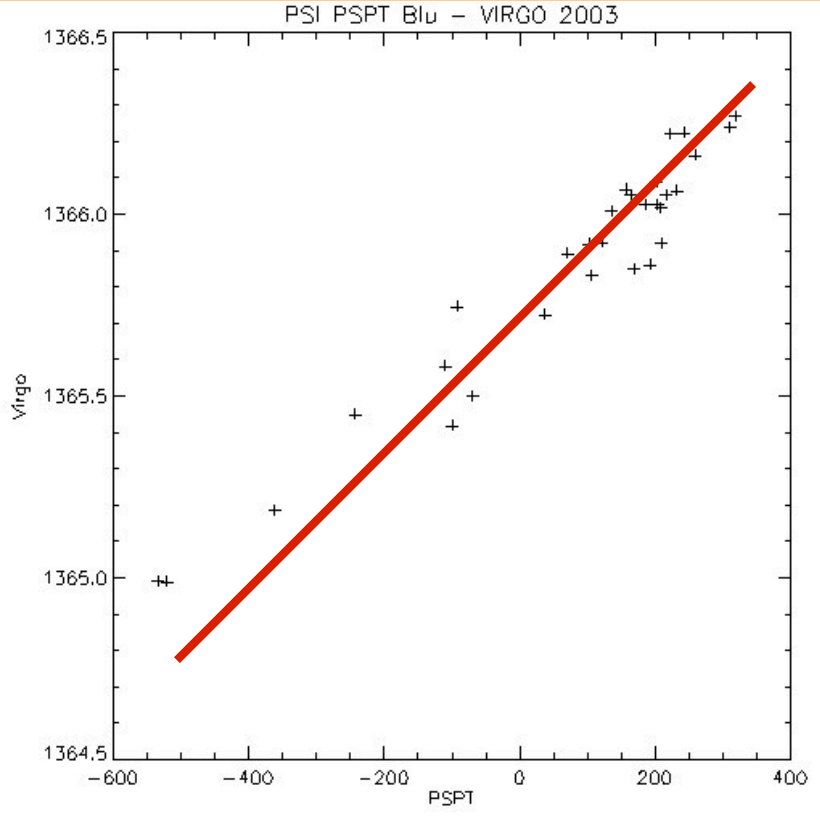
LA CORRELAZIONE LINEARE

La correlazione indica la tendenza che hanno due variabili (X e Y) a *variare insieme*, ovvero, a *covariare*. Ad esempio, si può supporre che vi sia una relazione tra l'insoddisfazione della madre e l'aggressività del bambino, nel senso che all'aumentare dell'una aumenta anche l'altra.

Quando si parla di correlazione bisogna prendere in considerazione due aspetti: *il tipo di relazione esistente* tra due variabili e *la forma della relazione*.

Per quanto riguarda il **tipo di relazione**, essa può essere *lineare* o *non lineare*

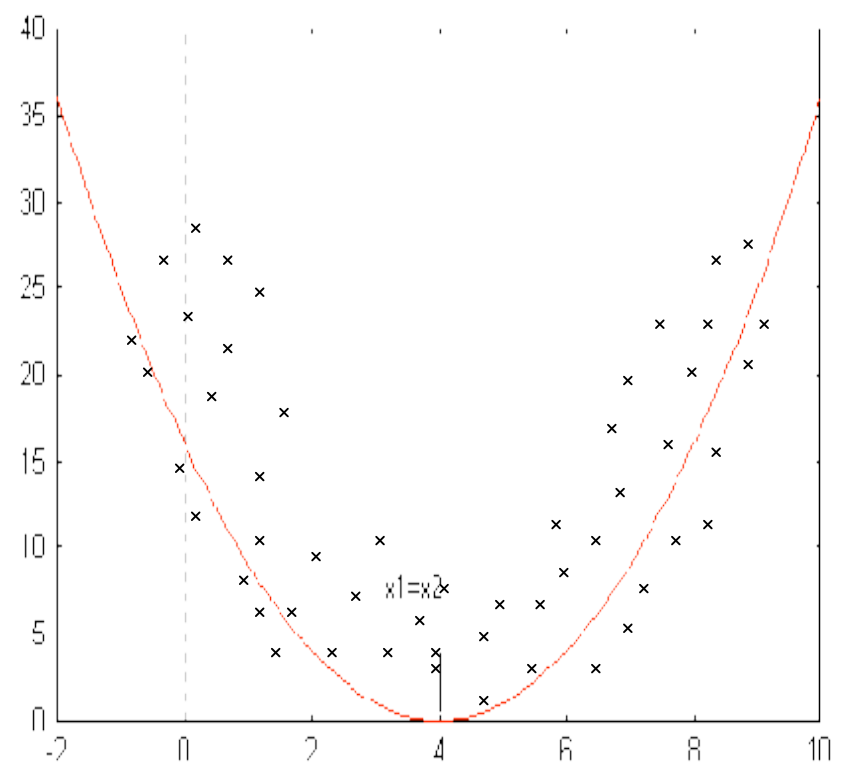
- La relazione è di tipo *lineare* se, rappresentata su assi cartesiane, si avvicina alla forma di una retta.



In questo caso, all'aumentare (o al diminuire) di X aumenta (diminuisce) Y.

Ad esempio, all'aumentare dell'altezza di una persona aumenta anche il suo peso.

- La relazione è di tipo *non lineare,* se rappresentata su assi cartesiane, ha un andamento curvilineo (parabola o iperbole).

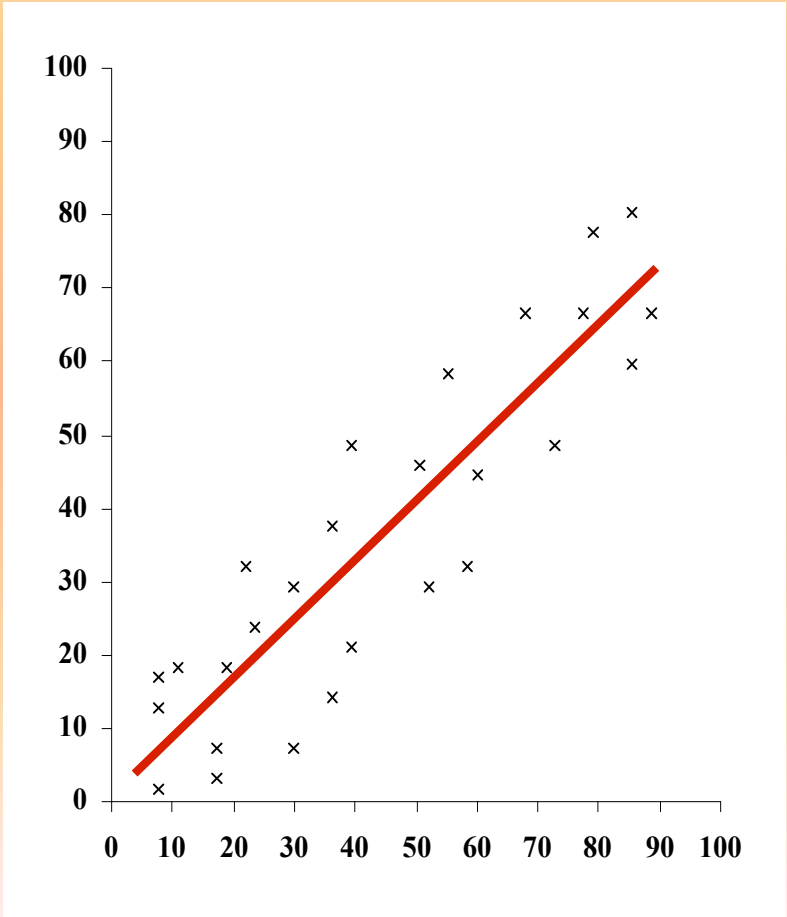


In questo caso a livelli bassi e alti di X corrispondono livelli bassi di Y; mentre a livelli intermedi di X corrispondono livelli alti di Y.

Ad esempio, il tempo impiegato per risolvere un problema è alto quando l'ansia è bassa o alta, è elevato quando l'ansia ha livelli medi.

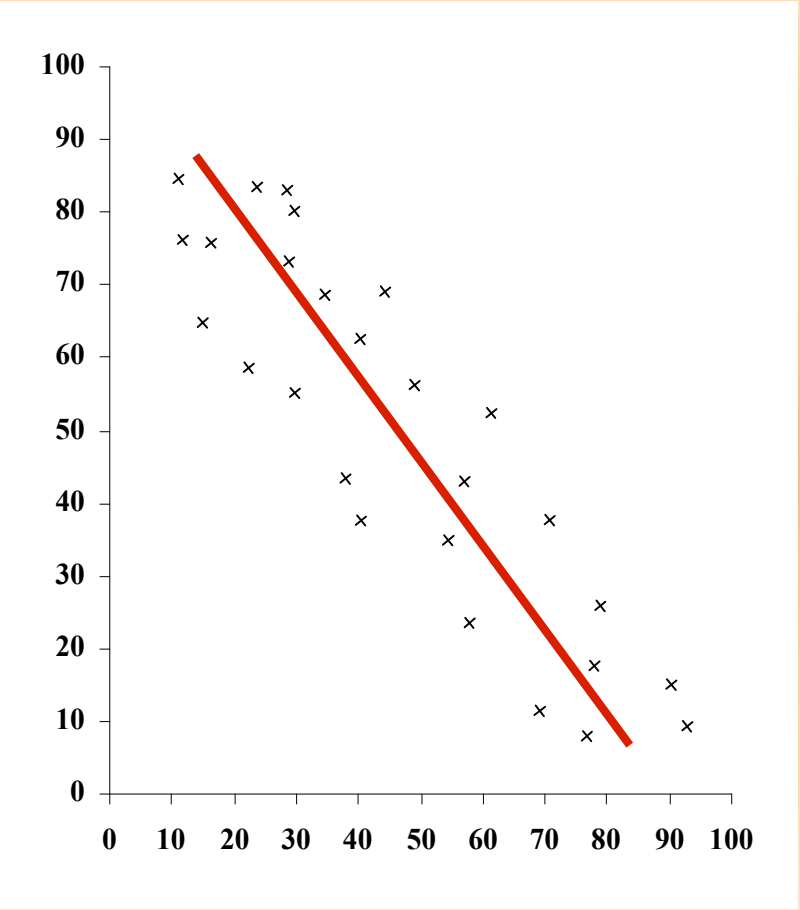
Per quanto riguarda la **forma della relazione**, si distinguono l'*entità* e la *direzione*.

La **direzione** può essere: *positiva*, se all'aumentare di una variabile aumenta anche l'altra.



Ad esempio, all'aumentare dell'identificazione con l'ingroup aumenta anche il pregiudizio.

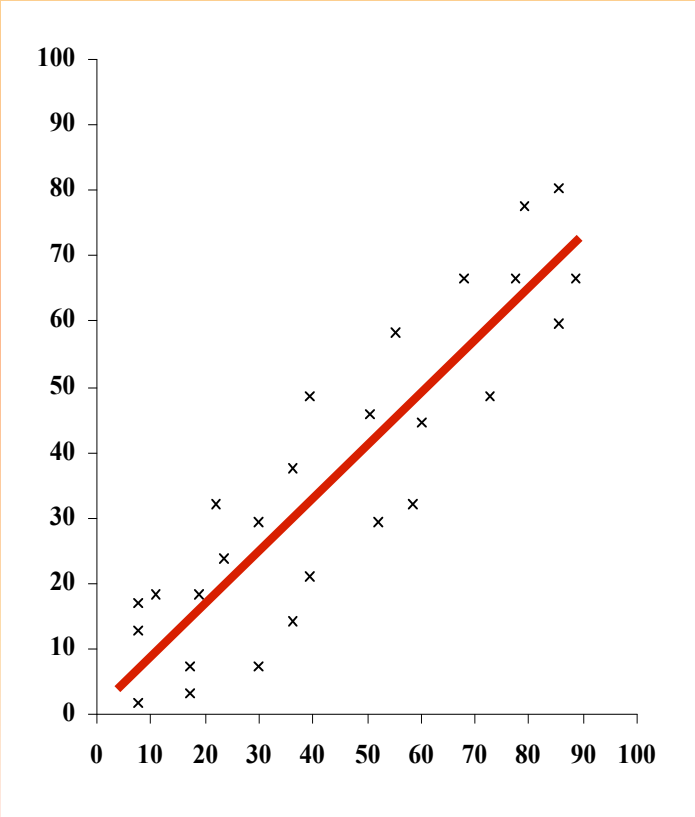
La direzione è *negativa* se all'aumentare di una variabile diminuisce l'atra.



Ad esempio, all'aumentare della qualità del contatto, diminuisce il pregiudizio nei confronti dell'outgroup.

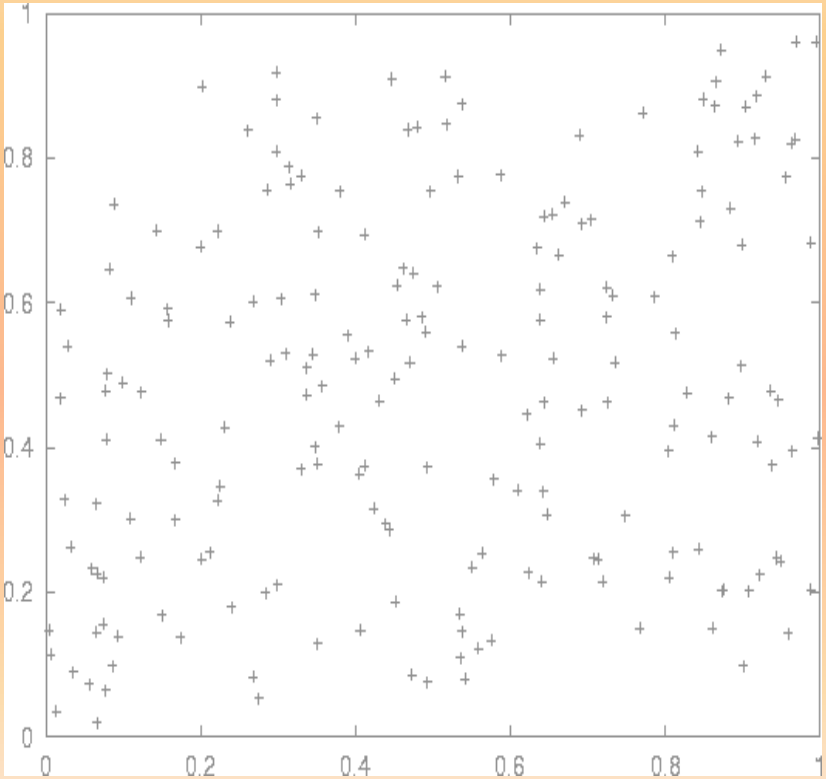
L’**entità** si riferisce alla forza della relazione esistente tra due variabili.

Quanto più i punteggi sono raggruppati attorno ad una retta, tanto *più forte* è la relazione tra due variabili.



Ad esempio, quanto più elevata è la temperatura, tanto più si suda.

Se i punteggi sono dispersi in maniera uniforme, invece, tra le due variabili *non esiste* alcuna relazione.



Ad esempio, non esiste alcuna relazione tra la temperatura e il livello di identificazione con l'ingroup.

Per esprimere la relazione esistente tra due variabili, in termini entità e direzione, si utilizza il **coefficiente di correlazione**.

Tale coefficiente è standardizzato e può assumere valori che vanno da **–1.00** (correlazione perfetta negativa) e **+1.00** (correlazione perfetta positiva). Una correlazione uguale a **0** indica che tra le due variabili non vi è alcuna relazione.

Nota. La correlazione non include il concetto di causa-effetto, ma solo quello di rapporto tra variabili. La correlazione ci permette di affermare che tra due variabili c'è una *relazione sistematica*, ma non che una causa l'altra.

Esistono vari tipi di coefficienti di correlazione a seconda del tipo di scala della variabile.

- Per le scale a **intervalli** o **rapporti** equivalenti si usa il coefficiente ***r* di Pearson**.
- Per le scale **ordinali** si usano il coefficiente r_s **di Spearman** o il coefficiente ***tau* di Kendall**.
- Per le scale **categoriali** (dicotomiche) si usano il coefficiente r_{phi} o il coefficiente r_{pbis}.

Il coefficiente di correlazione *r* di Pearson

Tale coefficiente serve a misurare la correlazione tra variabili a intervalli o a rapporti equivalenti. È dato dalla somma dei prodotti dei punteggi standardizzati delle due variabili ($z_x z_y$) diviso il numero dei soggetti (o delle osservazioni).

$$r = \frac{\sum z_x z_y}{N}$$

Tale coefficiente può assumere valori che vanno da –1.00 (tra le due variabili vi è una correlazione perfetta negativa) e + 1.00 (tra le due variabili vi è una correlazione perfetta positiva). Una correlazione uguale a 0 indica che tra le due variabili non vi è alcuna relazione.

Per effettuare i calcoli si utilizza la seguente formula, derivata dalla risoluzione della precedente.

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

Tale coefficiente può assumere valori che vanno da –1.00 (tra le due variabili vi è una correlazione perfetta negativa) e + 1.00 (tra le due variabili vi è una correlazione perfetta positiva). Una correlazione uguale a 0 indica che tra le due variabili non vi è alcuna relazione.

Per stabilire se una correlazione è significativa, si fa riferimento alla distribuzione campionaria di *r*, tabulata in apposite tavole, in corrispondenza dei gradi di libertà (*N* – 2) del coefficiente.

Esempio.
Verificare l'esistenza di una relazione tra l'identificazione con l'ingroup e il pregiudizio, nei seguenti 5 soggetti.

Soggetto	Identificazione	Pregiudizio
1	10	7
2	12	5
3	15	8
4	13	6
5	12	4

Ss	X	Y	X^2	Y^2	XY
1	10	7	100	49	70
2	12	5	144	25	60
3	15	8	225	64	120
4	13	6	169	36	78
5	12	4	144	16	48
Σ	62	30	782	190	376

In questo modo otteniamo:

$\Sigma X = 62$

$\Sigma Y = 30$

$\Sigma X^2 = 782$

$\Sigma Y^2 = 190$

$\Sigma XY = 376$

Applichiamo la formula:

$$r = \frac{5*376 - 62*30}{\sqrt{(5*782 - 62^2)(5*190 - 30^2)}}$$

$$r = \frac{1880 - 1860}{\sqrt{66 * 50}} \qquad r = \frac{20}{57.44} = 0.35$$

Per stabilire se la correlazione è significativa, calcoliamo il *t*, utilizzando la seguente formula.

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \qquad t = \frac{0.35}{\sqrt{\frac{1 - 0.35^2}{5 - 2}}} \qquad t = \frac{0.35}{\sqrt{\frac{0.88}{3}}}$$

$$t = \frac{0.35}{0.54} = 0.65$$

Confrontiamo il valore di t ottenuto con il valore critico relativo a n – 2 g.d.l., ovvero a 3 g.d.l.

Nota. Bisogna tenere in considerazione l'*ipotesi bidirezionale*.

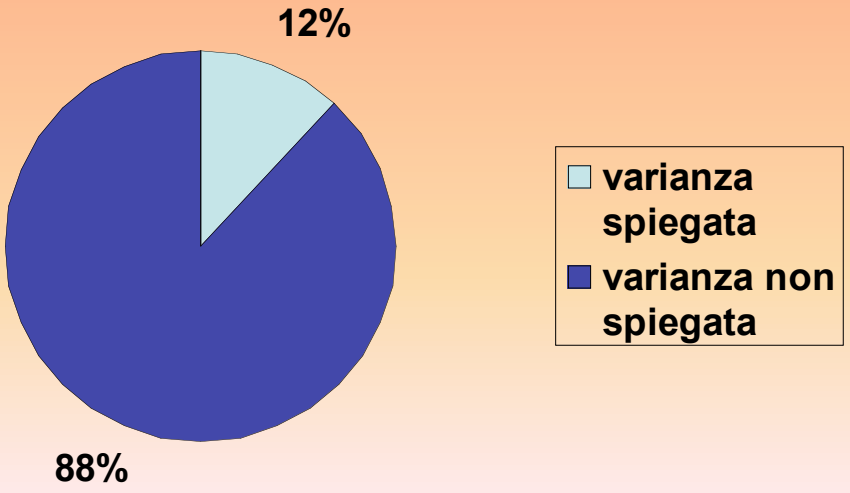
	Ipotesi monodirezionale					
	0.10	0.05	0.025	0.010	0.005	0.0005
gdl	Ipotesi bidirezionale					
	0.20	0.10	0.050	0.020	0.010	0.001
1	3.078	6.314	12.706	31.821	63.656	636.578
2	1.886	2.920	4.303	6.965	9.925	31.600
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.160	3.012	4.221
14	1.345	1.761	2.145	2.145	2.977	4.140

Il valore ottenuto (0.65) non supera il valore critico (3.184), quindi, la relazione tra identificazione e pregiudizio non è significativa (r = 0.35, *ns*).

Il coefficiente di determinazione r^2

Il coefficiente di determinazione misura l'ammontare di variabilità di una variabile spiegato dalla sua relazione con un'altra variabile. Nel caso specifico della correlazione il coefficiente r^2 indica la percentuale di varianza che hanno in comune due variabili.

Nell'esempio precedente, abbiamo trovato un r pari a 0.37, da cui ricaviamo $r^2 = 0.35^2 = 0.12$. Ovvero, abbiamo che l'identificazione e il pregiudizio condividono il 12% di variabilita.



Esercizio.

Verificare se esiste una relazione tra l'altezza e il peso e quale è l'ammontare della variabilità comune delle due variabili.

Soggetto	Altezza	Peso
1	155	47
2	176	68
3	164	53
4	170	64
5	157	57
6	162	60
7	169	63

$R = 0.87, p < .05 [t (5) = 3.92, p < .05]$

$R^2 = 0.76$, 76% di varianza comune

Il coefficiente r_s di Spearman

Tale coefficiente serve per misurare la correlazione tra due variabili di tipo ordinale.

Ad esempio, un ricercatore potrebbe chiedere ad un insegnante di mettere in ordine di rango gli studenti per profitto (dal più bravo al meno bravo) e per socievolezza (dal più socievole al meno socievole) e vedere, quindi, se tra le due variabili esiste una relazione.

Il coefficiente r_s di Spearman è un'approssimazione del coefficiente di Pearson e la formula è la seguente:

$$r_s = 1 - \frac{6\Sigma\, d_i^2}{N * (N^2 - 1)}$$

In cui d_i è la differenza tra i ranghi delle due variabili per il soggetto i-esimo.
La relazione tra X e Y è espressa tenendo conto delle concordanti o differenti posizioni di ciascun soggetto nelle due graduatorie.

Esempio.

I seguenti dati si riferiscono a due graduatorie, relative al profitto e alla socievolezza.

Ss	Rango Profitto	Rango Voti
1	2	1
2	5	2.5
3	3	2.5
4	7	6
5	1	4
6	4	5
7	6	7

Ss	Rango X	Rango Y	d	d^2
1	2	1	1	1
2	5	2.5	2.5	6.25
3	3	2.5	0.5	0.25
4	7	6	1	1
5	1	4	-4	9
6	4	5	-1	1
7	6	7	-1	1

Da dati calcoli, otteniamo $\Sigma d^2 = 19.5$

Applicando la formula otteniamo:

$$r_s = 1 - \frac{6*19.5}{7*(49-1)} = 1 - \frac{117}{336} = 1 - 0.35 = 0.65$$

Per la significatività di r_s si fa riferimento alle apposite tavole di r_s con N-2 gdl.
In questo caso, r_s = 0.65, è inferiore al valore critico (**0.714**), quindi, non vi è alcuna relazione tra le due variabili.

Anche r_s può assumere valori compresi tra –1.00 e +1.00, con gli stessi significati visti per *r*.

È evidente che, se i soggetti occupassero esattamente le stesse posizioni nelle due graduatorie, per X e per Y, le differenze *d* sarebbero tutte uguali a 0 e r_s sarebbe uguale a +1.00, massima correlazione positiva.

Se, invece, si verificasse una corrispondenza perfetta tra posizioni opposte in X e Y, r_s risulterebbe uguale a –1.00.

Il coefficiente r_s ha il difetto di dare una stima per eccesso della correlazione tra X e Y se, per almeno una variabile, si riscontrano molti ranghi uguali.

Esercizio.
Verificare l'esistenza di una relazione tra l'ordine di arrivo in una gara su 100m e l'ordine di arrivo in una gara su 1000m.

Ss	Rango 100	Rango 1000
1	5	1
2	9	5
3	6	2
4	2	6
5	4	7
6	1	3
7	3	4
8	7	10
9	10	9
10	8	8

$r_s = 0.47$, *ns*

Il coefficiente *tau* di Kendall

Anche questo coefficiente serve per misurare la correlazione tra due variabili di tipo ordinale, ma è esente dal difetto del coefficiente r_s. La formula è la seguente:

$$tau = \frac{S}{0.5 * N * (N - 1)}$$

In cui *S* si ottiene come somma nel modo seguente.
Dati i seguenti ranghi relativi a 7 soggetti, su due variabili.

Soggetto	Rango X	Rango Y
A	2	1
B	5	2.5
C	3	2.5
D	7	6
E	1	4
F	4	5
G	6	7

Prima si mettono in graduatoria i valori di X e si considerano i corrispondenti valori di Y.

Soggetto	Rango X	Rango Y
A	2	1
B	5	2.5
C	3	2.5
D	7	6
E	1	4
F	4	5
G	6	7

Quindi, si confronta ciascun valore di Y con tutti quelli che seguono e si segna +1 ogni volta che i due ranghi confrontati si trovano in ordine corretto rispetto alla graduatoria delle Y, si segna, invece, -1 ogni volta che si trovano in ordine errato.

Ss	X	Y
E	1	4
A	2	1
C	3	2.5
F	4	5
B	5	2.5
G	6	7
D	7	6

	E	A	C	F	B	G	D
E	/	-1	-1	+1	-1	+1	+1
A			+1	+1	+1	+1	+1
C				+1	0	+1	+1
F					-1	+1	+1
B						+1	+1
G							-1
H							/

S è la somma algebrica dei valori +1 e –1 assegnati. Il denominatore è il valore massimo di *S* che si otterrebbe se tutti gli Y si trovassero nell'ordine corretto.

	E	A	C	F	B	G	D
E	/	-1	-1	+1	-1	+1	+1
A			+1	+1	+1	+1	+1
C				+1	0	+1	+1
F					-1	+1	+1
B						+1	+1
G							-1
H							/

S = 15 x (+1) + 5 x (-1) = 15 – 5 = 10

Applicando la formula per il calcolo di *tau* si ottiene:

$$tau = \frac{10}{0.5 * 7 * (7 - 1)}$$

$$tau = \frac{10}{21} = 0.48$$

Confrontando il valore ottenuto, con il valore critico di significatività, ottenuto dall' incrocio tra il valore di S e la numerosità (in questo caso 10 e 7), si trova che *tau* = 0.48, non è significativo.

Anche *tau* può assumere valori compresi tra –1.00 e +1.00, con gli stessi significati visti per *r*.

Nota. È possibile utilizzare i coefficienti di correlazione per ranghi su scale a intervalli o rapporti equivalenti nei casi in cui non è possibile applicare *r* di Pearson. Per fare questo bisogna trasformare il livello di misura della variabile, calcolando gli ordini di rango sui punteggi originali.

L'ordine di rango è, comunque, meno sensibile della misura vera e propria.

Inoltre, le statistiche basate sui ranghi sono meno potenti di quelle basate su misure continue.

Esercizio.
Verificare se esiste una relazione tra la prestazione ottenuta ad un compito di matematica e quella ottenuta ad un compito di fisica, entrambe espresse su scala ordinale.

Soggetto	Matematica	Fisica
A	3	2
B	7	7
C	1	1
D	8	10
E	2	3
F	4	4
G	10	9
H	5	6
I	6	5
L	9	8

Tau = 0.82, $p < .05$