CAPITOLO XXI

TEST NON PARAMETRICI PER CORRELAZIONE, CONCORDANZA, REGRESSIONE MONOTONICA E REGRESSIONE LINEARE

21.1. La correlazione non parametrica ρ (rho) di Spearman, con la distribuzione di Hotelling-Pabst 1

21.2. Il coefficiente di correlazione τ (tau) di Kendall; il τ_a e τ_b di Kendall con i ties 11

21.3. Confronto tra ρ e τ; potenza del test e numero di osservazioni necessarie per la significativita' 20

21.4. Altri metodi per la correlazione non parametrica: test di Pitman con le permutazioni; test della mediana di Blomqvist 25

21.5. Il test di Daniels per il trend 34

21.6. Significativita' della regressione e della correlazione lineare parametrica con i test nonparametrici ρ e τ 41

21.7. Il coefficiente di correlazione parziale: $\tau_{12,3}$ di Kendall, $\rho_{12,3}$ di Spearman 46

21.8. Il coefficiente di concordanza tra valutatori: la w di Kendall; sue relazioni con la correlazione non parametrica e con il test di Friedman per k campioni dipendenti. Cenni sulla top-down concordance 53

21.9. Cenni sul coefficiente di concordanza u di Kendall, in confronti appaiati 63

21.10. La regressione lineare non parametrica 66

21.11. Calcolo della retta di regressione non parametrica con il metodo di Theil o test di Theil-Kendall 68

21.12. Confronto tra la retta parametrica e la retta di Theil 76

21.13. Significativita' di b con il τ di Kendall 78

21.14. La regressione lineare non parametrica con il metodo dei tre gruppi di Bartlett 86

21.15. Il test di Hollander per il confronto tra due coefficienti angolari 92

21.16. La regressione monotonica di Iman-Conover 98

21.17. Trend lineare di Armitage per le proporzioni e le frequenze 104

CAPITOLO XXI

TEST NON PARAMETRICI
PER CORRELAZIONE, CONCORDANZA,
REGRESSIONE MONOTONICA E REGRESSIONE LINEARE

21.1. LA CORRELAZIONE NON PARAMETRICA ρ (rho) DI SPEARMAN, CON LA DISTRIBUZIONE DI HOTELLING-PABST.

La correlazione è uno dei metodi statistici più antichi, diffuso già all'inizio del '900, almeno dieci anni prima del test t di Student e quasi trenta prima dell'analisi della varianza. La **metodologia non parametrica** proposto da

- C. **Spearman** nel 1904 (con l'articolo ***The proof and measurement of association between two things*** su **American Journal of Psychology** vol. 15, pp. 72 – 101 e con l'articolo ***A footrule for measuring correlation*** , pubblicato nel 1906 su **Brit. Journ. Psychol**. n. 2)

è una **correlazione basata sui ranghi,** che ricorre agli stessi concetti della correlazione parametrica **r** di **Pearson** presentata anch'essa poco prima, nel 1900, e successivamente chiamata ***Pearson's Product Moment Sample Correlation Coefficient***.

Questa metodologia non parametrica ha subito varie elaborazioni e modifiche. Era ancora discussa negli anni '20, come può dimostrare l'articolo di W. S. **Gosset** (**Student**) del 1921 ***An experimental determination of the probable error of Dr. Spearman's correlation coefficients*** (comparso su **Biometrika** vol. 13). Ora, dopo un secolo, è ancora uno dei test più ampiamente utilizzati, per lo studio dell'associazione tra due variabili quantitative.

In riferimento ai test di correlazione non parametrica, in letteratura sono ricorrenti i termini di **correlazione tra ranghi**, **cograduazione**, **associazione** e **concordanza.** Da molti utenti della statistica, sono usati come sinonimi; ma

- i primi due (correlazione e cograduazione) dovrebbero essere utilizzati in modo appropriato solo con **scale almeno di tipo ordinale**,
- mentre le ultime due (associazione e concordanza) con **scale qualitative o categoriali**.

Il coefficiente di correlazione di Spearman è sovente indicato con il simbolo greco ρ **(rho);** in altri testi è simboleggiato con $\mathbf{r_S}$, per evidenziare la sua affinità con il test **r** di **Pearson** dal quale è derivato. Come esso, può variare

- tra **+1 e –1** quando la **correlazione è massima**, con valore **positivo** oppure **negativo**;
- è **vicino a zero**, quando **non esiste correlazione**.

Il metodo richiede che **entrambe le variabili siano misurate su una scala almeno ordinale**, per cui ognuna delle due serie di misure, indicate sovente con **X** e **Y** anche se non esiste tra esse una relazione di causa effetto e dovrebbero essere correttamente indicate con X_1 e X_2, non dovrebbe avere valori uguali entro la stessa sequenza. Questo metodo può essere più potente del test **r** di **Pearson** anche per scale d'intervallo o di rapporto, quando le condizioni di validità del test parametrico non sono pienamente soddisfatte. Di conseguenza, come per altri test non parametrici, la sua utilizzazione è consigliabile insieme con il test parametrico, ad ulteriore dimostrazione e verifica delle conclusioni raggiunte. In particolare quando si disponga solo di pochi dati e pertanto non sia possibile dimostrare che le condizioni di validità del test parametrico sono soddisfatte in modo completo.

Il coefficiente di correlazione per ranghi di Spearman serve per verificare l'**ipotesi nulla dell'indipendenza tra due variabili,** nel senso che gli **N** valori della variabile **Y** hanno le stesse probabilità di associarsi con ognuno degli **N** valori di **X**.
L'ipotesi alternativa di esistenza di una associazione può prevedere un risultato **positivo oppure negativo**. Nel primo caso è detta **associazione diretta**: le coppie di valori sono contemporaneamente alti o bassi sia per **X** che per **Y**; nel secondo caso, chiamata anche **associazione indiretta**, a valori alti di **X** corrispondono valori bassi di **Y** o viceversa.
Per una illustrazione didattica chiara, i vari passaggi logici richiesti dal metodo proposto da **Spearman** possono essere suddivisi in 5 fasi, di seguito presentate nella dimostrazione di un caso.
e
- sia **bilaterale**

$$\mathbf{H_0: \rho = 0} \quad \text{contro} \quad \mathbf{H_1: \rho \neq 0}$$

- sia **unilaterale** in una direzione

$$\mathbf{H_0: \rho \leq 0} \quad \text{contro} \quad \mathbf{H_1: \rho > 0}$$

oppure nell'altra

$$\mathbf{H_0: \rho \geq 0} \quad \text{contro} \quad \mathbf{H_1: \rho < 0}$$

è utile riportare i dati come nella tabella seguente

Variabili	Coppie di valori osservati						
X	8	5	7	14	22	21	41
Y	12	3	2	10	25	19	22
Soggetti	A	B	C	D	E	F	G

2 - Successivamente, occorre **ordinare i ranghi della variabile X**, assegnando **1** al valore più piccolo e progressivamente valori interi maggiori, fino ad **N** per il valore più alto.
Se i dati della variabile **X** hanno due o più valori uguali, è necessario assegnare ad ognuno di essi come rango la media delle loro posizioni.

Variabili	Coppie di valori osservati						
X	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Y	3	2	12	10	25	16	22
Soggetti	**B**	**C**	**A**	**D**	**F**	**E**	**G**

Anche se ininfluente ai fini dei calcoli successivi, è utile alla comprensione della misura di correlazione porre nell'ordine naturale (da **1** a **N**) i ranghi della variabile **X** e spostare la collocazione dei valori di **Y** relativi al medesimo soggetto, come nella tabella sovrastante.

3 - Sostituire anche gli **N** valori di **Y** con i ranghi rispettivi; per valori di **Y** uguali, usare la media dei loro ranghi:

Variabili	Coppie di valori osservati						
X	1	2	3	4	5	6	7
Y	**2**	**1**	**4**	**3**	**7**	**5**	**6**
Soggetti	B	C	A	D	F	E	G

Si ottiene la riga Y, riportata in grassetto.

4 - Se le due distribuzioni (quella della serie delle X e quella della serie delle Y)
- sono **correlate in modo positivo** (**r = +1**), i valori della variabile **X** e della **Y** relativi allo stesso soggetto saranno uguali;
- sono **correlate in modo negativo** (**r = -1**), a valori alti di **X** saranno associati valori bassi di **Y** e viceversa;
- se tra le due variabili **non esiste correlazione** (**r = 0**), i valori di **X** e di **Y** relativi agli stessi soggetti saranno associati in modo casuale.

Per quantificare questo grado di correlazione o concordanza, Spearman ha proposto **la distanza tra le coppie dei ranghi** (d_{R_i})

Variabili	Coppie di valori osservati						
X	1	2	3	4	5	6	7
Y	2	1	4	3	7	5	6
R_i	**-1**	**+1**	**-1**	**+1**	**-2**	**+1**	**+1**
R^2_i	**1**	**1**	**1**	**1**	**4**	**1**	**1**

come calcolate nella terza riga (R_i); successivamente devono essere elevate al quadrato come riportate nella quarta riga (R^2_i)

L'indicatore di correlazione, da cui derivano i passaggi logici e metodologici successivi, è la somma di questi quadrati:

$$\sum d^2_{R_i}$$

Con i dati dell'esempio, la somma delle $\mathbf{d^2_{Ri}}$ è uguale a **10** (1 + 1 + 1 + 1 + 4 + 1 + 1 = 10)

5 - Quando **r = +1**, le coppie di osservazioni di **X** e **Y** hanno lo stesso rango e pertanto questa sommatoria è uguale a **0**.

Quando **r = -1**, se **X** è ordinato in modo crescente, **Y** è ordinato in modo decrescente: di conseguenza, le differenze sono massime e la sommatoria raggiunge un valore massimo determinato dal numero di coppie di osservazioni (**N**).

Quando **r = 0**, mentre i ranghi di **X** sono ordinati in modo crescente quelli di **Y** hanno una distribuzione casuale: la sommatoria delle $d^2_{R_i}$ tende ad un valore medio, determinato dal numero di coppie di osservazioni (**N**).

Il test è fondato sulla statistica

$$D = \sum d^2_{R_i}$$

ed è conosciuto anche come **test statistico di Hotelling-Pabst** (vedi l'articolo di H. **Hotelling** e M. R. **Pabst** del 1936 ***Rank correlation and tests of significance involving no assumption of normality***, in **Annals of Mathematical Statistics**, Vol. 7, pp. 429-443).

Per essa sono state proposte tavole di valori critici, al fine di valutare la significatività del test.

6 - Il **coefficiente di correlazione tra ranghi (ρ) di Spearman** è derivato dalla formula della correlazione di Pearson

$$r = \frac{cod_{xy}}{dev_x}$$

Applicata ai ranghi, dopo semplificazione diviene

$$\rho = \frac{\sum_i \left(Rxi - \frac{N+1}{2} \right) \cdot \left(Ryi - \frac{N+1}{2} \right)}{\frac{N(N^2-1)}{12}}$$

Il coefficiente di correlazione per ranghi di Spearman è semplicemente il coefficiente di correlazione di Pearson applicato ai ranghi.

Ritornando alla somma degli scarti tra i ranghi,

la formula abbreviata può essere scritta come

$$\rho = 1 - \frac{6 \cdot \sum d_{R_i}^2}{N^3 - N}$$

con **N** uguale al numero di coppie di osservazioni.

In vari testi, è scritto con la formula equivalente

$$\rho = 1 - \frac{6 \cdot \sum d_{R_i}^2}{N \cdot \left(N^2 - 1 \right)}$$

Quando due o più valori di X o di Y sono identici (ties) e pertanto hanno lo stesso rango, l'attribuzione dei punteggi medi riduce il valore della devianza. Con pochi valori identici, l'effetto è trascurabile. Con molti valori identici, è bene calcolare un **fattore di correzione T** sia per la variabile **X** (T_x) sia per la **Y** (T_y)

$$T = \sum_{i=1}^{g} \left(t_i^3 - t_i \right)$$

dove

g è il numero di raggruppamenti con punteggi identici e

t è il numero di ranghi identici entro ogni raggruppamento.

Con queste correzioni, nel caso di molti valori identici **la formula completa del ρ di Spearman** diventa

$$\rho = \frac{N^3 - N - 6 \cdot d^2 - \frac{T_x + T_y}{2}}{\sqrt{\left(N^3 - N\right)^2 - \left(T_x + T_y\right) \cdot \left(N^3 - N\right) + T_x \cdot T_y}}$$

Come in tutti i ties e già evidenziato, **la correzione determina**

- **una differenza sensibile quando uno stesso valore è uguale in molti casi**,
- **un effetto trascurabile o comunque ridotto quando si hanno molti valori ripetuti solo 2 volte.**

Di conseguenza, nonostante la correzione, questo test è da evitare e può essere utile ricorrere ad altri metodi, quando uno o più valori sono ripetuti con frequenza elevata, nella X e/o nella Y.

Nel caso di **piccoli campioni** (**N < 20-25**), la significatività di ρ è fornita dalle tabelle dei valori critici. Nella pagina successiva sono riportati i valori critici di **ρ,** sia per test a una coda che per test a due code.

Alla probabilità α prefissata, **si rifiuta l'ipotesi nulla se il valore calcolato è uguale o superiore a quello riportato nella tabella**.

Nel caso di **grandi campioni** (**N > 20-25**), quando è valida l'ipotesi nulla d'assenza di correlazione, il valore di ρ è distribuito con media **0** e deviazione standard **1**. Per la sua significatività è stato proposto

- sia il ricorso alla **distribuzione Z con la trasformazione**

$$Z = \rho \cdot \sqrt{N - 1}$$

- sia alla **distribuzione t di Student con gdl N - 2 con la trasformazione di ρ**

$$t_{(N-2)} = \rho \cdot \sqrt{\frac{N - 2}{1 - \rho^2}}$$

Tra **t** e **Z**,

- il test **t** sembra preferibile, in quanto giustamente più cautelativo ma pertanto meno potente, quando il campione ha **meno di 50 osservazioni**;
- per **campioni di dimensioni maggiori**, i due metodi risultano equivalenti poiché i valori critici sono quasi coincidenti,
- non diversamente da quanto avviene per il confronto tra due medie.

Valori critici del coefficiente ρ di Spearman

per test a 1 coda (1ª riga) **e test a 2 code** (2ª riga)

α

N	**0.05**	0.025	**0.01**	0.005	**0.001**	0.0005	**1 coda**
	0.10	**0.05**	0.02	**0.01**	0.002	**0.001**	**2 code**
4	1.000	---	---	---	---	---	
5	.900	1.000	1.000	---	---	---	
6	.829	.886	.943	1.000	---	---	
7	.714	.786	.893	.929	1.000	1.000	
8	.643	.738	.833	.881	.952	.976	
9	.600	.700	.783	.833	.917	.933	
10	.564	.648	.745	.794	.879	.903	
11	.536	.618	.709	.755	.845	.873	
12	.503	.587	.671	.727	.825	.860	
13	.484	.560	.648	.703	.802	.853	
14	.464	.538	.622	.675	.776	.811	
15	.443	.521	.604	.654	.754	.786	
16	.429	.503	.582	.635	.732	.765	
17	.414	.485	.566	.615	.713	.748	
18	.401	.472	.550	.600	.695	.728	
19	.391	.460	.535	.584	.677	.712	
20	.380	.447	.520	.570	.662	.696	
21	.370	.435	.508	.556	.648	.681	
22	.361	.425	.496	.544	.634	.667	
23	.353	.415	.486	.532	.622	.654	
24	.344	.406	.476	.521	.610	.642	

ESEMPIO. La concentrazione delle sostanze organiche presenti nell'acqua può essere misurata mediante il BOD (da Biological Oxygen Demand, la richiesta biochimica dell'ossigeno), il COD (da Chemical Oxygen Demand, la richiesta chimica dell'ossigeno) e il TOC (da Total Organic Carbon, il carbonio organico totale).
Lungo un corso d'acqua sono state fatte 16 rilevazioni del BOD_5 (a 5 giorni) e dell'azoto ammoniacale, con la successiva serie di misure.

Stazione	BOD_5	N
s_1	5	0,7
s_2	5	0,8
s_3	12	5,6
s_4	35	24,3
s_5	11	9,7
s_6	7	1,8
s_7	8	1,6
s_8	9	4,8
s_9	9	1,7
s_{10}	20	4,8
s_{11}	14	5,6
s_{12}	13	3,2
s_{13}	16	3,6
s_{14}	15	2,9
s_{15}	13	3,9
s_{16}	11	2,8

S'intende verificare se tra le due serie di valori esista una correlazione positiva significativa, nonostante la non normalità delle distribuzioni, come evidenzia la semplice lettura dei dati della stazione S_4.e S_5, ovviamente da confermare con le analisi relative.

Risposta. Il test è unilaterale, con

$$\mathbf{H_0: \rho \leq 0} \qquad \text{contro} \qquad \mathbf{H_1: \rho > 0}$$

Il metodo può essere suddiviso in 7 fasi; per le prime 5, qui elencate, i calcoli sono riportati nella tabella successiva:

1 - ordinare in modo crescente i valori del BOD_5 ed attribuire i ranghi relativi (**colonne 1a e 1b**);

2 - trasformare in ranghi i corrispondenti valori di **N** (**colonne 2a e 2b**);

3 - calcolare la differenza **d** tra i ranghi (**colonna 3**);

4 - elevare al quadrato tali differenze (**d^2 nella colonna 4**);

5 - calcolare la somma dei quadrati delle differenze (**somma delle d^2 nella colonna 4**);

Con i dati dell'esempio, ($\sum d_{R_i}^2$) è uguale a 220,5.

6 - Per **N** uguale a 16, calcolare il valore di **ρ**

$$\rho = 1 - \frac{6 \cdot \sum d_{R_i}^2}{N^3 - N} = 1 - \frac{6 \cdot 220{,}5}{4096 - 16} = 1 - \frac{1323}{4080} = 1 - 0{,}324 = \mathbf{0{,}676}$$

che risulta uguale a 0,676.

	1a	**2a**	**1b**	**2b**	**3**	**4**
Stazione	BOD_5	N	**R(BOD_5)**	**R(N)**	**d**	**d^2**
s1	5	0,7	**1,5**	**1**	**0,5**	**0,25**
s2	5	0,8	**1,5**	**2**	**-0,5**	**0,25**
s6	7	1,8	**3**	**5**	**-2**	**4**
s7	8	1,6	**4**	**3**	**1**	**1**
s9	9	1,7	**5,5**	**4**	**1,5**	**2,25**
s8	9	4,8	**5,5**	**11,5**	**-6**	**36**
s16	11	2,8	**7,5**	**6**	**1,5**	**2,25**
s5	11	9,7	**7,5**	**15**	**-7,5**	**56,25**
s3	12	5,6	**9**	**13,5**	**-4,5**	**20,25**
s12	13	3,2	**10,5**	**8**	**2,5**	**6,25**
s15	13	3,9	**10,5**	**10**	**0,5**	**0,25**
s11	14	5,6	**12**	**13,5**	**-1,5**	**2,25**
s14	15	4,8	**13**	**11,5**	**1,5**	**2,25**
s13	16	3,6	**14**	**9**	**5**	**25**
s10	20	2,9	**15**	**7**	**8**	**64**
s4	35	24,3	**16**	**16**	**0**	**0**

7 - Nella **tabella dei valori critici**, alla probabilità **$\alpha = 0.01$** per un test a **1** coda il valore riportato è 0.582. Il valore calcolato è superiore: si rifiuta l'ipotesi nulla e si accetta implicitamente l'ipotesi alternativa dell'esistenza di un'associazione positiva tra le due serie di dati rilevati.

Benché il numero di dati (**N** = 16) sia oggettivamente ridotto, la significatività può essere stimata sia con la distribuzione **Z** che con la distribuzione **t** di Student.

Con il **test Z** si ottiene

$$Z = 0{,}676 \cdot \sqrt{16 - 1} = 0{,}676 \cdot 3{,}873 = 2{,}62$$

un valore di **Z** uguale a 2,62.

Nella tabella della distribuzione **normale unilaterale**, a

- Z = 2,62 corrisponde la probabilità $\alpha = 0.0044$.

L'approssimazione con il risultato precedente è molto buona: con tabelle di ρ più dettagliate e ovviamente all'aumentare del numero di osservazioni, la differenza risulta trascurabile; in questo caso è di circa il 2/1000.

Con il **test t** si ottiene

$$t_{(15)} = 0{,}676 \cdot \sqrt{\frac{16-2}{1-0{,}676^2}} = 0{,}676 \cdot \sqrt{\frac{14}{1-0{,}457}}$$

$$t_{(15)} = 0{,}767 \cdot \sqrt{\frac{14}{0{,}543}} = 0{,}676 \cdot \sqrt{25{,}783} = 0{,}676 \cdot 5{,}078 = 3{,}433$$

un valore di **t** uguale a 3,433 con 15 gdl.

Nella tabella sinottica dei valori critici del **t di Student per un test unilaterale**,

- $t_{(15)}$ =3,433 si trova tra la probabilità α **= 0.005** ($t_{15} = 2{,}947$) e α **= 0.0005** ($t_{15} = 4{,}073$).

La conclusione non è molto differente da quella ottenuta con i due metodi precedenti.
I tre risultati sono approssimativamente equivalenti.

Un altro esempio di correlazione con il test $\mathbf{r_s}$ di Spearman è riportato nel successivo paragrafo dedicato al **test di Daniels**.

Sono stati proposti anche altri metodi, per stimare la significatività della regressione non parametrica ρ di Spearman. Tra i testi a maggior diffusione, quello di

- W. J. **Conover** del 1999 (***Practical nonparametric statistics***, 3rd ed. John Wiley & Soons, New York, 584) riporta i valori critici dei quantili, esatti quando X e Y sono indipendenti,
- calcolati da G. J. **Glasser** e R. F. **Winter** nel 1961 (nell'articolo ***Critical values of the coefficient of rank correlation for testing the hypothesis of indipendence,*** pubblicato su **Biometrika** Vol. 48, pp. 444-448).

21.2. IL COEFFICIENTE DI CORRELAZIONE τ (tau) DI KENDALL; IL τ_a E τ_b DI KENDALL CON I TIES.

Oltre 30 anni dopo il ρ (o $\mathbf{r_s}$) di Spearman,

- **M. G. Kendall** nel 1938 con l'articolo ***A new measure of rank correlation*** (pubblicato su **Biometrika** vol. 30, pp. 81-93) e in modo più dettagliato nel 1948 con la descrizione dettagliata della metodologia nel volume ***Rank correlation methods*** (edito a Londra da C. **Griffin**)

ha proposto il test τ **(tau).** Questo metodo

- ha le **stesse assunzioni**,
- può essere utilizzato nelle **medesime condizioni** e
- **sui medesimi dati del test ρ di Spearman**.

I **risultati tra i due test sono molto simili**, anche se matematicamente non equivalenti, per i motivi che saranno di seguito spiegati con l'illustrazione della metodologia. Tuttavia, da parte di molti autori il ρ di Spearman è preferito perché più semplice, meglio conosciuto e del tutto analogo al coefficiente parametrico r di Pearson.

Il vantaggio del test τ deriva dalla sua **estensione**

- sia all'analisi dei coefficienti di **correlazione parziale o netta** (illustrata nei paragrafi successivi), che tuttavia successivamente è stata estesa anche al ρ con **risultati equivalenti,**
- sia alla misura dell'**accordo tra giudizi multipli**.

La metodologia per stimare il **τ di Kendall** può essere suddivisa in 6 fasi: le prime due sono uguali a quelle del test **ρ di Spearman,** si differenzia per la misura dell'accordo tra le due distribuzioni.

1 - Dopo la presentazione tabellare dei dati con due misure per ogni oggetto d'osservazione

Variabili	Coppie di valori osservati						
X	8	5	7	14	22	21	41
Y	12	3	2	10	25	19	22
Oggetti	A	B	C	D	E	F	G

occorre ordinare per ranghi la variabile **X**, assegnando il rango 1 al valore più piccolo e progressivamente un rango maggiore, fino ad **N**, al valore più grande. Se sono presenti due o più valori uguali nella variabile **X**, assegnare ad ognuno come rango la media delle loro posizioni.

La **scala** comunque **dovrebbe essere continua, anche se di rango**, e quindi non avere valore identici, se non in casi eccezionali.

E' indispensabile collocare nell'ordine naturale (da 1 a N) i ranghi della variabile X, spostando di conseguenza i valori della **Y** relativi agli stessi soggetti

Variabili	Coppie di valori osservati						
X	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Y	3	2	12	10	25	16	22
Oggetti	**B**	**C**	**A**	**D**	**F**	**E**	**G**

2 - Sostituire gli **N** valori di **Y** con i ranghi rispettivi; per valori di **Y** uguali, come al solito usare la media dei ranghi.

I **ranghi di Y risultano distribuiti secondo il rango della variabile X,** come nella tabella seguente:

Variabili	Coppie di valori osservati						
X	1	2	3	4	5	6	7
Y	**2**	**1**	**4**	**3**	**7**	**5**	**6**
Oggetti	B	C	A	D	F	E	G

Il metodo proposto da Kendall utilizza le informazioni fornite dall'ordine della sola variabile Y.
E' un concetto che richiama il **metodo delle precedenze**, già utilizzate in vari test nn parametrici per il confronto tra le tendenze centrali.

3 - Se le due distribuzioni sono correlate
- in **modo positivo (r = +1)**, anche i ranghi della variabile Y sono ordinati in modo crescente, **concordanti con l'ordine naturale**;
- in **modo negativo (r = -1)**, i valori di Y risulteranno ordinati in modo decrescente e saranno **discordanti dall'ordine naturale**;

- se tra le due variabili **non esiste correlazione (r = 0), l'ordine della variabile Y risulterà casuale e il numero di ranghi concordanti e di quelli discordanti dall'ordine naturale tenderà ad essere uguale, con somma 0**.

Per quantificare il grado di correlazione o concordanza, Kendall ha proposto di **contare per la sola variabile Y**

Y	**2**	**1**	**4**	**3**	**7**	**5**	**6**

- **quante** sono le **coppie di ranghi** che sono **concordanti** e
- **quante** quelle **discordanti dall'ordine naturale**.

Per esempio, elencando in modo dettagliato tutte le singole operazioni,

- il valore 2 è seguito da 1: non è nell'ordine naturale e pertanto contribuirà con -1; inoltre è seguito da altri 5 valori maggiori, che contribuiranno insieme con +5: il contributo complessivo del valore 2 al calcolo delle concordanze è uguale a **+4**;
- il valore 1 è seguito da 5 valori maggiori e contribuirà con **+ 5**;
- il valore 4 contribuisce con -1, perché seguito dal 3, e con +3, in quanto i 3 successivi sono maggiori, per un valore complessivo di **+2**;
- il valore 3 contribuisce con **+3**;
- il valore 7 contribuisce con **-2**, in quanto seguito da 2 valori minori;
- il valore 5 contribuisce con **+1**.
- il valore 6 è l'ultimo e non fornisce alcun contributo al calcolo delle concordanze; con esso termina il calcolo delle differenze tra concordanze e discordanze.

Nella tabella seguente è riportato il conteggio dettagliato e complessivo delle concordanze (+) e delle discordanze (-)

2	**1**	**4**	**3**	**7**	**5**	**6**	**Totale**
	-	+	+	+	+	+	**+4**
		+	+	+	+	+	**+5**
			-	+	+	+	**+2**
				+	+	+	**+3**
					-	**-**	**-2**
						+	**+1**
Totale (concordanze meno discordanze)							**+13**

La misura della concordanza complessiva con la variabile X è dato dalla somma algebrica di tutte le concordanze e le discordanze.

Il totale di concordanze e discordanze dei 7 valori dell'esempio (+4, +5, +2, +3, -2, +1) è uguale a **+13.**

4 – Per ricondurre il valore calcolato a un campo di variazione compreso tra +1 e –1, il numero totale di concordanze e discordanze di una serie di valori deve essere rapportato al **massimo totale possibile**. Poiché i confronti sono fatti a coppie, con **N** dati il numero totale di confronti concordanti o discordanti è dato dalla combinazione di **N** elementi **2** a **2**

$$C_N^2$$

Con una serie di 7 dati come nell'esempio, il numero complessivo di confronti, quindi il massimo totale possibile di concordanze o discordanze, è

$$C_7^2 = \frac{7!}{(7-2)! \cdot 2!}$$

uguale a 21.

5 - Secondo il metodo proposto di **Kendall**, il grado di relazione o concordanza (τ) tra la variabile **X** e **Y** può essere quantificato dal rapporto

$$\tau = \frac{totale(concordanze - discordanze)}{massimo\ \ totale\ \ possibile} = \frac{totale(concordanze - discordanze)}{C_N^2}$$

Con i 7 dati dell'esempio,

$$\tau = \frac{+13}{21} = +0{,}619$$

τ è uguale a +0,619.

Il τ **di Kendall** varia in modo simile al coefficiente **r di Pearson**: è

- **+1**, quando la correlazione tra X e Y è massima e positiva,
- **-1**, quando la correlazione tra le due variabili è massima e negativa;
- **0,** quando non esiste alcuna correlazione.

La **formula abbreviata** è

$$\tau = \frac{2 \cdot totale(concordanze - discordanze)}{N \cdot (N-1)}$$

dove **N** è il numero di coppie di dati.

Nel caso in cui siano presenti **due o più valori identici nella successione delle Y**, il confronto con l'ordine naturale non determina né una concordanza né una discordanza: il loro confronto non contribuisce al calcolo di τ e si riduce il valore di **N**.
La mancata correzione comporterebbe che il rango di variazione non sarebbe più tra **-1** e **+1**.

Considerando la presenza di **valori identici sia nella variabile Y sia nella variabile X,** la formula corretta diventa

$$\tau = \frac{2 \cdot totale(concordanze - discordanze)}{\sqrt{N \cdot (N-1) \cdot T_x} \cdot \sqrt{N \cdot (N-1) \cdot T_y}}$$

dove

- **N** è il numero totale di coppie di dati delle variabili X e Y,
- $T_x = \sum (t_x^2 - t_x)$ dove
- t_x è il numero di osservazioni identiche di ogni gruppo di valori identici della variabile **X**,
- $T_y = \sum (t_y^2 - t_y)$ dove
- t_y è il numero di osservazioni identiche di ogni gruppo di valori identici della variabile **Y**.

Nel **caso di ties**, da L. A. **Goodman** e W. H. **Kruskal** nel 1963 (vedi l'articolo ***Measures of association for cross-classifications. III: Approximate sample theory***, pubblicato su **Journal of the American Statistical Association** Vol. 58, pp. 310 – 364) hanno proposto che τ sia stimato con la relazione

$$\tau = \frac{N_C - N_D}{N_C + N_D}$$

dove

- N_C = numero di concordanze
- N_D = numero di discordanze

Questo valore **τ** è strettamente correlato con il **coefficiente gamma** (***gamma coefficient***), tanto da poter essere identificato con esso, come sarà dimostrato nel paragrafo dedicato a tale indice; ha il grande vantaggio di variare tra +1 e –1 anche quando sono presenti dei ties.

Valori critici del coefficiente di correlazione semplice τ di Kendall per test a 1 coda e a 2 code

α

	0.05	0.025	**0.01**	0.005	**1 coda**
N	0.10	**0.05**	0.02	**0.01**	**2 code**
4	1.000				
5	.800	.800	1.000		
6	.733	.867	.867	1.000	
7	.619	.714	.810	.810	
8	.571	.643	.714	.786	
9	.500	.556	.667	.722	
10	.467	.511	.600	.644	
11	.418	.491	.564	.600	
12	.394	.455	.545	.576	
13	.359	.436	.513	.564	
14	.363	.407	.473	.516	
15	.333	.390	.467	.505	
16	.317	.383	.433	.483	
17	.309	.368	.426	.471	
18	.294	.346	.412	.451	
19	.287	.333	.392	.439	
20	.274	.326	.379	.421	
21	.267	.314	.371	.410	
22	.257	.296	.352	.391	
23	.253	.295	.344	378	
24	.246	.290	.341	.377	

Per **piccoli campioni**, i valori critici sono forniti dalla tabella relativa, riportata nella pagina precedente.
Il risultato dell'esempio, con N = 7, per un test ad 1 coda risulta significativo alla probabilità α =0.05.

Per **grandi campioni** la significatività del τ di Kendall può essere verificata
con la distribuzione normale Z

$$Z = \frac{\tau - \mu_\tau}{\sigma_\tau} \qquad (*)$$

Quando è vera l'ipotesi nulla (assenza di correlazione o d'associazione),
- per la media μ_τ vale l'uguaglianza

$$\mu_\tau = 0$$

(cioè l'ordine della variabile Y è casuale e la somma totale delle sue concordanze e discordanze è nulla),

- mentre la varianza σ_τ^2 è data da

$$\sigma_\tau^2 = \frac{2 \cdot (2N + 5)}{9N \cdot (N - 1)}$$

dove **N** è il numero di coppie di dati.

Sostituendo nella precedente relazione(*) per la normale Z e semplificando, con la **formula abbreviata** si ottiene
- una stima più rapida di **Z** mediante la relazione

$$Z = \frac{3\tau \cdot \sqrt{n \cdot (n-1)}}{\sqrt{2 \cdot (2n + 5)}}$$

Anche in questo caso sono stati proposti altri metodi per valutare la significatività di τ. Tra i testi a maggior diffusione, quello di
- W. J. **Conover** del 1999 (***Practical nonparametric statistics***, 3rd ed. John Wiley & Soons, New York, 584) riporta i valori critici dei quantili, esatti quando X e Y sono indipendenti,
- proposti da D. J. **Best** nel 1973 (nell'articolo ***Extended tables for Kendall's tau***, pubblicato su **Biometrika** Vol. 60, pp. 429-430) e nel 1974 (nella relazione ***Tables for Kendall's tau and an examination of the normal approximation***, pubblicato su **Division of Mathematical Statistics**,

Technical Paper n° 39, edito da **Commonwealth Scientific and Industrial Research Organization**, Australia)

ESEMPIO. Mediante il τ di Kendall, rispondere alla medesima domanda di verifica della significatività dell'associazione tra le variabili X e Y, utilizzando gli stessi dati dell'esercizio precedente sul ρ di Spearman.

	1	2	**1**	**2**	**3a**	**3b**	**4**
Stazione	BOD_5	N	**R(BOD_5)**	**R(N)**	**concord.**	**discord.**	**diff.**
s1	5	0,7	**1,5**	**1**	**+15**	**-0**	**+15**
s2	5	0,8	**1,5**	**2**	**+14**	**-0**	**+14**
s6	7	1,8	**3**	**5**	**+11**	**-2**	**+9**
s7	8	1,6	**4**	**3**	**+12**	**-0**	**+12**
s9	9	1,7	**5,5**	**4**	**+11**	**-0**	**+11**
s8	9	4,8	**5,5**	**11,5**	**+4**	**-4**	**0**
s16	11	2,8	**7,5**	**6**	**+9**	**-0**	**+9**
s5	11	9,7	**7,5**	**15**	**+1**	**-7**	**-6**
s3	12	5,6	**9**	**13,5**	**+1**	**-5**	**-4**
s12	13	3,2	**10,5**	**8**	**+5**	**-1**	**+4**
s15	13	3,9	**10,5**	**10**	**+3**	**-2**	**+1**
s11	14	5,6	**12**	**13,5**	**+1**	**-3**	**-2**
s14	15	4,8	**13**	**11,5**	**+1**	**-2**	**-1**
s13	16	3,6	**14**	**9**	**+1**	**-1**	**0**
s10	20	2,9	**15**	**7**	**+1**	**-0**	**+1**
s4	35	24,3	**16**	**16**	**-**	**-**	**-**
Totale differenze (concordanze – discordanze)							**63**

La metodologia del τ di Kendall richiede i seguenti passaggi (riportati nella tabella da colonna 1 a colonna 4):

1 - ordinare in modo crescente i valori del BOD_5 ed attribuire i ranghi relativi;

2 - trasformare in ranghi i corrispondenti valori di **N**;

3 - calcolare per ogni punteggio di **N** il numero di concordanze e di discordanze;

4 - calcolare la somma complessiva di tutte le concordanze e le discordanze.

La somma totale delle differenze tra concordanze e discordanze risulta positiva (+63).

5 - Tradotto nel corrispondente coefficiente mediante

$$\tau = \frac{2 \cdot (+63)}{16 \cdot 15} = \frac{+126}{240} = +0{,}525$$

si ottiene un valore di τ uguale a +0,525.

6 - Per un test unilaterale, la tabella dei valori critici del τ di Kendall
- con **N** = 16 e alla probabilità α **= 0.005**
- riporta un valore di τ uguale a 0,483.

Il valore calcolato (0,525) è superiore in modulo.

Di conseguenza, si rifiuta l'ipotesi nulla e si accetta l'ipotesi alternativa: esiste un'associazione o correlazione positiva tra le due serie di dati, con probabilità P < 0.005 di commettere un errore di I tipo.

Il campione utilizzato nell'esempio può essere ritenuto sufficientemente grande.
Pertanto, è possibile valutare la significatività del coefficiente τ = + 0,525
mediante il **test Z**:

$$Z = \frac{3\tau \cdot \sqrt{n \cdot (n-1)}}{\sqrt{2 \cdot (2n+5)}} = \frac{3 \cdot (+0{,}525) \cdot \sqrt{16 \cdot 15}}{\sqrt{2 \cdot (32+5)}} = \frac{+1{,}575 \cdot \sqrt{240}}{\sqrt{69}} = \frac{+24{,}340}{8{,}307} = +2{,}93$$

che risulta Z = +2,93.
Nella distribuzione normale, a **Z** uguale a 2,93 per un test ad una coda corrisponde una probabilità **P =** 0,0017. E' un risultato che non si discosta in modo rilevante da quello precedente, fornita dalle tabelle dei valori critici.

Alcuni testi di statistica presentano una procedura di calcolo delle precedenze che è più complessa di quella illustrata e propongono 2 misure differenti (τ_a, τ_b); la scelta tra τ_a e τ_b dipende dal numero di valori identici e quindi dalla continuità del tipo di scala utilizzato.
E' possibile determinare i casi concordi, discordi oppure a pari merito, confrontando simultaneamente i valori di **X** e **Y** in una coppia d'oggetti.

Una coppia di casi è
- **concorde (P),** se per un oggetto i valori di entrambi le variabili sono più bassi o più alti rispetto ai valori dell'altro caso;
- **discorde (Q)**, se per una variabile è maggiore e per l'altra minore, o viceversa;
- **pari merito** (**T**), se hanno lo stesso valore per la variabile **X** (T_X) o per la variabile Y (T_Y).

Il τ_a è la differenza tra coppie concordi e discordi (P-Q), **rapportata al numero totale di coppie d'oggetti**:

$$\tau_a = \frac{P-Q}{C_n^2}$$

Se **non esistono coppie con valori uguali**, questa misura varia tra -1 e +1.
Se **esistono coppie con valori uguali**, il campo di variazione è più limitato e dipende dal numero di valori pari merito presenti sia nella variabile **X** che nella variabile **Y**.

Il τ_b normalizza la differenza P-Q, prendendo in considerazione anche i valori pari merito delle due variabili in modo separato

$$\tau_b = \frac{P-Q}{\sqrt{\left(P+Q+T_X\right)\cdot\left(P+Q+T_Y\right)}}$$

L'associazione tra due variabili può essere valutate anche con altri metodi, che utilizzano tabelle di contingenza.

21.3. CONFRONTO TRA ρ E τ; POTENZA DEL TEST E NUMERO DI OSSERVAZIONI NECESSARIE PER LA SIGNIFICATIVITA'.

I coefficienti di correlazione non parametrica ρ di Spearman e τ di Kendall richiedono variabili almeno di tipo ordinale. Se i valori sono misurati su una scala ad intervalli o di rapporti, le osservazioni devono essere trasformate nei loro ranghi. Anche con ranghi, è possibile **calcolare il coefficiente di correlazione r di Pearson, utilizzandoli appunto al posto dei valori rilevati.** E' interessante osservare che il risultato della correlazione non parametrica **ρ di Spearman** coincide con quello ottenuto mediante il **metodo r di Pearson**, quando sono utilizzati i ranghi. E' una convergenza tra test parametrico e non parametrico corrispondente, già evidenziata per altri test:

- per l'ANOVA con la varianza non parametrica di Kruskall-Wallis,
- per il test t di Student con il test U di Mann-Whitney.

Nonostante questa coincidenza dei risultati, è importante comprendere che la correlazione parametrica e quella non parametrica analizzano **caratteristiche differenti della relazione** esistente tra le due variabili. Mentre

- la correlazione parametrica di **Pearson** valuta la significatività di una **correlazione di tipo lineare**,

- la correlazione non parametrica di **Spearman** e di **Kendall** valutano l'esistenza della **monotonicità**; è una condizione più generale, realizzata sempre quando esiste regressione lineare.

In altri termini, la correlazione non parametrica

- risulta + 1 quando all'aumentare della prima variabile aumenta anche la seconda,
- risulta – 1 quando all'aumentare della prima la seconda diminuisce,

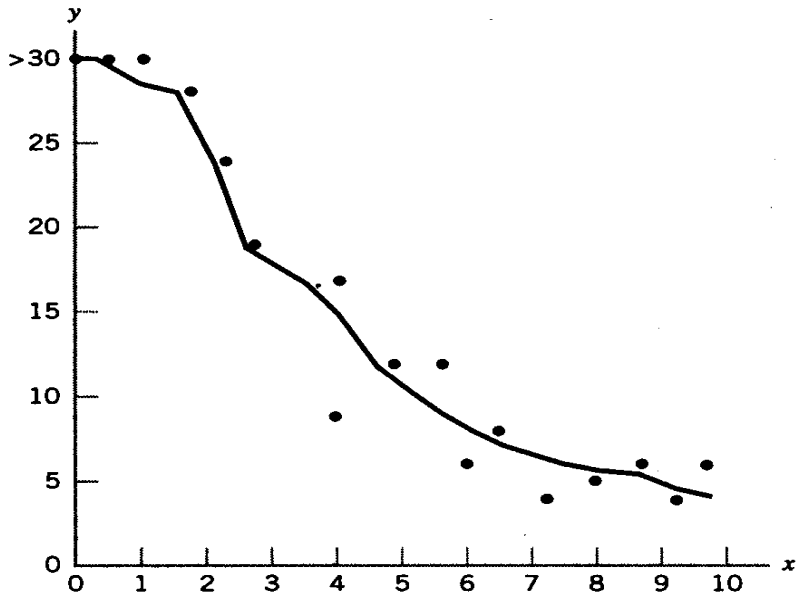
ma **senza richiedere che tali incrementi siano costanti**, come per la retta.

Con termini più tecnici, il concetto è che

- **se due variabili hanno una regressione monotonica**,
- **i loro ranghi hanno una relazione lineare**.

Il testo di W. J. **Conover** del 1999 (***Practical nonparametric statistics***, 3rd ed. John Wiley & Soons, New York, 584) mostra il diagramma di dispersione e il tipo di relazione tra le due variabili nella seguente serie di valori

X	0	0,5	1,0	1,8	2,2	2,7	4,0	4,0	4,9	5,6	6,0	6,5	7,3	8,0	8,8	9,3	9,8
Y	>30	>30	>30	28	24	19	17	9	12	12	6	8	4	5	6	4	6



CURVA DI REGRESSIONE MONOTONICA CON I DATI OSSERVATI

Nella tabella e nel grafico,

- in 17 contenitori e con una osservazione della durata di 30 giorni,
- X è la quantità di zucchero aggiunta al mosto d'uva,
- Y è il numero di giorni dopo i quali ha avuto inizio la fermentazione.

Nelle tre quantità minori, (cioè con X uguale a 0 poi 0,5 e 1,0) la fermentazione non aveva ancora avuto inizio dopo 30 giorni di osservazione (> 30).

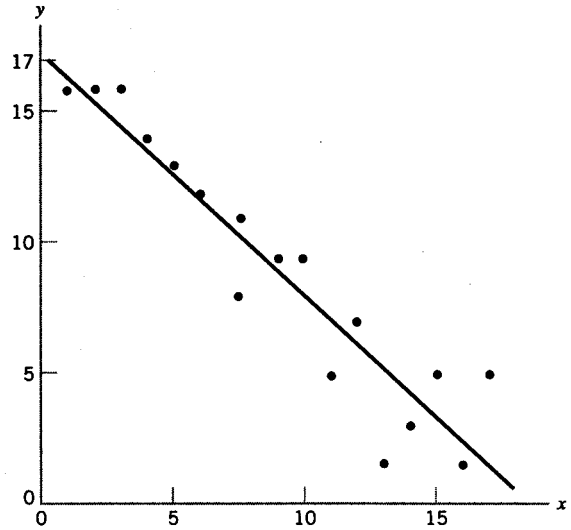
Con i dati originali, riportati nella tabella e nel grafico, per utilizzare la correlazione **r** di Pearson

- un primo problema è rappresentato dalla presenza di dati stimati con approssimazione, addirittura troncati "*censored*" come >30, per cui non è possibile il calcolo né delle medie né della codevianza e delle devianze;
- il secondo problema è una linearità dei dati molto approssimata, come la rappresentazione grafica evidenzia visivamente.

Quando al posto dei valori si utilizzano i ranghi relativi (nella tabella successiva i valori precedenti sono stati trasformati in ranghi)

Rx	1	2	3	4	5	6	7,5	7,5	9	10	11	12	13	14	15	16	17
Ry	16	16	16	14	13	12	11	8	9,5	9,5	5	7	1,5	3	5	1,5	5

la rappresentazione grafica evidenzia la differente distribuzione lineare dei punti



DISPOSIZIONE DEI PUNTI E RETTA DI REGRESSIONE OTTENUTA CON LA TRASFORMAZIONE DEI DATI IN RANGHI

Questi concetti, come calcolare i punti della curva segmentata della figura precedente e come calcolare la retta con in ranghi in quella sovrastane sono sviluppati nel paragrafo dedicato alla regressione monotonica di Iman-Conover.

Per quanto attiene la **potenza dei due test**, il **ρ di Spearman e il di τ Kendall hanno la stessa potenza nel rifiutare l'ipotesi nulla**, anche se i valori di ρ e τ sono numericamente differenti per lo stesso campione di dati.
Stime del**l'efficienza asintotica relativa di Pitman per il test τ di Kendall,** ovviamente rispetto al test parametrico **r di Pearson** e nel caso che l'ipotesi nulla sia vera, riportano che;
- quando la distribuzione dei dati è Normale, la potenza del τ è uguale a 0,912 $(3/\pi)^2$;
- quando la distribuzione dei dati è Rettangolare, la potenza del τ è uguale a 1;
- quando la distribuzione dei dati è Esponenziale Doppia, la potenza del τ è uguale a 1,266 (81/64).

Quando l'**ipotesi nulla H_0 è vera**, le probabilità α fornite dai due metodi sono **molto simili**; per grandi campioni distribuiti normalmente, esse tendono ad essere molto simili.
Ma quando l'**ipotesi nulla H_0 è falsa,** quindi si accetta come vera l'ipotesi alternativa H_1, i due differenti indici **sono diversamente sensibili alle distorsioni determinate dal diverso campo di variazione** (questi concetti sono sviluppati nel capito della correlazione parametrica e dell'intervallo di confidenza di **r**); di conseguenza, i risultati tendono a differire maggiormente.

Con i dati dell'esempio utilizzato nel paragrafo precedente, per un test unilaterale
- con il ρ **di Spearman** è stato ottenuto un valore di **Z** = 2,62 corrispondente alla probabilità α = 0.0044
- con il τ **di Kendall** è stato ottenuto un valore di **Z** = 2,93 che corrisponde a una probabilità α = 0.0017.

La differenza tra le probabilità stimate con i due diversi indici è in assoluto inferiore al 3/1000 e quindi oggettivamente molto limitata; ma è elevata (2,59 a 1), se considerata in rapporto alle piccole probabilità stimate. E' uno dei problemi che si pone nella valutazione dei risultati: se è più corretto fornire una stima in termini assoluti oppure in termini relativi.
Per il confronto tra ρ e τ, al momento non è noto quale indice in generale dia il valore più corretto.

Quanti dati è necessario raccogliere perché una regressione non parametrica sia significativa?

Secondo la proposta di G. E. **Noether** del 1987 (vedi articolo ***Sample size determination for some common nonprametric tests***, su **Journal of the American Statistical Association** Vol. 82, pp. 68-79), riportata nel testo di P. **Sprent** e N. C. **Smeeton** del 2001 (***Applied nonparametric statistical methods***, 3rd ed. Chapman & Hall/CRC, 461 p.) e in quello di M. **Hollander** e D. A. **Wolfe** del 1999 (***Nonparametric Statistical Methods***, 2nd ed., New York, , John Wiley & Sons) una stima approssimata del numero (**n**) di dati necessari affinché un valore τ_1 di correlazione non parametrica sia significativo alla probabilità α e con rischio β

è data dalla relazione

$$n \approx \frac{4 \cdot \left(Z_\alpha + Z_\beta\right)^2}{9 \cdot \tau_1^2}$$

dove

- τ_1 è il valore di τ che si vuole risulti significativo rispetto all'ipotesi nulla H_0: $\tau = 0$
- α è la probabilità o rischio di I Tipo, scelta per il test, la cui ipotesi alternativa può essere bilaterale oppure unilaterale
- β è la probabilità o rischio di II tipo di non trovare una differenza che in realtà esiste,
- ricordando che, per prassi e in accordo con l'indicazione di Cohen, la probabilità β è scelta con un rapporto di circa 5 a 1 rispetto a α.

ESEMPIO (CON TEST BILATERALE). Una analisi preliminare di una serie di rilevazioni ha permesso di stimare un valore di correlazione non parametrica $\tau = 0{,}3$.
Quanti dati (n) occorre raccogliere perché tale valore risulti significativamente differente da 0, in un test bilaterale alla probabilità $\alpha = 0.05$ e con un rischio $\beta = 0.20$ (quindi con una potenza $1 - \beta = 0.80$)?

Risposta. Dalla tabella della distribuzione normale, si ricava

- per α **= 0.05 bilaterale**, $\mathbf{Z_\alpha = 1{,}96}$
- per β **= 0,20** (sempre **unilaterale**), $\mathbf{Z_\beta = 0{,}84}$

Da essi risulta che

$$n \approx \frac{4 \cdot \left(Z_\alpha + Z_\beta\right)^2}{9 \cdot \tau_1^2} = \frac{4 \cdot \left(1{,}96 + 0{,}84\right)^2}{9 \cdot 0{,}3^2} = \frac{31{,}36}{0{,}81} = 38{,}7$$

che il numero di dati necessario è almeno 39.

Per rifiutare l'ipotesi nulla $\mathbf{H_0: \tau = 0}$ ed accettare implicitamente l'ipotesi alternativa bilaterale $\mathbf{H_1: \tau \neq 0}$ alla probabilità $\boldsymbol{\alpha = 0.05}$ e con una potenza $\mathbf{1-\beta = 0.20,}$ con $\boldsymbol{\tau_1 = 0,3}$ servono almeno 39 osservazioni.

ESEMPIO (CON TEST UNILATERALE). Nell'esempio precedente, quanti dati è necessario raccogliere se il test che si vuole utilizzare è unilaterale?

Risposta. Dalla tabella della distribuzione normale, si ricava

- per $\boldsymbol{\alpha = 0.05}$ **unilaterale**, $\mathbf{Z_\alpha = 1,645}$
- per $\boldsymbol{\beta = 0,20}$ (sempre **unilaterale**), $\mathbf{Z_\beta = 0,84}$

Da essi risulta che

$$n \approx \frac{4 \cdot \left(Z_\alpha + Z_\beta\right)^2}{9 \cdot \tau_1^2} = \frac{4 \cdot \left(1{,}645 + 0{,}84\right)^2}{9 \cdot 0{,}3^2} = \frac{24{,}70}{0{,}81} = 30{,}5$$

che il numero di dati necessario è almeno 31.

Se la potenza ($\mathbf{1 - \beta}$) è 0.90, quindi con $\boldsymbol{\beta = 0.10}$ il cui valore di $\mathbf{Z_\beta = 1,28}$ il numero minimo di dati necessari

$$n \approx \frac{4 \cdot \left(Z_\alpha + Z_\beta\right)^2}{9 \cdot \tau_1^2} = \frac{4 \cdot \left(1{,}645 + 1{,}28\right)^2}{9 \cdot 0{,}3^2} = \frac{34{,}22}{0{,}81} = 42{,}2$$

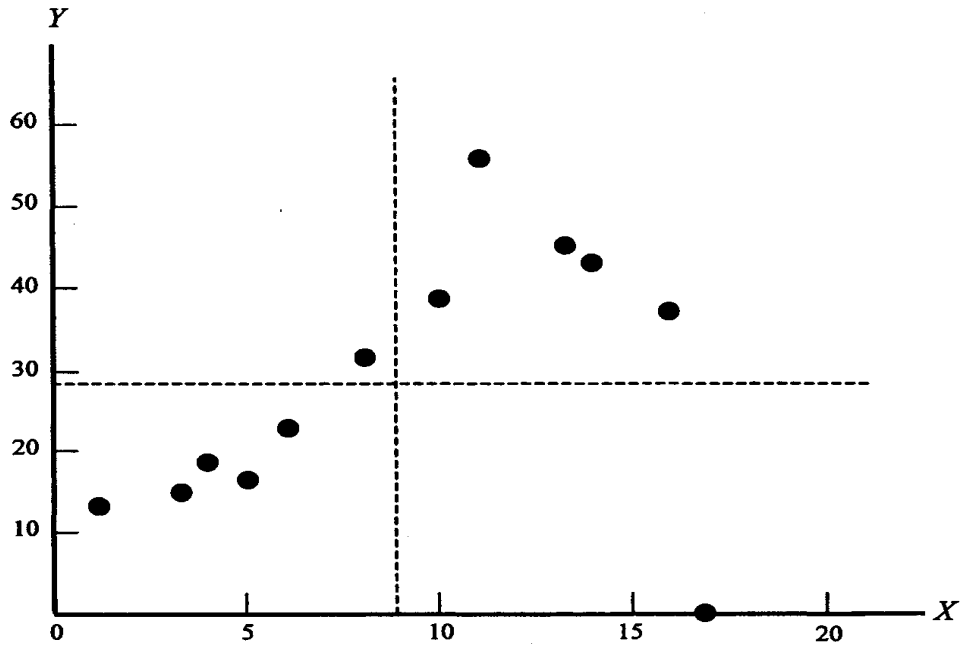
diventa $n \geq 43$.

21.4. ALTRI METODI PER LA CORRELAZIONE NON PARAMETRICA: TEST DI PITMAN CON LE PERMUTAZIONI; TEST DELLA MEDIANA DI BLOMQVIST.

Per l'analisi della correlazione non parametrica, sono stati proposti altri metodi molto meno noti poiché riportati raramente sia nei testi di statistica applicata sia nelle librerie informatiche a grande diffusione internazionale. Di norma, sono test meno potenti del ρ e del τ, in quanto fondati su condizioni di validità più generali; altri, in situazioni contingenti, offrono alcuni vantaggi pratici.
Nella ricerca applicata è quindi utile la conoscenza di alcuni di questi metodi.

Tra essi, riportati nell'ultima versione nel testo di P **Sprent** e N. C. **Smeeton** del 2001 dal titolo ***Applied Nonparametric Statistical Methods*** (Chapman & Hall/CRC, London, 461 p.) possono essere ricordati:

- il **test di Pitman**, quando si dispone di un numero di dati particolarmente ridotto;
- il **test della mediana di Blomqvist**, quando le due serie di misure sono approssimate o contengono valori con attendibilità differente, con tutti i vantaggi e gli svantaggi propri dei test della mediana già presentati.



Si supponga che questo diagramma di dispersione, tratto dal testo di Sprent e Smeeton citato, sia la rappresentazione grafica del numero di errori commessi da 12 allievi impegnati prima in un compito di matematica (M) e successivamente in uno di lingua straniera (L):

Matematica	(X)	1	3	4	5	6	8	10	11	13	14	16	17
Lingua	(Y)	13	15	18	16	23	31	39	56	45	43	37	0

Che tipo di relazione esiste tra le due serie di dati? Può essere vera la teoria che afferma che gli studenti migliori in matematica sono i migliori anche nell'apprendimento delle lingue?
L'interpretazione, sempre necessaria dal punto di vista disciplinare seppure non richiesta dall'analisi statistica, potrebbe essere che i migliori in matematica sono tali perché più diligenti, logici e studiosi; quindi, con poche eccezioni, anche i migliori in tutte le altre discipline, tra cui lo studio della lingua.

Ma può essere ugualmente convincente anche la teoria opposta.
Chi è portato alla logica matematica ha poca attitudine per l'apprendimento alle lingue; inoltre la conoscenza delle lingue straniere richiedono attività e impegni, come i viaggi, i soggiorni all'estero e i contatti con le persone, che male si conciliano con lo studio e la riflessione richiesti dalla matematica.

I risultati della tabella e la loro rappresentazione nel diagramma di dispersione sembrano complessivamente (all'impressione visiva che tuttavia deve essere tradotta nel calcolo delle probabilità con un test) deporre a favore della prima teoria; ma il punto anomalo, l'allievo che ha commesso più errori in matematica e nessuno in lingua, forse per condizioni familiari particolari, è un dato importante a favore della seconda teoria.
Come tutti i valori anomali, questo dato da solo sembra in grado di contraddire l'analisi fondata su tutti gli altri, almeno di annullarne le conclusioni. I metodo parametrici, che ricorrono al quadrato degli scarti, danno un peso rilevante a questi dati anomali; ma la loro presenza evidenzia una condizione di non validità di tale analisi.

Per quanto attiene la verifica statistica di queste teorie mediante l'analisi della regressione, i dati riportati, costruiti ad arte ma verosimili nella ricerca applicata come affermano i due autori, rappresentano un esempio didattico che bene evidenzia quattro caratteristiche dei dati, da discutere sempre nella scelta del test più adatto per analisi con la correlazione:
- il tipo di scala,
- la normalità della distribuzione,
- l'omogeneità della varianza,
- la presenza di un valore anomalo.

Anche questa breve introduzione evidenzia quanto sia utile avere un quadro ampio delle opportunità offerte dai test, per scegliere sempre quello con la potenza maggiore, in rapporto alle caratteristiche dei dati e nel pieno rispetto dei presupposti di validità.

Nel 1937, E. J. G. **Pitman ha proposto un test di correlazione non parametrica** (vedi l'articolo ***Significance tests that may applied to samples fron any population, II: The correlation coefficient***

test, pubblicato su **Journal of the Royal Statistical Society**, Suppl. 4, pp. 225-232), che utilizza il calcolo combinatorio già illustrato nei test di casualizzazione per due campioni dipendenti e per due campioni indipendenti.

Per verificare l'ipotesi nulla sulla correlazione (**H_0: $\rho = 0$)** contro un'ipotesi alternativa sia bilaterale (**H_1: $\rho \neq 0$),** che può essere anche unilaterale in una delle due direzioni (**H_1: $\rho < 0$** oppure **H_1: $\rho > 0$)**, propone un test esatto, fondato sulle **permutazioni**.

Sviluppando un esempio didattico con soli 4 coppie di dati,

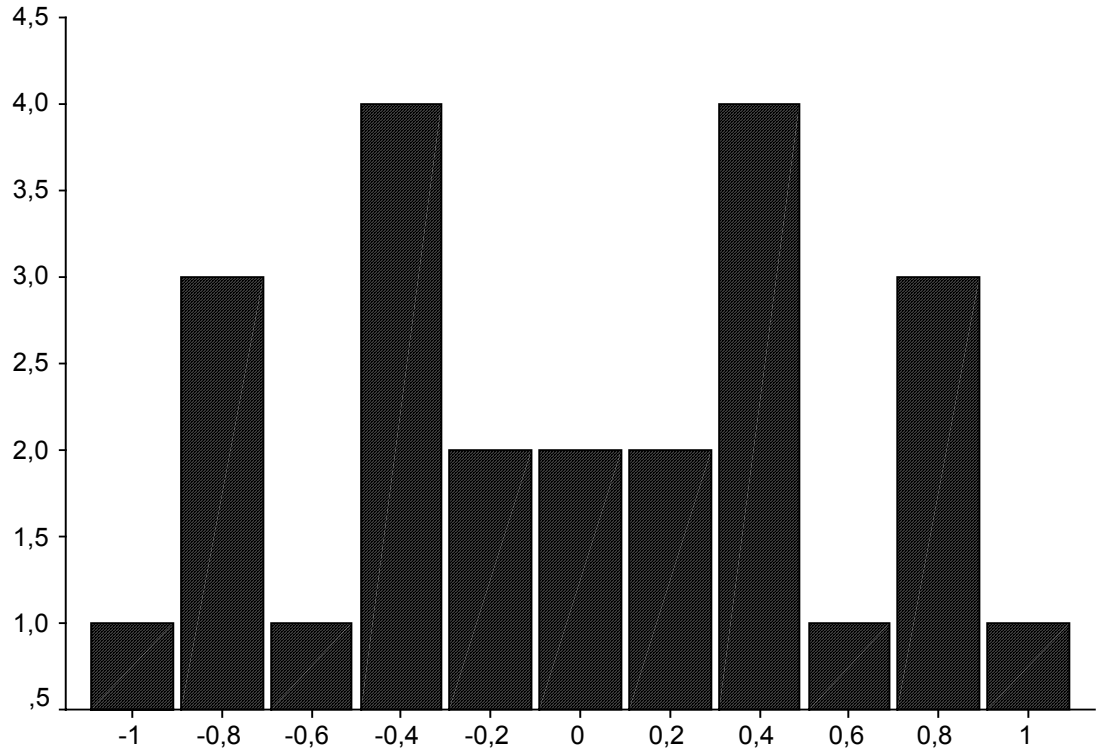
- dopo aver ordinato i valori della variabile X in modo crescente, come nei metodi di Spearman e di Kendall,

- prende in considerazione la variabile Y, stimandone tutte le possibili permutazioni, che con **n** dati sono **n!**

Con n = 4 esse sono 4! = 24 come nella tabella successiva, nella quale sono organizzate in modo logico e con valori di ρ tendenzialmente decrescente.

Elenco dei casi	Rango 1	Rango 2	Rango 3	Rango 4	ρ
1	1	2	3	4	+1,0
2	1	2	4	3	+0,8
3	1	3	2	4	+0,8
4	1	3	4	2	+0,4
5	1	4	2	3	+0,4
6	1	4	3	2	+0,2
7	2	1	3	4	+0,8
8	2	1	4	3	+0,6
9	2	3	1	4	+0,4
10	2	4	1	3	0,0
11	2	3	4	1	-0,2
12	2	4	3	1	-0,4
13	3	1	2	4	+0,4
14	3	2	1	4	+0,2
15	3	1	4	2	0,0
16	3	2	4	1	-0,4
17	3	4	1	2	-0,6
18	3	4	2	1	-0,8
19	4	1	2	3	-0,2
20	4	1	3	2	-0,4
21	4	2	1	3	-0,4
22	4	2	3	1	-0,8
23	4	3	1	2	-0,8
24	4	3	2	1	-1,0

Tale distribuzione può essere riassunta in una tabella, qui organizzata in modo crescente per il valore di ρ



DISTRIBUZIONE DEI VALORI DI CORRELAZIONE

Valori di ρ	-1,0	-0,8	-0,6	-0,4	-0,2	0,0	+0,2	+0,4	+0,6	+0,8	+1,0
N	1	3	1	4	2	2	2	4	1	3	1
Freq. Rel.	0,042	0,125	0,042	0,166	0,083	0,084	0,083	0,166	0,042	0,125	0,042

(Le frequenze relative sono arrotondate alla terza cifra, affinché il totale sia 1,00)

Con la distribuzione di frequenza di tutte le permutazioni, **si rifiuta l'ipotesi nulla H_0 quando la serie osservata degli Y campionari è collocata agli estremi della distribuzione**, nella zona di rifiuto

Nel caso dell'esempio, se l'ipotesi alternativa fosse stata H_1: $\rho > 0$ si sarebbe potuto rifiutare l'ipotesi nulla solo se la distribuzione campionaria fosse stata quella estrema, riportata per ultima nella tabella con l'ordine crescente dei valori, cioè quella con i ranghi 1, 2, 3, 4 in ordine naturale.

Infatti la probabilità di trovarla per caso, nella condizione che H_0 sia vera, è $P < 0.05$ (esattamente $P = 0.042$, avendo frequenza 1/24).
La diffusione dei computer e di programmi informatici appropriati permette l'uso di questo metodo.
Con pochi dati, come con 4 ranghi, la distribuzione delle probabilità ha forma simmetrica ma non normale, evidenziata visivamente dalla figura precedente. Per un numero più alto di osservazioni, la distribuzione delle probabilità tende alla normale, ricordando che il numero delle permutazioni cresce rapidamente all'aumentare di n!
Ad esempio con 8! è già 40320 e con 10! è addirittura 3628800.

Nel caso qui presentato, il metodo delle permutazioni è stato applicato ai ranghi per semplicità didattica. In realtà **questo test di permutazione dovrebbe essere applicato ai valori osservati**, richiedendo le stesse condizioni di validità di tutti i test di permutazione, cioè la **distribuzione normale dei dati**. In tale condizione il valore di correlazione varia da –1 a +1 e la sua efficienza asintotica relativa o efficienza di **Pitman** è stimata uguale a 1. Ovviamente offre gli stessi vantaggi dei test di permutazione, con campioni piccoli:
- la possibilità di calcolare direttamente la significatività, con un numero minimo di dati;
- la stima delle probabilità esatte.

Per grandi campioni è inapplicabile, anche limitando l'analisi ai valori collocati nella zona di rifiuto, per il numero rapidamente crescente delle permutazioni.

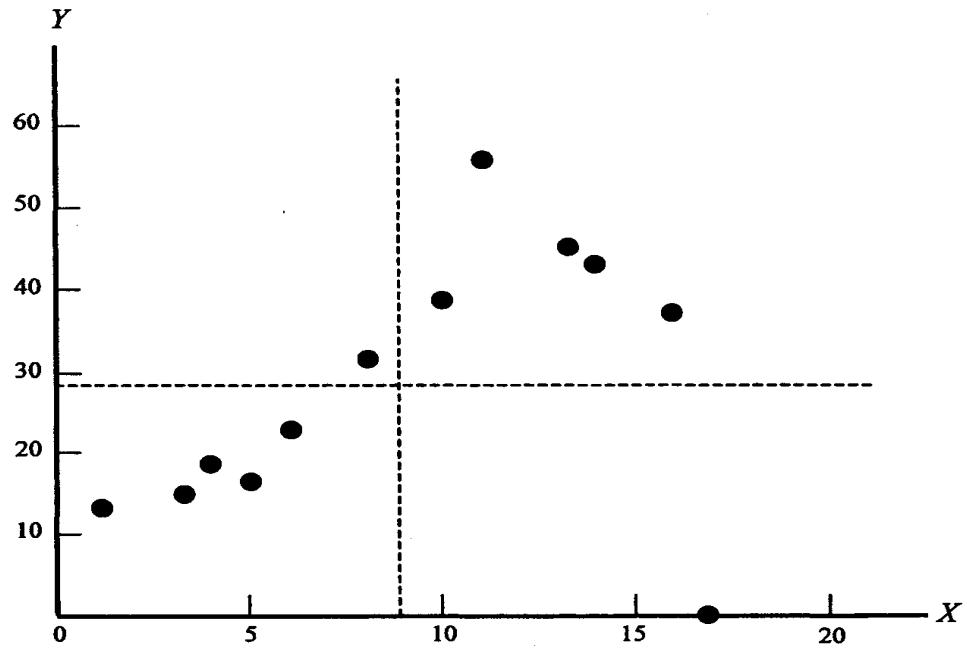
Il **test della mediana per la correlazione**, analogo al test della mediana per due campioni, in realtà è un **test di associazione;** nell'articolo di presentazione, dall'autore (Blomqvist) è stato chiamato **test di dipendenza tra due variabili casuali**.
Il metodo è stato proposto nel 1950 da N. **Blomqvist** con l'articolo ***On a mesure of dependence between two random variables*** (pubblicato su **The Annals of Mathematical Statistics**, Vol. 21, pp. 593-600); nel 1951 è stato ripreso sulla stessa rivista in un elenco di test analoghi, nell'articolo ***Some tests based on dichotomization*** (**The Annals of Mathematical Statistics**, Vol. 22, pp. 362-371).
Questo test è utile in particolare quando la distribuzione dei dati si allontana dalla normalità, come nell'esempio riportato in precedenza.

I dati rappresentati nel diagramma di dispersione possono essere riportati in un tabella di contingenza 2 x 2, considerando

- la collocazione dei punti sopra o sotto la mediana,
- congiuntamente per la variabile X e la variabile Y
- in modo da rispettare la loro collocazione grafica.

I punti del diagramma di dispersione, nel quale le due rette perpendicolari rappresentano la mediana di X e la mediana di Y, permettono di costruire la tabella 2 x 2 con facilità:

Dal grafico



è semplice ricavare la tabella di contingenza 2 x 2:

		Variabile X		
		< Mediana	> Mediana	Totale
Variabile Y	> Mediana	***1***	***5***	6
	< Mediana	***5***	***1***	6
	Totale	6	6	12

Se l'ipotesi nulla H_0 fosse vera, nel grafico e nella tabella da esso ricavata, i valori della variabile X e della variabile Y dovrebbero essere distribuiti in modo indipendente; perciò avere frequenze simili nei quattro quadranti e quindi nelle quattro caselle.

La casualità della distribuzione osservata può essere verificata con il metodo esatto di Fisher.

Esso permette anche di stimare la probabilità esatta in un test unilaterale. L'uso dei computer ha esteso con facilità il calcolo a grandi campioni.

Riprendendo i concetti fondamentali di tale metodo applicati a questo caso, per verificare l'ipotesi

$$\mathbf{H_0: \rho \leq 0} \quad \text{contro} \quad \mathbf{H_1: \rho > 0}$$

- dapprima si calcola la probabilità di ottenere la risposta osservata (dove il valore minimo in una casella è 1)

$$P_1 = \frac{6!\cdot 6!\cdot 6!\cdot 6!}{1!\cdot 5!\cdot 5!\cdot 1!\cdot 12!} = 0{,}03896$$

- successivamente quella delle risposte più estreme nella stessa direzione, che in questo caso è solamente una:

		Variabile X		
		< Mediana	> Mediana	Totale
Variabile Y	> Mediana	***0***	***6***	6
	< Mediana	***6***	***0***	6
	Totale	6	6	12

$$P_0 = \frac{6!\cdot 6!\cdot 6!\cdot 6!}{0!\cdot 6!\cdot 6!\cdot 0!\cdot 12!} = 0{,}00108$$

- Infine, per semplice somma, si ricava la probabilità di avere la risposta osservata e tutte quelle più estreme nella stessa direzione (in questo caso solo una), nella condizione che H_0 sia vera

$$P = P_1 + P_0 = 0{,}03896 + 0{,}00108 = 0{,}04004$$

Con i dati dell'esempio, si ricava una probabilità totale inferiore al 5%, che indica di rifiutare l'ipotesi nulla e quindi implicitamente di accettare quella alternativa.

Se il test fosse stato bilaterale, tale probabilità dovrebbe essere raddoppiata.

Per le indicazioni che ne derivano sulla scelta del test più appropriato, è interessante confrontare questo risultato con quello ottenuto applicando il test **r** di Pearson, ρ di Sperman e τ di Kendall agli stessi dati del grafico, cioè alla serie bivariata:

M	(X)	1	3	4	5	6	8	10	11	13	14	16	17
L	(Y)	13	15	18	16	23	31	39	56	45	43	37	0

Utilizzando un programma informatico a grande diffusione, sono stati ottenuti i risultati dei tre test.

A) Con il **test r di Pearson** il valore di correlazione è risultato **r = 0,373**.
In un test unilaterale, ad esso corrisponde una probabilità **P = 0.116**.
Non solo non permette di rifiutare l'ipotesi nulla, ma indurrebbe a ritenere che l'ipotesi nulla sia vera, dato il valore elevato della probabilità P stimata.

B) Con il **test ρ di Spearman** si è ottenuto un valore di correlazione **ρ = 0, 434**.
In un test unilaterale, ad esso corrisponde una probabilità **P = 0.080** che non permette di rifiutare l'ipotesi nulla.

C) Con il **test τ di Kendall** si è ottenuto un valore di correlazione τ = 0, 424; più esattamente il programma riporta **τ_b = 0, 424**.
In un test unilaterale, ad esso corrisponde una probabilità **P = 0.027** che permette di rifiutare l'ipotesi nulla.

Questo confronto tra i tre metodi che analizzano la correlazione con gli stessi dati evidenzia la forte differenza, attesa a causa del tipo di distribuzione, tra i due test non parametrici e quello parametrico. Ma evidenzia anche una marcata differenza, inattesa anche se nota ed effetto della forte anomalia della distribuzione, tra ρ e τ.
Purtroppo non è stato ancora proposto un metodo per raccordare logicamente due probabilità così differenti e trarne una decisione finale condivisa.

In conclusione, con i dati dell'esempio, il test più semplice, fondato sui segni (P = 0.04) e quindi con condizioni di validità più generali e rispettate anche in questo caso, si dimostra il più potente e il più corretto.

Per una esatta comprensione dei metodi e una dizione corretta delle conclusioni, è importante ricordare ancora che i tre tipi di test verificano l'esistenza di rapporti differenti tra le due variabili. Se un test risulta significativo,

- nella **correlazione parametrica** dimostra l'esistenza di una **relazione lineare,**
- mentre nella **correlazione non parametrica** dimostra l'esistenza di una **relazione monotonica**;
- con il **test della mediana** dimostra solamente che esiste **associazione** (se positiva o negativa dipende dal segno) tra valori alti e bassi delle due variabili.

21.5. IL TEST DI DANIELS PER IL TREND

Nel 1950, H. E. **Daniels** (con l'articolo ***Rank correlation and population models***, pubblicato su **Journal of the Royal Statistical Society** (B), vol. 12, pp. 171-181) ha proposto di utilizzare il **test ρ di Spearman**

- per verificare se **nel tempo una variabile cambia in modo monotonico**, con ipotesi sia unilaterale che bilaterale come è possibile per la correlazione,
- contro l'ipotesi nulla che essa si mantenga costante, cioè che **il tempo e l'altra variabile rilevata siano mutuamente indipendenti**.

Il test può essere facilmente esteso ad una successione spaziale.

Per gli stessi scopi, vari ricercatori utilizzano anche il **test τ di Kendall**.

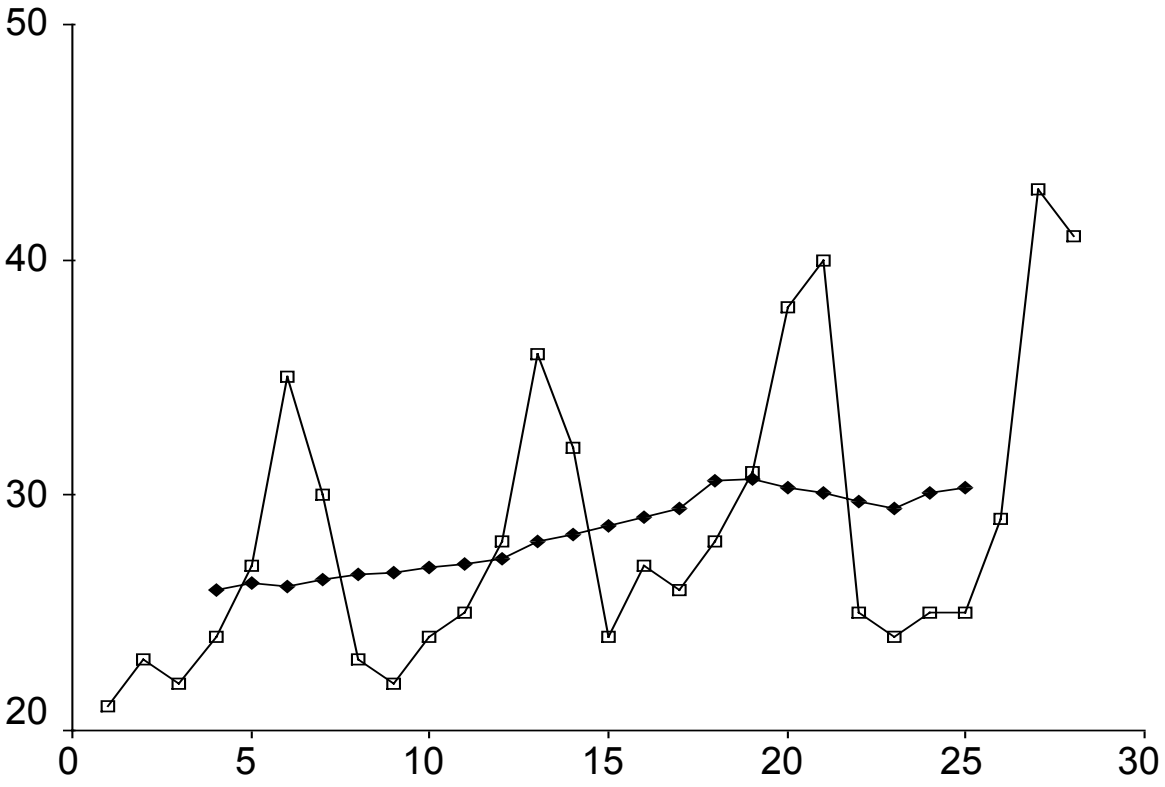
I risultati sono del tutto analoghi a quelli del ρ, seppure non coincidenti, per i motivi illustrati nei paragrafi precedenti. Come negli altri casi, il vantaggio del test ρ rispetto al test τ è di essere stimato in modo più semplice e rapido. E' un aspetto sempre importante nella pratica della statistica, quando i calcoli sono svolti manualmente.

Questi due test **possono essere applicati a dati per i quali è già stato proposto** il test di **Cox e Stuart**, cioè ad una successione temporale di valori. Pertanto, con la finalità di presentare l'applicazione del metodo ρ e di analizzare le situazioni nei quali **scegliere quello più adeguato**, è utile riprendere gli stessi dati utilizzati per il test di Cox e Stuart.

Si supponga di avere la successione temporale di 28 osservazioni, riportate nella tabella e nel grafico successivi. Si vuole valutare se è confermata l'ipotesi di una tendenza significativa all'aumento dei valori medi.

1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	11°	12°	13°	14°
I Settimana							II Settimana						
L	M	M	G	V	S	D	L	M	M	G	V	S	D
21	**23**	**22**	**24**	**27**	**35**	**30**	**23**	**22**	**24**	**25**	**28**	**36**	**32**

15°	16°	17°	18°	19°	20°	21°	22°	23°	24°	25°	26°	27°	28°
III Settimana							IV Settimana						
L	M	M	G	V	S	D	L	M	M	G	V	S	D
24	**27**	**26**	**28**	**31**	**38**	**40**	**25**	**24**	**25**	**25**	**29**	**43**	**41**



Rappresentazione grafica dei dati e della loro media mobile a 7 elementi

A differenza dei test parametrici, che richiedono espressamente i dati delle singole osservazioni e non medie, in quanto indispensabili per calcolare la covarianza e la stima della varianza d'errore, l'analisi non parametrica può essere condotta anche sulle medie, sulle mediane o altri quantili ritenuti importanti (come il 25° e il 75° percentile). Infatti usando le medie si ottiene direttamente la deviazione standard delle medie, cioè l'errore standard.

Nel caso dell'esempio, per valutare la tendenza di fondo del periodo servendosi delle singole osservazioni raccolte, si pone il problema della grande variabilità presente nell'arco di una settimana, come evidenzia la successione dei dati e soprattutto mostra visivamente la rappresentazione grafica. La forte oscillazione di periodo (in questo caso settimanale) tende a nascondere il cambiamento sistematico ipotizzato (una differenza monotonica tra l'inizio e la fine mese). Inoltre, sono presenti molti valori uguali, che rendono il calcolo dei ranghi più complesso e soprattutto fanno diventare il risultato del test approssimato, dovendo ricorrere a medie dei ranghi.
Di conseguenza, per applicare il test di Daniels appare conveniente utilizzare la successione delle medie mobili, già calcolate nel test di Cox e Stuart. Il test di Daniels non richiede attenzione alle variazioni cicliche e quindi non richiede che il periodo sia suddiviso in 2 fasi corrispondenti.

1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	11°	12°	13°	14°
---	**---**	**---**	**26,0**	**26,3**	**26,1**	**26,4**	**26,6**	**26,7**	**26,9**	**27,1**	**27,3**	**28,0**	**28,3**

15°	16°	17°	18°	19°	20°	21°	22°	23°	24°	25°	26°	27°	28°
28,7	**29,1**	**29,4**	**30,6**	**30,7**	**30,3**	**30,1**	**29,7**	**29,4**	**30,1**	**30,3**	**---**	**---**	**---**

Da questi 22 valori medi, si deriva la seguente tabella, che riporta tutti i calcoli necessari per il test:
In essa

- nella colonna 1 è riportato il rango dei tempi, che ovviamente sono sempre in successione ordinata,
- nella colonna 2 è riportata la media mobile della variabile analizzata,
- nella colonna 3 il rango di questo ultimo valore,
- nella colonna 4 la differenza in valore assoluto $|d|$ tra i due ranghi (colonna 1 – colonna 3),
- nella colonna 5 il quadrato di tale differenza d^2.

(1)	(2)	(3)	(4)	(5)
Tempo	Valore	Rango	$\lvert d \rvert$	d^2
1	26,0	1	0	0
2	26,3	3	1	1
3	26,1	2	1	1
4	26,4	4	0	0
5	26,6	5	0	0
6	26,7	6	0	0
7	26,9	7	0	0
8	27,1	8	0	0
9	27,3	9	0	0
10	28,0	10	0	0
11	28,3	11	0	0
12	28,7	12	0	0
13	29,1	13	0	0
14	29,4	14,5	0,5	0,25
15	30,6	21	6	36
16	30,7	22	6	36
17	30,3	19,5	2,5	6,25
18	30,1	17,5	0,5	0,25
19	29,7	16	3	9
20	29,4	14,5	5,5	30,25
21	30,1	17,5	3,5	12,25
22	30,3	19,5	2,5	6,25
			$\sum d^2 =$	138,5

Dall'ultima colonna, si ricava la somma di tali quadrati

$$\sum d^2 = 138{,}5.$$

Successivamente, per **N** uguale a 22, si calcola il valore di **ρ**

$$\rho = 1 - \frac{6 \cdot \sum d_{R_i}^2}{N^3 - N} = 1 - \frac{6 \cdot 138{,}5}{22^3 - 22} = 1 - \frac{831}{10626} = 1 - 0{,}078 = 0{,}922$$

che risulta uguale a 0,922.

La sua significatività può essere determinata mediante il test **t di Student**

$$t_{(N-2)} = \rho \cdot \sqrt{\frac{N-2}{1-\rho^2}} = t_{(20)} = 0{,}922 \cdot \sqrt{\frac{22-2}{1-0{,}922^2}} = 0{,}922 \cdot \sqrt{\frac{20}{0{,}15}} = 10{,}65$$

che risulta t = 10,65 con 20 gdl.

Poiché era stata ipotizzata una tendenza alla crescita, quindi la verifica dell'ipotesi unilaterale

$$\mathbf{H_0: \rho \leq 0} \quad \text{contro} \quad \mathbf{H_1: \rho > 0}$$

si confronta il valore ricavato con la distribuzione unilaterale.

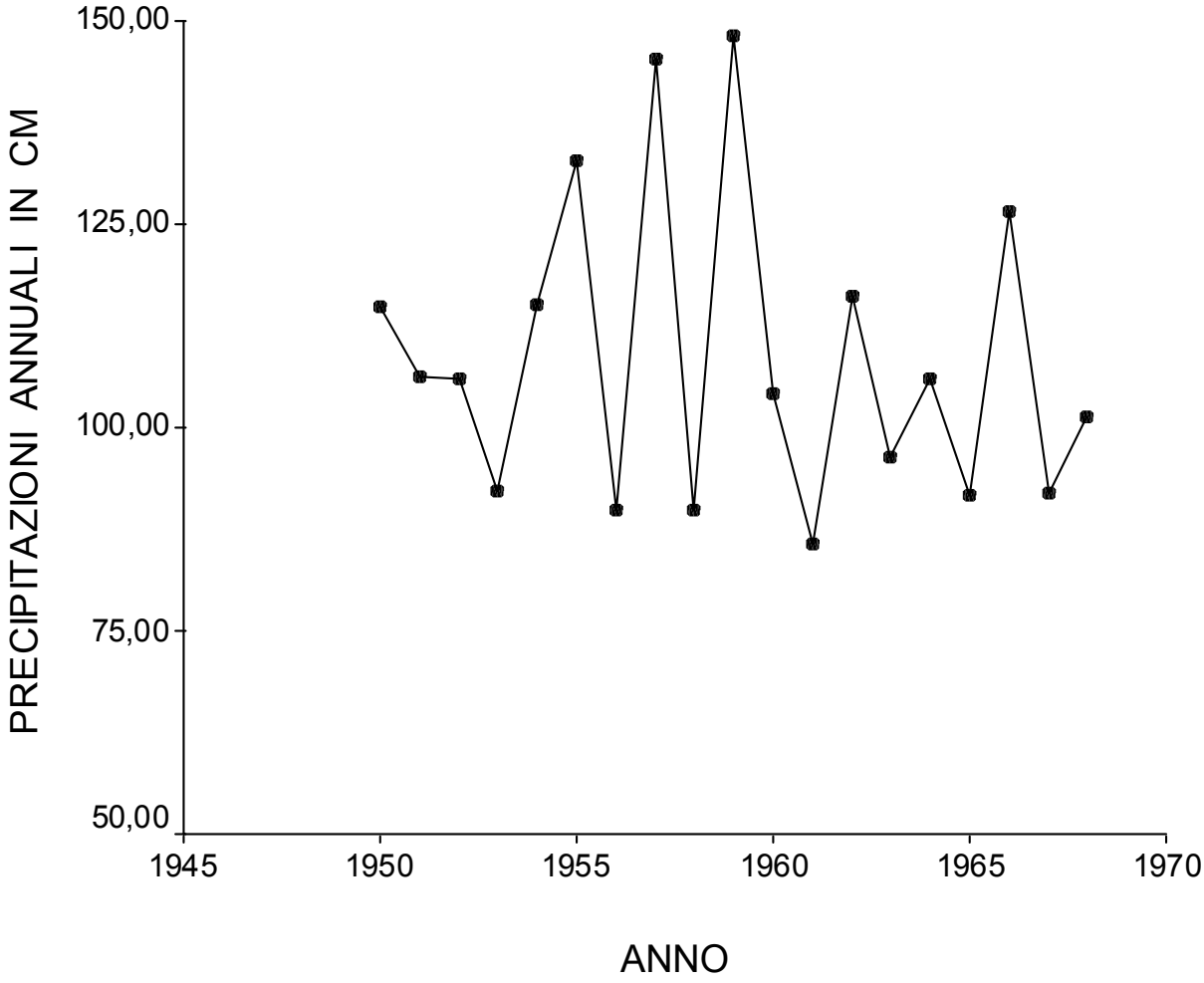
Il risultato è altamente significativo, dato che il valore critico di t con gdl = 20 è uguale a 3,850 per una probabilità unilaterale $\alpha = 0.0005$ mentre il valore calcolato è 10,65. Si rifiuta H_0 con probabilità $P < 0.0005$.

Se il test fosse stato bilaterale, si sarebbe rifiutata l'ipotesi nulla con probabilità $P < 0.001$.

ESEMPIO. Nel testo di W. J **Conover** del 1999 (a pag. 323 di ***Practical Nonparametric Statistics***, 3rd ed., **John Wiley & Sons**, New York, 584 p.) si propone l'analisi della quantità di pioggia (Y, qui riportata in cm.) annuale, per il periodo 1950-1968:

(1)	(2)	(3)	(4)	(5)	(6)
Anno Xi	Valore Yi	Rango Xi	Rango Yi	$\lvert d \rvert$	d^2
1950	114,9	1	12	11	121
1951	116,4	2	15	13	169
1952	106,1	3	11	8	64
1953	92,1	4	6	2	4
1954	115,0	5	13	8	64
1955	132,7	6	17	11	121
1956	89,8	7	2,5	4,5	20,5
1957	145,2	8	18	10	100
1958	89,8	9	2,5	6,5	42,25
1959	148,1	10	19	9	81
1960	104,3	11	9	2	4
1961	85,6	12	1	11	121
1962	116,2	13	14	1	1
1963	96,3	14	7	7	49
1964	106,0	15	10	5	25
1965	91,6	16	4	12	144
1966	126,6	17	16	1	1
1967	92,0	18	5	13	169
1968	101,3	19	8	11	121
				$\sum d^2 =$	1421,5

La sua rappresentazione grafica mette in evidenza una distribuzione dei dati che non mostra la ciclicità del caso precedente, ma che è caratterizzata da forti variazioni casuali anche tra anni contigui.



Per **N** uguale a 19, si calcola il valore di **ρ**

$$\rho = 1 - \frac{6 \cdot \sum d_{R_i}^2}{N^3 - N} = 1 - \frac{6 \cdot 1421{,}5}{19^3 - 19} = 1 - \frac{8529}{6840} = 1 - 1{,}247 = -0{,}247$$

che risulta uguale a -0,247.

Indica una correlazione di tipo negativo, cioè una diminuzione della quantità di pioggia durante il periodo.

In assenza di una teoria specifica, il test dovrebbe essere bilaterale.
La significatività del valore di ρ, calcolata con il t di Student,

$$t_{(N-2)} = \rho \cdot \sqrt{\frac{N-2}{1-\rho^2}} = t_{(17)} = -0{,}247 \cdot \sqrt{\frac{19-2}{1-(-0{,}247)^2}} = -0{,}247 \cdot \sqrt{\frac{17}{0{,}939}} = -1{,}051$$

fornisce un valore di t = -1,051 con 17 gdl. E' un valore basso, corrispondente ad una probabilità P superiore a

- $\alpha = 0{,}20$ in una distribuzione bilaterale,
- $\alpha = 0{,}10$ in una distribuzione unilaterale.

In modo più preciso, la probabilità P è vicina al 30% in una distribuzione bilaterale e al 15% in una distribuzione unilaterale. La probabilità che il valore di ρ calcolato sia stato determinato solamente dal caso è elevata.
Di conseguenza, non solo non si rifiuta l'ipotesi nulla H_0, ma si può affermare che, durante il periodo considerato, non si è realizzata una variazione sistematica nella quantità di pioggia.

Nel testo citato,
- Conover scrive che i test per il trend, basati sul **ρ di Spearman** e sul **τ di Kendall,** in generale sono considerati **più potenti** del test di **Cox e Stuart**.
Come già evidenziato dallo stesso A. **Stuart** nel 1956 (vedi ***The efficiencies of test of randomness against normal regression***, pubblicato su **Journal of the American Statistical Association**, Vol. 51, pp. 285-287),
quando la distribuzione dei dati è normale,
- rispetto al test **r di Pearson** **l'efficienza asintotica relativa** (A.R.E. da **A**symptotic **R**elative **E**fficiency) del test **ρ di Spearman** e del **τ di Kendall** è uguale a **0,98**
- mentre l'efficienza o potenza del test di **Cox e Stuart** è **0,78.**

Meno potente dei test che utilizzano la correlazione non parametrica, in quanto utilizza i segni e non i ranghi, il test di **Cox e Stuart è applicabile in condizioni molto più generali,** anche in presenza di valori fortemente anomali. In particolare, anche rispetto al test della mediana, è utile nel caso di una ciclicità accentuata e quando si confrontano tra loro i valori medi o mediani di un periodo relativamente lungo, separabile in due serie con la stessa variazione ciclica.

21.6. SIGNIFICATIVITA' DELLA REGRESSIONE E DELLA CORRELAZIONE LINEARE PARAMETRICA CON I TEST NONPARAMETRICI ρ E τ.

Nei paragrafi dedicati alla regressione lineare parametrica, è stato discusso come la verifica dell'ipotesi nulla sulla linearità con

$$H_0: \beta = 0$$

attraverso la significatività del valore **r** di correlazione, per il quale l'ipotesi nulla è

$$H_0 : \rho = 0$$

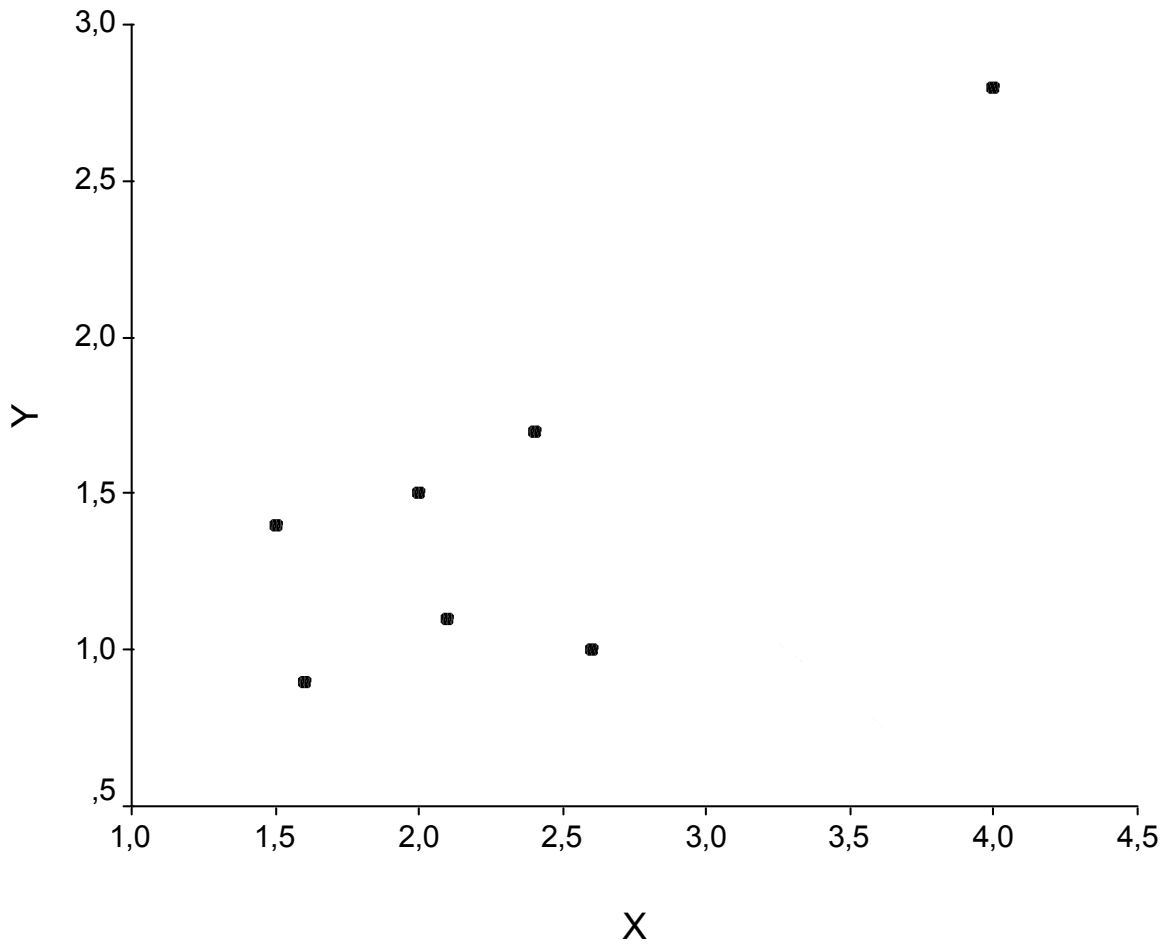
sia un metodo concettualmente errato, anche se conduce alla stima di una probabilità α identica a quella ottenuta con il test che utilizza il valore del coefficiente angolare **b**.

I motivi fondamentali evidenziati, a causa dei quali i testi non confondano mai le due procedure, sono
- la diversa ipotesi che i test per la regressione e quelli per la correlazione verificano e
- le differenze nelle condizioni di validità.

Quando uno o più punti sono anomali e quindi la distribuzione dei dati è fortemente asimmetrica, una condizione di non validità della regressione lineare che è frequente nella pratica sperimentale, questi due metodi parametrici (che forniscono risultati di α coincidenti) spesso risultano significativi. Ma dimostrare che i dati non sono distribuiti in modo normale neppure approssimativamente, quindi che l'analisi condotta non è attendibile, non è semplice; in particolare, è quasi impossibile rifiutare l'ipotesi nulla sulla normalità, quando i dati sono pochi.
Un metodo pratico e semplice di tale verifica può essere l'uso della correlazione lineare non parametrica.

Questi concetti possono essere espressi con termini più rigorosi:
- **la funzione di distribuzione del test parametrico r di Pearson dipende dalla funzione di distribuzione bivariata, cioè della X e della Y, mentre i test non parametrici per l'analisi della correlazione sono indipendenti da essi**.

La **verifica della validità** della regressione e della correlazione lineare, seppure calcolate con metodi parametrici, può quindi essere fornita dalla **correlazione non parametrica**. Se la risposta della correlazione non parametrica è significativa o vicina alla probabilità del test parametrico, il risultato di questo ultimo può essere ritenuto corretto. In caso contrario, quando il test di correlazione non parametrica non risulta significativo, si può dedurre che facilmente il test parametrico è stato applicato senza rispettare la condizione di normalità della distribuzione dei dati.



A dimostrazione empirica di questa procedura sperimentale, si assuma di voler calcolare il coefficiente di regressione lineare semplice, per la seguente serie di dati bivariati:

X	1,5	1,6	2,0	2,1	2,4	2,6	4,0
Y	1,4	0,9	1,5	1,1	1,7	1,0	2,8

La sua rappresentazione grafica evidenzia visivamente la presenza di un punto anomalo: un valore di X nettamente più alto degli altri, associato ad un valore di Y ugualmente elevato e distante dalla distribuzione di tutti gli altri punti.

Dai dati sperimentali, dopo aver calcolato le quantità

$X \cdot Y$	2,10	1,44	3,00	2,31	4,08	2,60	11,20	$\sum X \cdot Y = 26{,}73$
X^2	2,25	2,56	4,00	4,41	5,76	6,76	16,0	$\sum X^2 = 41{,}74$
Y^2	1,96	0,81	2,25	1,21	2,89	1,00	7,84	$\sum Y^2 = 17{,}96$

si ricava il valore di **b**

$$b = \frac{\sum X \cdot Y - \frac{\sum X \cdot \sum Y}{n}}{\sum X^2 - \frac{\left(\sum X\right)^2}{n}} = \frac{26{,}73 - \frac{16{,}20 \cdot 10{,}40}{7}}{41{,}74 - \frac{(16{,}20)^2}{7}} = \frac{2{,}66}{4{,}25} = 0{,}626$$

e successivamente il valore di **a**

$$a = \overline{Y} - b \cdot \overline{X} = 1{,}486 - 0{,}626 \cdot 2{,}314 = 0{,}037$$

Infine, si scrive la retta

$$\hat{Y}_i = a + b \cdot X_i = 0{,}037 + 0{,}626 \cdot X_i$$

Per testare l'esistenza di una relazione lineare tra X e Y con ipotesi bilaterale, cioè per verificare l'ipotesi bilaterale

H_0: $\beta = 0$ contro H_1: $\beta \neq 0$

1- dopo aver calcolato **la devianza totale**

$$S.Q._{Totale} = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n} = 17{,}960 - \frac{(10{,}40)^2}{7} = 2{,}509$$

2 - la **devianza dovuta alla regressione**

$$S.Q._{regressione} = \frac{\left(\sum X \cdot Y - \frac{\sum X \cdot \sum Y}{n}\right)^2}{\sum X^2 - \frac{\left(\sum X\right)^2}{n}} = \frac{\left(26{,}73 - \frac{16{,}20 \cdot 10{,}40}{7}\right)^2}{41{,}74 - \frac{(16{,}20)^2}{7}} = \frac{(2{,}66)^2}{4{,}25} = 1{,}665$$

è possibile costruire la tabella dell'ANOVA

Fonte di variazione	$S.Q.$	$D.F.$	S^2
Totale	2,509	6	---
Regressione	1,665	1	1,665
Errore	0,844	5	0,169

completando i dati relativi alla devianza e ai df dell'errore con la proprietà additiva.

Il test F

$$F_{(1,5)} = \frac{1{,}665}{0{,}169} = 9{,}85$$

stima F = 9,85 con df 1 e 5.

Poiché il valore critico per un test bilaterale alla probabilità $\alpha = 0.05$ è uguale a 10,0 e il valore calcolato è leggermente minore (9,85),a causa del numero limitato di dati si può affermare che il test è tendenzialmente significativo.

L'analisi della **correlazione parametrica r di Pearson** conduce alle stesse conclusioni.

Dal valore di R^2, ricavato per semplicità e in modo rapido dal rapporto tra le devianze,

$$R^2 = \frac{S.Q._{\text{Regressione}}}{S.Q._{Totale}} = \frac{1{,}665}{2{,}509} = 0{,}663$$

si ottiene facilmente quello di **r**

$$r = \sqrt{R^2} = \sqrt{0{,}663} = 0{,}8143$$

Per verificare la significatività dell'ipotesi bilaterale

H_0: $\rho = 0$ contro H_1: $\rho \neq 0$

si può utilizzare il test F

$$F_{(1,5)} = \frac{r^2 \cdot (n-2)}{1-r^2} = \frac{(0{,}814)^2 \cdot 5}{1-(0{,}814)^2} = \frac{3{,}313}{0{,}336} = 9{,}86$$

che ovviamente fornisce lo stesso identico risultato della regressione lineare, a meno delle approssimazioni introdotte nel calcolo come numero di decimali; quindi determina la stessa probabilità α, essendo sempre un valore di F con gdl 1 e 5.

Questa **stessa ipotesi**

H_0: $\rho = 0$ contro H_1: $\rho \neq 0$

e implicitamente quella sulla regressione lineare possono essere verificate mediante **la correlazione non parametrica.**

Con i ranghi

X	1	2	3	4	5	6	7	
Y	4	1	5	3	6	2	7	
$\lvert d \rvert$	3	1	2	1	1	4	0	
d^2	9	1	4	1	1	16	0	$\sum d^2 = 32$

secondo il metodo di Spearman si ricava

$$\rho = 1 - \frac{6 \cdot \sum d^2}{n^3 - n} = 1 - \frac{6 \cdot 32}{7^3 - 7} = 1 - \frac{192}{336} = 0{,}429$$

la stima ρ **= 0,429**. E' un risultato **molto lontano dalla significatività**. Infatti nella tabella dei valori critici per un test bilaterale con **df = 5** e α **= 0.10** il valore critico è ρ **= 0,900**.

Il confronto tra due risultati così differenti, quello della probabilità α stimata con un metodo parametrico e quella della probabilità α ottenuta mediante la correlazione non parametrica, permette di dedurre che **la significatività ottenuta con il test parametrico è del tutto imputabile alla presenza di un valore anomalo; di conseguenza, in realtà tale regressione lineare non esiste.**

Non esiste una regressione lineare tra i ranghi e quindi non può esistere nemmeno tra i valori osservati

Applicando il test τ di Kendall invece del test ρ di Spearman si perviene alle stesse conclusioni.

21.7. IL COEFFICIENTE DI CORRELAZIONE PARZIALE: $\tau_{12,3}$ DI KENDALL, $\rho_{12,3}$ DI SPEARMAN

Quando si analizza la correlazione tra 2 variabili, in varie occasioni può sorgere il dubbio che la causa principale di un valore elevato, non importa se positivo o negativo, possa essere attribuita non ad una **correlazione effettiva** tra loro,

- ma ad una **correlazione o associazione di ognuna con una terza variabile**.

Può anche succedere il fenomeno opposto.

L'inatteso basso valore di correlazione tra due variabili, che notoriamente sono tra loro correlate in modo positivo

- sia imputabile all'**azione di segno opposto di una terza variabile**, alla quale entrambe sono correlate.

Come già discusso nel caso della correlazione parametrica, gli effetti delle variazioni congiunte tra due variabili dovute al loro legame con una terza possono essere eliminati, quando si tenga costante quest'ultima; ma è un procedimento e che limita le conclusioni a quel caso specifico. Non è detto che, per altri valori della terza variabile, le prime due mantengano lo stesso grado di correlazione calcolata per alcuni. Analizzare tutti i casi possibili richiede molto tempo. E' un caso frequente nella ricerca ambientale, biologica e farmacologica, che sovente utilizzano le correlazioni statistiche come indicazione preliminare e come supporto all'indagine causale.

Le sostanze contaminanti rilasciate in fiumi e laghi determinano effetti nocivi sulle popolazioni sia animali sia vegetali e sulle comunità acquatiche. Gli effetti principali possono essere raggruppati in alcune categorie: eutrofizzanti, deossigenanti, tossici, fisici come quelli sulla temperatura, chimici come quelli sul pH, patogeni.

Temperatura, eutrofizzazione e deossigenazione risultano tra loro correlati. Per misurare in modo corretto la relazione tra eutrofizzazione e deossigenazione in un gruppo di laghi, occorrerebbe fare rilevazioni a temperatura costante; ma le conclusioni sarebbero limitate a quella specifica temperatura. Nello stesso modo, dovrebbero essere verificate le relazioni sia tra temperatura e deossigenazione con un livello di eutrofizzazione costante, sia le relazioni tra temperatura ed eutrofizzazione con un indice di deossigenazione mantenuto stabile.

Per rendere generali le conclusioni sulla correlazione tra due variabili occorrerebbe un numero molto alto di esperimenti, estesi a tutti i livelli o modalità della terza variabile.

La correlazione parziale tra k variabili permette di valutare il grado di correlazione esistente tra ogni coppia di variabili, utilizzando solo una serie di dati. Il caso più semplice e frequente è quello

tra 3 variabili, indicate con 1, 2 e 3; ad esso viene limitata la presentazione della correlazione parziale non parametrica τ di Kendall, come è già stato fatto con l'indice **r** di Pearson nel capitolo precedente sulla correlazione parziale parametrica.

Il τ (leggesi: ***valore di correlazione tra le variabili 1 e 2, indipendente da 3***) **di Kendall**

- misura quanto le misure delle variabili 1 e 2 siano correlate, indipendentemente dal loro accordo con quelle di 3.

La correlazione parziale $\tau_{12,3}$ può essere derivata dai valori delle 3 correlazioni semplici ($\boldsymbol{\tau_{12}}$, $\boldsymbol{\tau_{13}}$ e $\boldsymbol{\tau_{23}}$), in accordo con la relazione

$$\tau_{12,3} = \frac{\tau_{12} - \tau_{13} \cdot \tau_{23}}{\sqrt{(1-\tau_{13}^2) \cdot (1-\tau_{23}^2)}}$$

L'applicazione e l'interpretazione sono identiche a quelle della correlazione netta parametrica; di conseguenza, per tali argomenti si rinvia ad essa.

Il test può essere unilaterale oppure bilaterale.

Nel primo caso, l'ipotesi nulla H_0 è che non esista correlazione, mentre l'ipotesi alternativa H_1 unilaterale può supporre la presenza di una correlazione sia positiva che negativa, che deve sempre essere chiaramente espressa.

Nel secondo caso, l'ipotesi alternativa H_1 bilaterale verifica la semplice presenza di una correlazione, senza alcuna indicazione di segno.

Con più variabili, il numero di confronti possibili diventa elevato. Si pone lo stesso problema di significatività già discusso per i confronti tra più medie.

Si ricorre a concetti analoghi al **t** di **Bonferroni** per confronti multipli a posteriori, se i confronti non sono già prestabiliti e limitati a quelli ritenuti importanti per la ricerca.

La significatività è stimata attraverso gli stessi metodi e gli stessi valori della correlazione semplice.

Per **piccoli campioni**, la significatività è fornita dalla **tabella dei valori critici** (vedi pagina successiva).

Valori critici del coefficiente di correlazione parziale $\tau_{12,3}$ di Kendall

per test a 1 coda e a 2 code

α

	0.05	0.025	**0.01**	0.005	**0.001**	**1 coda**
N	0.10	**0.05**	0.02	**0.01**	0.002	**2 code**
4	.707	1.000				
5	.667	.802	.816	1.000		
6	.600	.667	.764	.866	1.000	
7	.527	.617	.712	.761	.901	
8	.484	.565	.648	.713	.807	
9	.443	.515	.602	.660	.757	
10	.413	.480	.562	.614	.718	
11	.387	.453	.530	.581	.677	
12	.365	.430	.505	.548	.643	
13	.347	.410	.481	.527	.616	
14	.331	.391	.458	.503	.590	
15	.319	.377	.442	.485	.570	
16	.305	.361	.423	.466	.549	
17	.294	.348	.410	.450	.532	
18	.284	.336	.395	.434	.514	
19	.275	.326	.382	.421	.498	
20	.268	.318	.374	.412	.488	

Per **grandi campioni**, si ricorre alla **distribuzione normale**, con formula identica a quella della correlazione semplice

$$Z = \frac{3\tau \cdot \sqrt{N \cdot (N-1)}}{\sqrt{2 \cdot (2N+5)}}$$

dove

- τ è il valore calcolato con la formula presentata per la correlazione netta, cioè $\tau_{12,3}$
- **N** è il numero di dati, uguale per tutte le variabili a confronto.

Anche il test ρ di Spearman è stato esteso alle misure di correlazione parziale, nello stesso modo descritto per il τ di Kendall. Come già ricordato in altre situazioni, il vantaggio dell'uso del **ρ di Spearman** è determinato dal fatto che si ottiene lo stesso risultato dei coefficienti della **correlazione parziale r di Pearson**, usando **i ranghi al posto dei dati**.

Poiché i programmi informatici per il calcolo dell'indice **r di Pearson** sono più diffusi di quelli per i test equivalenti non parametrici, è sufficiente questa semplice sostituzione per applicare il test desiderato.

Come nella correlazione lineare semplice parametrica e in quella non parametrica, è necessario tenere presente che mentre **la distribuzione dei valori parametrici $r_{12,3}$ dipende dalla funzione di distribuzione multivariata delle variabili 1, 2 e 3**, la distribuzione dei valori non parametrici $\tau_{12,3}$ e $\rho_{12,3}$ sono ***distribution free***, cioè **sono indipendenti dalla forma di distribuzione** dei dati originari, ma solamente quando le tre variabili considerate sono tra loro mutuamente indipendenti. Sono concetti illustrati da vari studiosi e discussi in particolare da G. **Simon** in due articoli del 1977; ad essi si rimanda per approfondimenti **(**il primo: ***A nonparametric test of total indipendence based on Kendall's tau***, pubblicato su **Biometrika** Vol. 64, pp. 277-282; il secondo: ***Multivariate generalization of Kendall's tau with application to data reduction***, pubblicato su **Journal of the American Statistical Association** Vol. 72, pp. 367 - 376).

Kendall ha proposto anche una metodologia utile per **stimare direttamente la correlazione netta, servendosi dei dati originari delle tre variabili**. Il metodo è lungo e complesso; in pratica, può essere applicato con successo ricorrendo ai calcoli manuali, solo nel caso in cui **N** sia limitato a poche osservazioni.

Per **comprendere in modo dettagliato ed operativo questo metodo di Kendall**, si supponga di avere misurato tre variabili (X, Y, Z), su un gruppo formato da 5 rilevazioni campionarie, con i seguenti risultati:

	Variabili		
Campione	X	Y	Z
I	5	18	7
II	9	12	10
III	12	13	2
IV	21	15	11
V	54	31	36

Calcolare la correlazione parziale o netta ($\tau_{YZ,X}$) tra Y e Z, al netto di X.

La metodologia richiede alcuni passaggi:

1- La serie dei valori deve essere ordinata in modo crescente per la variabile X (già fatto nella tabella di presentazione dei dati).

2 - Trasformare i valori in ranghi, entro ogni variabile, ordinando separatamente i ranghi delle altre due variabili (**Y** e **Z**), ma senza spostare la loro collocazione in riferimento alla rilevazione della variabile **X**.
Con i dati del campione, dopo queste due operazioni si ottiene una nuova tabella:

	Variabili		
Campione	X	Y	Z
I	1	3	2
II	2	1	3
III	3	2	1
IV	4	4	4
V	5	5	5

3 - Per ognuna delle 3 variabili, tradurre l'ordine dei ranghi in concordanze e discordanze,
- assegnando + ad ogni coppia in ordine naturale o crescente e
- assegnando - ad ogni coppia in ordine non naturale o decrescente;

ovviamente per la variabile X esistono solo concordanze e quindi segni +

Coppie di dati	Variabile		
	X	Y	Z
1 - I,II	+	-	+
2 - I,III	+	-	-
3 - I,IV	+	+	+
4 - I,V	+	+	+
5 - II,III	+	-	-
6 - II,IV	+	+	+
7 - II,V	+	+	+
8 - III,IV	+	-	+
9 - III,V	+	+	+
10 - IV,V	+	+	+

Ad esempio,
per la variabile Y la coppia di ranghi per la coppia I e II è decrescente (3 e 1) e pertanto è stato assegnato -,
mentre per la variabile Z la stessa coppia di ranghi (2 e 3) è crescente e pertanto è stato assegnato +.

4 - Riassumere in una tabella 2 x 2, come quella sottoriportata, l'informazione delle concordanze e delle discordanze delle variabili Y e Z, in relazione all'ordine di X.
Per esempio, nella coppia di rilevazioni I e II,
il rango della Y (-) è in disaccordo con X,
mentre quello di Z (+) è in accordo con X.

		Variabile Y		
		Concord. con X	Discord. Con X	Totale
Variabile Z	Concord. Con X	**A 6**	**B 2**	n_1 8
	Discord. Con X	**C 0**	**D 2**	n_2 2
	Totale	n_3 6	n_4 4	**N 10**

Con i dati dell'esempio, le coppie di rilevazioni sono complessivamente **10**, come riportato nella casella contrassegnata dal simbolo **N**.

Nella **casella A,** il numero **6** indica che in 6 coppie di rilevazioni (I,IV; I,V; II,IV; II,V; III,V; IV,V) le variabili **Y** e **Z** sono tra loro concordanti come con la variabile **X**; in altri termini, in 6 casi quando si ha il segno + nella colonna **X** si ha + sia nella colonna **Y** sia in quella **Z**.

Nella **casella B,** il numero **2** indica che in 2 coppie (I,II; III,IV) la variabile **Y** è discordante da **X**, mentre la variabile **Z** è concordante con **X**; in altri termini, mentre si ha il segno + nella colonna **X** si ha contemporaneamente il segno **-** nella colonna **Y** e quello + nella colonna **Z**.

Nella **casella C,** il numero **0** indica che non si ha alcuna coppia in cui variabile **Y** è concordante con **X**, mentre la variabile **Z** è discordante da **X**; in altri termini, non è presente alcun caso in cui si ha + nella colonna **X**, mentre si ha anche + nella colonna **Y** e contemporaneamente il segno **-** nella colonna **Z**.

Nella **casella D,** il numero **2** indica che in 2 coppie (I,II; II,III) la variabile **Y** e la variabile **Z** discordano simultaneamente dalla variabile **X**; in altri termini, nella casella D è riportato il numero di casi in cui si ha il segno **-** contemporaneamente sia nella variabile **Y** sia nella variabile **Z**, mentre ovviamente si ha il segno + nella variabile **X**.

Le **caselle $\mathbf{n_1}$, $\mathbf{n_2}$, $\mathbf{n_3}$, $\mathbf{n_4}$** riportano i totali parziali.

Nella casella $\mathbf{n_1}$ è riportato il numero di concordanze (8) tra la variabile **X** (+) e la variabile **Z** (+), senza considerare la variabile **Y**.

Nella casella $\mathbf{n_2}$ è riportato il numero di discordanze (2) tra la variabile **X** (+) e la variabile **Z** (-), ignorando la variabile **Y**.

Nella casella $\mathbf{n_3}$ è riportato il numero di concordanze (6) tra la variabile **X** (+) e la variabile **Y** (+), senza considerare la variabile **Z.**

Nella casella $\mathbf{n_4}$ è riportato il numero di discordanze (4) tra la variabile **X** (+) e la variabile **Y (-)**, senza considerare la variabile **Z**.

Una volta individuate concordanze e discordanze tra le tre variabili, il coefficiente di correlazione parziale $\tau_{YZ,X}$ **di Kendall, (la correlazione tra Y e Z tenendo costante X)**, è calcolata rapidamente mediante la relazione

$$\tau_{YZ,X} = \frac{A \cdot D - B \cdot C}{\sqrt{n_1 \cdot n_2 \cdot n_3 \cdot n_4}}$$

Con i dati dell'esempio,

$$\tau_{YZ,\,X} = \frac{6 \cdot 2 - 2 \cdot 0}{\sqrt{8 \cdot 2 \cdot 6 \cdot 4}} = \frac{+12}{\sqrt{384}} = \frac{+12}{19{,}60} = +0{,}61$$

si ottiene un valore di correlazione netta uguale a +0,61.

Trattandosi di un **campione piccolo**, per la sua significatività si utilizza la tabella precedente.

Nel caso di **grandi campioni**, data la complessità delle operazioni descritte, il metodo diventa di difficile applicazione, quando non si dispone di programmi informatici.
La significatività è fornita dal test **Z**; se è riportato solo il suo valore e non il numero di dati o dei gdl, la probabilità relativa è data dalla distribuzione normale.

21.8. IL COEFFICIENTE DI CONCORDANZA TRA VALUTATORI: LA W DI KENDALL. SUE RELAZIONI CON LA CORRELAZIONE NON PARAMETRICA E CON IL TEST DI FRIEDMAN PER K CAMPIONI DIPENDENTI. CENNI SULLA TOP-DOWN CONCORDANCE

I coefficienti di correlazione ρ di Spearman e τ di Kendall sono applicato **a due variabili**, cioè a **due serie di ranghi** (o di valori trasformati in ranghi), relativi a **N** oggetti o individui. Con gli **indici di concordanza**, è possibile verificare l'accordo complessivo tra **più variabili** quando
- si dispone di **k** serie di ranghi,
- riportati per **N** valutazioni.

I test proposti per queste **misure di associazione - correlazione,** definite con il termine tecnico di **concordanza,** sono numerosi. Alcuni sono pubblicati nel volume di Sir Maurice George **Kendall** (1907-1983) divulgato nel 1970 (***Ranks correlation methods***, 4th ed. stampato a Londra da Griffin) e nella sua edizione più recente, sempre di Sir M. G. **Kendall** ma con J. D. **Gibbons** del 1980 (***Ranks Correlation Methods,*** 5th ed. stampato a Londra da Edward Arnold).
Tra le misure di concordanza che è possibile trovare nella letteratura statistica, quella più frequentemente proposta nei programmi informatici e nei testi internazionali a maggior diffusione è il **coefficiente di concordanza W di Kendall** (***Kendall's Coefficient of Concordance***).

La metodologia è stata proposta in modo indipendente con due articoli pubblicati quasi contemporaneamente nel 1939:

- il primo da M. G. **Kendall** e B. **Babington-Smith** (vedi ***The problem of m rankings***, su **The Annals of Mathematical Statistics** Vol. 10, pp. 275-287),
- il secondo da W. A. **Wallis** (vedi ***The correlation ratio for ranked data***, su **Journal of the American Statistical Association**, Vol. 34, pp. 533-538).

Le **misure di associazione e di concordanza non sono test inferenziali: hanno solamente un valore descrittivo** della intensità della relazione. E' quindi sempre importante **verificare la significatività del valore calcolato mediante test inferenziali**.
Il **coefficiente di concordanza W** è costruito in modo tale da assumere solamente valori che variano tra 0 e +1:

$$0 \leq W \leq 1$$

Quando
- esiste **totale accordo** tra le **N** serie di **k** ranghi, si ha **W = 1**
- le **N** serie di **k** ranghi sono **puramente casuali**, si ha **W = 0**.

Il valore di **W non può essere negativo**, in quanto con **N** serie di k ranghi non è possibile avere tra esse disaccordo completo.

Il coefficiente di concordanza **W di Kendall** può essere visto con due ottiche diverse:
- una **generalizzazione del test** ρ **e del test** τ: infatti esso **misura la divergenza** nella valutazione **tra N serie di k misure ordinali**,
- una **analisi della varianza non parametrica a due criteri di classificazione**; infatti può essere utilizzato nelle stesse condizioni del **test di Friedman**, in quanto entrambi sono fondati sullo stesso modello matematico: pertanto la significatività può essere determinata nello stesso modo mediante il χ^2 o il test F.

Queste due relazioni, in particolare quella con il **test di Friedman**, sono presentati in modo più approfondito nella seconda parte del paragrafo.

L'**indice di divergenza W** può essere calcolato direttamente da una serie di dati.

Si supponga che 4 ricercatori (I, II, II, IV) debbano stabilire una classifica tra 5 situazioni ambientali (A, B, C, D, E), per valutare il loro livello di degrado:

RICERCATORI (N)	SITUAZIONI AMBIENTALI (k)				
	A	B	C	D	E
I	2	1	4	5	3
II	1	2	5	4	3
III	1	2	3	5	4
IV	2	1	4	5	3
R_i	**6**	**6**	**16**	**19**	**13**
$\overline{R_i}$	**1,50**	**1,50**	**4,00**	**4,75**	**3,25**

Successivamente o al momento della graduatoria, i punteggi attribuiti dagli **N** ricercatori alle **k** situazioni ambientali (A, B, C, D, E) sono trasformati in ranghi entro la stessa riga, attribuendo **1** al valore minore ed **k** a quello maggiore.
Per esempio,
- secondo il ricercatore I la situazione B è quella meno degradata e la D quella maggiormente degradata,
- mentre il ricercatore II valuta la situazione A come migliore e la C come quella peggiore.

Se fosse vera l'ipotesi nulla $\mathbf{H_0}$ dell'assenza totale d'accordo tra i ricercatori (cioè essi hanno fornito valutazioni di rango sulla base di principi totalmente differenti), le somme dei ranghi per colonna (R_i) sarebbero tra loro uguali e le medie per colonna ($\overline{R_i}$) uguali alla media generale.

Viceversa, se fosse vera l'ipotesi alternativa H_1 di pieno accordo tra i ricercatori (essi forniscono la stessa valutazione sulle k situazioni), le somme (R_i) e le medie relative ($\overline{R_i}$) avrebbero differenze massime.

Se l'ipotesi nulla (H_0) fosse vera, l'indice di divergenza dovrebbe essere W = 0.
Nel caso opposto (H_0 falsa) di massima divergenza, l'indice dovrebbe essere W = 1.

L'ottimo testo di statistica applicata di David H. **Sheskin** pubblicato nel 2000 (***Handbook of PARAMETRIC and NONPARAMETRIC STATISTICAL PROCEDURES***, 2nd ed. Chapman & Hall/CRC, London, 982 p.) presenta in modo dettagliato la procedura, qui ulteriormente chiarita in tutti i suoi passaggi logici e metodologici.

Si supponga, come nella tabella successiva, che sei esperti (indicati da I a VI; quindi N = 6) abbiano espresso un giudizio su 4 prodotti o situazioni (indicati con A, B, C, D; quindi k = 4). La loro valutazione, espressa direttamente in ranghi o per trasformazione successiva, è stata

	Prodotti				
Esperti	A	B	C	D	
I	3	2	1	4	
II	3	2	1	4	
III	3	2	1	4	
IV	4	2	1	3	
V	3	2	1	4	
VI	4	1	2	3	
$\sum R_i$	20	11	7	22	T = 60
$\left(\sum R_i\right)^2$	400	121	49	484	G = 1054

Vi vuole verificare se i sei esperti concordano globalmente nella loro valutazione, in modo significativo.

Risposta. In termini più tecnici,

- dopo aver fornito una misura della concordanza (W) degli N valutatori, dove

$$W = \frac{Varianza - calcolata - della \sum R_i}{Varianza - massima - possibile - della \sum R_i}$$

- si intende verificare la sua significatività, cioè testare l'ipotesi

H_0: W = 0 contro H_1: W ≠ 0

Con metodi del tutto analoghi a quelli della varianza tra trattamenti, di cui è riportata la formula abbreviata,

- dapprima si calcolano i totali

$$\sum_{j=1}^{k} \sum_{i=1}^{N} R_{ij} = T = 60$$

$$\sum_{j=1}^{k}\left(\sum_{i=1}^{N} R_{ij}\right)^2 = G = 1054$$

- successivamente si ricava il **coefficiente di concordanza W** con

$$Varianza - calcolata - della \sum R_i = \frac{k \cdot G - T^2}{k} = \frac{4 \cdot 1054 - 60^2}{4} = \frac{4216 - 3600}{4} = 154$$

$$Varianza - massima - possibile - della \sum R_i = \frac{N^2 \cdot k \cdot \left(k^2 - 1\right)}{12} = \frac{6^2 \cdot 4 \cdot \left(4^2 - 1\right)}{12} = \frac{36 \cdot 60}{12} = 180$$

ottenendo

$$W = \frac{154}{180} = 0{,}8556$$

Con **formula ulteriormente semplificata** che, come molte di esse, ha il difetto di nascondere i concetti, è possibile il calcolo più rapido

$$W = \frac{\left(12 \cdot G\right) - \left[3 \cdot N^2 \cdot k \cdot \left(k+1\right)^2\right]}{N^2 \cdot k \cdot \left(k^2 - 1\right)} = \frac{\left(12 \cdot 1054\right) - \left[3 \cdot 6^2 \cdot 4 \cdot \left(4+1\right)^2\right]}{6^2 \cdot 4 \cdot \left(4^2 - 1\right)} = \frac{12648 - 10800}{2160} = 0{,}8556$$

Nel caso di **piccoli campioni** (k da 3 a 7; N da 3 a 20), sono stabiliti valori critici ricavati da quelli proposti da M. **Friedman** nel 1940 per il suo test (in ***A comparison of alternative tests of significance for the problem of m rankings***, pubblicato su **Annals of Mathematical Statistics**, Vol. 11, pp. 86-92).

Nel caso dell'esempio, con N = 6 e k = 4 il valore critico alla probabilità $\alpha = 0.01$ è 0,553. Poiché il valore calcolato è superiore a quello critico, si rifiuta l'ipotesi nulla con probabilità P inferiore a 0.01. Esiste un accordo molto significativo tra i 6 esperti nell'attribuzione della graduatoria ai 4 prodotti.

Per **grandi campioni,** ma con limiti non chiaramente definibili come in tutti questi casi, una buona approssimazione è data dalla distribuzione χ^2 con gdl = k - 1 dopo la trasformazione di W mediante la relazione

$$\chi^2_{(k-1)} = N \cdot (k-1) \cdot W$$

Valori critici del Coefficiente di Concordanza W di Kendall

α = 0.05					
	K				
N	3	4	5	6	7
3	---	---	0,716	0,660	0,624
4	---	0,619	0,552	0,512	0,484
5	---	0,501	0,449	0,417	0,395
6	---	0,421	0,378	0,351	0,333
8	0,376	0,318	0,287	0,267	0,253
10	0,300	0,256	0,231	0,215	0,204
15	0,200	0,171	0,155	0,145	0,137
20	0,150	0,129	0,117	0,109	0,103

α = 0.01					
	K				
N	3	4	5	6	7
3	---	---	0,840	0,780	0,737
4	---	0,768	0,683	0,629	0,592
5	---	0,644	0,571	0,524	0,491
6	---	0,553	0,489	0,448	0,419
8	0,522	0,429	0,379	0,347	0,324
10	0,425	0,351	0,309	0,282	0,263
15	0,291	0,240	0,211	0,193	0,179
20	0,221	0,182	0,160	0,146	0,136

Ulteriori valori per K = 3		
N	α = 0.05	α = 0.01
9	0,333	0,469
12	0,250	0,359
14	0,214	0,311
16	0,187	0,274
18	0,166	0,245

Con i dati dell'esempio,

$$\chi^2_{(3)} = 6 \cdot (4-1) \cdot 0{,}8556 = 15{,}40$$

si ottiene un valore del chi quadrato uguale a 15,40 con 3 gdl.

Poiché nella tabella dei valori critici con $\alpha = 0.01$ il valore riportato è 11,34 si rifiuta l'ipotesi nulla con probabilità di errare $P < 0.01$.

La **corrispondenza di questo test con il test di Friedman** offre altre soluzioni per valutare la significatività del valore W calcolato.

Caso di **piccoli campioni.**

Applicato ai dati dell'esempio, il test di Friedman serve per decidere se i totali dei ranghi ($\mathbf{T_i}$ osservati), sommati per colonna, sono significativamente differenti dell'atteso.

Per il test, si calcola la statistica **Fr**

$$\mathbf{Fr} = \sum_{i=1}^{k} \left(T_i - \frac{N(k+1)}{2} \right)^2$$

E' ovvio che tale valore di **Fr** tenderà

- a **0** nel caso di accordo tra totali osservati e totali attesi ($\mathbf{H_0}$ vera e casualità della distribuzione dei ranghi),
- a un valore alto al crescere dello scarto tra essi ($\mathbf{H_0}$ falsa e attribuzione sistematicamente differente dei ranghi ai fattori riportati in colonna)

Con i dati dell'esempio (T_i = 20, 11, 7, 22), poiché N = 6 e k = 4 la somma attesa dei ranghi per colonna è

$$6 \text{ x } (4+1)/2 = 15$$

ovviamente corrisponde alla somma totale dei ranghi (60) diviso k (4).

Con la formula presentata, si ottiene

$$Fr = (20-15)^2 + (11-15)^2 + (7-15)^2 + (22-15)^2 = 5^2 + 4^2 + 8^2 + 7^2 = 25 + 16 + 64 + 49 = 154$$

un valore di **Fr** uguale a 154.

Poiché nella tabella di Friedman per piccoli campioni i valori critici **Fr** riportati sono

- **Fr** = 102 alla probabilità $\alpha = 0.01$
- **Fr** =128 alla probabilità $\alpha = 0.001$

è possibile rifiutare l'ipotesi nulla H_0: W = 0 e accettare implicitamente H_1: $W \neq 0$ con probabilità $P < 0.001$.

Nel caso di **grandi campioni**, come già presentato nel paragrafo dedicato al test di Friedman, si può calcolare il chi quadrato relativo

$$\chi^2_F = \frac{12}{N \cdot k \cdot (k+1)} \sum_{i=1}^{k} \left(T_i - \frac{N \cdot (k+1)}{2} \right)^2$$

in cui

- la seconda parte è data dagli scarti al quadrato tra somma osservata ed attesa,
- mentre la prima dipende dall'errore standard, determinato numero di dati, trattandosi di ranghi.

La formula abbreviata che ricorre con frequenza maggiore nei testi di statistica è

$$\chi^2_F = \frac{12 \cdot \sum_{i=1}^{k} T_i^2}{N \cdot k \cdot (k+1)} - 3N \cdot (k+1)$$

dove:

- **N** è il numero di righe od osservazioni in ogni campione (tutte con il medesimo numero di dati),
- **k** è il numero di colonne o campioni a confronto,
- $\mathbf{T_i}$ è la somma dei ranghi della colonna i e la sommatoria Σ è estesa a tutte le colonne.

Sempre con i dati dell'esempio (T_i = 20, 11, 7, 22), poiché N = 6 e k = 4

$$\chi^2_F = \frac{12 \cdot \left(20^2 + 11^2 + 7^2 + 22^2\right)}{6 \cdot 4 \cdot (4+1)} - 3 \cdot 6 \cdot (4+1) = \frac{12 \cdot 1054}{120} 90 = 105{,}4 - 90 = 15{,}4$$

si ottiene χ^2_F = 15,4

E' un risultato che fornisce una probabilità α del tutto coincidente con quello ottenuto mediante la W (W = 0,8556).

Infatti è possibile passare dall'uno all'altro, sulla base delle due relazioni:

- da W a χ^2_F

$$\chi^2_F = N \cdot (k-1) \cdot W = 6 \cdot (4 - 1) \cdot 0{,}8556 = 15{,}4$$

- da χ^2_F a W

$$W = \frac{\chi^2_F}{N \cdot (k-1)} = \frac{15{,}4}{6 \cdot (4-1)} = 0{,}8556$$

La **corrispondenza tra coefficiente di concordanza W di Kendall e coefficiente di correlazione per ranghi di Spearman** è importante per i concetti implicati; meno dal punto di vista pratico. Per tale motivo si rinvia a testi che lo presentano in modo dettagliato. Tra essi, quello David H. **Sheskin** pubblicato nel 2000 (***Handbook of PARAMETRIC and NONPARAMETRIC STATISTICAL PROCEDURES***, 2nd ed. Chapman & Hall/CRC, London, 982 p.). Il concetto di base, che è possibile dimostrare in modo semplice con un esempio, è che con **N** valutatori, mediante il ρ di Spearman è possibile calcolare tutte le correlazioni semplici tra loro, pari alle combinazioni 2 a 2 dei **k** oggetti. La media ($\overline{\rho}$) di tutti questi coefficienti di correlazione ρ è in relazione diretta con il valore di W, mediante il rapporto

$$\overline{\rho} = \frac{(WN) - 1}{N - 1}$$

TIES

Come tutte le misure fondate sui ranghi, anche nel caso della W di Kendall si richiede che la scala utilizzata per attribuire i punteggi sia continua, in modo tale da non avere valori identici. Non sempre è possibile, poiché in realtà la scala che spesso viene usata è di fatto limitata e quindi si determinano ***ties.***

Quando i ties sono pochi, è possibile apportare una correzione, il cui effetto è sempre quella di aumentare il valore di W, poiché ne riduce la varianza.

Il seguente esempio dove N = 4 e k = 4

	Prodotti				
Esperti	A	B	C	D	
I	1	3	3	3	
II	1	4	2	3	
III	2	3	1	4	
IV	1,5	1,5	3,5	3,5	
$\sum R_i$	5,5	11,5	9,5	13,5	T = 40
$\left(\sum R_i\right)^2$	30,25	132,25	90,25	182,25	G = 435

utilizza un campione molto piccolo, che ha finalità esclusivamente didattiche anche se è riportato nelle tabelle dei valori critici come caso possibile nella ricerca applicata. In esso si osserva che

- nella prima riga é presente un ties con 3 valori identici,
- nella quarta riga sono presenti due ties, ognuno con 2 valori identici.

Per la correzione si deve stimare

$$\sum_{i=1}^{N}\left(\sum_{j=1}^{k}\left(t_j^3 - t_j\right)\right)$$

dove con i dati dell'esempio si ha

- per la riga 1 $t_j = 3$,
- per la riga 2 $t_j = 0$,
- per la riga 3 $t_j = 0$,
- per la riga 4 $t_j = 2$, due volte

Applicando la formula indicata, si ottiene

$$\sum_{i=1}^{N}\left(\sum_{j=1}^{k}\left(t_j^3 - t_j\right)\right) = \left(3^3 - 3\right) + \left(0\right) + \left(0\right) + \left((2^3 - 2) + (2^3 - 2\right) = 24 + 0 + 0 + 12 = 36$$

e il **coefficiente di concordanza W** che, senza correzione, sarebbe stato

$$W = \frac{\left(12 \cdot G\right) - \left[3 \cdot N^2 \cdot k \cdot \left(k+1\right)^2\right]}{N^2 \cdot k \cdot \left(k^2 - 1\right)} = \frac{\left(12 \cdot 435\right) - \left[3 \cdot 4^2 \cdot 4 \cdot \left(4+1\right)^2\right]}{4^2 \cdot 4 \cdot \left(4^2 - 1\right)} = \frac{5220 - 4800}{960} = 0{,}4375$$

W = 0,4375

mentre con la correzione diviene

$$W = \frac{\left(12 \cdot G\right) - \left[3 \cdot N^2 \cdot k \cdot \left(k+1\right)^2\right]}{N^2 \cdot k \cdot \left(k^2 - 1\right) - N \cdot \sum_{i=1}^{N}\left(\sum_{j=1}^{k}\left(t_j^3 - t_j\right)\right)}$$

$$W = \frac{\left(12 \cdot 435\right) - \left[3 \cdot 4^2 \cdot 4 \cdot \left(4+1\right)^2\right]}{\left(4^2 \cdot 4 \cdot \left(4^2 - 1\right)\right) - \left(4 \cdot 36\right)} = \frac{5220 - 4800}{960 - 144} = \frac{420}{816} = 0{,}5147$$

W = 0,5147.

Poiché per N = 4 e k = 4 alla probabilità $\alpha = 0.05$ il valore critico è 0,619 con W = 0,5147 non è possibile rifiutare l'ipotesi nulla.
E' tuttavia evidente l'effetto della correzione per i ties, (aumento del valore W di concordanza da 0,4375 a 0,5147) tanto più marcato quanto più ampio è il ties.

Il coefficiente di concordanza W di Kendall valuta l'intensità di gradimento come nei casi illustrati; ma è utilizzato anche per misurare la **concordanza complessiva fra tre o più variabili**. E' infatti chiamato anche ***Rank Correlation among Several Variables***. Con le modalità qui illustrate, è applicato spesso ai casi descritti nel paragrafo dedicato alla correlazione parziale.
Il testo di Jarrold H. **Zar** del 1999 (**Biostatistical Analysis**, 4th ed. Prentice – Hall, Inc. Ney Jersey, 663 p + App. 212) sviluppa in particolare esempi di questo tipo. Per approfondimenti sull'argomento si rimanda ad esso.

Nello stesso testo è spiegata anche la **Top-Down Concordance**, chiamata anche ***Weighted rank/top-down concordance***. Con essa, si prendono in considerazione le situazioni caratterizzate dai punteggi estremi. Nell'esempio delle valutazioni forniti da 6 esperti su 4 prodotti, serve per verificare se coloro che concordano nel dare la loro preferenza (rango 1) al prodotto C sono concordi anche nell'attribuire il punteggio minimo (rango 4) al prodotto D o viceversa.
Nella ricerca ambientale e industriale, dove si svolgono indagini sulle opinioni o sui consumi, può essere utilizzato per valutare se il gradimento massimo dato a una situazione è strettamente correlato con il livello di gradimento minimo espresso per un'altra situazione o prodotto
Nell'assunzione di personale, in cui 4 dirigenti (N = 4) danno una valutazione in ranghi di k canditati, oltre a valutare se essi concordano globalmente nel giudizio con il test illustrato nella prima parte del paragrafo, è possibile valutare se quando concordano all'attribuire il punteggio più alto a un candidato concordano pure nell'attribuzione del punteggio minore. Nell'esempio riportato da Zar, dove 3 ragazzi esprimono il loro gradimento a 6 differenti gusti di gelato, si vuole valutare se coloro che preferiscono un certo gusto concordano anche nella bassa preferenza da essi attribuita a un altro sapore.
Anche in questo caso si rinvia a questo testo per approfondimenti.

21.9. CENNI SUL COEFFICIENTE DI CONCORDANZA u DI KENDALL, IN CONFRONTI APPAIATI

Sempre attribuito allo statistico inglese **Sir Maurice George Kendall** (1907 – 1983) e riportato nelle varie edizioni del volume ***Ranks correlation methods*** (vedi la 4th ed. stampata a Londra da Griffin, più recentemente sempre di M. G. **Kendall** ma in collaborazione con J. D. **Gibbons** del 1980 ***Ranks***

Correlation Methods, 5th ed. Edward Arnold, London), un altro **coefficiente di concordanza è la u di Kendall**, per confronti tra coppie.
Tra i testi a maggior diffusione internazionale, questo test è riportato in quello di Sidney **Siegel** e N. John **Castellan** Jr. del 1988 ***Nonparametric Statistics for the Behavioral Sciences***, pubblicato in italiano con la traduzione a cura di Ettore **Caracciolo** nel 1992 (***Statistica non parametrica***, seconda edizione, Mc-Graw-Hill, Libri Italia, 477 p.)

E' una tecnica che può essere utile in particolare nella ricerca etologica, nello studio di preferenze di animali. Sono ugualmente possibili le applicazioni nelle indagini a carattere ambientale e ecologico, psicologico e di marketing, che come punto di riferimento hanno l'uomo.
Davanti a **N** oggetti, un animale difficilmente riesce ad esprimere preferenze complesse e stabilire un ordine di priorità tra più oggetti. Per esempio, davanti a 5 cibi diversi, un animale potrebbe alimentarsi solo con uno, ignorando gli altri dopo averli solo annusati; posto di fronte alla scelta tra 5 femmine, un maschio potrebbe preferirne solo una. In questi casi, al ricercatore non è fornito nessun elemento per stabilire un rango tra le altre 4 situazioni. Di conseguenza, nel suo piano sperimentale, il ricercatore deve impostare la ricerca in modo più complesso: ricorrere a una serie di confronti a coppie e porre l'individuo ogni volta davanti alla scelta tra due soli oggetti, annotando la preferenza, per poi ripetere l'esperimento per tutte le coppie possibili di **N** oggetti.

In problemi di marketing, a un gruppo di persone, invece di chiedere una graduatoria tra 5 oggetti, si presentano solo 2 oggetti e si chiede di indicare la preferenza tra uno dei due. E' un modo per evitare le incongruenze delle scelte, sempre possibili ma di difficile spiegazione logica. Ad esempio, confrontando 3 oggetti (indicati in A, B e C), un individuo può preferire A a B, successivamente B a C, ma nell'ultimo confronto scegliere (illogicamente rispetto al comportamento precedente) C rispetto ad A. Affinché le **preferenze** attribuite siano **consistenti e stabili**, quindi godano della **proprietà transitiva** esposta in precedenza, è utile ricorrere a questo metodo.

Docenti di psicologia, gli autori del testo Siegel e Castellan sottolineano come le **preferenze possono non essere transitive**. Affermano che classificare un gruppo di compagni a partire da quello con il quale preferirebbe giocare, non rientra tra i comportamenti "naturali" di un bambino, che non ha alcun problema nell'indicare la sua preferenza entro ogni coppia. Con **N** uguale a 5 oggetti, l'esperimento deve essere ripetuto 10 volte, come è semplice calcolare con le combinazioni di 5 elementi 2 a 2 (C_N^2).
Se con la ricerca s'intende verificare se **k** individui concordano nelle loro preferenze, occorre poi ripetere i 10 esperimenti per ognuno d'essi e riassumere tutti i risultati in una **matrice di preferenza**, per i **k** individui assieme.

Per esempio, con 5 cibi (A, B, C, D, E) differenti si è posto ognuno dei 4 animali (I, II, II, IV) davanti alla scelta in confronti appaiati, con i seguenti risultati

	Scelte dei 4 soggetti			
Coppie di 5 oggetti	I	II	III	IV
A,B	A	A	B	A
A,C	A	C	A	A
A,D	D	D	D	A
A,E	A	A	E	A
B,C	B	C	B	B
B,D	D	D	D	D
B,E	B	E	E	E
C,D	C	D	D	D
C,E	C	C	E	E
D,E	D	D	D	E

I dati raccolti possono essere riassunti in una tabella quadrata **N x N**, nella quale sulle righe sono riportate le preferenze relative od ogni oggetto nei confronti appaiati.
Ovviamente, non esistono dati sulla diagonale.

OGGETTO PREFERITO

	Secondo elemento della coppia				
Oggetto preferito	A	B	C	D	E
A	---	3	3	1	3
B	1	---	3	0	1
C	1	1	---	1	2
D	3	4	3	---	3
E	1	3	2	1	---

Per esempio,

- nel confronto **A,C** 3 soggetti hanno dato la loro preferenza ad **A** (riassunto nel 3 riportato nella casella di riga **A** e colonna **C**) e 1 la sua preferenza a **C** (1 riportato all'incrocio tra la riga **C** e la colonna **A**);

- nel confronto **B,D** i 4 soggetti hanno dato tutti la loro preferenza a **D** (nella tabella sovrastante, è riportato 4 all'incrocio tra la riga **D** e la colonna **B** e quindi è stato riportato 0 all'incrocio tra la riga **B** e la colonna **D)**.

Questa tabella riassuntiva, di forma quadrata e non simmetrica, fornisce informazioni sulla concordanza dei **k** valutatori nell'attribuzione delle preferenze.
Se fosse vera l'ipotesi nulla, in ogni cella si avrebbe 2 (**k/2** con k = 4 valutatori).
Se fosse vera l'ipotesi alternativa di totale concordanza tra i valutatori, teoricamente metà delle caselle avrebbero 4 (k con k valutatori) e metà 0, escludendo la diagonale.
Il coefficiente di concordanza u di Kendall
- è u = 0, in assenza di accordo,
- è u = 1, in presenza di totale accordo
variando in continuità tra questi due estremi.
Per il calcolo di u e la stima della sua significatività, si rinvia al testo di Siegel e Castellan, citato all'inizio del paragrafo.

21.10. LA REGRESSIONE LINEARE NON PARAMETRICA

Calcolata la retta di regressione con il metodo dei minimi quadrati,

$$\hat{Y}_i = a + b \cdot X_i$$

per la sua significatività è possibile ricorrere a test non parametrici, come illustrato in precedenza, quando si ha il sospetto che non siano state rispettate le condizioni di validità.
E' il caso in cui
- i valori di **Y** non hanno una distribuzione normale (come succede per la presenza di uno o più valori anomali),
- i dati della Y non hanno una varianza d'errore costante, cioè scarti uguali dalla retta, al variare delle X,
- i dati sono espressi in percentuale e/o come rapporti rispetto a quantità fortemente variabili (quindi non hanno lo stesso intervallo fiduciale) oppure sono un punteggio (o comunque una scala di rango).

Resta sempre il problema che, con pochi dati, diventa impossibile dimostrare la normalità e omoschedasticità della distribuzione, se non già confermata da altre ricerche.
In tale situazione di incertezza sulla non validità della regressione con il metodo parametrico dei minimi quadrati, è sempre conveniente, quando non necessario, calcolare **una retta di regressione non parametrica.**

Tra le metodologie, quella più diffusa è la **proposta da H. Theil,** pubblicata nel 1950 in tre articoli sullo stesso volume della rivista scientifica ***Indagationes Mathematicae***.

Successivamente, per la **significatività** sia del coefficiente angolare **b** sia dell'intercetta **a**, calcolati in qualsiasi modo, è possibile ricorrere a test non parametrici.

Tra i più diffusi, è da ricordare il **test di Theil (a distribution-free test for the slope coefficient)**, proposto con vari articoli a partire dal 1950, che può essere utilizzato per verificare:

- la significatività di una retta **b** di regressione, con ipotesi nulla

$$\mathbf{H_0: \beta = 0}$$

ed ipotesi alternativa $\mathbf{H_1}$ sia bilaterale che unilaterale;

- la significatività della differenza tra una retta campionaria **b** ed un coefficiente angolare β_0 atteso, con ipotesi nulla

$$\mathbf{H_0: \beta = \beta_0} \text{ (prefissato)}$$

ed ipotesi alternativa $\mathbf{H_1}$ che anche in questo caso può essere sia unilaterale che bilaterale.

Questa seconda situazione è semplicemente una generalizzazione del caso precedente. Le due metodologie sono quindi identiche, con la sola eccezione di un passaggio preliminare.

Un altro test utile all'inferenza sulle rette è il **test di Hollander** (**a distribution-free test for the parallelism of two regression lines**), proposto dalla seconda metà degli anni '60, per verificare

- il parallelismo tra due coefficienti angolari $\mathbf{b_1}$ e $\mathbf{b_2}$, e quindi l'ipotesi nulla

$$\mathbf{H_0: \beta_1 = \beta_2}$$

e l'ipotesi alternativa sempre sia unilaterale che bilaterale.

Con altri test non parametrici sulla regressione lineare semplice, inoltre è possibile:

- con il **test di Maritz,** valutare la significatività dell'intercetta **a**

$$\mathbf{H_0: \alpha = \alpha_0}$$

- con il test di **Lancaster-Quade,** verificare congiuntamente ed in modo indipendente la significatività sia di **a** sia di **b** nelle loro 3 combinazioni:

(significatività sia di **a** che di **b**, solo di **a**, solo di **b**)**,**

- avere stime ed intervalli di confidenza per i parametri α e β.

Nella letteratura sulla regressione non parametrica, si trovano altri test, seppure attualmente meno diffusi nei programmi informatici e nei testi internazionali, per verificare le ipotesi precedenti.

E' possibile ricordare

- il test di Brown e Mood, quello di Hajek, quello di Adichie e quello di Sievers, tutti sulla linearità e quindi alternativi al test di Theil;

- altri test, proposti uno ancora da Brown e Mood e un altro ancora da Adichie, possono essere usati in alternativa a quello di Maritz;
- inoltre Daniels, Konijn e Quade separatamente hanno proposto metodi alternativi a quello di Lancaster-Quade.

21.11. CALCOLO DELLA RETTA DI REGRESSIONE NON PARAMETRICA CON IL METODO DI THEIL O TEST DI THEIL-KENDALL.

Nel 1950, H. **Theil** (1950a – ***A rank-invariant method of linear and polynomial regression analysis, I,*** pubblicato su **Proc. Kon. Nederl. Akad. Wetensch** A. 53, pp. 386-392; 1950b – ***A rank-invariant method of linear and polynomial regression analysis, II,*** pubblicato su **Proc. Kon. Nederl. Akad. Wetensch** A. 53, pp. 521-525; 1950c – ***A rank-invariant method of linear and polynomial regression analysis, III,*** pubblicato su **Proc. Kon. Nederl. Akad. Wetensch** A. 53, pp. 1397-1412) ha proposto un metodo per **calcolare una retta di regressione non parametrica** (***Theil's regression method)***.

La significatività è testata con il test della correlazione **τ di Kendall**, come proposto da P. K. **Sen** nel 1968 (vedi l'articolo ***Estimates of the regression coefficient.based on Kendall's tau*** su **Journal of the American Statistical Association**, vol. 63, pp. 1379-1389) da cui il nome di **Theil – Kendall**, utilizzato in vari testi.

La procedura del calcolo della retta non parametrica si fonda sulla **mediana di tutte le rette, calcolate sulle possibili coppie di punti**.
Per ognuna di esse, identificate dalle coppie di variabili ($\mathbf{X_i}$, $\mathbf{Y_i}$) e ($\mathbf{X_j}$, $\mathbf{Y_j}$), si stima il coefficiente angolare $\mathbf{b_{ij}}$, con la relazione

$$b_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$$

Poiché $\mathbf{b_{ij}}$ è uguale a $\mathbf{b_{ji}}$ con la sola inversione del segno, si devono **quantificare tutte le possibili combinazioni** delle **N** osservazioni in cui **j** è maggiore di **i**. Il valore di **β** è stimato dalla mediana (indicata con **b*** per distinguerlo da **b** calcolato con la media) di questi valori che, con **N** coppie d'osservazioni, sono **N(N-1)/2**.
Per l'inferenza si assume che

$$\beta = b^*$$

In assenza di un programma informatico (pochissime librerie statistiche fino ad ora riportano il metodo della regressione non parametrica di Theil), con il calcolo manuale questa procedura richiede molto

tempo, quando il numero di punti diventa alto. Per esempio, già con 13 coppie di dati (**N = 13**) il numero di $\mathbf{b_{ij}}$ sui quali stimare la mediana diventa 72 (13 x 12 / 2).

Per N maggiore di 12, lo stesso **Theil** ha proposto una metodologia che richiede meno tempo, appunto il **metodo abbreviato di Theil** (***the abbreviated Theil method***), che richiede un numero d'operazioni nettamente minore. Essi non fornisce lo stesso risultato del metodo precedente.
E' quindi da utilizzare nell'impossibilità pratica di ricorrere al primo, poiché sfrutta solo in modo parziale l'informazione contenuta nell'insieme dei dati.

Per avvalersi della procedura abbreviata, dopo aver posto i dati in ordine crescente per la variabile **X**, si conta il numero **N** di coppie di valori. Il metodo differisce leggermente, se **N** è pari o dispari.

Per **N** pari, dopo aver separato i dati in due metà esatte, si devono calcolare **N/2** differenze
sia per la variabile **X** con

$$\mathbf{X_{ij} = X_{(i + N/2)} - X_i}$$

sia per la variabile **Y** con

$$\mathbf{Y_{ij} = Y_{(i + N/2)} - Y_i}$$

Successivamente, occorre individuare
- sia la mediana delle **N/2** differenze della **X**,
- sia la mediana delle **N/2** differenze della **Y**;

Il valore della retta **b*** è il rapporto tra queste due mediane

$$\mathbf{b^* = \text{mediana } Y_{ij} \,/\, \text{mediana } X_{ij}}$$

Per **N** dispari, sempre
- dopo aver ordinato la serie dei dati originari in modo crescente per X,
- si elimina una coppia di valori a caso.

Alcuni autori suggeriscono la coppia di valori X e Y corrispondenti alla mediana di X; di conseguenza, **il numero di punti diventa pari e si ricade nel metodo precedente**.

Ottenuto il valore di **b* come miglior stima di β**, è possibile calcolare il valore dell'intercetta **a** come miglior stima di **α,** con due metodi che seguono logiche diverse:

1 - in modo analogo al primo metodo descritto per **b*,** dapprima si calcolano tutte le $\mathbf{a_i}$, che in questo caso sono **N** come i punti ($\mathbf{X_i}$, $\mathbf{Y_i}$) rilevati

$$\mathbf{a_i = Y_i - b^* X_i}$$

e successivamente si stima **a*** come mediana delle **N** intercette $\mathbf{a_i}$ calcolate;

2 - in modo analogo alla statistica parametrica, in cui

$$\mathbf{a} = \overline{Y} - b \cdot \overline{X}$$

si stima **a**, indicata appunto con $\hat{a}$ per distinguerla dal valore parametrico, sostituendo nella formula precedente la mediana delle **X** e quella delle **Y** alle medie rispettive

$$\hat{a} \;=\; \textbf{mediana } (Y_i) \;-\; \textbf{b}^* \bullet \textbf{mediana}(X_i)$$

Con questo metodo, più conveniente nel caso di grandi campioni perché più rapido, si ottiene una **retta** che passa **non per l'incrocio delle medie** ma per **l'incrocio delle mediane,** considerato **il baricentro non parametrico** del **diagramma di dispersione** dei punti.

I due metodi possono dare risultati differenti.

La metodologia descritta per il **calcolo della retta di regressione lineare non parametrica** può essere spiegata in modo più semplice, comprensibile anche a non esperti di statistica, illustrando un esempio in tutti i suoi passaggi.
Il metodo esteso, applicabile quando si dispone di piccoli campioni, sarà presentato separatamente da quello abbreviato con un secondo esempio.

<u>Metodo per campioni piccoli</u> (N ≤ 12).
Si supponga di voler valutare gli effetti di 7 dosaggi di una sostanza tossica (**X**), su vari campioni di una popolazione animale. La dose 0 (zero), detto anche campione bianco, corrisponde all'assenza del principio attivo; sovente serve come controllo. I risultati sono stati misurati come percentuale d'individui morti (**Y**) su una serie di somministrazioni, ottenendo i seguenti dati:

Rango delle X	1	2	3	4	5	6	7
Valori di X	0	1	2	3	4	5	6
Valori di Y	2,9	3,1	3,4	4,0	4,6	5,1	12,4

Per calcolare il coefficiente angolare **b** mediante il metodo proposto da **Theil,** è utile seguire le procedure di seguito descritte:

1 – ordinare i valori di **X** in modo crescente (operazione già effettuata nella tabella di presentazione dei dati, trattandosi di dosi crescenti);

2 – quantificare le possibili combinazioni (**N (N – 1) / 2**),

che in questo caso sono 21 (7 x 6 / 2)

e per ognuna calcolare il coefficiente angolare $\mathbf{b_{ij}}$ mediante la relazione

$$b_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$$

con **j** maggiore di **i.**

Di seguito, sono riportati tutti i risultati e i calcoli relativi nei loro passaggi:

$b_{12} = (Y_2 - Y_1) / (X_2 - X_1) = (3{,}1 - 2{,}9) / (1 - 0) = 0{,}200$

$b_{13} = (Y_3 - Y_1) / (X_3 - X_1) = (3{,}4 - 2{,}9) / (2 - 0) = 0{,}250$

$b_{14} = (Y_4 - Y_1) / (X_4 - X_1) = (4{,}0 - 2{,}9) / (3 - 0) = 0{,}367$

$b_{15} = (Y_5 - Y_1) / (X_5 - X_1) = (4{,}6 - 2{,}9) / (4 - 0) = 0{,}425$

$b_{16} = (Y_6 - Y_1) / (X_6 - X_1) = (5{,}1 - 2{,}9) / (5 - 0) = 0{,}440$

$b_{17} = (Y_7 - Y_1) / (X_7 - X_1) = (12{,}4 - 2{,}9) / (6 - 0) = 1{,}583$

$b_{23} = (Y_3 - Y_2) / (X_3 - X_2) = (3{,}4 - 3{,}1) / (2 - 1) = 0{,}300$

$b_{24} = (Y_4 - Y_2) / (X_4 - X_2) = (4{,}0 - 3{,}1) / (3 - 1) = 0{,}450$

$b_{25} = (Y_5 - Y_2) / (X_5 - X_2) = (4{,}6 - 3{,}1) / (4 - 1) = 0{,}500$

$b_{26} = (Y_6 - Y_2) / (X_6 - X_2) = (5{,}1 - 3{,}1) / (5 - 1) = 0{,}500$

$b_{27} = (Y_7 - Y_2) / (X_7 - X_2) = (12{,}4 - 3{,}1) / (6 - 1) = 1{,}860$

$b_{34} = (Y_4 - Y_3) / (X_4 - X_3) = (4{,}0 - 3{,}4) / (3 - 2) = 0{,}600$

$b_{35} = (Y_5 - Y_3) / (X_5 - X_3) = (4{,}6 - 3{,}4) / (4 - 2) = 0{,}600$

$b_{36} = (Y_6 - Y_3) / (X_6 - X_3) = (5{,}1 - 3{,}4) / (5 - 2) = 0{,}567$

$b_{37} = (Y_7 - Y_3) / (X_7 - X_3) = (12{,}4 - 3{,}4) / (6 - 2) = 2{,}250$

$b_{45} = (Y_5 - Y_4) / (X_5 - X_4) = (4{,}6 - 4{,}0) / (4 - 3) = 0{,}600$

$b_{46} = (Y_6 - Y_4) / (X_6 - X_4) = (5{,}1 - 4{,}0) / (5 - 3) = 0{,}550$

$b_{47} = (Y_7 - Y_4) / (X_7 - X_4) = (12{,}4 - 4{,}0) / (6 - 3) = 2{,}800$

$b_{56} = (Y_6 - Y_5) / (X_6 - X_5) = (5{,}1 - 4{,}6) / (6 - 5) = 0{,}500$

$b_{57} = (Y_7 - Y_5) / (X_7 - X_5) = (12{,}4 - 4{,}6) / (7 - 5) = 3{,}900$

$b_{67} = (Y_7 - Y_6) / (X_7 - X_6) = (12{,}4 - 5{,}1) / (7 - 6) = 7{,}300$

Questi risultati solitamente sono pubblicati in modo più sintetico, sotto forma di una matrice triangolare come la seguente:

X Y	0 2,9	1 3,1	2 3,4	3 4,0	4 4,6	5 5,1	6 12,4
X = 0 ; Y = 2,9	---	0,200	0,250	0,367	0,425	0,440	1,583
X = 1 ; Y = 3,1	---	---	0,300	0,450	0,500	0,500	1,860
X = 2 ; Y = 3,4	---	---	---	0,600	0,600	0,567	2,250
X = 3 ; Y = 4,0	---	---	---	---	0,600	0,550	2,800
X = 4 ; Y = 4,6	---	---	---	---	---	0,500	3,900
X = 5 ; Y = 5,1	---	---	---	---	---	---	7,300
X = 6 ; Y = 12,4	---	---	---	---	---	---	---

3 – stimare la mediana di questi **N(N-1)/2** valori b_{ij} **;** è facilmente identificata dalla sua serie ordinata per ranghi:

Rango	1	2	3	4	5	6	7	8	9	10	**11**
b_{ij}	0,200	0,250	0,300	0,367	0,425	0,440	0,450	0,500	0,500	0,500	**0,550**

Rango	12	13	14	15	16	17	18	19	20	21
b_{ij}	0,567	0,600	0,600	0,600	1,583	1,860	2,250	2,800	3,900	7,900

In questo esempio, la mediana risulta uguale a 0,550 corrispondendo alla 11ª posizione sulle 21 misure stimate; di conseguenza, si assume

$$b^* = 0{,}550$$

4 – con **N** uguale a 7,
dal valore di **b*** e mediante la relazione

$$a_i = Y_i - b^* X_i$$

si calcolano altrettanti valori delle intercette $\mathbf{a_i}$.

Di seguito, sono riportati tutti i risultati con i dati dell'esempio:

$$a_1 = 2{,}9 - 0{,}550 \times 0 = 2{,}90$$
$$a_2 = 3{,}1 - 0{,}550 \times 1 = 2{,}55$$
$$a_3 = 3{,}4 - 0{,}550 \times 2 = 2{,}30$$
$$a_4 = 4{,}0 - 0{,}550 \times 3 = 2{,}35$$
$$a_5 = 4{,}6 - 0{,}550 \times 4 = 2{,}40$$
$$a_6 = 5{,}1 - 0{,}550 \times 5 = 2{,}35$$
$$a_7 = 12{,}4 - 0{,}550 \times 6 = 9{,}10$$

5 – la mediana di questi **N** valori $\mathbf{a_i}$ è identificata dalla sua serie ordinata per rango:

Rango	1	2	3	**4**	5	6	7
a_i	2,30	2,35	2,35	**2,40**	2,55	2,90	9,10

Coincidendo con il quarto dei 7 valori, è uguale a 2,40.

Di conseguenza,

$$\mathbf{a^* = 2{,}40}$$

6 – la retta calcolata in modo non parametrico, nella sua forma estesa, per i 7 punti rilevati è

$$\hat{Y}_i = \mathbf{2{,}4 + 0{,}55} \bullet X_i$$

<u>Un modo alternativo per calcolare l'intercetta a</u> è:

1- individuare nei dati originari

Valori di X	0	1	2	**3**	4	5	6
Valori di Y	2,9	3,1	3,4	**4,0**	4,6	5,1	12,4

la mediana delle X_i, che risulta uguale a 3, e

la mediana delle Y_i, che risulta uguale a 4,0;

2 – dalla relazione

$$\hat{a} = \textbf{mediana } (Y_i) - b^* \bullet \textbf{mediana}(X_i)$$

calcolare il valore di $\hat{a}$

$$\hat{a} = 4{,}0 - 0{,}55 \times 3 = 4{,}0 - 1{,}65 = 2{,}35$$

che risulta uguale a **2,35.**

<u>Metodo per grandi campioni</u> (N > 12).

Si supponga di voler valutare la crescita media di una specie animale con l'aumentare dell'età. A questo scopo, sono stati raccolti campioni d'individui dall'età 4 all'età 20, stimando per ognuno la lunghezza media del campione:

Età X	4	5	6	7	8	9	10	11	12
Lungh. Y	40	45	51	55	60	67	68	65	71

Età X	13	14	15	16	17	18	19	20
Lungh. Y	74	76	76	78	83	82	85	89

Con 17 osservazioni, il metodo di Theil nella versione estesa richiederebbe il calcolo di 136 (17 x 16 / 2) coefficienti angolari $\mathbf{b_{ij}}$. Per effettuare l'operazione in tempi non eccessivamente lunghi, è conveniente utilizzare

il **<u>metodo abbreviato</u>** (che segue le procedure seguenti):

1 – Si ordinano i dati della variabile **X** per rango (spesso è un'operazione già effettuata nella tabella di presentazione, come n questo caso).

2 - Si individua la mediana delle **X**: su 17 dati è il nono valore e corrisponde al punto (X = 12; Y = 71);

successivamente si calcolano le 8 differenze

- sia per la variabile **X** tra i valori $\mathbf{X_{(i + N/2)}}$ e $\mathbf{X_i}$:

$$X_{10} - X_1 = 13 - 4 = 9$$
$$X_{11} - X_2 = 14 - 5 = 9$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots..$$
$$X_{17} - X_8 = 20 - 11 = 9$$

(nel caso specifico sono tutte uguali a 9)

- sia per la variabile **Y** tra i valori $\mathbf{Y_{(i+N/2)}}$ e $\mathbf{Y_i}$:

$$Y_{10} - Y_1 = 74 - 40 = 34$$
$$Y_{11} - Y_2 = 76 - 45 = 31$$
$$Y_{12} - Y_3 = 76 - 51 = 25$$
$$Y_{13} - Y_4 = 78 - 55 = 23$$
$$Y_{14} - Y_5 = 83 - 60 = 23$$
$$Y_{15} - Y_6 = 82 - 67 = 15$$
$$Y_{16} - Y_7 = 85 - 68 = 17$$
$$Y_{17} - Y_8 = 89 - 65 = 24$$

3 – Nelle 8 differenze di **X** e di **Y** si scelgono le 2 mediane:

- per $\mathbf{X_{ij}}$, essendo in questo caso tutte uguali a 9, ovviamente la mediana è 9;
- per $\mathbf{Y_{ij}}$ conviene ordinare gli 8 valori in modo crescente

15 17 23 23 24 25 31 34

e dalla serie ordinata per rango emerge che la mediana cade tra 23 (4° valore) e 24 (5° valore) e quindi è 23,5.

4 – Utilizzando la formula

$$\mathbf{b^* = mediana\ Y_{ij} / mediana\ X_{ij}}$$

si ottiene un valore di **b***

$$\mathbf{b^* = 23{,}5 / 9 = 2{,}611}$$

uguale a 2,611.

5 – Da **b*** si stima $\hat{a}$, ovviamente con il metodo più breve: dopo aver identificato, sulla serie dei dati originari,

la mediana delle **X** che risulta uguale a 12 e

la mediana delle **Y** che risulta uguale a 71,

attraverso la relazione

$$\hat{a} = \mathbf{mediana\ (Y_i) - b^* \bullet mediana(X_i)}$$

si calcola il valore di $\hat{a}$

$$\hat{a} = 71 - 2{,}611 \cdot 12 = 71 - 31{,}332 = 39{,}668$$

che risulta uguale a **39,668**

6 – **La retta di regressione lineare semplice non parametrica con il metodo abbreviato di Theil** risulta

$$\hat{Y}_i = 39{,}668 + 2{,}661 \cdot X_i$$

21.12. CONFRONTO TRA LA RETTA PARAMETRICA E LA RETTA DI THEIL

Ai fini di un'interpretazione corretta dei risultati della regressione, è sempre utile costruire il diagramma di dispersione dei punti campionari, con la retta relativa. Per una scelta ragionata, è necessario comprendere esattamente

- sia le caratteristiche distintive della retta di regressione lineare semplice non parametrica, calcolata con il metodo di Theil,
- sia le differenze rispetto a quella parametrica, calcolata con il principio dei minimi quadrati.

A questo scopo, dopo aver ripreso i dati dell'esempio precedente

Valori di X	0	1	2	**3**	4	5	6
Valori di Y	2,9	3,1	3,4	**4,0**	4,6	5,1	12,4

- sui quali è stata calcolata la regressione di Theil nella versione estesa:

$$\hat{Y}_i = 2{,}4 + 0{,}55 \cdot X_i$$

- si calcola anche la retta di regressione parametrica, che con i parametri stimati dai dati della tabella,

$\sum X \cdot Y = 140{,}2$ $\sum X = 21$ $\sum Y = 35{,}5$ $\sum X^2 = 91$ n = 7 $\overline{X} = 3$ $\overline{Y} = 5{,}07$

applicati alla formula per il coefficiente angolare

$$b = \frac{\sum X \cdot Y - \dfrac{\sum X \cdot \sum Y}{N}}{\sum X^2 - \dfrac{\left(\sum X\right)^2}{N}}$$

si ottiene un valore di **b**

$$b = \frac{140{,}2 - \frac{21 \cdot 35{,}5}{7}}{91 - \frac{21^2}{7}} = \frac{140{,}2 - 106{,}5}{91 - 63} = \frac{33{,}7}{28} = 1{,}20$$

uguale a 1,20 e da esso un valore di **a**

$$a = \overline{Y} - b \cdot \overline{X}$$

$$a = 5{,}07 - 1{,}20 \cdot 3 = 5{,}07 - 3{,}60 = 1{,}47$$

uguale a 1,47

per cui la retta parametrica è

$$\hat{Y}_i = 1{,}47 + 1{,}20 \cdot X_i$$

Sia il confronto visivo

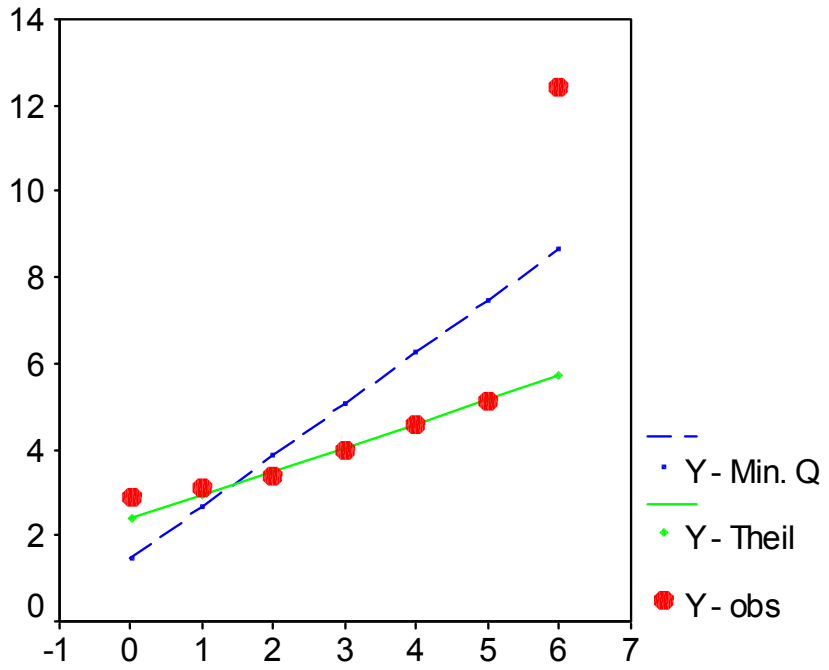
- tra i punti sperimentali e le due rette (figura successiva),
- sia la differenza tra i valori di Y osservati (riga 1 nella tabella successiva)
- e quelli calcolati con la retta di Theil (riga 2 nella tabella successiva)
- e quella dei minimi quadrati (riga 3 nella tabella successiva)

1	Valori di Y osservati	2,90	3,10	3,40	4,00	4,60	5,10	12,40
2	Y calcolati con Theil	2,40	2,95	3,50	4,05	4,60	5,15	5,70
3	Y calcolati con minimi quadrati	1,47	2,67	3,87	5,07	6,27	7,47	8,67

evidenzia come l'ultima osservazione sia anomala, rispetto alle altre 5; inoltre come la retta non parametrica si avvicini molto ai primi 5 punti ignorando praticamente l'ultimo dato, mentre la retta parametrica sia da esso attratta, allontanandosi dagli altri.

E' una stima simile a quello della mediana rispetto alla media, in presenza di un valore anomalo. In questo caso, l'effetto di distorsione della stima parametrica rispetto ai dati reali è accentuato dal fatto che **la retta parametrica minimizza la somma dei quadrati degli scarti**.

Oltre che dalle condizioni di validità, la scelta dipende quindi dal valore che si vuole attribuire all'osservazione anomala, rispetto a tutte le altre.



Nel grafico,

- i punti sono i 7 valori osservati,

- la retta che incrocia i 6 punti e si avvicina al valore anomalo è la retta parametrica calcolata con il metodo dei minimi quadrati,

- la retta che passa per i 6 punti ed ignora il valore anomalo è quella non parametrica, calcolata con il metodo di Theil.

21.13. SIGNIFICATIVITA' DI b CON IL τ DI KENDALL.

Il **test di Theil** per la significatività del coefficiente **b** di regressione lineare semplice verifica

- l'**ipotesi bilaterale**

$$\mathbf{H_0: \beta = 0} \quad \text{contro} \quad \mathbf{H_1: \beta \neq 0}$$

- oppure **una delle due ipotesi unilaterali**

$$\mathbf{H_0: \beta \leq 0} \quad \text{contro} \quad \mathbf{H_1: \beta > 0}$$

$$\mathbf{H_0: \beta \geq 0} \quad \text{contro} \quad \mathbf{H_1: \beta < 0}$$

in funzione della conoscenza del problema e quindi della domanda al quale il ricercatore è interessato.

Come già riportato nel paragrafo dedicato alla regressione, la significatività del coefficiente angolare (b) calcolato è verificata con il test di correlazione **τ di Kendall**. E' la proposta originaria di H. **Theil,** nell'articolo del 1950 ***A rank-invariant method of linear and polynomial regression analysis*** (pubblicato su **Indagationes Mathematicae** Vol. 12, pp. 85-91) e generalizzato da P. K. **Sen** nel 1968 con l'articolo ***Estimates of the regression coefficient based on Kendall's tau*** (su **Journal of the American Statistical Association**, vol. 63, pp. 1379-1389). Da qui il nome di **Theil – Kendall** dato al metodo per la stima della regressione lineare non parametrica, comprendendo nella metodologia anche il test di significatività per il coefficiente angolare.

La procedura può essere spiegata in modo semplice, con un esempio. Al fine di evidenziare le analogie con il test parametrico e per un successivo confronto dei risultati, l'esempio utilizza gli stessi dati che sono serviti per il calcolo della regressione lineare semplice parametrica.

1 - Dopo aver individuata la variabile dipendente o effetto (Y) e la variabile indipendente o causale (X),

Individui	A	B	C	D	E	F	G
Peso (Y)	52	68	75	71	63	59	57
Altezza (X)	160	178	183	180	166	175	162

2 - si stima il coefficiente angolare **b**, che risulta

$$b = \frac{\sum(x \cdot y) - \frac{\sum x \cdot \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{76945 - \frac{1204 \cdot 445}{7}}{207598 - \frac{1204^2}{7}} = 0,796$$

uguale a 0,796

Anche se priva di significato in questo problema specifico (non esiste alcuna persona di altezza 0), l'intercetta **a**

$$a = \overline{Y} - b\overline{X} = 63,571 - 0,796 \cdot 172 = -73,354$$

risulta uguale a -73,354

ed è utile per scrivere in modo completo la retta di regressione

$$\hat{Y}_i = -73{,}354 + 0{,}796 \cdot X_i$$

3 - Per valutare la significatività del coefficiente angolare **b**, con ipotesi nulla

$$\mathbf{H_0: \beta = 0}$$

e, in questo caso, con ipotesi alternativa bilaterale

$$\mathbf{H_1: \beta \neq 0}$$

i dati devono essere ordinati secondo il rango della variabile **X** (l'altezza),

Individui	A	G	E	F	B	D	C
Peso (Y)	**52**	**57**	**63**	**59**	**68**	**71**	**75**
Altezza (X)	160	162	166	175	178	180	183

riportando gli Y_i relativi, che successivamente devono essere trasformati in ranghi entro la variabile Y:

Individui	A	G	E	F	B	D	C
Peso (Y)	52	57	63	59	68	71	55
Y in ranghi	**1**	**2**	**4**	**3**	**5**	**6**	**7**

4 - Ponendo l'attenzione solo sui valori della Y, se all'aumentare di X

- i valori di Y tendono ad aumentare (quindi i ranghi di Y sono in ordine naturale), la regressione tende ad essere significativa, con coefficiente angolare positivo,
- il valore di Y resta approssimativamente costante (quindi i dati di Y sono in ordine casuale), la regressione è assente o non significativa,
- il valore di Y tende a diminuire (quindi i ranghi di Y sono in ordine decrescente), la regressione tende ad essere significativa, con coefficiente angolare negativo.

Per quantificare il grado di correlazione o concordanza dei ranghi di Y con l'ordine naturale, si può utilizzare la proposta di Kendall: contare **quante sono le coppie di ranghi che sono concordanti e quante quelle discordanti dall'ordine naturale**.

Per un calcolo corretto, facilmente verificabile, è utile riportare

- il conteggio dettagliato delle concordanze (+) e delle discordanze (-)
- e il loro totale generale.

1	2	4	3	5	6	7	Totale
	+	+	+	+	+	+	**+6**
		+	+	+	+	+	**+5**
			-	+	+	+	**+2**
				+	+	+	**+3**
					+	+	**+2**
						+	**+1**
Totale (concordanze meno discordanze)							**+19**

La misura della concordanza complessiva con la variabile X è dato dalla somma algebrica di tutte le concordanze e le discordanze.
Il totale di concordanze e discordanze con i 7 valori dell'esempio è **+19.**

5 - Il numero totale di concordanze e discordanze di una serie di valori deve essere rapportato al numero **massimo totale possibile**. Poiché i confronti sono fatti a coppie, con **N** dati il numero totale di confronti concordanti o discordanti è dato dalla combinazione di **N** elementi **2** a **2**

$$C_N^2$$

Con una serie di 7 dati come nell'esempio, il numero complessivo di confronti, quindi il massimo totale possibile di concordanze o discordanze, è

$$C_7^2 = \frac{7!}{(7-2)! \cdot 2!} = \mathbf{21}$$

Secondo il metodo proposto di Kendall, il grado di relazione o concordanza (τ) tra la variabile X e Y può essere quantificato dal rapporto

$$\tau = \frac{totale(concordanze - discordanze)}{massimo\ \ totale\ \ possibile} = \frac{totale(concordanze - discordanze)}{C_N^2}$$

Con i 7 dati riportati,

$$\tau = \frac{+19}{21} = +0{,}905$$

è uguale a +0,905.

6 - La **scala** dovrebbe essere **continua**, ma sono **accettati valori discreti**.
In caso di due o più valori identici, il confronto tra due punteggi di Y uguali non determina né una concordanza né una discordanza: il loro confronto non contribuisce al calcolo di τ, abbassando il valore al numeratore. Di conseguenza, deve essere diminuito anche il valore al denominatore.
La formula, corretta per la presenza di valori identici (ties), diventa

$$\tau = \frac{2 \cdot totale(concordanze - discordanze)}{\sqrt{N \cdot (N-1) \cdot T_x} \cdot \sqrt{N \cdot (N-1) \cdot T_y}}$$

dove

- **N** è il numero totale di coppie di dati delle variabili **X** e **Y,**
- T_x è dato da $T_x = \sum(t_x^2 - t_x)$ dove t_x è il numero di osservazioni identiche di ogni gruppo di valori identici della variabile X,
- T_y è dato da $T_y = \sum(t_y^2 - t_y)$ dove t_y è il numero di osservazioni identiche di ogni gruppo di valori identici della variabile Y.

Per **piccoli campioni**, i valori critici sono forniti dalla tabella relativa, riportata nel paragrafo della correlazione non parametrica τ di **Kendall**.
Con 7 dati, alla probabilità α = 0.005 per un test ad una coda e 0.01 per un test a 2 code, il valore critico riportato è 0.810.
Il valore calcolato è superiore a quello della tabella: si rifiuta l'ipotesi nulla. In un test bilaterale la risposta dovrebbe essere: il coefficiente di regressione lineare b si discosta significativamente da 0.
Se la domanda fosse stata unilaterale positiva, la risposta dovrebbe essere: all'aumentare dell'altezza il peso aumenta in modo significativo.

Per **grandi campioni** la significatività del τ di Kendall può essere verificata con la distribuzione normale

$$Z = \frac{\tau - \mu_\tau}{\sigma_\tau} \quad (*)$$

dove

$$\mu_\tau = 0$$

e

$$\sigma_\tau^2 = \frac{2 \cdot (2N + 5)}{9N \cdot (N - 1)}$$

e **N** = numero di dati.

Sostituendo e semplificando, si ottiene una stima più rapida di **Z** mediante la relazione

$$Z = \frac{3\tau \cdot \sqrt{n \cdot (n-1)}}{\sqrt{2 \cdot (2n + 5)}}$$

Per la **verifica dell'ipotesi nulla più generale**

$$\mathbf{H_0: \beta = \beta_0}$$

(con ipotesi alternativa sia unilaterale che bilaterale) dove β_0 è un coefficiente angolare qualsiasi, quindi anche 0, il metodo di **Theil** ha una leggera modifica nelle prime fasi.

E' proposto il medesimo esempio e si evidenziano le differenze tra i due metodi nei vari passaggi.

1 – Si supponga di avere già calcolato il coefficiente angolare **b** (uguale a 0,796) e di volere verificare l'ipotesi se esso non si discosti dal valore $\beta_0 = 0{,}9$ (aumento di Kg 0,9 in peso per l'aumento di 1 cm. in altezza)

$$\mathbf{H_0: \beta \geq 0{,}9}$$

contro l'ipotesi alternativa che la crescita media in peso sia minore

$$\mathbf{H_1: \beta < 0{,}9}$$

2 – Il passo successivo consiste nel calcolare i valori attesi di **Y**, secondo la relazione

$$\hat{Y}_i = a + 0{,}9 \cdot X_i$$

Tuttavia, poiché l'interesse è rivolto non ai singoli valori ma al loro rango, i valori attesi possono essere più semplicemente calcolati come

$$\hat{Y}_i = 0{,}9 \cdot X_i$$

Con formula più generale si utilizza

$$\hat{Y}_i = \beta_0 \cdot X_i$$

Con i dati dell'esempio, si ottiene

Individui	A	B	C	D	E	F	G
Peso (Y osservati)	52	68	75	71	63	59	57
Y teorici o attesi	**144,0**	**160,2**	**164,7**	**162,0**	**149,4**	**157,5**	**145,8**
Altezza (X)	160	178	183	180	166	175	162

3 – Calcolare le differenze **D_i** tra **Y_i attesi** e **Y_i osservati**

Individui	A	B	C	D	E	F	G
Di (Y_i teorici – Y_i attesi)	**92,0**	**92,2**	**89,7**	**91,0**	**86,4**	**98,5**	**88,8**
Altezza (X)	160	178	183	180	166	175	162

4 – Ordinare i dati secondo il rango di **X** (l'altezza),

Individui	A	G	E	F	B	D	C
D_i	**92,0**	**88,8**	**86,4**	**98,5**	**92,2**	**91,0**	**89,7**
Altezza (X)	160	162	166	175	178	180	183

riportare i dati delle differenze **D_i** (nella riga centrale) e successivamente trasformarle in ranghi (come nella tabella seguente)

Individui	A	G	E	F	B	D	C
D_i	92,0	88,8	86,4	98,5	92,2	91,0	89,7
D_i in ranghi	**5**	**2**	**1**	**7**	**6**	**4**	**3**

5 – Porre l'attenzione sui soli valori delle differenze **D_i (Y_i attesi – Y_i osservati)**:

- se il valore di β è statisticamente uguale a β_0, il rango delle **D_i** dipenderà solo dalle variazioni casuali in peso e sarà indipendente dal coefficiente angolare **b** (l'intercetta **a** è una costante),
- se il valore di β è statisticamente maggiore di β_0, il rango delle **D_i** tenderà ad essere in ordine inverso a quello delle **X_i**,
- se il valore di β è statisticamente minore di β_0, il rango delle **D_i** tenderà ad avere lo stesso ordine di **X_i**.

6 – Contare **quante sono le coppie di ranghi delle D_i che sono concordanti e quante quelle discordanti dall'ordine naturale**:

5	**2**	**1**	**7**	**6**	**4**	**3**	**Totale**
	-	-	+	+	-	-	**-2**
		-	+	+	+	+	**+3**
			+	+	+	+	**+4**
				-	-	-	**-3**
					-	-	**-2**
						-	**-1**
Totale (concordanze meno discordanze)							**-1**

La misura della concordanza complessiva tra Y osservati e Y attesi, quindi di β con β_0, è data dalla somma totale delle concordanze e discordanze:
essa risulta uguale a –1.

7 – Il valore di τ, dato dal rapporto del numero totale delle concordanze meno le discordanze rispetto al numero massimo possibile, risulta

$$\tau = -1 / 21 = - 0{,}047$$

eguale a **– 0,047**.

Il valore calcolato non solo è inferiore a quello riportato nella tabella, ma è molto vicino a quello atteso nell'ipotesi nulla; di conseguenza, si può affermare non solo che **b non si discosta significativamente da β_0**, ma anche che **β è molto vicino a β_0.**

Per una scelta adeguata tra il test parametrico e il corrispondente non parametrico, è importante conoscere la loro efficienza. Nel caso di grandi campioni, dipende dalla forma di distribuzione dei dati, di solito schematizzata in tre situazioni: normale, rettangolare ed esponenziale doppia.

<u>L'efficienza asintotica</u> del test di Theil rispetto al t di Student per il test parametrico

- è uguale a 0,95 ($3/\pi$), quando la distribuzione è Normale;
- è uguale a 1, quando la distribuzione è Rettangolare;
- è uguale a 1,5 (3/2), quando la distribuzione è Esponenziale Doppia.

In termini elementari, quando la distribuzione è perfettamente normale, i due test hanno efficienza molto simile; ma quando la forma della distribuzione è lontana dalla normalità, il test non parametrico è più efficiente di quello parametrico.

La regressione lineare semplice non parametrica può essere utilizzata in modo appropriato anche per

- **analizzare una serie storica o geografica di dati,**
- **cioè gli effetti della distanza da un'origine**

che può essere di natura qualsiasi, da geografica a temporale.

Già nel 1945, prima ancora della proposta di Theil, H. B. **Mann** (con l'articolo ***Non parametric test against trend***, sulla rivista **Econometrica** vol. 13, pp. 245-259) affermava che quando X è una misura temporale, che può essere espressa in anni, mesi, giorni, ore o secondi, la regressione non parametrica con ipotesi H_0: $\beta = 0$ ed ipotesi alternativa H_1 sia unilaterale che bilaterale, può essere impiegata per **verificare se esiste un trend,** cioè una tendenza alla diminuzione o all'aumento del carattere Y.

Analogo a questo test è **quello proposto da Daniel,** cioè della correlazione non parametrica per il trend, già illustrato in questo capitolo. E' utlizzato più diffusamente del test di Mann.

21.14. LA REGRESSIONE LINEARE NON PARAMETRICA CON IL METODO DEI TRE GRUPPI DI BARTLETT

Oltre al **metodo di Theil,** la cui significatività è analizzata mediante la correlazione non parametrica τ **di Kendall e** pertanto il metodo è chiamato sia **Theil-Kendall**, sia **metodo robusto di Kendall (*Kendall's robust line-fit method*)**, un altro metodo non parametrico ancora più semplice e rapido, ma molto meno diffuso, è riportato nei testi di

- Robert R. **Sokal** e F. James **Rohlf** del 1995 ***Biometry. The principles and practice of statistics in biological research*** (3^{rd} ed. W. H. Freeman and Company, New York, XIX, + 887 p.).

- Owen L. **Davis** e Peter L. **Goldsmith** del 1980 ***Statistical Methods in Research and Production***, ***with special reference to Chemical Industry*** (4th Revised Edition, published for Imperial Chemical Industries Limited, Longman Group Limited, London , XIII + 478 pp.)

E' il **metodo di Bartlett** o più estesamente

- **metodo dei tre gruppi di Bartlett** (***Bartlett's three-group method***),

proposto appunto da M. S. **Bartlett** nel 1949 con l'articolo ***Fitting a Straight Line when Both variables are Subject to Error*** (pubblicato su **Biometrics** Vol. 5 , pp.: 207-212).

La procedura del test può essere illustrata sviluppando un esempio già utilizzato per presentare il metodo di Theil, sia per mostrarne con maggiore evidenza i differenti approcci sia per favorire il confronto tra i risultati.

Si supponga di voler valutare la crescita media di una specie animale con l'aumentare dell'età, in una fase della vita in cui la successione dei valori nel tempo può essere considerata lineare. Ma, in questo esempio, per l'analisi si dispone solamente dei valori medi di un gruppo individui.

Ne consegue che non è possibile ricorrere alla statistica parametrica, poiché nei test d'inferenza parametrici si richiede sempre la variabilità d'errore sia valutata a partire dalle singole osservazioni.

Per il problema presentato, sono stati raccolti campioni di individui con età variabile da 4 a 20 giorni (X) e stimando per ogni età la dimensione media del campione (Y):

Età X	4	5	6	7	8	9	10	11	12
Lungh. Y	40	45	51	55	60	67	68	65	71

Età X	13	14	15	16	17	18	19	20
Lungh. Y	74	76	76	78	83	82	85	89

La procedura richiede la seguente serie di passaggi logici:

1 - Ordinare i valori della variabile X, in modo crescente. In questo caso, essendo X il tempo o l'età, la variabile X è già ordinata.

2 - Prendendo i valori secondo l'ordine della X, costruire tre gruppi di dimensioni possibilmente uguali, in particolare devono avere lo stesso numero di osservazioni il primo e il terzo gruppo.
Con i dati dell'esempio (n = 17), appare logico formare i tre gruppi con un numero di dati uguale a 6, 5 e 6 rispettivamente

	Gruppo 1		Gruppo 2		Gruppo 3	
	X_1	Y_1	X	Y	X_3	Y_3
	4	40	10	68	15	76
	5	45	11	65	16	78
	6	51	12	71	17	83
	7	55	13	74	18	82
	8	60	14	76	19	85
	9	67			20	89
Medie	6,50	53,00			17,50	82,17

3 – Da questi dati occorre **calcolare le medie** del **gruppo 1** e del **gruppo 3**

- $\overline{X}_1 = 6{,}50 \quad \overline{Y}_1 = 53{,}00$
- $\overline{X}_3 = 17{,}50 \quad \overline{Y}_3 = 82{,}17$

e le due **medie generali** considerando tutti i 17 dati

$$\overline{X} = \frac{204}{17} = 12{,}00 \qquad \overline{Y} = \frac{1.165}{17} = 68{,}53$$

4 – Il **coefficiente angolare** b è

$$b = \frac{\overline{Y}_3 - \overline{Y}_1}{\overline{X}_3 - \overline{X}_1} = \frac{82{,}17 - 53{,}00}{17{,}50 - 6{,}50} = \frac{29{,}17}{11{,}00} = 2{,}65$$

5 – L'**intercetta** a è

$$a = \overline{Y} - b\overline{X} = 68{,}53 - 2{,}65 \cdot 12{,}00 = 68{,}53 - 31{,}80 = 36{,}73$$

6 – Infine **la retta non parametrica di Bartlett** è

$$\hat{Y}_i = a + bX_i = 36{,}73 + 2{,}65 \cdot X_i$$

7 – La **retta non parametrica con il metodo abbreviato di Theil,** presentata nei paragrafi precedenti, risultava

$$\hat{Y}_i = a + bX_i = 39{,}668 + 2{,}661 \cdot X_i$$

Tra i due metodi,

- la preferenza degli statistici è attribuita al metodo di Theil,

in quanto utilizza le mediane. Ne deriva che è meno sensibile alla eventuale non normalità della distribuzione, in particolare alla asimmetria determinata dalla presenza di dati anomali: la dizione di **metodo robusto** attribuito da vari statistici al metodo di **Theil – Kendall (*Kendall's robust line-fit method*)** dipende appunto da queste osservazioni.

Per la **significatività della regressione lineare**, anche con **questo metodo di Bartlett**

- si può ricorrere alla **correlazione non parametrica** (il **τ di Kendall** oppure il **ρ di Spearman,** indifferentemente).

Se il test di correlazione risulta significativo, si può affermare che è significativa anche la retta non parametrica. Vale a dire che

- la serie dei **dati trasformati in ranghi** cresce in modo lineare
- e quindi la **serie dei valori reali cresce in modo monotonico** (concetto già illustrato nei paragrafi precedenti)

Una applicazione di questo test è richiesta nel caso della Regressione Modello II, anche se forse risulta ancora più appropriato il metodo di Theil - Kendall.

L'esempio successivo utilizza dati che richiedono appunto una analisi **Model II Regression**, discussa in un capitolo precedente.

ESEMPIO (TRATTO DA **SOKAL** E **ROHLF** PAG. 547-549). Si supponga di voler valutare la relazione lineare tra il peso di alcune femmine di un pesce della California *Scorpaenichthys marmoratus* (espresso in 100 g) e il numero di uova prodotto (in migliaia), con un campione di 11 misure, nel quale entrambe le variabili sono soggette ai medesimi errori di rilevazione o determinazione delle dimensioni:

Peso X	14	17	24	25	27	33	34	37	40	41	42
Uova Y	61	37	65	69	54	93	87	89	100	90	97

Calcolare la retta con il metodo dei tre gruppi di Bartlett.

Risposta. Dai dati della tabella è bene ricavare la organizzazione in gruppi, dopo aver ordinato i dati in modo crescente per la variabile X (in questo caso, nel testo sono già ordinate)

1 - Si ricava la tabella

	Gruppo 1		Gruppo 2		Gruppo 3	
	X_1	Y_1	X	Y	X_3	Y_3
	14	61	27	54	37	89
	17	37	33	93	40	100
	24	65	34	87	41	90
	25	69			42	97
Medie	20,0	58,00			40,0	94,0

2 - e da essa **le medie** del **gruppo 1** e del **gruppo 3**

- $\overline{X}_1 = \frac{80}{4} = 20{,}0 \qquad \overline{Y}_1 = \frac{232}{4} = 58{,}0$

- $\overline{X}_3 = \frac{160}{4} = 40{,}0 \qquad \overline{Y}_3 = \frac{376}{4} = 94{,}0$

e le **due medie generali** considerando tutti gli 11 dati

$$\overline{X} = \frac{334}{11} = 30{,}36 \qquad \overline{Y} = \frac{842}{11} = 76{,}55$$

3 – Il **coefficiente angolare** b è

$$b = \frac{\overline{Y}_3 - \overline{Y}_1}{\overline{X}_3 - \overline{X}_1} = \frac{94{,}0 - 58{,}0}{40{,}0 - 20{,}0} = \frac{36{,}0}{20{,}0} = 1{,}80$$

4 – L'**intercetta** a è

$$a = \overline{Y} - b\overline{X} = 76{,}55 - 1{,}80 \cdot 30{,}36 = 76{,}55 - 54{,}65 = 21{,}90$$

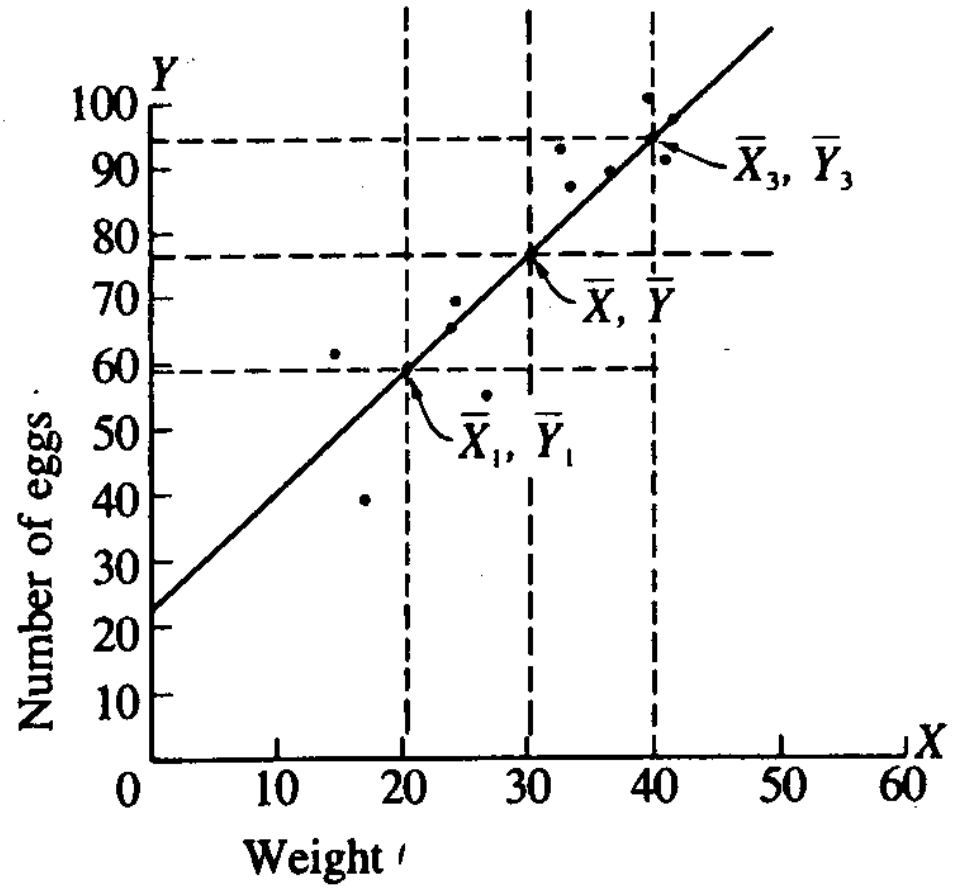
5 – Infine la **retta non parametrica di Bartlett** è

$$\hat{Y}_i = a + bX_i = 21{,}90 + 1{,}80 \cdot X_i$$

Se su questi dati fosse stata calcolata la **retta di regressione lineare parametrica** (da verificare come esercizio), secondo il testo si sarebbe trovata

$$\hat{Y}_i = a + bX_i = 19{,}77 + 1{,}87 \cdot X_i$$

In tutti i casi di analisi della regressione lineare e della correlazione, è sempre utile riportare il grafico o diagramma di dispersione, con la retta relativa



In essa si evidenziano le caratteristiche della distribuzione dei dati, in particolare la presenza di **eventuali valori anomali**, mostra se i punti sono distribuiti

- in **modo casuale** e **approssimativamente bilanciato** (situazione corretta)
- oppure in **modo regolare** o con la maggioranza dei punti da una sola parte della retta (situazione errata).

Inoltre è possibile osservare come la retta

- attraversi il baricentro complessivo dei dati

di coordinate

$$\overline{X} = 30{,}36 \quad \text{e} \quad \overline{Y} = 76{,}55$$

in quanto su di esse è stata calcolata l'intercetta $a = 21{,}9$

- e si avvicini molto ai punti individuati dalle medie del gruppo 1 e del gruppo 3

di coordinate

$$\overline{X}_1 = 20{,}0 \quad \text{e} \quad \overline{Y}_1 = 58{,}0$$

$$\overline{X}_3 = 40{,}0 \quad \text{e} \quad \overline{Y}_3 = 94{,}0$$

in quanto con esse è stato ricavato il coefficiente angolare $b = 1{,}8$.

Se è presente un dato fortemente anomalo, tale da alterare sensibilmente una delle tre medie, la retta non attraversa i punti individuati dalle medie dei gruppi 1 e 3.

21.15. IL TEST DI HOLLANDER PER IL CONFRONTO TRA DUE COEFFICIENTI ANGOLARI

Quando in due campioni indipendenti bivariati $\mathbf{C_1}$ e $\mathbf{C_2}$ sono stati calcolati i due coefficienti angolari $\mathbf{b_1}$ e $\mathbf{b_2}$, può sorgere il problema di verificare se essi siano paralleli:

$$\mathbf{H_0:\ \beta_1 = \beta_2}$$

Tra i test proposti (come quello di R. F. **Potthoff** con l'articolo del 1965 ***A non parametric test of whether two simple regression lines are parallel*** nel volume University of North Carolina, Institute of Statistics Mimeno series 445, oppure quello di P. K. **Sen** per la verifica simultanea del parallelismo tra più coefficienti angolari con l'articolo del 1969 ***On a class of rank order test for the parallelism of several regression lines*** pubblicato in **The Annals of Mathematical Statistics**, vol. 40 pp. 1668-1683), attualmente il più diffuso nella ricerca sperimentale è quello illustrato da **M. Hollander** in modo completo nel 1970 (con l'articolo ***A distribution-free test for parallelism*** pubblicato su **Journal of the American Statistical Association** vol. 65, pp.387-394).

Per la significatività, il test di Hollander **utilizza il test dei ranghi con segno di Wilcoxon per due campioni dipendenti**; di conseguenza richiede le stesse condizioni di validità e come ipotesi alternativa accetta

- sia test bilaterali

$$\mathbf{H_0: \beta_1 = \beta_2} \quad \text{contro} \quad \mathbf{H_1: \beta_1 \neq \beta_2}$$

- sia unilaterali nelle due differenti direzioni

$$\mathbf{H_0 : \beta_1 \leq \beta_2} \quad \text{contro} \quad \mathbf{H_1: \beta_1 > \beta_2}$$

$$\mathbf{H_0 : \beta_1 \geq \beta_2} \quad \text{contro} \quad \mathbf{H_1: \beta_1 < \beta_2}$$

secondo la domanda posto dal problema che si analizza.
E' quindi analogo al test parametrico t di Student, per 2 campioni dipendenti.

Per un'illustrazione semplice e facilmente comprensibile, è utile scomporre la metodologia di **Hollander** in sette passaggi logici fondamentali.

1 – Per l'applicazione del test di Hollander, è indispensabile il rispetto di una **prima condizione preliminare**: in ognuno dei due campioni $\mathbf{C_1}$ e $\mathbf{C_2}$, **il numero di osservazioni non può essere di poche unità, ma deve essere abbastanza grande, superiore almeno ad una decina**; infatti, nella fase finale entro lo stesso campione si devono formare due campioni dipendenti di dimensioni **n**, per cui il numero di osservazioni **N** si dimezza.

2 – La **seconda condizione preliminare**, insita nei concetti appena espressi sull'appaiamento dei dati, riguarda le dimensioni dei due campioni $\mathbf{C_1}$ e $\mathbf{C_2}$, dei quali si vogliono confrontare i relativi coefficienti angolari: **entrambi devono avere un numero N d'osservazioni pari (N = 2n)**. Nel caso in cui il numero **N** d'osservazioni sia dispari, si elimina un'osservazione a caso.

3 – La **terza condizione preliminare** è che **i due campioni devono avere lo stesso numero d'osservazioni**; di conseguenza, se si deve eliminare un'osservazione l'operazione deve essere eseguita su entrambi i campioni.

4 – Nel campione C_1 si stimano **n** coefficienti angolari $\mathbf{b_{1,k}}$ accoppiando i valori di $\mathbf{Y_{1,k}}$ e $\mathbf{X_{1,k}}$ rispettivamente **c**on quelli di $\mathbf{Y_{1,k+n}}$ e $\mathbf{X_{1,k+n}}$

$$b_{1,k} = \frac{Y_{1,k+n} - Y_{1,k}}{X_{1,k+n} - X_{1,k}}, \qquad k = 1, 2, \ldots, n.$$

5 – nello stesso modo, per il campione C_2 si stimano **n** coefficienti angolari $\mathbf{b_{2,k}}$ accoppiando i valori di $\mathbf{Y_{2,k}}$ e $\mathbf{X_{2,k}}$ rispettivamente **c**on quelli di $\mathbf{Y_{2,k+n}}$ e $\mathbf{X_{2,k+n}}$

$$b_{2,k} = \frac{Y_{2,k+n} - Y_{2,k}}{X_{2,k+n} - X_{2,k}}, \qquad k = 1, 2, \ldots, n.$$

6 – Si formano coppie a caso tra un valore $\mathbf{b_{1,k}}$ e di un valore $\mathbf{b_{2,k}}$, calcolando **n** differenze $\mathbf{d_k}$

$$\mathbf{d_k = b_{1,i} - b_{2,j}}$$

con il loro segno.

7 – Su queste **n** differenze $\mathbf{d_k}$ con segno, si applica il test di Wilcoxon dei ranghi con segno.
Infatti dopo aver trasformato i valori ottenuti nei loro ranghi, mantenendo il segno,
- se è vera l'ipotesi nulla

$$\mathbf{H_0:}\ \beta_1 = \beta_2$$

la somma dei ranghi positivi e di quelli negativi tenderà ad essere uguale e quindi quella minore tenderà al valore medio,
- mentre se è vera l'ipotesi alternativa bilaterale

$$\mathbf{H_1:}\ \beta_1 \neq \beta_2$$

oppure una delle 2 ipotesi unilaterali prefissate

$$\mathbf{H_1:}\ \beta_1 < \beta_2 \quad \text{oppure} \quad \mathbf{H_1:}\ \beta_1 > \beta_2$$

la somma minore tenderà a 0.

Come esempio d'applicazione di questa metodologia, in vari testi di statistica non parametrica è riportato un esempio pubblicato da A. C. **Wardlaw** e G. **van Belle** nel 1964 (con l'articolo ***Statistical aspect of the mouse diaphragm test for insulin***, sulla rivista **Diabetes** n. 13, pp. 622-633).
Il caso è interessante anche per la ricerca biologica e ambientale, poiché illustra il confronto tra due rette calcolate con due soli punti o meglio sue soli valori di X per i quali si abbiano misure ripetute della Y. E' una situazione sperimentale che ricorre con frequenza nelle ricerche di ecotossicologia o comunque nella valutazione degli effetti di un dosaggio con due soli punti, per i quali non è possibile applicare la regressione parametrica.

ESEMPIO. E' stata valutata la capacità di un ormone, somministrato a 2 dosi differenti ($X_1 = 0,3$ e $X_2 = 1,5$), di stimolare la sintesi di glicogeno, misurata in termini di densità ottica (Y_i), in presenza di insulina di 2 tipi differenti ($\mathbf{C_1}$ = insulina standard; $\mathbf{C_2}$ = campione di insulina da saggiare). Si vuole verificare se, nelle due differenti condizioni sperimentali, le due rette dose-risposta sono parallele; in termini biologici, se si ha una risposta simile all'aumentare della dose, pure considerando gli effetti di una diversa concentrazione del principio attivo.

A questo scopo,

- sono state ottenute 12 misure nella condizione C_1 e 12 nella condizione C_2,
- delle quali 6 alla dose 0,3 e 6 alla dose 1,5 (riportati nella tabella seguente).

Prova	C_1		C_2	
	$X_{1,j}$	$Y_{1,j}$	$X_{2,j}$	$Y_{2,j}$
1	Log 0,3	230	log 0,3	310
2	Log 0,3	290	log 0,3	265
3	Log 0,3	265	log 0,3	300
4	Log 0,3	225	log 0,3	295
5	log 0,3	285	log 0,3	255
6	log 0,3	280	log 0,3	280
7	log 1,5	365	log 1,5	415
8	log 1,5	325	log 1,5	375
9	log 1,5	360	log 1,5	375
10	log 1,5	300	log 1,5	275
11	log 1,5	360	log 1,5	380
12	log 1,5	385	log 1,5	380

Com'è prassi in questi casi, per rendere lineare la risposta, in sostituzione del valore della dose è stato utilizzato il suo logaritmo (log dose).

Il metodo di **Hollander** per la **verifica del parallelismo** tra le due rette segue questi passaggi logici.

1 – Dapprima si verifica che nella programmazione dell'esperimento siano state rispettate le tre condizioni preliminari richieste:

- numero di osservazioni per campione superiore a 10: sono $\mathbf{N} = 12$);
- numero di osservazioni entro ogni campione pari e uguale per le due dosi: sono 6 per dose;
- numero di osservazioni uguali per i due campioni: sono 12 in entrambi.

2 – Successivamente, si calcolano misure ripetute del coefficiente angolare $\mathbf{b_{i,k}}$ in entrambi i campioni

$$b_{i,k} = \frac{Y_{i,k+n} - Y_{i,k}}{X_{i,k+n} - X_{i,k}}$$

abbinando

- la prima osservazione del dosaggio 0,3 con la prima osservazione del dosaggio 1,5
- la seconda osservazione del dosaggio 0,3 con la seconda osservazione del dosaggio 1,5

e così fino a

- la sesta osservazione del dosaggio 0,3 con la sesta osservazione del dosaggio 1,5

ottenendo i seguenti risultati:

- per il campione C_1

$$b_{1,1} = \frac{Y_{1,7} - Y_{1,1}}{X_{1,7} - X_{1,1}} = \frac{365 - 230}{\log 1{,}5 - \log 0{,}3} = \frac{135}{0{,}699} = 193{,}1$$

$$b_{1,2} = \frac{Y_{1,8} - Y_{1,2}}{X_{1,8} - X_{1,2}} = \frac{325 - 290}{\log 1{,}5 - \log 0{,}3} = \frac{35}{0{,}699} = 50{,}1$$

$$b_{1,3} = \frac{Y_{1,9} - Y_{1,3}}{X_{1,9} - X_{1,3}} = \frac{360 - 265}{\log 1{,}5 - \log 0{,}3} = \frac{95}{0{,}699} = 135{,}9$$

$$b_{1,4} = \frac{Y_{1,10} - Y_{1,4}}{X_{1,10} - X_{1,4}} = \frac{300 - 225}{\log 1{,}5 - \log 0{,}3} = \frac{75}{0{,}699} = 107{,}3$$

$$b_{1,5} = \frac{Y_{1,11} - Y_{1,5}}{X_{1,11} - X_{1,5}} = \frac{360 - 285}{\log 1{,}5 - \log 0{,}3} = \frac{75}{0{,}699} = 107{,}3$$

$$b_{1,6} = \frac{Y_{1,12} - Y_{1,6}}{X_{1,12} - X_{1,6}} = \frac{385 - 280}{\log 1{,}5 - \log 0{,}3} = \frac{105}{0{,}699} = 150{,}2$$

- per il campione C_2

$$b_{2,1} = \frac{Y_{2,7} - Y_{2,1}}{X_{2,7} - X_{2,1}} = \frac{415 - 310}{\log 1{,}5 - \log 0{,}3} = \frac{105}{0{,}699} = 150{,}2$$

$$b_{2,2} = \frac{Y_{2,8} - Y_{2,2}}{X_{2,8} - X_{2,2}} = \frac{375 - 265}{\log 1{,}5 - \log 0{,}3} = \frac{110}{0{,}699} = 157{,}4$$

$$b_{2,3} = \frac{Y_{2,9} - Y_{2,3}}{X_{2,9} - X_{2,3}} = \frac{375 - 300}{\log 1{,}5 - \log 0{,}3} = \frac{75}{0{,}699} = 107{,}3$$

$$b_{2,4} = \frac{Y_{2,10} - Y_{2,4}}{X_{2,10} - X_{2,4}} = \frac{275 - 295}{\log 1{,}5 - \log 0{,}3} = -\frac{20}{0{,}699} = -28{,}6$$

$$b_{2,5} = \frac{Y_{2,11} - Y_{2,5}}{X_{2,11} - X_{2,5}} = \frac{380 - 255}{\log 1{,}5 - \log 0{,}3} = \frac{125}{0{,}699} = 178{,}8$$

$$b_{2,6} = \frac{Y_{2,12} - Y_{2,6}}{X_{2,12} - X_{2,6}} = \frac{380 - 280}{\log 1{,}5 - \log 0{,}3} = \frac{100}{0{,}699} = 143{,}1$$

3 – Si accoppiano casualmente, con estrazione di numeri random, i 6 valori $\mathbf{b_{i,j}}$ del campione $\mathbf{C_1}$ con i 6 valori del campione $\mathbf{C_2}$ e si calcolano le differenze; nell'esempio riportato da **Wardlaw** e **van Belle** sono stati ottenuti gli accoppiamenti e le differenze

$$d_k = b_{1,i} - b_{2,j}$$

di seguito riportati, con tutti i calcoli

$\mathbf{d_1 = b_{1,1} - b_{2,1}} = 193{,}1 - 150{,}2 = \mathbf{42{,}9}$

$\mathbf{d_2 = b_{1,2} - b_{2,2}} = 50{,}1 - 157{,}4 = \mathbf{-\ 107{,}3}$

$\mathbf{d_3 = b_{1,3} - b_{2,4}} = 135{,}9 - -28{,}6 = \mathbf{164{,}5}$

$\mathbf{d_4 = b_{1,4} - b_{2,5}} = 107{,}3 - 178{,}8 = \mathbf{-\ 71{,}5}$

$\mathbf{d_5 = b_{1,5} - b_{2,6}} = 107{,}3 - 143{,}1 = \mathbf{-\ 35{,}8}$

$\mathbf{d_6 = b_{1,6} - b_{2,3}} = 150{,}2 - 107{,}3 = \mathbf{42{,}9}$

4 – Infine a queste 6 differenze si applica il test T di Wilcoxon per 2 campioni dipendenti:

a) dopo averle ordinate in modo crescente, considerando il loro valore assoluto

- 35,8 42,9 42,9 -71,5 -107,3 164,5

b) si attribuisce il rango, considerando che due valori (42,9) sono uguali

1 2,5 2,5 4 5 6

c) si riporta il segno dei valori originari

-1 +2,5 +2,5 -4 -5 +6

d) calcolando la somma dei ranghi positivi, che risulta uguale a 11

e) e la somma dei negativi, che risulta uguale a 10.

Il valore di T è la somma minore e quindi è uguale a 10.

5 –Dopo aver stimato T, per calcolare la significatività occorre stabilire la direzione del confronto: se bilaterale oppure unilaterale.

Questo test sul parallelismo, come presentato dal problema, è bilaterale.
Poiché sulla tabella dei valori critici per il test **T** di Wilcoxon per 2 campioni dipendenti, con **N** uguale a 6 e per un test a due code, il valore riportato è 2, non si può rifiutare l'ipotesi nulla. Per una conclusione più articolata, è conveniente mettere in evidenza che la somma dei negativi (10) e la somma dei positivi (11) sono tra loro molto vicine, come atteso quando l'ipotesi nulla è vera; pertanto, in termini di parallelismo tra le due rette, si deve concludere che **non solo non si può rifiutare l'ipotesi nulla ma la probabilità che le due rette siano parallele è molto elevata**.

Della metodologia presentata, un **punto di debolezza**, criticato da vari autori,
- è **l'accoppiamento casuale delle stime** delle $\mathbf{b_{i,k}}$ nei due campioni, per calcolare le **n** differenze $\mathbf{d_k}$.

Appunto perché generato in modo casuale, può fornire risposte diverse a partire dagli stessi dati. Come in tutti questi casi, ovviamente **non è assolutamente accettabile la ripetizione fino ad ottenere la risposta desiderata** o più vicina a quanto atteso.

21.16. LA REGRESSIONE MONOTONICA DI IMAN-CONOVER

In una distribuzione bivariata, nella quale
- una **variabile** è identificata con la **causa** e è detta variabile **indipendente**, indicata con **X**,
- l'altra è identificata con l'**effetto** e è detta **variabile dipendente**, indicata con **Y**,

può sorgere il problema di verificare **se al crescere della prima variabile la seconda cresce o diminuisce**, **senza richiedere che il rapporto sia di tipo lineare**, cioè costante.
E' la **regressione monotonica**.

Tra i metodi presenti in letteratura, quello proposto da R. L. **Iman** e W. J. **Conover** nel 1979 (vedi ***The use of the rank transform in regression*** su pubblicato su **Technometrics** vol. 21 pp. 499-509) è il più diffuso; inoltre, è presentato nel testo di **Conover** del 1999 (***Practical Nonparametric Statistics***, 3rd ed. John Wiley & Sons, New York, 584 p.), indubbiamnete da annoverare tra quelli internazionali più noti.

Da esso è tratto l'esempio seguente, qui illustrato con una presentazione più dettagliata della metodologia, una esposizione di tutti i passaggi logici e con la correzione di alcuni risultati.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Campioni	X_i	Y_i	$R_{(Xi)}$	$R_{(Yi)}$	$R_{(Xi)} \cdot R_{(Yi)}$	$R^2_{(Xi)}$	$\hat{R}_{(Yi)}$	$\hat{Y}_i$
A	0,0	>30	1	16	16	1	16,4800	>30
B	0,5	>30	2	16	32	4	15,5450	29,54
C	1,0	>30	3	16	48	9	14,6100	28,61
D	1,8	28	4	14	56	16	13,6750	26,68
E	2,2	24	5	13	65	25	12,7400	22,70
F	2,7	19	6	12	72	36	11,8050	18,60
G	4,0	17	7,5	11	82,5	56,25	10,4025	15,00
H	4,0	9	7,5	8	60	56,25	10,4025	15,00
I	4,9	12	9	9,5	85,5	81	9,0000	11,00
L	5,6	12	10	9,5	95	100	8,0650	9,13
M	6,0	6	11	5	55	121	7,1300	8,13
N	6,5	8	12	7	84	144	6,1950	7,20
O	7,3	4	13	1,5	19,5	169	5,2600	6,26
P	8,0	5	14	3	42	196	4,3250	5,67
Q	8,8	6	15	5	75	225	3,3900	5,20
R	9,3	4	16	1,5	24	256	2,4550	4,64
S	9,8	6	17	5	85	289	1,5200	4,02
				$\sum_{i=1}^{n}$	996,5	1784,5		

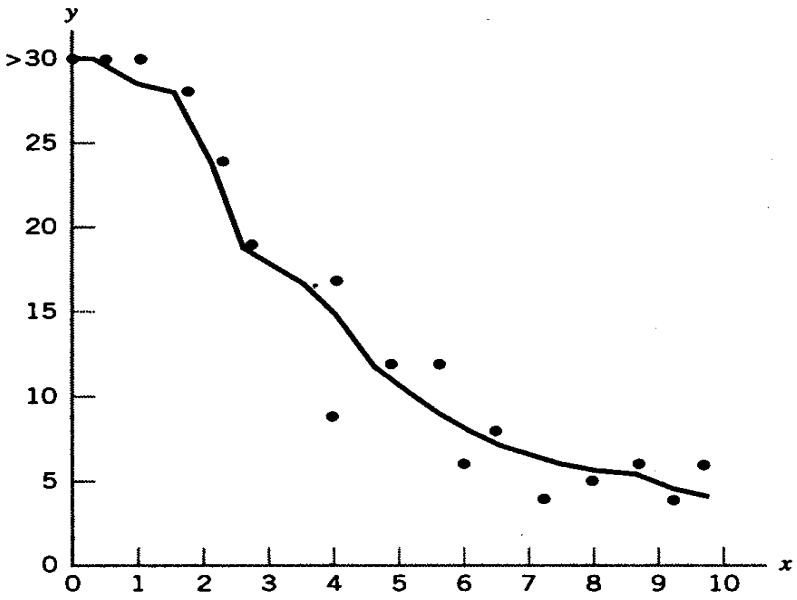
La **teoria** sottostante a tale approccio, già presentato nel paragrafo dedicato alla **correlazione non parametrica**, è che

- quando tra i **ranghi** di due variabili esiste **regressione lineare**,
- tra i loro **valori osservati** (su scale ad intervalli o di rapporti) esiste **regressione monotonica**.

Di conseguenza, è utile **calcolare la retta di regressione lineare sui ranghi e la sua significatività**. Per questo ultimo test, cioè per valutare l'ipotesi nulla $\mathbf{H_0: \beta = 0}$ si ritorna alla correlazione non parametrica già illustrata, per un concetto del tutto analogo alla verifica della regressione lineare di Theil.

Per valutare se l'aggiunta di zucchero al mosto d'uva favorisce la fermentazione, in 17 esperimenti indipendenti alla stessa quantità di mosto è stata aggiunta una quantità differente di zucchero (X, misurata in libbre); successivamente per 30 giorni è stato valutato se la fermentazione era iniziata (Y,

misurata in giorni trascorsi). Dopo 30 giorni l'esperimento è stato interrotto. Nei tre contenitori ai quali erano state aggiunte le quantità di zucchero minori (cioè X uguale a 0,0; 0,5; 1,0 libbre) la fermentazione non era ancora iniziata: a essi è stata attribuita la misura approssimata Y >30.



CURVA DI REGRESSIONE MONOTONICA CON I DATI OSSERVATI

I dati dei 17 **campioni indipendenti** (individuati nella colonna 1 dalle lettere da A a S) sono riportati nelle colonne 2 (X_i) e 3 (Y_i) della tabella precedente. Nel grafico sottostante, sono rappresentati i punti (X_i, Y_i) che identificano le 17 osservazioni (i segmenti che descrivono la tendenza sono spiegati nella parte finale del paragrafo).

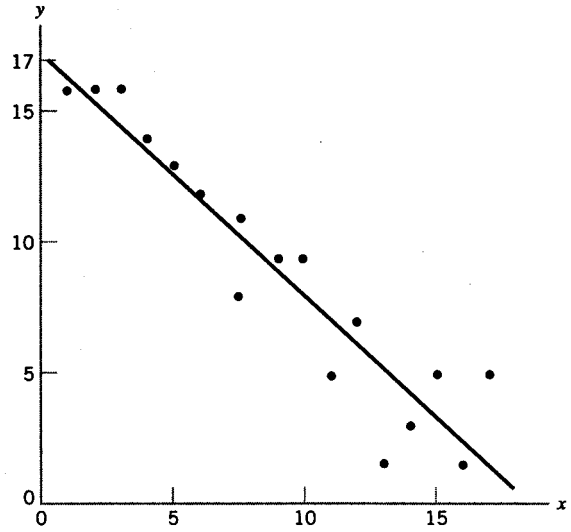
Il metodo di **Iman-Conover** richiede **due serie** di passaggi:

- la prima per **calcolare la retta di regressione sui ranghi dei valori**;
- la seconda per ritornare da questa retta ai valori originali, **trasformadola in una serie di segmenti** che descrivono la regrassione monotonica di **Y** su X, nella scala effettivamente utilizzata.

Questi passaggi logici e metodologici sono:

1 - Trasformare i valori della variabile X e Y (riportati nelle colonne 2 e 3 della tabella precedente) nei loro ranghi (come nelle colonne 4 e 5).

Il grafico sottostante, come evidenzia anche la differente scala riportata in ascissa e in ordinata, è la rappresentazione dei punti mediante i loro ranghi.



RETTA DI REGRESSIONE OTTENUTA CON LA TRASFORMAZIONE DEI DATI IN RANGHI

2 – Calcolare la retta sui ranghi (quella rappresentata nella figura). Dapprima si stima il coefficiente angolare b con

$$b = \frac{\sum_{i=1}^{n} R_{(Xi)} \cdot R_{(Yi)} - \frac{n \cdot (n+1)^2}{4}}{\sum_{i=1}^{n} \left(R_{(Xi)}\right)^2 - \frac{n \cdot (n+1)^2}{4}}$$

una formula corrispondente a quella abbreviata della retta parametrica.

Utilizzando i dati dell'esempio, dove

- $\sum_{i=1}^{n} R_{(Xi)} \cdot R_{(Yi)} = 996{,}5$ (colonna 6)
- $\sum_{i=1}^{n} \left(R_{(Xi)}\right)^2 = 1784{,}5$ (colonna 7)
- $n = 17$

$$b = \frac{996{,}5 - \dfrac{17 \cdot 18^2}{4}}{1784{,}5 - \dfrac{17 \cdot 18^2}{4}} = \frac{996{,}5 - 1377}{1784 - 1377} = \frac{-380{,}5}{407} = -0{,}935$$

si ottiene il coefficiente angolare b = **-0,935**.

Successivamente, si stima l'intercetta a con

$$a = \frac{(1-b)\cdot(n+1)}{2}$$

Utilizzando i dati dell'esempio

$$a = \frac{(1-(-0{,}935))\cdot(17+1)}{2} = \frac{1{,}935 \cdot 18}{2} = \frac{34{,}83}{2} = 17{,}415$$

risulta a = **17,415**.

Di conseguenza, la retta di regressione stimata mediante i ranghi è

$$\hat{R}_{(Yi)} = a + b \cdot R_{(Xi)} = 17{,}415 - 0{,}935 \cdot R_{(Xi)}$$

3 – Questa formula permette di calcolare **i valori attesi per ogni rango di Y** a partire dai ranghi di X, cioè gli $\hat{R}_{(Yi)}$ riportati nella colonna 8, anche se per tracciare la retta è sufficiente calcolarne solo due.

Ad esempio,

- per il campione A con X_i = 0,0 e quindi rango $R_{(Xi)}$ = 1 si ottiene la stima del rango di Y

$$\hat{R}_{(Yi)} = 17{,}415 - 0{,}935 \cdot 1 = 16{,}48$$

- per i campioni G e H con X_i = 4,0 e rango $R_{(Xi)}$ = 7,5 si ottiene la stima del rango di Y

$$\hat{R}_{(Yi)} = 17{,}415 - 0{,}935 \cdot 7{,}5 = 10{,}4025$$

La retta è costruita con i punti $(R_{(Xi)}, \hat{R}_{(Yi)})$, cioè utilizzando i dati della colonna 4 e quelli riportati nella colonna 8.

4 – Per valutare la significatità della retta così calcolata (**H_0: $\beta = 0$**), quindi **se esiste regressione monotonica sui dati originali**, è sufficiente valutare la significatività della correlazione non parametrica (**H_0: $\rho = 0$**); può essere ottenuta indifferentemente con il test **ρ di Spearman** oppure con il **τ di Kendall**.

5 - Dalla retta calcolata sui ranghi, **si ritorna alla scala originale di X e Y** calcolando gli $\hat{Y}_i$ riportati nella colonne 9. Unendo i punti individuti dai valori X_i osservati della colonna 1 e quelli $\hat{Y}_i$ stimati riportati nella colonna 9 (cioè i punti $X_i, \hat{Y}_i$), si ottiene la linea spezzata, rappresentata nella prima figura.

6 - Per identificare tutti i valori $\hat{Y}_i$ di questa **regressione monotonica**, la procedura è complessa e richiede alcune scelte, che **dipendono dal valore ottenuto del rango stimato** per Y (cioè $\hat{R}_{(Yi)}$ della colonna 8):

a) se $\hat{R}_{(Yi)}$ è uguale al rango reale della stessa osservazione Y, cioè $R_{(Yi)}$ della colonna 5, si attribuisce a $\hat{Y}_i$ il valore osservato Y_i;

b) se $\hat{R}_{(Yi)}$ è compreso tra il rango di due osservazioni adiacenti di Y, cioè Y_k e Y_{k+1} della colonna 3, con Y_k minore in valore di Y_{k+1}, per ottenere il valore stimato di Y, cioè $\hat{Y}_i$ della colonna 9, si usa l'interpolazione

$$\hat{Y}_i = Y_{k+1} + \frac{\hat{R}_{(Yi)} - R_{(Yk+1)}}{R_{(Yk+1)} - R_{(Yk)}} \cdot \left(Y_{k+1} - Y_k\right)$$

c) se $\hat{R}_{(Yi)}$ è minore del rango osservato $R_{(Yi)}$ più piccolo (colonna 4), il valore stimato di Y (cioè $\hat{Y}_i$ della colonna 9) è uguale a quel valore osservato minore;

d) se $\hat{R}_{(Yi)}$ è maggiore del rango osservato $R_{(Yi)}$ più grande (colonna 4), il valore stimato di Y (cioè $\hat{Y}_i$ della colonna 9) è uguale a quel valore osservato maggiore.

Ad esempio,

- per il **campione A** il valore stimato del rango è $\hat{R}_{(Yi)}$ = 16,48 (colonna 8); poiché è maggiore del rango più grande ($R_{(Yi)}$ = 16 di colonna 4) , il valore stimato di Y è $\hat{Y}_i$ >30;
- per il **campione B** il valore stimato del suo rango è $\hat{R}_{(Yi)}$ = 15,55 (colonna 8); è compreso tra il rango 16 e il rango 14; quindi

$$\hat{Y}_i = 30 + \frac{15{,}54 - 16}{16 - 14} \cdot (30 - 28) = 30 + \frac{-0{,}46}{2} \cdot 2 = 29{,}54$$

- per il **campione C**

$$\hat{Y}_i = 30 + \frac{14{,}61 - 16}{16 - 14} \cdot (30 - 28) = 30 + \frac{-1{,}39}{2} \cdot 2 = 28{,}61$$

- per il **campione D**

$$\hat{Y}_i = 28 + \frac{13{,}67 - 14}{14 - 13} \cdot (28 - 24) = 28 + \frac{-0{,}33}{1} \cdot 4 = 26{,}68$$

- per il **campione E**

$$\hat{Y}_i = 24 + \frac{12{,}74 - 13}{13 - 12} \cdot (24 - 19) = 24 + \frac{-0{,}26}{1} \cdot 5 = 22{,}70$$

- per il **campione F**

$$\hat{Y}_i = 19 + \frac{11{,}80 - 12}{12 - 11} \cdot (19 - 17) = 19 + \frac{-0{,}2}{1} \cdot 2 = 18{,}60$$

- per il **campione G**

$$\hat{Y}_i = 17 + \frac{10{,}40 - 11}{11 - 9{,}5} \cdot (17 - 22) = 17 + \frac{-0{,}60}{1{,}5} \cdot 5 = 15{,}00$$

I valori stimati sono riportati nella colonna 9 (gli altri sono ripresi da Conover).
Nell'esempio utilizzato, tre valori sono molto approssimati (>30) e tra loro identici, quindi che introducono bias nei calcoli. Il metodo dimostra di essere abbastanza robusto da riuscire ugualmente a fornire stime della regressione monotonica. Tuttavia, lo stesso Conover raccomanda di usare scale continue, quindi definite con precisione e senza valori identici

21.17. TREND LINEARE DI ARMITAGE PER LE PROPORZIONI E LE FREQUENZE

Le tabelle di contingenza **2 x k** riportano i risultati di risposte binarie in k campioni, come è stato presentato nel capitolo III per il test χ^2 o il test G. Con questi test, ad esempio, possono essere confrontate

- le proporzioni di persone affette da malattie polmonari in k aree con livelli d'inquinamento atmosferico differenti,
- le k proporzioni di analisi della qualità dell'acqua con una quantità di nitrati superiore ai livelli di attenzione,
- il numero di cavie decedute o che non hanno raggiunto la maturità sessuale, nel confronto degli effetti di k sostanze tossiche.

Come già illustrato nel caso del confronto tra più proporzioni, con k campioni l'ipotesi nulla è

$$\mathbf{H_0: \pi_A = \pi_B = \pi_C = \quad = \pi_k}$$

contro l'ipotesi alternativa

$$\mathbf{H_1: \text{ non tutte le } \pi \text{ sono uguali}}$$

Altri casi che ricorrono con frequenza nella ricerca applicata è la verifica di
- un **trend nel tempo, nello spazio, tra dosi crescenti di un farmaco**,
- tra l'incidenza di patologie in **gruppi d'individui appartenenti a classi d'età progressivamente più anziane**.

Per scale di tipo ordinale, Peter **Armitage**
- con un articolo del 1955 (***Test for linear trends in proportion and frequencies***, pubblicato dalla rivista **Biometrics** vol. 11 pp.: 375-386)
- e in un paragrafo del suo testo di Statistica del 1971 (vedi ***Statistical Methods in Medical Research***. John Wiley and Sons, New York, 504 pp. oppure la traduzione italiana ***Statistica Medica. Metodi statistici per la ricerca in Medicina***, quarta edizione, marzo 1981 Feltrinelli, Milano, pp. 493)

illustra un metodologia di **scomposizione del χ^2 e dei rispettivi gradi di libertà nei suoi componenti, per la verifica della linearità**.

L'ipotesi nulla è la presenza di un gradiente lineare, contro l'ipotesi alternativa di un allontanamento da esso.
Il metodo è simile a quanto già noto per la scomposizione della regressione con l'analisi della varianza.

Questa tecnica offre due vantaggi rispetto ai tradizionali test χ^2 e G:
- permette la **verifica di un trend lineare**, quindi una informazione ulteriore rispetto alla tradizionale ipotesi uguaglianza delle proporzioni;
- agendo sulla scomposizione dei gdl, **è la procedura χ^2 più potente per il rifiuto dell'ipotesi nulla sulle differenze tra proporzioni**.

La devianza totale, cioè il valore del χ^2 totale con k-1 gdl, è scomposta in

- una devianza dovuta alla regressione, cioè un χ^2 che analizza il trend lineare con 1 gdl
- una devianza dovuta agli scarti dalla regressione, cioè un χ^2 che comprende gli allontanamenti dalla linearità con gdl **k-2**.

Il test richiede un numero minimo di gruppi pari a tre (altrimenti tra due punti passa sempre una retta).

La metodologia può essere spiegata in modo semplice e chiaro illustrando una sua applicazione al contesto ambientale (sono stati utilizzati gli stessi dati riportati nel testo di Armitage, versione italiana, pp. 353-355).

I vari quartieri di una città sono stati classificati in zone a inquinamento atmosferico basso, medio e alto, sulla base dei valori medi mensili degli ultimi cinque anni. Una visita medica ai bambini iscritti nelle scuole dei vari quartieri ha dato i seguenti risultati:

	Grado di inquinamento			
	Basso	Medio	Alto	TOTALE
Bambini **con** malattie polmonari	**19**	**29**	**24**	72
Bambini **senza** malattie polmonari	497	560	269	1326
Totale bambini visitati	516	589	293	1398
Proporzione di bambini ammalati	**0,0368**	**0,0492**	**0,0819**	**0,0515**

Nella tabella, in aggiunta alle informazioni classiche necessarie per il calcolo del χ^2, per analizzare il trend è conveniente riportare la proporzione di bambini con malattie polmonari, in funzione del livello d'inquinamento (come nella riga 4 della tabella).

Le tre proporzioni rilevate

- **0,0368** per le aree a inquinamento basso
- **0,0492** per le aree a inquinamento medio
- **0,0819** per le aree a inquinamento alto

sono in accordo con un possibile trend lineare?

Per rispondere a tale quesito, Armitage propone una metodologia che può essere schematizzata in alcuni passaggi:

1 - Calcolare il valore del chi quadrato totale, che avrà k-1 df

Di conseguenza, dopo aver stimato i valori attesi

	Grado di inquinamento			
	Basso	Medio	Alto	TOTALE
Bambini **con** malattie polmonari	**26,6**	**30,3**	**15,1**	72
Bambini **senza** malattie polmonari	**489,4**	**558,7**	**277,9**	1326
Totale bambini visitati	516	589	293	1398

con la solita formula

$$\chi^2_{(g.d.l.)} = \sum_{i=1}^{k} \frac{\left(f_i^{oss} - f_i^{att}\right)^2}{f_i^{att}}$$

si ottiene

$$\chi^2_{(2)} = \frac{(19-26,6)^2}{26,6} + \frac{(29-30,3)^2}{30,3} + \frac{(24-15,1)^2}{15,1} + \frac{(497-489,4)^2}{489,4} + \frac{(560-558,7)^2}{558,7} + \frac{(269-277,9)^2}{277,9}$$

$$\chi^2_{(2)} = \frac{57,76}{26,6} + \frac{1,69}{30,3} + \frac{79,21}{15,1} + \frac{57,76}{489,4} + \frac{1,69}{558,7} + \frac{79,21}{277,9}$$

$$\chi^2_{(2)} = 2,171 + 0,056 + 5,246 + 0,118 + 0,003 + 0,285 = 7,879$$

un χ^2 **= 7,879** con **2** gdl.

Poiché con **2** gdl il valore critico per

- $\alpha = 0.05$ è 5,991
- $\alpha = 0.025$ è 7,378
- $\alpha = 0.01$ è 9,210

si può affermare che le tre proporzioni sono tra loro significativamente differenti, con probabilità minore del 2,5% di errare.

Il valore del χ^2 totale dipende dagli scarti delle **k** proporzioni dalla proporzione totale; quando non supera il valore critico, esso indica che le tre proporzioni sono statisticamente uguali.

2 - Successivamente, dopo aver attribuito un punteggio o rango alla posizione dei **k** gruppi, sulla base delle informazioni riportate nella tabella successiva

	Grado di inquinamento			
	Basso	Medio	Alto	TOTALE
Bambini **con** malattie polmonari ($\mathbf{f_i}$)	19	29	24	72 (**R**)
Totale bambini visitati ($\mathbf{C_i}$)	516	589	293	1398 (N)
Proporzione di bambini ammalati	**0,0368**	**0,0492**	**0,0819**	
Rango o punteggio ($\mathbf{z_i}$)	**1**	**2**	**3**	

e utilizzando la simbologia inserita ($\mathbf{f_i}$, $\mathbf{C_i}$, $\mathbf{z_i}$, **R**, **N**), con

$$\chi^2_{trend} = \frac{N\left(N\sum_{i=1}^{k} f_i z_i - R\sum_{i=1}^{k} C_i z_i \right)^2}{R(N-R)\cdot\left[N\sum_{i=1}^{k} C_i z_i^2 - \left(\sum_{i=1}^{k} C_i z_i \right)^2 \right]}$$

si ottiene

$$\chi^2_{trend} = \frac{1398\cdot\left[1398\cdot(19\cdot 1+29\cdot 2+24\cdot 3)-72\cdot(516\cdot 1+589\cdot 2+293\cdot 3)\right]^2}{72\cdot(1398-72)\cdot\left[1398\cdot(516\cdot 1^2+589\cdot 2^2+293\cdot 3^2)-(516\cdot 1+589\cdot 2+293\cdot 3)^2\right]}$$

$$\chi^2_{trend} = \frac{1398\cdot(1398\cdot 149-72\cdot 2573)^2}{72\cdot 1326\cdot\left[(1398\cdot 5509)-(2573)^2\right]}$$

$$\chi^2_{trend} = \frac{1398\cdot(23046)^2}{95472\cdot 1081253} = \frac{742.503.126.168}{103.229.386.416} = 7{,}193$$

un valore del χ^2 per il trend uguale a **7,193** con **1** gdl.

Poiché con **1** gdl il valore critico per

- $\alpha = 0.05$ è 3,841
- $\alpha = 0.025$ è 5,024
- $\alpha = 0.01$ è 6,635

si può affermare che la varianza spiegata dalla retta rispetto alla proporzione totale (la media generale delle proporzioni) è altamente significativa.

3 - Per sottrazione di questo secondo valore del primo, si ricava il valore del chi quadrato per l'allontanamento dalla linearità (χ^2_d)

$$\chi^2_d = \chi^2 - \chi^2_{trend}$$

con gdl k-2

Con i dati dell'esempio, il chi quadrato per l'allontanamento dalla linearità χ^2_d

$$\chi^2_d = 7{,}879 - 7{,}193 = 0{,}686$$

risulta uguale a 0,686 con 1 gdl.

Poiché i valori critici sono identici a quelli appena riportati, il chi quadrato di errore dalla retta risulta trascurabile.

4 - Infine, per una presentazione completa dei risultati e facilitare la loro interpretazione è vantaggioso riportare i tre χ^2 (totale, per il trend e per l'allontanamento dalla linearità) in una tabella

χ^2		DF	P
Totale	7,879	2	< 0.025
Trend lineare	7,193	1	<0.01
Allontanamento dalla linearità	0,686	1	>0.05

In conclusione,

- **le tre proporzioni sono significativamente differenti;**
- **sono distribuite lungo una retta in modo altamente significativo,**

In termini discorsivi, si può concludere che passando all'aumentare del gardo di inquinamento si ha un incremento quasi costante, cioè lineare, nella proporzione di bambini con malattie polmonari.

Se il chi quadrato per l'allontanamento dalla linearità fosse risultato significativo, si sarebbe dovuto concludere che tra le **k** proporzioni esiste una differenza significativa, ma con un trend differente dalla linearità.

Raffinamenti successivi del metodo considerano coefficienti differenti,

- in funzione del valore di X
- e dove Y è la proporzione ottenuta in ogni gruppo.

Ma, in queste condizioni, i gruppi non sono più in una scala ordinale bensì in una scala di rapporti o ad intervalli. Di conseguenza questa metodologia viene superata dalla possibilità di calcolare la retta con la regressione parametrica.

Mantenendo l'informazione di tipo ordinale per la X, è possibile anche utilizzare una regressione lineare non parametrica, come il metodo di Theil.

Infine, se la domanda fosse solamente quella di un **incremento della frequenza di malattie** all'**aumentare del grado di inquinamento**, è possibile utilizzare anche una **regressione monotonica**.