressione precedente nella seguente

$$r = \frac{M_1(XY) - M_1(X)M_1(Y)}{\sqrt{M_1(X^2) - [M_1(X)]^2} \cdot \sqrt{M_1(Y^2) - [M_1(Y)]^2}}$$

Coefficiente di correlazione r

Il coefficiente di correlazione appartiene alla classe degli indici relativi e il suo campo di variazione è compreso tra:

-1 e $+1$ estremi inclusi.

La caratteristica di maggior interesse del coefficiente di correlazione risiede nel fatto che esprime, oltre che l'intensità della correlazione esistente tra i caratteri X e Y , anche il verso di tale correlazione, individuato dal segno dell'indice r

Coefficiente di correlazione r

- Il coefficiente di correlazione lineare è dunque una misura della interdipendenza fra due caratteri a due livelli:
- *l'ordine di grandezza numerica* misura quanto la X dipende linearmente da Y e ne subisce l'effetto;
- *il segno* distingue la concordanza (al crescere di Y cresce anche la X) dalla discordanza (al crescere di Y la X decresce).

Coefficiente di correlazione r

Passiamo al calcolo del coefficiente di correlazione:
La formula precedentemente vista possiamo anche scriverla

$$r = \frac{\textit{codevianza}(x, y)}{\sqrt{\textit{devianza}(x) \cdot \textit{devianza}(y)}}$$

Coefficiente di correlazione r

- La formula precedente equivale a:

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

e , come è noto, possiamo anche scriverla

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{\left(n \sum x^2 - (\sum x)^2 \right) \cdot \left(n \sum y^2 - (\sum y)^2 \right)}}$$

Coefficiente di correlazione r

La *codevianza* (somma dei prodotti degli scarti di ciascuna coppia di variabili dalla rispettiva media) mostra una evidente analogia con la devianza, considerando il prodotto degli scarti di due variabili al posto di una presi due volte (al quadrato);

Per tale motivo, mentre la devianza è sempre positiva, la codevianza può essere negativa.

La variabilità congiunta non può comunque essere maggiore alla somma della variabilità delle 2 variabili.

Coefficiente di correlazione r

Per comprendere meglio il il significato del concetto di correlazione, passiamo ad esaminare i dati relativi ad un campione di 15 pazienti nefropatici.

I risultati ottenuti misurando cinque variabili di interesse fisiopatologico sono riportati di seguito.

Coefficiente di correlazione r

Paz.	Azoto Ureico (mg/dl)	Pressione Diastolica (mm _{Hg})	Creatinina (mg/dl)	Acido Urico (mg/dl)	Colesterolo (mg/dl)	Azoto Ureico (mg/dl)
1	30	90	1,5	7,5	190	30
2	32	95	1,6	7,8	195	32
3	28	85	1,7	7,9	185	28
4	27	85	1,4	7,0	255	27
5	26	75	1,3	9,4	275	26
6	25	80	0,8	7,5	300	25
7	24	85	0,9	9,0	285	24
8	40	95	1,5	7,7	195	40
9	26	85	1,8	6,9	185	26
10	23	80	1,9	8,5	180	23
11	20	70	2,0	9,0	170	20
12	30	95	1,2	7,0	280	30
13	40	95	1,6	8,5	200	40
14	36	90	1,5	8,6	190	36
15	40	105	2,1	9,0	205	40

Coefficiente di correlazione r

Supponiamo di essere interessati inizialmente ad una coppia di variabili, ad es. la *pressione diastolica* e l'*azoto ureico*

Vediamo come procedere:

- Prima di eseguire l'analisi, possiamo disegnare il diagramma di dispersione
- [vai a "Diagrammi di dispersione"](#)
- *Il diagramma mostra chiaramente che la pressione diastolica tende ad aumentare al crescere dei valori di azoto ureico*
- Verifichiamo ora attraverso il calcolo del *coefficiente di correlazione r*

Coefficiente di correlazione r

Pz	Azoto Ureico (mg/dl)	Pressione Diastolica mm Hg				
	X	Y	x^2	y^2	xy	
1	30	90	900	8.100	2.700	
2	32	95	1.024	9.025	3.040	
3	28	85	784	7.225	2.380	
4	27	85	729	7.225	2.295	
5	26	75	676	5.625	1.950	
6	25	80	625	6.400	2.000	
7	24	85	576	7.225	2.040	
8	40	95	1.600	9.025	3.800	
9	26	85	676	7.225	2.210	
10	23	80	529	6.400	1.840	
11	20	70	400	4.900	1.400	
12	30	95	900	9.025	2.850	
13	40	95	1.600	9.025	3.800	
14	36	90	1.296	8.100	3.240	
15	40	105	1.600	11.025	4.200	

447	1.310	13.915	115.550	39.745
Σx	Σy	Σx^2	Σy^2	Σxy
585.570				
$\Sigma x \Sigma y$				

$$\begin{aligned}
 r &= \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2) \cdot (n \sum y^2 - (\sum y)^2)}} = \frac{15 \cdot 39.745 - 585.570}{\sqrt{(15 \cdot 13.915 - (447)^2) \cdot (15 \cdot 115.550 - (1.310)^2)}} = \\
 &= \frac{596.575 - 585.570}{\sqrt{(208.725 - 199.809) \cdot (1.733.250 - 1.716.100)}} = \frac{10.605}{\sqrt{152.909.400}} = \frac{10.605}{12.365,65} = 0,86
 \end{aligned}$$

*Coefficiente di correlazione r_s per **ranghi** di Spearman*

Nei casi in cui non è possibile applicare le rigide condizioni del coefficiente di correlazione "r" di Pearson, o nei casi in cui si presentino dei valori "anomali" (che come è noto influenzano notevolmente le tecniche parametriche), si può utilizzare il coefficiente di correlazione per "**ranghi**" di Spearman. Tale coefficiente utilizza infatti i ranghi delle osservazioni secondo le variabili x e y invece che i dati grezzi.

Coefficiente di correlazione r_s per ranghi di Spearman

L'approccio è quello di assegnare separatamente dei ranghi alle osservazioni x ed y e poi calcolare il coefficiente di correlazione secondo la seguente formula:

$$r_s = 1 - \frac{6 \sum_i d_i^2}{n^3 - n}$$

Dove:

- n è il numero di coppie di valori x_i e y_i
- d è la differenza tra i ranghi di x_i e y_i
- 6 è una costante peculiare di tale formula

Coefficiente di correlazione r_s per ranghi di Spearman

Percentuale di nascite assistite	Ranghi di X	Tasso di mortalità delle madri	Ranghi di Y		
X		Y		d	d²
5	1	582	11	-10,0	100,0
24	2	450	10	-8,0	64,0
27	3	195	8	-5,0	25,0
29	4	284	9	-5,0	25,0
40	5	762	12	-7,0	49,0
57	6	90	6	0,0	0,0
70	7	25	4	3,0	9,0
82	8,5	61	5	3,5	12,3
82	8,5	124	7	1,5	2,3
87	10	12	3	7,0	49,0
96	11	11	2	9,0	81,0
99	12	5	1	11,0	121,0

$\sum d^2 = 537,5$

$$r_s = 1 - \frac{6 \sum d_i^2}{n^3 - n} = 1 - \frac{6 \cdot 537,5}{12^3 - 12} = 1 - \frac{3.225,0}{1.716,0} = 1 - 1,88 = -0,88$$

Coefficiente di correlazione r_s per ranghi di Spearman

Come altre tecniche non parametriche, il coefficiente di correlazione dei ranghi di Spearman presenta vantaggi e svantaggi.

È molto meno (o affatto) sensibile alle osservazioni "atipiche" del coefficiente di correlazione di Pearson. Inoltre può essere usato quando una o entrambe le variabili sono ordinali.

Poiché si basa su ranghi e non su osservazioni reali, però, il metodo non parametrico NON utilizza tutte le "informazioni" note di una distribuzione.

Coefficiente di correlazione r la significatività

Ai fini di una corretta interpretazione delle relazioni esistenti tra i dati campionari non è sufficiente fermarsi al calcolo del coefficiente "r".

Occorre infatti sempre verificare se la correlazione così misurata sia espressione di una reale associazione tra le variabili o il frutto della variabilità casuale.

Coefficiente di correlazione r la significatività

Pertanto è necessario verificare l'ipotesi zero di nessuna correlazione tra le due variabili nella popolazione:

$$H_0: \rho = 0$$

Anche per il coefficiente "r" esistono delle tavole che, in rapporto ai gradi di libertà, forniscono il valore critico al desiderato livello di significatività.

I gradi di libertà per la correlazione sono pari a:

$$\mathbf{g.l. = numero di coppie - 2 = n - 2}$$

Valori critici del coefficiente di correlazione "r"

<i>g.d.l.</i>	<i>$\alpha = 0,05$</i>	<i>$\alpha = 0,01$</i>	<i>g.d.l.</i>	<i>$\alpha = 0,05$</i>	<i>$\alpha = 0,01$</i>
1	0.99692	0.999877	25	0.3809	0.4869
2	0.95000	0.990000	30	0.3494	0.4487
3	0.8783	0.95873	35	0.3246	0.4182
4	0.8114	0.91720	40	0.3044	0.3932
5	0.7545	0.8745	45	0.2875	0.3721
6	0.7067	0.8343	50	0.2732	0.3541
7	0.6664	0.7977	60	0.2500	0.3248
8	0.6319	0.7646	70	0.2319	0.3017
9	0.6021	0.7348	80	0.2172	0.2830
10	0.5760	0.7079	90	0.2050	0.2673
11	0.5529	0.6835	100	0.1946	0.2540
12	0.5324	0.6614	110	0.1857	0.2425
13	0.5139	0.6411	120	0.1779	0.2324
14	0.4973	0.6226	130	0.1710	0.2235
15	0.4821	0.6055	140	0.1648	0.2155
16	0.4683	0.5897	150	0.1593	0.2083
17	0.4555	0.5751	190	0.1417	0.1855
18	0.4438	0.5614	200	0.1381	0.1809
19	0.4329	0.5487			
20	0.4227	0.5368			
25	0.3809	0.4869			

Coefficiente di correlazione r la significatività

Pertanto riprendendo l'esempio precedente il valore di $r = 0,86$ riscontrato tra l'Azoto Ureico e Pressione va verificato con il valore critico di $0,6411$ (per un livello di significatività pari a $\alpha=0,01$) con 13 ($n-2=15-2$) gradi di libertà.

Poiché $0,86 > 0,6411$ possiamo rifiutare l'ipotesi nulla di nessuna correlazione delle variabili nella popolazione e affermare che esiste una associazione fra i dati (con una probabilità di errore pari all' 1%)

Coefficiente di correlazione r la significatività

In mancanza della tabella dei valori critici per il coefficiente "r" è possibile fare ricorso alla distribuzione della "t di Student" applicando il seguente algoritmo:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} \qquad t_s = r_s \cdot \sqrt{\frac{n-2}{1-r_s^2}}$$

che permette di definire per ogni valore di r il corrispondente livello di probabilità utilizzando le tavole della t con $n-2$ gradi di libertà.

Coefficiente di correlazione r la significatività

Considerando sempre l'esempio precedente:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} = 0,86 \cdot \sqrt{\frac{15-2}{1-(0,86)^2}} = 6,076$$

che supera il valore critico della t pari a 3,012 e pertanto (come d'altronde era prevedibile) anche in questo caso possiamo rifiutare l'ipotesi nulla.

Distribuzione "t"

d.f.	$P=0.1$	0.05	0.02	0.01	0.002	0.001
1	6.314	12.706	31.821	63.657	318.31	636.62
2	2.920	4.303	6.965	9.925	22.327	31.598
3	2.353	3.182	4.541	5.841	10.214	12.924
4	2.132	2.776	3.747	4.604	7.173	8.610
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.895	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318
13	1.771	2.160	2.650	3.012	3.852	4.221
14	1.761	2.145	2.624	2.977	3.787	4.140
15	1.753	2.131	2.602	2.947	3.733	4.073
16	1.746	2.120	2.583	2.921	3.686	4.015
17	1.740	2.110	2.567	2.898	3.646	3.965
18	1.734	2.101	2.552	2.878	3.610	3.922
19	1.729	2.093	2.539	2.861	3.579	3.883
20	1.725	2.086	2.528	2.845	3.552	3.850
21	1.721	2.080	2.518	2.831	3.527	3.819
22	1.717	2.074	2.508	2.819	3.505	3.792
23	1.714	2.069	2.500	2.807	3.485	3.767
24	1.711	2.064	2.492	2.797	3.467	3.745
25	1.708	2.060	2.485	2.787	3.450	3.725
26	1.706	2.056	2.479	2.779	3.435	3.707
27	1.703	2.052	2.473	2.771	3.421	3.690
28	1.701	2.048	2.467	2.763	3.408	3.674
29	1.699	2.045	2.462	2.756	3.396	3.659
30	1.697	2.042	2.457	2.750	3.385	3.646
40	1.684	2.021	2.423	2.704	3.307	3.551
60	1.671	2.000	2.390	2.660	3.232	3.460
120	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.645	1.960	2.326	2.576	3.090	3.291