CAPITOLO XV

TEST NON PARAMETRICI PER PIU' CAMPIONI

15.1. I test non parametrici piu' utilizzati, per k campioni 1

15.2. Estensione del test della mediana 2

15.3. Cenni sul test di Nemenyi e altri per la mediana in k gruppi 10

15.4. Analisi della varianza per ranghi ad un criterio di classificazione: il test di Kruskal-Wallis 13

15.5. Confronto con il test F e confronti multipli con i ranghi 26

15.6. Test per l'eterogeneita' della varianza con k campioni 33

15.7. Confronti tra piu' proporzioni e confronti multipli relativi 38

15.8. Il test Q di Cochran 47

15.9. Estensione del test di McNemar o test di Bowker 52

15.10. Test di Friedman o analisi della varianza per ranghi a 2 criteri di classificazione, con una e con k repliche 57

15.11. I confronti multipli tra medie di ranghi nell'analisi della varianza non parametrica, a due criteri di classificazione 76

15.12. Test di Quade 85

15.13. L'esempio di Koch: uso di metodi non parametrici, nell'analisi statistica di un esperimento complesso con k fattori 92

CAPITOLO XV

TEST NON PARAMETRICI
PER PIU' CAMPIONI

15.1. I TEST NON PARAMETRICI PIU' UTILIZZATI, PER K CAMPIONI.

Tutti i test non parametrici hanno in comune una sola caratteristica: non richiedono la normalità della distribuzione. Presentano quindi un vantaggio indubbio rispetto a quelli parametrici, sia quando la non-normalità è evidente, sia quando la normalità della distribuzione dei dati è solo probabile. In esperimenti nuovi e con pochi dati, non è possibile dimostrare che la condizione di normalità è rispettata, data la bassa potenza dei test relativi. Il test per l'omogeneità della varianza (Hartley, Cochran, Bartlett, Levene) può non risultare significativo, in particolare con campioni piccoli; ma non essere in grado di rifiutare l'ipotesi nulla sulla normalità ed omoschedasticità non significa che essa sia vera.

Il test F ed il test t bidirezionale sono metodi statisticamente robusti, rispetto a questa condizione: la non-normalità, pure quando causata da forte asimmetria, determina errori ridotti, nel calcolo della significatività. Alcuni autori stimano che una forte asimmetria può fissare la variazione delle probabilità **P** tra **0.04** e **0.07** quando quella nominale α è 0.05, mentre essa varia tra 0.005 e 0.02 quando α nominale è 0.01. E' un'approssimazione che permette di accettare i risultati, di non essere ipercritici sulla loro validità, poiché anche i test non parametrici non sempre conducono a stime più precise.

Le trasformazioni dei dati (presentate e discusse nella parte finale del capitolo precedente) possono rappresentare una soluzione. Ma se, dopo la loro applicazione, le caratteristiche della distribuzione non migliorano in modo sensibile, avvicinandosi sufficientemente a quelle richieste dai test parametrici, i metodi **non parametrici sono indubbiamente da preferire.**

Le situazioni più frequenti, in cui è consigliato scegliere un test non parametrico, si realizzano quando:

1 - **i campioni mantengono una distribuzione estremamente asimmetrica**;

2 - **il test è a una coda e l'alternativa sarebbe il test t unilaterale** (che è più sensibile all'asimmetria);

3 - **i campioni a confronto hanno un numero di dati molto differente** (poiché soprattutto in statistica parametrica influisce sulla potenza del test);

4 - **le distribuzioni non sono omoschedastiche** (condizione d'**invalidità severa**, per i test parametrici);

5 - **sono presenti valori fortemente anomali** (***outliers***) (perdita di valore della media e quindi delle misure da essa derivate, come la varianza);
6 - **è stata utilizzata una scala ordinale oppure qualitativa** (e quindi è doveroso utilizzare l'informazione realmente fornita dai dati).

Per campioni sufficientemente grandi, i test parametrici possono essere utilizzati anche con dati rappresentati da conteggi di variabili binarie: abitualmente, **hanno una distribuzione non normale in piccoli campioni, ma la approssimano abbastanza bene con campioni sufficientemente grandi**, come già ripetutamente dimostrato per la distribuzione binomiale, la poissoniana, la ipergeometrica e con il test χ^2 in tabelle 2 x 2.

Anche i test non parametrici per più campioni, proposti in tempi e situazioni diverse da autori differenti per affrontare problemi specifici, possono essere didatticamente classificati sulla base dell'organizzazione dell'esperimento. Nei testi di statistica applicata, è molto frequente la suddivisione in

- **test per k campioni indipendenti**,
- **test per k campioni dipendenti,**

analoghi rispettivamente al**l'analisi della varianza ad un criterio** e al**l'analisi a due criteri di classificazione**.

I test non parametrici per **<u>k campioni indipendenti</u>** maggiormente utilizzati nella ricerca e più frequentemente riportati nei programmi informatici sono:
- il χ^2 **e il test G in tabelle M x N** (già presentati in modo dettagliato nel capitolo specifico),
- l'estensione del **test della mediana** e il **test di Nemenyi,** per la tendenza centrale
- il test di **Kruskal-Wallis**, detto anche analisi non parametrica della varianza per un criterio di classificazione, che può essere visto come l'estensione del test di Wilcoxon–Mann-Whitney, fondato sui ranghi;
- i **test per l'omogeneità della varianza.**

I test non parametrici per **<u>k campioni dipendenti</u>** più diffusi sono:
- il test **Q di Cochran** e il test di **Bowker** (o estensione del test di McNemar)
- il test di **Friedman**, denominato pure analisi della varianza per ranghi a due criteri di classificazione.

15.2. ESTENSIONE DEL TEST DELLA MEDIANA

L'estensione a k campioni (con k > 2) del test della mediana (***the median test for several samples***) serve per **verificare l'ipotesi se le tendenze centrali sono significativamente differenti**. Anche

questo è concettualmente molto semplice e non ha un autore definito, in quanto diffuso fin dalle prime proposte sul chi quadrato.

L'ipotesi nulla, che prende in considerazione **le mediane** (me) delle popolazioni dalle quali sono state estratti i vari campioni (A, B, …, K), è che esse siano tutte uguali

$$\mathbf{H_0: me_A = me_B = \ldots = me_K}$$

mentre l'ipotesi alternativa, solo bilaterale come in tutti i confronti multipli effettuati simultaneamente, è

H_1: le mediane dei k gruppi non sono tutte uguali

E' fondato sulle tecniche di analisi di una tabella **k x 2** con il χ^2 o il test **G** quando il numero di osservazioni è sufficientemente grande; ricorre alla distribuzione ipergeometrica quando il numero di osservazioni è limitato o quando, come ora, è possibile disporre di un computer che svolga la rilevante quantità di calcoli richiesta in un campione grande.
Rispetto ai vari test sulla tendenza centrale, è utilizzato in modo appropriato **quando le misure sono approssimate e possono essere ordinate per rango solo in modo parziale, con molti valori identici, in particolare se collocati agli estremi.**. In altri termini, quando la distribuzione è lontana dalla normalità e presenta molti ties.

Nella raccolta dei dati ambientali, biologici e medici, può succedere di utilizzare strumenti tarati con alta precisione per i valori centrali, più frequenti; ma che non siano in grado di valutare le misure collocate verso gli estremi, più rare, troppo piccole per essere rilevate o troppo grandi per essere determinate con la stessa precisione di quelle intorno alla norma.
Si deve ricorrere necessariamente al test della mediana anche nell'**analisi di distribuzioni in cui le misure estreme siano state raggruppate in classi aperte**, formate da valori minori (< X) e/o maggiori (> X) di una quantità determinata.

Per i calcoli manuali, il fattore limitante all'uso del test della mediana per **k** campioni indipendenti è quello del χ^2 o del test **G**; il **numero di osservazioni deve essere sufficientemente elevato**.
Nel caso di **k** campioni con **un numero di osservazioni molto ridotto**, si può utilizzare il **metodo esatto, fondato sulla distribuzione ipergeometrica**, analogo al metodo esatto di Fisher in tabelle 2x2; ma richiede molti calcoli e quindi è praticamente applicabile solo con programmi informatici.

Il metodo dell'estensione del test della mediana per il confronto tra le tendenze centrali di più campioni può essere schematizzato in alcuni punti fondamentali, seguendo la sua applicazione ad un esempio.

1 – Dopo aver raccolto i dati di **k** campioni indipendenti, con un numero **r** di osservazioni che può essere diverso, come nell'analisi della varianza ad un criterio di classificazione,

GRUPPO		
A	B	C
< 1	3,7	2,1
<1	2,8	2,5
3,8	0,9	2,9
2,1	2,2	>10
- -	2,5	8,7
- -	- -	1,6

2 - ordinare per ranghi tutte le osservazioni dei **k** gruppi a confronto, come se fossero un gruppo unico, mantenendo per ogni valore l'informazione del gruppo di appartenenza

<1	<1	0,9	1,6	2,1	2,1	2,2	**2,5**	2,5	2,8	2,9	3,7	3,8	8,7	>10
A	A	B	C	A	C	B	**B**	C	B	C	B	A	C	C

3- Identificare la mediana di questa distribuzione unica,
che con dati dell'esempio è l'8° valore, il 2,5 del gruppo B (in grassetto).
Poiché il dato successivo ha un valore identico, collocare la mediana tra il 7° e l'8° valore.

4 - Contare quante sono le osservazioni di ogni gruppo che hanno valore inferiore (**n_1**) e quante quelle che hanno valore superiore (od uguale, in questo caso) (**n_2**) alla mediana:

GRUPPI	< mediana	≥ mediana
A	3	1
B	2	3
C	2	4

L'eventuale valore corrispondente alla mediana può essere classificato in uno dei due sottogruppi indifferentemente (in questo caso è stato contato con quelli superiori alla mediana).
Con **k** gruppi si ottiene una distribuzione di frequenza in una tabella **k x 2**, come quella appena riportata.

5 - **Se è vera l'ipotesi nulla** (H_0: i vari gruppi a confronto sono estratti dalla stessa popolazione o da campioni con la stessa tendenza centrale), **ogni gruppo dovrebbe avere lo stesso numero di osservazioni prima e dopo la mediana**; se i due gruppi non hanno la stessa frequenza, la stima delle frequenze attese può essere fatta come nella tabella riportata

GRUPPI	< mediana	≥ mediana	Totale
A	**1,87** (4 x 7 / 15)	**2,13** (4 x 8 / 15)	**4**
B	**2,33** (5 x 7 / 15)	**2,67** (5 x 8 / 15)	**5**
C	**2,80** (6 x 7 / 15)	**3,20** (6 x 8 / 15)	**6**
Totale	**7**	**8**	**15**

Se l'ipotesi nulla è falsa, almeno un gruppo dovrebbe avere una prevalenza significativa di osservazioni con valore minore o maggiore della mediana.

6 - Nel caso di **campioni grandi**, l'accordo tra la distribuzione osservata e la distribuzione attesa può essere analizzata con il test χ^2 corrispondente (o il test G), con **gdl** uguali a **k – 1** (k = numero di gruppi).

7- Se **<u>i campioni sono piccoli</u>** si ricorre al metodo esatto, disponendo di un programma informatico adeguato.

Indicando con

GRUPPI	< mediana	≥ mediana	Totale
A	a	b	n_1
B	c	d	n_2
C	e	f	n_3
Totale	n_4	n_5	**N**

dove

- **a, b, c, d, e, f**, sono le frequenze inferiori e superiori alla mediana nei **k** gruppi,
- **n_1, n_2 , n_3 , n_4 , n_5** , **N** sono rispettivamente i totali marginali e il totale generale

la formula generale per calcolare la probabilità esatta della risposta specifica ottenuta, derivata dalla **distribuzione ipergeometrica** è

$$\text{Pe} = \frac{n_1!n_2!n_3!n_4!n_5!}{a!b!c!d!e!f!N!}$$

nel caso di tre gruppi;

è **facilmente estensibile a k gruppi** mediante

$$\text{Pe} = \frac{\text{Prodotto dei fattoriali dei totali marginali!}}{\text{Prodotto dei fattoriali delle frequenze di casella e del totale generale}}$$

Con questa formula, si stima la probabilità di avere solo la risposta sperimentale.

Come nel metodo esatto di Fisher, **<u>la probabilità calcolata deve essere sommata con quelle di tutte le risposte più estreme</u>**. Poiché in questo caso si tratta di un test bilaterale (come sempre con **k** campioni), **le risposte più estreme possono essere individuate con facilità dal valore della loro**

probabilità esatta: <u>sono risposte più estreme tutte quelle che hanno una probabilità inferiore a quella calcolata per la risposta sperimentale</u>.

ESEMPIO 1. Si intende verificare se esiste una differenza significativa nella densità (numero di individui entro una superficie unitaria) di 5 specie vegetali (A, B, C, D, E). A questo scopo è stato raccolto un numero variabile di campioni, in aree di dimensione diversa, stimandone la concentrazione media in modo approssimato.

A	4	3	<1	7	<1	1	2	3	1
B	2	9	7	6	9	8	7	---	---
C	3	4	2	4	5	2	<1	3	---
D	8	7	9	>9	7	---	---	---	---
E	5	4	6	5	5	7	---	---	---

Le 5 aree hanno una concentrazione mediana significativamente differente?

Risposta.

1 - I rapporti sono stati calcolati su **k** (con **k > 2**) campioni di superficie non costanti e sono espressi con misure approssimate, che forniscono solo la dimensione del fenomeno, non una misura di una scala ad intervalli o di rapporti; di conseguenza, il test appropriato è l'estensione della mediana per **k** campioni.

L'ipotesi nulla è che tutti i campioni abbiano la stessa mediana

$$\mathbf{H_0: me_A = me_B = me_C = me_D = me_E}$$

con ipotesi alternativa che almeno una sia differente o

$$\mathbf{H_1: \text{non tutte le mediane sono uguali}}$$

2 - Come prima elaborazione dei dati, è utile costruire una distribuzione ordinata di tutti i valori, conservando per ognuno l'informazione del gruppo di appartenenza, allo scopo di

- determinare la mediana comune,
- contare per ognuno dei 5 gruppi a confronto quanti sono i valori inferiori e quelli superiori

<1	<1	<1	1	1	2	2	2	2	3	3	3	3	4	4	4	4	**5**	5	5	5	
A	A	C	A	A	A	B	C	C	A	A	C	C	A	C	C	E	**C**	E	E	E	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	**18**	19	20	21	Segue

6	6	7	7	7	7	7	7	8	8	9	9	9	>9
B	E	A	B	B	D	D	E	B	D	B	B	D	D
22	23	24	25	26	27	28	29	30	31	32	33	34	35

Nell'esempio, i dati sono 35: la mediana è il 18° valore e quindi 5. Poiché i valori identici alla mediana sono più di uno ed appartengono a vari gruppi, risulta utile aggregare il valore della mediana al gruppo dei valori maggiori di essa (>).

3 - Per ognuno dei 5 gruppi a confronto, si contano quanti sono i valori minori e quanti sono uguali o maggiori alla mediana. Si ottiene la tabella (5 x 2) seguente.

Gruppi	<	≥	Totale
A	**8**	**1**	9
B	**1**	**6**	7
C	**7**	**1**	8
D	**0**	**5**	5
E	**1**	**5**	6
Totale	17	18	35

4 - Se è vera l'ipotesi nulla che tutti i gruppi hanno la stessa tendenza centrale, la distribuzione attesa è facilmente intuibile: ogni gruppo dovrebbe avere metà (nell'esempio esattamente 17/35) dei suoi valori prima della mediana e l'altra metà (esattamente 18/35) dei suoi valori dopo la mediana.

La significatività delle differente distribuzione dei 5 gruppi è verificata con un test χ^2 con 4 df.

Con i dati dell'esempio, il numero di dati potrebbe essere ritenuto insufficiente per una sua applicazione valida per il test χ^2 o il test G, poiché tutte le 10 caselle delle frequenze attese hanno valori inferiori a 5. Sarebbe quindi più corretto l'uso della distribuzione ipergeometrica.

E' una ulteriore conferma del fatto che, quando si utilizza in modo ridotto la quantità d'informazione contenuta nei dati raccolti, per raggiungere con il test la potenza sufficiente è necessario disporre di campioni di dimensioni relativamente grandi.

Nel caso del test chi quadrato, quando si hanno gruppi con un numero di osservazioni particolarmente ridotto è utile procedere ad una aggregazione delle classi; con i dati dell'esempio, sarebbe conveniente unire i 4 gruppi minori B e C, D e E in due soli gruppi secondo le loro affinità. Si ricostruiscono le condizioni di validità, che nella distribuzione attesa richiedono almeno 5 osservazioni per casella, anche se tale operazione comporta un'ulteriore perdita di informazione e non permette l'analisi dettagliata per ogni singola specie.

I programmi di computer di norma forniscono almeno tre risultati, tra loro non identici:
- il valore del **χ^2 con gdl 4, (sovente, un secondo χ^2 con la correzione di Yates),**
- il valore del **test G sempre con 4 gdl, (sovente, una seconda versione con la correzione di Williams o di Mantel-Haenzel),**
- la probabilità esatta, stimata con la distribuzione ipergeometrica.

ESEMPIO 2. Tra i testi internazionali, il test della mediana per k campioni indipendenti è riportato nel volume di P. **Sprent** del 1993 (***Applied Nonparametric Statistical Methods*****,** 2nd ed., Chapman & Hall, London), dal quale sono tratti l'esempio successivo e i commenti.
Sono stati raccolti 6 campioni di dati, contenenti rispettivamente 4, 7, 5, 4, 2, 6, osservazioni.
Il confronto con la mediana comune a tutti i dati ha fornito il seguente risultato

Campione	<	>	Totale
I	**4**	**0**	4
II	**2**	**5**	7
III	**2**	**3**	5
IV	**3**	**1**	4
V	**2**	**0**	2
VI	**1**	**5**	6
TOTALE	14	14	28

Verificare alla probabilità **α = 0.05** l'ipotesi nulla

H_0: i sei campioni sono tutti estratti da popolazioni con la stessa mediana

contro l'ipotesi alternativa bilaterale

H_1: le sei mediane non sono tutte uguali

Risposta. E' possibile utilizzare il metodo classico

$$\chi^2_{(DF)} = \frac{(Oss. - Att.)^2}{Att.}$$

per il quale è necessario stimare la distribuzione attesa, nella condizione che H_0 sia vera.

Poiché ogni gruppo dovrebbe avere metà delle osservazioni prima e metà dopo la mediana, la **distribuzione attesa** è

Campione	<	>	Totale
I	**2**	**2**	4
II	**3,5**	**3,5**	7
III	**2,5**	**2,5**	5
IV	**2**	**2**	4
V	**1**	**1**	2
VI	**3**	**3**	6
TOTALE	14	14	28

Il valore del chi quadrato risulta

$$\chi^2_{(5)} = 2 \cdot \left(\frac{2^2}{2} + \frac{1{,}5^2}{3{,}5} + \frac{0{,}5^2}{2{,}5} + \frac{1^2}{2} + \frac{1^2}{1} + \frac{2^2}{3} \right) = 11{,}15$$

uguale a **11,15** con **5** df, mentre il valore critico alla probabilità **α = 0.05** è **11,07**.

Si rifiuta l'ipotesi nulla, con probabilità P < 0.05.

Secondo vari autori, i sei campioni dell'esempio sono piccoli e quindi sarebbe corretto utilizzare la distribuzione ipergeometica, invece di quella chi quadrato. **Sprent** osserva che

- la probabilità calcolata con il $\chi^2_{(5)}$ è esattamente P = 0.048,
- mentre con un test esatto risulta P = 0.046.

Un risultato molto vicino., tale da confutare eventuali obiezioni di validità nonostante il non rispetto delle condizioni di validità, in totale (N < 30) per casella nella distribuzione attesa (n_i <5).

15.3. CENNI SUL TEST DI NEMENYI E ALTRI PER LA MEDIANA IN K GRUPPI

Il **test di Nemenyi per la mediana (**a lui si devono altri test tra cui i confronti multipli non parametrici**)** non è particolarmente potente; ma è riportato in qualche pacchetto statistico a grande diffusione internazionale e quindi è citato in alcune pubblicazioni scientifiche al posto del test della mediana. Di conseguenza, può essere utile conoscerne i concetti fondamentali. A grandi linee, **può essere visto come una diversa elaborazione del test della mediana e serve per verificare le stesse ipotesi**.

Nel confronto tra più gruppi, l'ipotesi nulla resta

$$\mathbf{H_0: me_A = me_B = \ldots = me_K}$$

e l'ipotesi alternativa, solo bilaterale, è ancora

H_1: le mediane non sono tutte uguali

Il test di Nemenyi richiede due condizioni sperimentali più restrittive del precedente test sulla mediana, che ne limitano sensibilmente le possibilità di utilizzazione:

- **i k gruppi a confronto devono avere lo stesso numero d'osservazioni,**
- **i campioni devono essere di grandi dimensioni**.

Ha pure il **grave svantaggio di richiedere tabelle di valori critici difficilmente reperibili** (qui non riportati).

Il lavoro originale di **P. Nemenyi** (***Distribution-free multiple comparisons***, unpublished doctoral thesis, Princeton University, 1963) è citato in alcuni testi di statistica non parametrica, ma non è mai stato pubblicato. Il test è da ricordare, in quanto **presente in alcuni programmi informatici a partire dalla metà degli anni '60**, in particolare in quelli prodotti dal Dipartimento di Statistica della Università di Princeton, dotato di una struttura di ricerca di fama mondiale nella statistica applicata.

La metodologia, differente dal test della mediana solo nella parte conclusiva, richiede che:

1- nella programmazione dell'esperimento si raccolga un numero d'osservazioni identico in tutti i gruppi a confronto

$$n_1 = n_2 = \ldots = n_k$$

come nell'esempio seguente con 4 gruppi (che in realtà hanno un numero d'osservazioni troppo limitato, per un'applicazione corretta)

GRUPPO			
A	**B**	**C**	**D**
< 1	3,7	2,1	3,8
<1	2,8	2,5	2,2
3,8	0,9	2,9	5,2
2,1	2,2	>10	>10
3,2	2,5	8,7	>10

2 - Dopo aver definito l'ipotesi nulla

$$\mathbf{H_0: me_A = me_B = \ldots = me_K}$$

con ipotesi alternativa bilaterale

H_1: le mediane non sono tutte uguali

3 - ordinare per ranghi tutte le osservazioni dei **k** gruppi a confronto come se fossero un gruppo unico, mantenendo per ogni valore l'informazione del gruppo di appartenenza

<1	<1	0,9	2,1	2,1	2,2	2,2	2,5	2,5	2,8	**mediana**
A	A	B	A	C	B	D	B	C	B	segue

2,9	3,2	3,7	3,8	3,8	5,2	8,7	>10	>10	>10
C	A	B	A	D	D	C	C	D	D

4 – Successivamente, **identificare la mediana** di questa distribuzione unica (nella tabella precedente è a metà tra il 10° e l'11° valore) e, per ogni gruppo,

GRUPPI	< mediana	**> mediana**	Totale
A	3	$\mathbf{f_A = 2}$	5
B	4	$\mathbf{f_B = 1}$	5
C	2	$\mathbf{f_C = 3}$	5
D	1	$\mathbf{f_D = 4}$	5

contare quante sono le osservazioni che hanno valore superiore alla mediana (l'eventuale valore corrispondente alla mediana è classificato in uno dei due gruppi, anche se non dovrebbero risultare sbilanciati).

Nella distribuzione dei valori sopra e sotto la mediana, costruita come nel test relativo, **si prendono in considerazione solo le frequenze superiori alla mediana**; si utilizza solo una serie di frequenze f_1, f_2, , f_k (quelle riportate il grassetto nella tabella precedente).

5 - Con **k** gruppi e quindi **k** frequenze di valori superiori alla mediana, sono possibili **k(k – 1)/2** confronti 2 a 2, tra le frequenze assolute $\mathbf{f_i}$.

Da esse si deriva l'indice **h,** determinato dalla **differenza massima in valore assoluto**

$$\mathbf{h = \max |f_i - f_j|}$$

Con i dati dell'esempio, fra i 6 possibili confronti 2 a 2, si sceglie quello di B con D

$$\mathbf{h = |1 - 4| = 3}$$

che fornisce la differenza assoluta massima **h = 3**.

6 - Attraverso **tavole dei valori critici**, è possibile stimare la probabilità di ottenere per caso, nella condizione che l'ipotesi nulla sia vera, uno scarto uguale o superiore a quello stimato.
I programmi informatici riportano la probabilità **P**. Tale valore, confrontato con quello α prescelto, permette di decidere se è possibile rifiutare l'ipotesi nulla.

Per gli stessi scopi, in letteratura si trovano altri test. Fra questi, è possibile ricordare:
- il test di **Rijkoort**, proposto nel 1952, che si fonda su modalità simili alla somma dei ranghi utilizzata nel test di T Wilcoxon;
- il test proposto congiuntamente da **Brown e Mood** nel 1951, affine a quello della mediana;
- il test di **Bhapkar** del 1961;
- il test di **Deshpandé** ed il test di **Sugiura**, elaborati in modo indipendente nel 1965 e riproposti insieme con correzioni nel 1968, che utilizzano il calcolo delle precedenze e quindi sono una estensione del test U.

Poiché sono test per ora riportati i in pochissimi programmi informatici ed analoghi a quelli già illustrati in modo dettagliato, per la loro presentazione si rimanda a testi specifici.

15.4. ANALISI DELLA VARIANZA PER RANGHI, A UN CRITERIO DI CLASSIFICAZIONE: IL TEST DI KRUSKAL-WALLIS

Quando si utilizzano misure rilevate con una scala continua, seppure ordinale, quindi tutti i dati possono essere disposti in ranghi con un numero nullo o comunque ridottissimo di valori uguali, è utile ricorrere ad **un test più potente del test della mediana**. La quantità di informazione contenuta in ogni osservazione è superiore a quella utilizzata nel test della mediana, che si limita a classificare i valori in alti e bassi; di conseguenza, diviene più probabile verificare la significatività della differenza nella tendenza centrale, pure disponendo di un numero inferiori di dati. E' lo stesso concetto espresso nel confronto tra il test dei segni e il test di Wilcoxon-Mann-Whitney, nel caso di due campioni.
Il test proposto nel 1952 da W.H. **Kruskal** e W. A. **Wallis**, chiamato **Kruskal-Wallis One-Way ANOVA by Ranks** o più semplicemente **the Kruskal-Wallis test,** (con l'articolo ***Use of ranks in one criterion variance analysis*** pubblicato su **Journal of the American Statistical Association** vol. 47, pp. 583–621 e con quello del solo **Kruskal** sempre del 1952 ***A non parametric test for the several sample problem*** pubblicato su **Annals of Mathematical Statistics** vol. 23, pp. 525-540) è l'equivalente non parametrico dell'analisi della varianza ad un criterio di classificazione.
E' uno dei test più potenti per verificare l'ipotesi nulla $\mathbf{H_0}$, cioè se **k** gruppi indipendenti provengano dalla stessa popolazione e/o da popolazioni che abbiano **la medesima mediana**.

Anche la **metodologia del test di Kruskal-Wallis** è molto semplice e può essere schematizzata in alcuni passaggi.

1 – Per verificare l'ipotesi nulla che tutti i campioni hanno la stessa mediana

$$\mathbf{H_0: me_A = me_B = me_C = me_D = me_E}$$

con ipotesi alternativa che almeno una è differente o

$$\mathbf{H_1: \text{non tutte le mediane sono uguali}}$$

come nell'analisi della varianza ad un criterio di classificazione,

i dati dei **k** gruppi a confronto

GRUPPO			
A	**B**	**C**	**D**
76	47	55	36
85	52	46	18
22	63	71	29
67	---	42	46
72	---	---	---
81	---	---	---

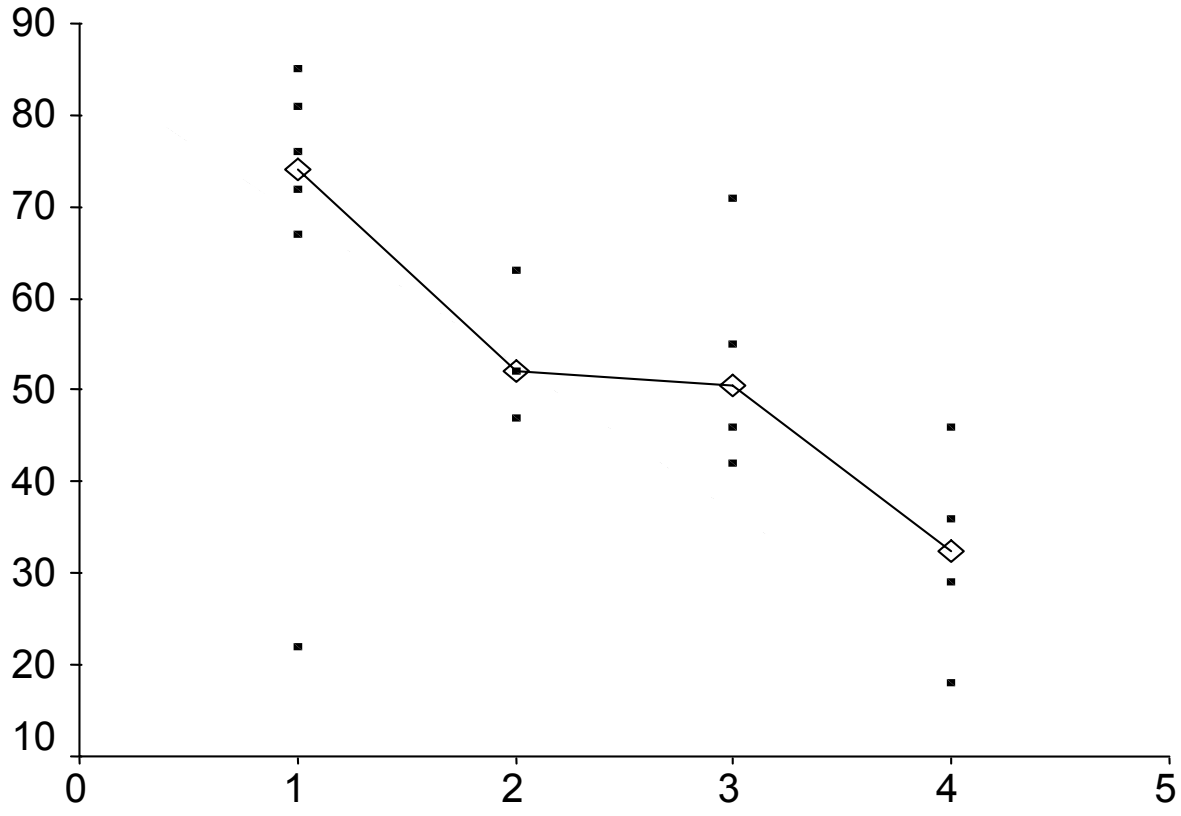
possono essere riportati in una tabella.

I gruppi a confronto possono avere un diverso numero d'osservazioni.

2 - Tutte le osservazioni dei **k** gruppi devono essere considerate come una serie unica e convertite in ranghi, mantenendo la stessa forma della tabella;

GRUPPO			
A	**B**	**C**	**D**
15	8	10	4
17	9	6,5	1
2	11	13	3
12	---	5	6,5
14	---	---	---
16	---	---	---

Se sono presenti misure uguali, a ciascuna di esse deve essere assegnato il loro rango medio.



Le misure originali, rappresentate su un diagramma cartesiano, evidenziano la maggiore variabilità dei dati del primo gruppo rispetto agli altri, in particolare di quella del secondo; inoltre questo gruppo ha un numero di dati sensibilmente minore, troppo piccolo per un confronto reale tra le varianze, che abbia una potenza sufficiente da rendere il risultato credibile.
Sono caratteristiche che devono indurre all'uso del test non parametrico.

3 – Calcolare,

	GRUPPO			
	A	**B**	**C**	**D**
	15	8	10	4
	17	9	6,5	1
	2	11	13	3
	12	- -	5	6,5
	14	- -	- -	- -
	16	- -	- -	- -
R_i	**76**	**28**	**34,5**	**14,5**
N_i	**6**	**3**	**4**	**4**
$\bar{r}_i$	**12,67**	**9,33**	**8,63**	**3,63**

come riportato nella tabella,

- la somma dei ranghi di ogni gruppo (R_i) e quella totale (**R**),
- il numero di osservazioni di ogni gruppo (n_i) e totale (**N**),
- da cui la media di ogni gruppo ($\bar{r}_i$)

Con

$$\mathbf{R} = 76 + 28 + 34{,}5 + 14{,}5 = \mathbf{153}$$

e

$$N = 6 + 3 + 4 + 4 = 17$$

la media generale $\bar{r}$

$$\bar{r} = \mathbf{R / N = 153 / 17 = 9}$$

risulta uguale a 9.

4 - Se i campioni provengono dalla stessa popolazione o da popolazioni con la stessa tendenza centrale (**H_0 vera**), queste medie aritmetiche dei ranghi di ogni gruppo ($\bar{r}_i$ = **R_i / n_i**) dovrebbero essere statisticamente simili sia tra loro sia alla media generale.

Da questo concetto è possibile derivare la formula per il calcolo di un indice (**g**), che dipende dalle differenze tra le medie dei gruppi e la media generale.
(Nella formula sottostante, la differenza tra la media di ogni gruppo e la media generale è nascosta dall'uso della somma totale dei ranghi, che ovviamente dipende da **N**)

$$\mathbf{g} = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left[\frac{\bar{r}_i - (N+1)}{2} \right]^2$$

La quantità

N (N + 1) / 12

è la varianza (riportata al denominatore nelle formule generali), che dipende solo da **N**,
mentre la media degli n_i ranghi (r_{ij}) è

$$\bar{r}_i = \frac{1}{n_i} \sum_{i=1}^{ni_i} r_{ij}$$

Come nell'analisi della varianza, il parametro **g** (vari testi lo indicano con **KW**, iniziali dei due autori) può essere calcolato a partire dalle somme, con **<u>una formula abbreviata</u>** che offre anche il vantaggio di evitare le approssimazioni dovute alle medie

$$g = \left(\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i}\right) - 3(N+1)$$

dove:

- n_i = numero di dati del campione o gruppo i-esimo,
- N = numero totale di osservazioni dei k campioni,
- k = numero di campioni a confronto,
- R_i = somma dei ranghi del campione o gruppo i-esimo,

e la sommatoria è estesa a tutti i k gruppi.

Di conseguenza, con una formula che evidenzia in modo più evidente le sue componenti, anche se i calcoli richiedono un tempo maggiore, l'ipotesi di uguaglianza fra mediane è basata sulla funzione

$$g = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i (\bar{r}_i - \bar{r})^2 = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{r}_i - \frac{N+1}{2}\right)^2$$

dove

- $\bar{r}$ = media generale dei ranghi che, con N dati, corrisponde a $\frac{N+1}{2}$
- $\bar{r}_i$ = media dei ranghi del gruppo i

Nel caso di campioni piccoli e con valori ripetuti, è conveniente dare la preferenza a questa ultima formula.

Il **parametro g** si distribuisce approssimativamente come la **distribuzione χ^2 con gdl k-1** (dove **k** è il numero di gruppi a confronto), quando le dimensioni del campione rispettano le condizioni minime richieste per la validità del chi quadrato (numero totale di osservazioni **N** non eccessivamente ridotto e numero minimo di dati per gruppo $\mathbf{n_i}$ non inferiore a 5).

L'approssimazione alla distribuzione chi quadrato è tanto migliore quanto maggiore è il numero (k) di gruppi e il numero di osservazioni entro ogni gruppo è alto, maggiore di 5.

Quando il **numero di gruppi a confronto è ridotto (uguale a 3)** ed il **numero di osservazioni entro ogni gruppo è basso (inferiore a 5)** la distorsione dalla distribuzione χ^2 è elevata; di conseguenza, **per la significatività di g (o KW) si fa ricorso a tabelle specifiche**, predisposte da Kruskal e Wallis (1952).

Sono tavole limitate a casi molto particolari, in quanto valgono solo per analisi con 3 gruppi e dimensioni non superiori a 5 osservazioni in ogni gruppo.

Esse iniziano da dimensioni minime di 2, 1 e 1, nei 3 gruppi. **Sono dimensioni nettamente inferiori a quelle richieste per il test F**: per campioni molto piccoli, è quindi preferibile ricorrere a questo test non parametrico.

Si rifiuta l'ipotesi nulla, alla probabilità α riportata nella tabella, quando il valore **g** (oppure **KW**) calcolato è uguale o superiore a quello critico riportato nella tabella. L'ultima riga coincide con i valori critici della distribuzione chi quadrato per gradi di libertà 2.

A differenza dell'analisi della varianza, il test di Kruskal-Wallis può essere utilizzato anche quando un gruppo ha una sola osservazione. Dal testo ***Non Parametric Statistical Methods*** di M. **Hollander** e D. **Wolfe** (John Wiley & Sons, New York) del 1973 è tratto questo esempio, sulle percentuali di acqua contenuta in cinque sostanze diverse:

Tabella dei valori critici di g (o KW) del test di Kruskal -Wallis, per confronti fra 3 campioni con un numero ridotto di osservazioni(≤ 5).

I 3 campioni devono essere ordinati per dimensioni in modo decrescente

(l'ultima riga coincide con il χ^2 per 2 df)

n_i per campione			Valori critici alle probabilità riportate				
n_1	n_2	n_3	**α =.10**	**α =.05**	**α =.01**	**α =.005**	**α =.001**
2	2	2	4,57				
3	2	2	4,50	4,71			
3	3	2	4,56	5,36			
3	3	3	4,63	5,60		7,20	
4	2	2	4,46	5,33			
4	3	2	4,51	5,45	6,45	7,00	
4	3	3	4,71	5,73	6,75	7,32	8,02
5	2	2	4,36	5,16	6,53		
5	3	2	4,65	5,25	6,82	7,18	
5	3	3	4,53	5,65	7,08	7,51	8,24
5	4	2	4,54	5,27	7,12	7,57	8,11
5	4	3	4,55	5,63	7,44	7,91	8,50
5	4	4	4,62	5,62	7,76	8,14	9,00
5	5	2	4,62	5,34	7,27	8,13	8,68
5	5	3	4,54	5,71	7,54	8,24	9,06
5	5	4	4,53	5,64	7,77	8,37	9,32
5	5	5	4,56	5,78	7,98	8,72	9,68
N grande e k = 3			**4,61**	**5,99**	**9,21**	**10,60**	**13,82**

Percentuali di acqua contenuta in 5 campioni

A	B	C	D	E
7,8	5,4	8,1	7,9	7,1
8,3	7,4	6,4	9,5	---
7,6	7,1	---	10,0	---
8,4	---	---	---	---
8,3	---	---	---	---

Dopo trasformazione in ranghi ed aver calcolato sia la somma dei ranghi sia n_i per colonna

	A	B	C	D	E
	7	1	9	8	3,5
	10,5	5	2	13	
	6	3,5		14	
	12				
	10,5				
Rj	46	9,5	11	35	3,5
n_i	5	3	2	3	1

con N = 14 attraverso

$$g = (\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i}) - 3(N+1)$$

si stima

$$g = \left(\frac{12}{14 \cdot (14+1)} \cdot (\frac{46^2}{5} + \frac{9{,}5^2}{3} + \frac{11^2}{2} + \frac{35^2}{3} \frac{3{,}5^2}{1}) \right) - 3 \cdot (14+1)$$

$$g = \left(\frac{12}{210} \cdot (423{,}2 + 30{,}1 + 60{,}5 + 408{,}3 + 12{,}2) \right) - 3 \cdot 15 = 0{,}057 \cdot 934{,}3 - 45 = 8{,}25$$

g = 8,25 mentre il valore critico del $\chi^2_{(4)}$ alla probabilità α = 0.05 è 9,49 e a α = 0.10 è 7,78.

Non è possibile rifiutare l'ipotesi nulla. Ma, poiché il valore calcolato è vicino a quello critico e corrisponde a una probabilità P leggermente superiore a 0.05, si può concludere che è **tendenzialmente significativo**. Con un numero leggermente maggiore di dati il test probabilmente sarebbe significativo.

Recentemente è stato evidenziato, come mostra una lettura attenta dei **valori critici** riportati nella tabella, che il valore di **g** non è monotono: il suo andamento

- è prima crescente e poi decrescente,
- è asimmetrico e caratterizzato da numerosi valori modali.

Pertanto, sono state evidenziate perplessità sulla sua effettiva capacità di permettere la verifica di ipotesi sull'uguaglianza di mediane; di conseguenza sono state proposte alcune modifiche, che tuttavia sono ancora poco utilizzate.

Il test di **K**ruskal-**W**allis, fondato sui ranghi, è analogo all'analisi della varianza ad un criterio di classificazione, come il test di **W**ilcoxon-**M**ann-**W**hitney, fondato ugualmente sui ranghi, è analogo al test t di Student. Lo stesso confronto vale per la sua efficienza asintotica relativa.

L'**efficienza asintotica relativa** del test **KW** rispetto al test **F** è quindi identica a quelle del test **WMW** rispetto al **t**:

- quando la distribuzione dei dati è Normale ha un valore uguale a 0,95 ($3/\pi$),
- quando la distribuzione dei dati è Rettangolare ha un valore uguale a 1,
- quando la distribuzione dei dati è Esponenziale Doppia ha un valore uguale a 1,50 (3/2).

ESEMPIO 1**.** L'ozono si forma da O_2 in presenza di NO_2 e di radiazione solare. A concentrazioni elevate, causa congestione polmonare; il limite di accettabilità in Italia è fissato dalla legge in 200 gm^{-3} (0,1 ppm).

Durante una giornata estiva, in quattro zone di una città (A, B, C, D) si sono rilevate le concentrazioni di O_3.

A	B	C	D
150	120	200	195
140	115	190	210
145	30	185	220
160	155	180	205
165	130	- -	175
170	- -	- -	430
125	- -	- -	- -

Esiste una differenza significativa tra le mediane della concentrazione di O_3 nelle quattro zone?

Risposta. E' noto che i valori di concentrazione di una sostanza nell'aria sovente hanno valori anomali, a causa delle correnti e della disposizione delle fonti. Con pochi dati e in una ricerca nuova, sono ignote le caratteristiche statistiche della popolazione da cui sono estratti i dati campionari.
Nell'esempio riportato, anche la semplice lettura e la rappresentazione grafica dei dati sono in grado di evidenziare la non-normalità dei dati di alcune zone e la loro non omoscedasticità. Nel gruppo D, la presenza del valore 430 determina una varianza sensibilmente maggiore ed una distribuzione lontana dalla normalità (come tuttavia è necessario dimostrare con test adeguati, riportati in paragrafi successivi). Non è quindi possibile applicare l'analisi della varianza parametrica, ma si impone il ricorso al test di Kruskal-Wallis.

1 - I valori devono essere sostituiti dal loro rango, calcolato su tutte le osservazioni dei **k** gruppi a confronto. Da essi, si calcola la somma dei ranghi (**R_i**) ed il numero di osservazioni (**n_i**) di ogni gruppo o campione.

	A	B	C	D
	8	3	18	17
	6	2	16	20
	7	1	15	21
	10	9	14	19
	11	5	- -	13
	12	- -	- -	22
	4	- -	- -	- -
R_i	**58**	**20**	**63**	**112**
n_i	**7**	**5**	**4**	**6**

2 - Con **N** = 22 e **k** = 4 si ottiene un valore di **g**

$$g = \frac{12}{22 \cdot 23} \cdot \left(\frac{58^2}{7} + \frac{20^2}{5} + \frac{63^2}{4} + \frac{112^2}{6}\right) - 3 \cdot 23$$

$$g = 0{,}0237 \cdot (480{,}6 + 80 + 992{,}3 + 2090{,}7) - 69 = 17{,}35$$

uguale a 17,35.

3 - La tabella dei valori critici con **3 gdl** riporta

- 7,82 alla probabilità $\alpha = 0.05$,
- 11,34 alla probabilità $\alpha = 0.01$,
- 16,27 alla probabilità $\alpha = 0.001$.

Pertanto, si può rifiutare l'ipotesi nulla, con una probabilità di commettere un errore di I° tipo inferiore a 0.001.

Per la sua applicazione corretta, **il test di Kruskal-Wallis richiede che la misura utilizzata sia continua**. Di conseguenza, non si dovrebbero avere valori identici; ma nella pratica sperimentale, per l'approssimazione della scala o dello strumento, può succedere che alcune siano uguali. In questo caso, **con valori identici che occupano lo stesso rango la varianza campionaria è ridotta e diviene opportuno correggere il valore di g** .

La correzione per misure ripetute (ties) aumenta il valore di **g**; quindi incrementa la probabilità di trovare differenze significative tra le mediane dei gruppi a confronto. Tuttavia l'effetto della correzione è quasi sempre trascurabile, quando **le misure identiche sono meno di un quarto** delle osservazioni e **sono distribuite tra più ranghi**.

Per ottenere il valore di **g'** corretto, si deve dividere la quantità **g** calcolata per un fattore di correzione **C**

$$C = 1 - \frac{\sum_{i=1}^{p} c_i (c_i^2 - 1)}{N \cdot (N^2 - 1)}$$

dove:

- **p** è il numero di raggruppamenti con ranghi ripetuti,
- **c** è il numero di ranghi ripetuti nel raggruppamento i-esimo,
- **N** è il numero totale di osservazioni nei **k** campioni a confronto.

Un altro metodo di correzione per i ties è quello di Hinkley (che apporta variazioni maggiori sul risultato); è spiegato in un paragrafo successivo.

ESEMPIO 2. In una ricerca sulla qualità della vita, in tre quartieri (X, Y, Z) della stessa città sono stati ottenuti i punteggi di seguito riportati, con la medesima impostazione tabellare di un'analisi della varianza ad 1 criterio di classificazione.

X	Y	Z
7	4	2
7	4	2
6	4	1
5	7	3
8	5	--

Esistono differenze significative tra le loro mediane?

Risposta. Si devono sostituire i punteggi con i ranghi relativi e calcolare le somme, come nella tabella sottostante:

	X	Y	Z
	12	6	2,5
	12	6	2,5
	10	6	1
	8,5	12	4
	14	8,5	--
R_i	**56,5**	**38,5**	**10**
n_i	**5**	**5**	**4**

Da esse, con la formula

$$g = \left(\frac{12}{N(N+1)}\sum_{i}^{k}\frac{R_i^2}{n_i}\right) - 3(N+1)$$

si stima il valore di **g**

$$= \left[\frac{12}{14\cdot(14+1)}\cdot\left(\frac{56{,}5^2}{5}+\frac{38{,}5^2}{5}+\frac{10^2}{4}\right)\right] - 3\cdot(14+1)$$

$$= \left[\frac{12}{210}\cdot\left(638{,}45+296{,}45+25\right)\right] - 45 \;=\; \mathbf{(0{,}05714\ x\ 959{,}9) - 45 \;=\; 9{,}848}$$

che risulta uguale a **9,848**.

Per 3 campioni di dimensioni 5, 5, 4 alla probabilità **α = 0.001** il valore critico riportato nella tabella è uguale a 9,32. Di conseguenza, con probabilità **P** inferiore a **0.001** si rifiuta l'ipotesi nulla: esiste una differenza significativa tra le 3 mediane a confronto.

I dati presentano un numero elevato di ripetizioni:

- il valore **2** e il valore **5** compaiono 2 volte;
- il valore **4** e il valore **7** compaiono 3 volte.

Il fattore di correzione C

$$C = 1 - \frac{\sum_{i=1}^{g} c_i (c_i^2 - 1)}{N \cdot (N^2 - 1)}$$

con i dati dell'esempio risulta

$$1 - \frac{\left[2 \cdot (2^2 - 1) + 2 \cdot (2^2 - 1) + 3 \cdot (3^2 - 1) + 3 \cdot (3^2 - 1)\right]}{14 \cdot (14^2 - 1)} = 1 - \frac{2 \cdot 3 + 2 \cdot 3 + 3 \cdot 8 + 3 \cdot 8}{14 \cdot 195}$$

$$1 - \frac{60}{2730} = \mathbf{1\text{-}0{,}02198 = 0{,}97802}$$

uguale a 0,97802

e il valore corretto di **g (g')**

g' = 9,848 / 0,97802 = 10,069

diviene **10,069** (il precedente era **9,848**).

Anche questo esempio dimostra che **il fattore di correzione aumenta il valore di g**, ma per entità trascurabili. **Con questi dati la correzione non era necessaria, poiché il valore stimato era già significativo**; tuttavia è stato applicato, per fornire una stima corretta di **g**.

Come in tutte le correzioni per i ***ties,*** nell'indice di correzione **un solo valore ripetuto molte volte ha un peso relativo maggiore di molti valori ripetuti poche volte**.

Come semplice dimostrazione si può stimare che, se nelle 14 osservazioni vi fosse stato un solo valore ripetuto 5 volte, il fattore di correzione C sarebbe stato

$$C = 1 - \frac{5 \cdot (5^2 - 1)}{14 \cdot (14^2 - 1)} = 1 - \frac{5 \cdot 24}{14 \cdot 195} = 1 - \frac{120}{2730} = \mathbf{1 - 0{,}04396 = 0{,}95604}$$

uguale a 0,95604

e il valore corretto di **g**

g' = 9,848 / 0,95604 = 10,301

uguale a **10,301** fornendo uno scarto superiore al precedente.

L'esempio è solo teorico, in quanto **con tanti dati identici come nel caso appena ipotizzato la validità del test è molto dubbia** perché modifica la distribuzione delle probabilità, come evidenziato nei capitoli precedenti

- per la distribuzione T nel caso di un campione,
- per la distribuzione U nel caso di due campioni indipendenti.

15.5. CONFRONTO CON IL TEST F E CONFRONTI MULTIPLI CON I RANGHI.

Secondo alcuni autori di testi di Statistica non parametrica, tra i quali G. Landenna e D. Marasini, (vedi ***Metodi statistici non parametrici***, edito da il Mulino, Bologna, 1990, pag. 234) "*un diverso modo per verificare l'ipotesi nulla* $\boldsymbol{H_0: me_A = me_B = \ldots = me_K}$, *in alternativa all'ipotesi* H_1 *in cui non tutte le mediane sono uguali, è quello di ricorrere ai confronti multipli che, nel caso di rifiuto dell'ipotesi nulla, consentono anche la identificazione delle popolazioni con mediane diverse ovvero dei trattamenti i cui effetti hanno provocato il rifiuto medesimo"*.

A parere della maggioranza degli autori, questo concetto non è espresso con la dovuta chiarezza e può indurre in errore.

Anche nella statistica non parametrica per **k** campioni,

- è possibile ricorrere ai **confronti multipli,** per individuare quali sono i gruppi che hanno una tendenza centrale tra loro significativamente differente,
- **<u>solo se</u> è stata rifiuta l'ipotesi nulla** con il test di Kruskal-Wallis (più avanti si vedrà anche il test di Jonchkeere e il test Umbrella di Mack-Wolfe). E' un principio di cautela, al quale molti testi consigliano di attenersi.

Il motivo è che le due risposte potrebbero non coincidere, date le differenze esistenti tra i vari metodi e la stima non coincidente della probabilità experiment-wise rispetto a quella dei comparison-wise. Quindi anche con il test di Kruskal-Wallis si potrebbe non rifiutare l'ipotesi nulla, quando i confronti multipli evidenziano almeno una differenza significativa.

Nel caso di confronti multipli non parametrici, una procedura molto lunga, possibile solo **con programmi informatici** ma concettualmente molto semplice, è fondata sugli stessi principi del test di **casualizzazione** (o **permutation test**), già illustrato sia per due campioni dipendenti che indipendenti.

Con **N** dati suddivisi in **k** gruppi, è teoricamente facile

- analizzare tutte le possibili risposte, determinate dalle diverse collocazioni dei valori nei **k** gruppi;
- stabilire la probabilità **α con l'applicazione del principio del Bonferroni**: il valore della probabilità α di ogni confronto è ottenuto dividendo la probabilità totale per il numero di confronti possibili. Ad esempio, scegliendo la probabilità complessiva di $\alpha_T = 0.05$, per ognuno dei k confronti il valore di α

$$\alpha = 0.05 / [k \cdot (k - 1)]$$

- sono collocati nella zona di rifiuto dell'ipotesi nulla le distribuzioni dei ranghi che danno le differenze massime tra i ranghi secondo questa nuova probabilità.

La probabilità utilizzata per ogni confronto è così bassa da impedire di superare quella experiment-wise prefissata. Ma, appunto per questo, sono test poco potenti.

Metodi operativi concettualmente più complessi, ma operativamente più rapidi, sono quelli analoghi alle procedure presentate nei confronti multipli.

Secondo la logica già esposta nella presentazione del **t di Bonferroni**, il valore della probabilità deve essere diviso per il numero di confronti possibili. Ad esempio, con campioni abbastanza grandi è possibile

- scegliere la probabilità complessiva di $\alpha = 0.05$;
- successivamente, per effettuare tutti i confronti tra **k** gruppi, occorre utilizzare il valore di **Z** (o di **t** quando i campioni sono molto piccoli) corrispondente ad un valore di α uguale a

$$\alpha = 0.05 / [k \cdot (k - 1)]$$

Per evitare di effettuare tutti i possibili confronti tra coppie di medie con **k** gruppi (pari a $\mathbf{C_k^2}$),

- con una procedura analoga a quella del test parametrico **T** di Tukey,
- si calcolano le differenze in valore assoluto $|\bar{r}_a - \bar{r}_b|$ tra le medie dei ranghi di tutti i gruppi.

Di esse sono significative, alla probabilità α, tutte quelle che sono uguali o maggiori della quantità **D,** data da

$$\mathbf{D} = Z_{\alpha,c} \cdot \sqrt{\frac{n \cdot (n+1)}{12} \cdot (\frac{1}{n_a} + \frac{1}{n_b})}$$

dove

- **N** è il numero complessivo di dati considerando tutti i gruppi,
- $\mathbf{n_a}$ e $\mathbf{n_b}$ sono il numero di dati nei due gruppi (chiamati a e b) a confronto,
- **c** è il numero di possibili confronti, che con **k** gruppi e uguale a $\frac{k \cdot (k-1)}{2}$,
- **Z** alla probabilità $\boldsymbol{\alpha/c}$ è tratta dalla distribuzione normale.

Se i gruppi a confronto hanno tutti lo stesso numero d'osservazioni, è corretto e conveniente calcolare un solo valore (D), chiamato **differenza minima significativa (*least significant difference*)**. Sono significative tutte le differenze tra coppie di medie di ranghi che risultano superiori alla quantità **D** calcolata.

ESEMPIO. In alcuni tratti di 5 corsi d'acqua è stata misurata la quantità di tensioattivi anionici (misurata in mgl^{-1}) presenti

A	B	C	D	E
0,50	0,58	0,43	0,90	1,06
0,86	0,42	0,70	1,13	1,98
0,90	0,62	0,75	0,80	1,42
0,23	0,60	0,58	0,95	1,48
0,55	0,48	0,89	0,82	0,85
0,75	0,60	0,75	0,80	0,90
- -	- -	0,75	0,60	1,90

Dopo aver verificato se esistono differenze significative tra le mediane dei diversi corsi d'acqua, in caso positivo, individuare tra quali corsi tali differenze sono significative.

Risposta. I primi passi dell'analisi sono:

- trasformare i valori nei loro ranghi,
- annotare quanti sono i valori identici e quante le loro repliche.

	A	B	C	D	E
	5	7,5	3	25	28
	22	2	13	29	33
	25	12	15,5	18,5	30
	1	10	7,5	27	31
	6	4	23	20	21
	15,5	10	15,5	18,5	25
	- -	- -	15,5	10	32
$\mathbf{R_i}$	74,5	45,5	93,0	148,0	200
n_i	6	6	7	7	7
$\bar{r}_i$	12,42	7,58	13,29	21,14	28,57

E' inoltre utile ricordare che, quando sono presenti valori identici, il rango da attribuire è il loro valore medio.

Con un numero elevato di gruppi e di osservazioni entro ogni gruppo è facile incorrere in errori nell'attribuzione dei ranghi. **Una verifica rapida dei calcoli** è data dalla corrispondenza tra somma totale dei ranghi e numero totale di osservazioni: tra essi esiste l'uguaglianza

$$\text{Somma di tutti i ranghi} = \frac{N \cdot (N + 1)}{2}$$

Nell'esempio, l'operazione di attribuzione dei ranghi e le loro somme per gruppo sono state effettuate in modo corretto, poiché,

$$(74{,}5 + 45{,}5 + 93 + 148 + 200) = 561 = \frac{33 \cdot 34}{2}$$

Per il calcolo di **g** è utile ricordare che **k** = 5 e **N** = 33; pertanto si ottiene un valore

$$g = \frac{12}{33 \cdot 34} \cdot \left(\frac{74{,}5^2}{6} + \frac{49{,}5^2}{6} + \frac{93^2}{7} + \frac{148^2}{7} + \frac{200^2}{7}\right) - 3 \cdot 34$$

$$g = \frac{12}{1122} \cdot (925{,}0 + 345{,}0 + 1235{,}6 + 3129{,}1 + 5714{,}3) - 102$$

$$g = 0{,}0107 \cdot 11349 - 102 = 21{,}704$$

di **g** uguale a 21,704.

Con 5 gruppi, la significatività è fornita dalla tabella sinottica dei valori critici del χ^2 per 4 gdl:

- alla probabilità α = 0.05 è uguale a 9,49,
- alla probabilità α = 0.01 è uguale a 13,28,
- alla probabilità α = 0.001 è uguale a 20,52.

A causa della presenza di misure ripetute, può rivelarsi vantaggioso **correggere il valore di g** (non in questo caso, in quanto già significativo ad una probabilità inferiore a 0.001).

Con i dati dell'esempio, poiché sono presenti i seguenti valori identici riportati nella tabella:

Per	2	Volte Compare il Valore	0,58	Che ha Rango Medio	7,5
"	3	"	0,60	"	10
"	4	"	0,75	"	15,5
"	2	"	0,80	"	18,5
"	3	"	0,90	"	25

il fattore di correzione C è

$$C = 1 - \frac{2 \cdot (2^2 - 1) + 3 \cdot (3^2 - 1) + 4 \cdot (4^2 - 1) + 2 \cdot (2^2 - 1) + 3 \cdot (3^2 - 1)}{33 \cdot (33^2 + 1)}$$

$$C = 1 - \frac{2 \cdot 3 + 3 \cdot 8 + 4 \cdot 15 + 2 \cdot 3 + 3 \cdot 8}{33 \cdot 1088} = 0{,}996658$$

uguale a 0,996658.

Il **valore corretto di g** diventa

$$\frac{21,704}{0,996658} = 21,777$$

uguale a 21,777 che, ovviamente, è ancor più significativo del valore stimato in precedenza.

Rifiutata l'ipotesi nulla (tutti i gruppi o campioni siano estratti dalla stessa popolazione o da popolazioni con la medesima mediana) **e quindi accettata l'ipotesi alternativa** (non tutte le mediane dei gruppi a confronto sono uguali), si può mettere in evidenza quali sono le medie dei ranghi che hanno una differenza significativa.

Per confronti semplici tra tutte le medie dei ranghi, con 5 campioni il numero di differenze può essere stimato con

$$C_2^5 = \frac{5!}{(5-2)!2!} = 10 \qquad \text{oppure} \qquad \frac{k \cdot (k-1)}{2} = \frac{5 \cdot 4}{2} = 10$$

e risulta uguale a 10.

Esse possono essere riportate in valore assoluto in una matrice triangolare con il relativo numero di osservazioni

		A(6)	B(6)	C(7)	D(7)	E(7)
		12,42	7,58	13,29	21,14	28,57
A	12,42					
B	7,58	**4,84**				
C	13,29	**0,87**	**5,71**			
D	21,14	**8,72**	**13,56**	**7,85**		
E	28,57	***16,15****	***20,99****	***15,28****	***7,43***	

Alla probabilità complessiva α_T **= 0.05** per **10** confronti simultanei, la probabilità α di ogni confronto è uguale a **0.005**; per un test bilaterale (quindi alla probabilità α = 0.0025 in una coda della distribuzione) sulla tavola della distribuzione normale ad essa corrisponde un valore di **Z** uguale a **2,81** (esattamente 2,807).

Per un test a due code, **la differenza minima significativa D** (con **N** = 33) per il confronto tra A e B che hanno entrambe 6 osservazioni

$$D = 2{,}807 \cdot \sqrt{\frac{33 \cdot 34}{12} \cdot (\frac{1}{6} + \frac{1}{6})} = 2{,}807 \cdot \sqrt{93{,}5 \cdot 0{,}33} = 15{,}66$$

è uguale a 15,66.

Per i 6 confronti tra le medie dei ranghi dei campioni A e B, che hanno 6 osservazioni, con i campioni C, D e E che hanno 7 osservazioni,

$$D = 2{,}807 \cdot \sqrt{\frac{33 \cdot 34}{12} \cdot (\frac{1}{6} + \frac{1}{7})} = 2{,}807 \cdot \sqrt{93{,}5 \cdot 0{,}309} = 15{,}10$$

D è uguale a 15,10.

Per i 3 confronti tra i gruppi C, D e E che hanno 7 osservazioni,

$$D = 2{,}807 \cdot \sqrt{\frac{33 \cdot 34}{12} \cdot (\frac{1}{7} + \frac{1}{7})} = 2{,}807 \cdot \sqrt{93{,}5 \cdot 0{,}286} = 14{,}51$$

D è uguale a 14,51.

Confrontando i valori D stimati con le differenze riportate nella tabella triangolare precedente, **alla probabilità complessiva α_T = 0.05 risultano significative solamente 3 delle 10 differenze calcolate tra i cinque gruppi** (i valori in corsivo e con un asterisco):

- la differenza tra la media dei ranghi del gruppo A e quella del gruppo E,
- la differenza tra la media dei ranghi del gruppo B e quella del gruppo E,
- la differenza tra la media dei ranghi del gruppo C e quella del gruppo E.

Osservando i tre diversi intervalli calcolati (15,66 - 15,10 - 14,51) è didatticamente utile sottolineare come l'intervallo diminuisca in modo non trascurabile anche all'aumento di una sola osservazione, quando i campioni sono numericamente così ridotti.

Con lo stesso numero di gruppi, **<u>l'intervallo minimo significativo</u> varia anche in rapporto al numero di confronti che si vogliono fare**.

Se, in modo analogo al test di Scheffé, si intende effettuare oltre ai confronti semplici anche **confronti complessi tra le medie combinate di alcuni gruppi**, il valore di **Z** cambia (cresce in rapporto all'aumentare del numero totale di confronti possibili, come evidenziato dalla tabella relativa).

Ad esempio, se si intendono fare 15 confronti

- mantenendo costante la probabilità complessiva di α_T = 0.05,
- la probabilità α di ogni confronto diventa (0.05/15) uguale a 0.0033 per un test unilaterale
- e a 0.00167 per un test bilaterale, al quale corrisponde un valore di Z uguale a 2,94.

Nel caso di **confronti tra alcuni trattamenti ed un controllo** (per esempio, se si intendesse confrontare le medie dei trattamenti B, C, D, E solamente con la media del campione A, ritenuto la situazione normale o standard) il numero di confronti si riduce notevolmente; è pari a **k-1**, quando **k** è il numero totale di gruppi.
Il metodo è utilizzato in medicina quando si intendono valutare gli effetti di alcuni farmaci rispetto al solo placebo; è frequente nella ricerca tossicologica ed ambientale, per confrontare più situazioni a rischio con quella normale.
Con 5 gruppi o campioni, di cui 1 è il placebo o controllo e 4 sono i farmaci o le situazioni da verificare, il numero di confronti è 4.
Dopo aver applicato l'analisi della varianza di Kruskal-Wallis per dimostrare l'esistenza di almeno una eventuale differenza tra i 5 gruppi, è possibile verificare la significatività di ognuna delle differenze tra le mediane dei **k-1** trattamenti rispetto alla mediana o tendenza centrale del controllo.

Il procedimento è identico al precedente, con 2 sole variazioni:

1 - Alla medesima probabilità di rifiutare l'ipotesi nulla, il valore di **Z** è minore, perché inferiore è il numero di possibili confronti; di conseguenza, sarà minore anche il valore della differenza minima significativa, a parità del numero di osservazioni impiegate nei rispettivi gruppi.
Per esempio alla probabilità complessiva $\alpha_T = 0.05$ la probabilità α di ogni confronto diviene:
- 0.05/4 uguale a 0.0125 e il valore di **Z** corrispondente è uguale a 2,24 per un test ad una coda,
- 0.025/4 uguale a 0.00625 e il valore di **Z** corrispondente è uguale a 2,50 per un test bilaterale.

2 - Con lo stesso numero **N** di osservazioni totali, è possibile ottenere una maggiore efficienza - potenza dei 4 confronti, non programmando gruppi con lo stesso numero di dati ma ponendo un maggiore numero di osservazioni nel gruppo o campione di controllo.
Infatti, il suo numero di osservazioni viene utilizzato nei calcoli di tutte le **k-1** differenze minime significative.
Ad esempio, con 30 osservazioni e 5 gruppi invece di attribuire 6 dati per gruppo risulta più conveniente assegnare 5 osservazioni ai 4 trattamenti e 10 osservazioni al controllo.
I 4 confronti saranno tutti tra due gruppi rispettivamente di 10 (il controllo) e 5 (un trattamento) osservazioni; essi risultano più potenti di 4 confronti effettuati sempre tra due gruppi di 6 osservazioni (sia il controllo che un trattamento).

15.6. TEST PER L'ETEROGENEITA' DELLA VARIANZA CON K CAMPIONI

Nella ricerca biologica, medica, ecologica ed ambientale, sono frequenti le situazioni in cui l'attenzione del ricercatore è rivolta alla **variabilità dei dati**, più che alla loro tendenza centrale. E' il caso di misure d'inquinamento che in zone differenti possono avere una variabilità diversa, pure con una tendenza centrale simile; anche se per l'inquinamento le mediane sono tutte sotto i limiti di legge, dove la varianza risulta maggiore è più urgente intervenire, poiché singole osservazioni possono superarli con frequenza più alta. E' il caso di farmaci, dove è importante la riposta media allo stimolo di una dose, ma ancor più la variabilità tra individui: un farmaco con una media peggiore può essere preferito, se garantisce una maggiore omogeneità di risposta dei pazienti. Il controllo di qualità di un prodotto industriale è fondato sulle modalità di riduzione della varianza, per garantire che tutte le confezioni siano uguali. Inoltre, come discorso più generale di confronto tra popolazioni, è possibile affermare che **k serie di dati campionari appartengono a popolazioni differenti, se hanno varianze statisticamente non uguali**.

E' un problema già discusso nel caso di 2 campioni indipendenti, che può essere **facilmente esteso a k campioni,** nello stesso modo con il quale il **test WMW** (Wilcoxon-Mann-Whitney) ha la sua generalizzazione nel **test KW** (Kruskal-Wallis).

Nella statistica parametrica è il test di Levene, di cui questo test può essere interpretato come il corrispondente non parametrico.

Da una distribuzione di dati classificati in **k gruppi**,

- si calcola la **media** di **ogni gruppo**
- e entro essi **le differenze in valore assoluto** di ogni dato dalla sua media;
- a questi **k gruppi di differenze** si applica il **test di Kruskal-Wallis sulle mediane**;
- se il **test risulta significativo**, vuol dire che **le mediane delle differenze sono significative;**
- in altri termini, **le varianze dei k gruppi sono significativamente differenti**.

In modo più dettagliato, prendendo come esempio d'applicazione un articolo pubblicato sulla rivista **Applied Statistics** nel 1989 (di D. V. **Hinkley**, ***Modified profile likelihood in trasformed linear models***, Vol. 38, pp. 495-506), la procedura presentata da **P. Sprent** nel volume ***Applied nonparametric statistical methods***, (second Edition, Chapman & Hall, London, 1993, pp. 155-156) prevede che

1 - per la verifica dell'ipotesi nulla

$$\mathbf{H_0: \sigma^2_A = \sigma^2_B = \ldots = \sigma^2_K}$$

contro l'ipotesi alternativa

$$\mathbf{H_1: \text{le } \sigma^2 \text{ dei gruppi a confronto non sono tutte uguali}}$$

in un caso **con pochi dati** (meno di quelli richiesti dal metodo di Moses per 2 campioni indipendenti, già illustrato), situazione non rara nella ricerca ambientale e biologica,

A	2	8	8	4
B	8	7	14	- -
C	33	59	48	56
D	60	101	67	- -

2 - dopo il **calcolo delle medie ($\bar{x}_i$) di ogni gruppo**

Gruppo	Xij				**medie $\bar{x}_i$**
A	2	8	8	4	**5,50**
B	8	7	14	- -	**9,67**
C	33	59	48	56	**49,00**
D	60	101	67	- -	**76,00**

3 - si debbano stimare **le deviazioni, in valore assoluto**, di ogni dato dalla media del suo gruppo

$$\left| x_{ij} - \bar{x}_i \right|$$

ottenendo una nuova tabella di scarti assoluti come la seguente

	$\left\| x_{ij} - \bar{x}_i \right\|$			
A	3,50	2,50	2,50	1,50
B	1,67	2,67	4,33	- -
C	16,00	10,00	1,00	7,00
D	16,00	25,00	9,00	- -

A essi si applica il test di Kruskal-Wallis, per valutare se hanno dimensioni medie (cioè le mediane se si parla dei valori, le medie se si parla dei loro ranghi) differenti.

4 – Di conseguenza, i dati riportati nell'ultima tabella devono essere ulteriormente **modificati nei ranghi** relativi, considerando tutto il campione e ottenendo

A	**7**	**4,5**	**4,5**	**2**
B	**3**	**6**	**8**	**- -**
C	**12,5**	**11**	**1**	**9**
D	**12,5**	**14**	**10**	**- -**

5 - Se **l'ipotesi nulla è vera** (variabilità uguale in ogni gruppo), i ranghi di ogni gruppo dovrebbero essere distribuiti casualmente e quindi avere medie uguali, sia tra loro, sia alla media generale.
Se **l'ipotesi nulla è falsa**, la media dei ranghi di almeno un gruppo dovrebbe essere significativamente differente da quella dagli altri.
E' **la stessa condizione** (sulle medie dei ranghi e mediane dei valori) **verificata dal test di Kruskal-Wallis** (che può essere applicato sui ranghi dell'ultima tabella).

6 - Per giungere alla stima di **g** con la formula abbreviata

$$g = \left(\frac{12}{N(N+1)} \sum_{i}^{k} \frac{R_i^2}{n_i}\right) - 3(N+1)$$

dapprima si calcolano i totali ($\mathbf{R_i}$) dei ranghi e il numero di osservazioni ($\mathbf{n_i}$) entro ogni gruppo

	A	B	C	D
$\mathbf{R_i}$	**18**	**17**	**33,5**	**36,5**
n_i	**4**	**3**	**4**	**3**

e con N = 14

$$g = \left[\frac{12}{14 \cdot (14+1)} \cdot \left(\frac{18^2}{4} + \frac{17^2}{3} + \frac{33{,}5^2}{4} + \frac{36{,}5^2}{3} \right) \right] - 3 \cdot (14+1)$$

$$g = \left[\frac{12}{210} \cdot \left(81{,}00 + 96{,}33 + 280{,}56 + 444{,}08\right) \right] - 45 \;=\; \mathbf{(0{,}05714\ \ x\ \ 901{,}97) - 45 \;=\; 6{,}54}$$

si ottiene **g** = 6,54 .

Il risultato deve essere confrontato con i valori critici del χ^2.

Per **k** = 4 gruppi, i **gdl** sono 3; il valore critico alla probabilità α = 0.05 è uguale a 7,815. Non è possibile rifiutare l'ipotesi nulla: i vari gruppi non hanno una variabilità significativamente differente.

<u>Poiché esistono valori identici, è possibile apportare la correzione relativa</u>.

Ricorrendo alla formula già illustrata

$$C = 1 - \frac{\sum_{i=1}^{p} c_i (c_i^2 - 1)}{N \cdot (N^2 - 1)}$$

dove:

- **p** è il numero di raggruppamenti con ranghi ripetuti,
- **c** è il numero di ranghi ripetuti nel raggruppamento i-esimo,
- **N** è il numero totale di osservazioni nei **k** campioni a confronto.

Ad esempio, con **p** = 2 e **c** = 2 si ottiene un termine di correzione **C**

$$\mathbf{C = 1 - \frac{(2x3) + (2x3)}{14x195} = 1 - 0{,}0044 = 0{,}\ 9956}$$

uguale a 0,9956 che non modifica sostanzialmente il valore di **g corretto**

g corretto = 6,54 / 0,9956 = 6,57

risultando uguale a 6,57 con arrotondamento, rispetto al 6,54 precedente.

Poiché **<u>il campione è piccolo ed esistono valori identici</u>**, **<u>come pubblicato anche da Hinkley</u>**, è conveniente **usare un'altra correzione** che, scritta nei suoi passaggi operativi,

è

$$g = \frac{(N-1) \cdot (Sp - C)}{Sr - C}$$

dove

$$Sp = \sum_{i=1}^{k} \left(\frac{s_i^2}{n_i} \right)$$

con $s_i = \sum r^2_{ik}$ (sommatoria del quadrato dei ranghi dei **k** gruppi) e

$\mathbf{n_i}$ = numero di dati di un gruppo

$$Sr = \sum_{i,k} r^4_{ik}$$

(sommatoria di tutti i ranghi elevati alla quarta)

$$C = \sum_{i=1}^{k} s_i^2 \ / N$$

Con i dati dell'esempio,

$$\mathbf{s_1} = 7^2 + 4{,}5^2 + 4{,}5^2 + 2^2 = 93{,}5$$

$$\mathbf{s_2} = 3^2 + 6^2 + 8^2 = 109{,}0$$

$$\mathbf{s_3} = 12{,}5^2 + 11^2 + 1^2 + 9^2 = 359{,}25$$

$$\mathbf{s_4} = 12{,}5^2 + 14^2 + 10^2 = 452{,}25$$

con

$\mathbf{s_1 + s_2 + s_3 + s_4 = N{\cdot}(N+1){\cdot}(2N+1) / 6}$ solo **quando non esistono valori identici**

infatti 93,5 + 109,0 + 359 25 + 452,25 = 1014

mentre 14 x 15 x 29 / 6 = 1015

Dai valori $\mathbf{s_1}$, $\mathbf{s_2}$, $\mathbf{s_3}$, $\mathbf{s_4}$ si ottiene **Sp** mediante

$$Sp = (93{,}5^2 / 4) + (109^2 / 3) + (359{,}25^2 / 4) + (452.25^2 / 3) = 106.587{,}724$$

che risulta uguale a 106.587,724

mentre **Sr** è dato **dalla somma di tutti i 14 ranghi alla quarta**

$$\mathbf{Sr = 7^4 + 4{,}5^4 + 4{,}5^4 + 2^4 + \quad + 12{,}5^4 + 14^4 + 10^4 = 127.157{,}25}$$

e risulta uguale a 127.157,25

e **C** è dato dal **quadrato della somma dei k** $\mathbf{s_i}$ **diviso N**

$$\mathbf{C = 1014^2 / 14 = 73.442{,}571}$$

e risulta uguale a 73.442,571.

Da essi si stima il valore corretto di **g**

$$\mathbf{g} = \frac{(N-1)\cdot(Sp-C)}{Sr-C} = \frac{(14-1)\cdot(106.587{,}724-73.442{,}571)}{127.157{,}25-73.442{,}571} = \frac{430.886{,}989}{53.714{,}679} = 8{,}0217$$

che risulta uguale a 8,0217.

Il nuovo valore risulta significativo alla probabilità α = 0.05 e rovescia le conclusioni precedenti fondate su altre modalità di correzione; con questo metodo, si evidenzia una differenza significativa nella variabilità dei 4 gruppi di dati.

Come il test di Levene è più potente dei test tradizionali fondati sui rapporti tra le varianze (Hartley e Cochran ai quali può essere aggiunto Bartlett, anche se ha una metodologia differente) questo test **risulta più potente di quelli fondati su misure dirette di variabilità**. E' quindi consigliato nei testi più recenti e è stato inserito nei programmi informatici più diffusi.

15.7. CONFRONTI TRA PIU' PROPORZIONI E CONFRONTI MULTIPLI RELATIVI.

Il confronto simultaneo tra più proporzioni è già stato illustrato nel capitolo III, presentando i metodi dei test χ^2 e **G** in tabelle di contingenza **2 x N**.

E' già stato spiegato come in esperimenti che danno risposte binarie, ad esempio in un esperimento di ecotossicologia nel quale si confrontano gli effetti letali di 4 sostanze tossiche (A, B, C, D), per facilitare i calcoli successivi e meglio comprendere i risultati del test, i dati devono essere riportati in una tabella come la seguente

	FREQUENZE OSSERVATE				
Risposte	A	B	C	D	TOTALE
Deceduti	**32**	**40**	**17**	**10**	99
Sopravvissuti	**56**	**70**	**66**	**21**	213
TOTALE	88	110	83	31	312

Successivamente si stimano le frequenze attese, nella condizione che l'ipotesi nulla sia vera.

Per ogni casella, la frequenza attesa è stimata facilmente con

Freq. attesa = Totale di riga x Totale di colonna / Totale generale

ottenendo

	FREQUENZE ATTESE				
Risposte	A	B	C	D	TOTALE
Deceduti	**27,9**	**34,9**	**26,3**	**9,9**	99
Sopravvissuti	**60,1**	**75,1**	**56,7**	**21,1**	213
TOTALE	88	110	83	31	312

Infine si possono calcolare

- il valore del χ^2 con 3 gdl mediante

$$\chi^2_{(g.d.l.)} = \sum_{i=1}^{n} \frac{\left(f_i^{oss} - f_i^{att}\right)^2}{f_i^{att}}$$

- il valore del **G** con gli stessi gdl mediante

$$G_{(gdl)} = G_{(gdl)} = 2\sum_{i=1}^{k}\left[Oss.\ln(Oss./Att.)\right]$$

oppure mediante

$$G_{(gdl)} = 2\left[\sum_{i=1}^{k} Oss.(\ln Oss.) - \sum_{i=1}^{k} Oss.(\ln Att.)\right]$$

Con gli stessi dati, per verificare l'ipotesi nulla

$$\mathbf{H_0: \pi_A = \pi_B = \pi_C = \pi_D}$$

contro l'ipotesi alternativa

H_1: non tutte le π sono uguali

oppure

H_1: almeno due π sono tra loro diverse

quando <u>il campione è grande e le proporzioni non sono prossime a **1** oppure a **0**</u> è possibile utilizzare il χ^2 con una modalità che offre altre opportunità di analisi.

A questo scopo, è vantaggioso presentare gli stessi in una tabella differente, che meglio evidenzia

Risposte	SOSTANZE				
	A	B	C	D	TOTALE
Deceduti	**32**	**40**	**17**	**10**	**99**
Campione (n_i)	88	110	83	31	312
Proporzioni (p_i)	**0,364**	**0,364**	**0,205**	**0,323**	**0,317**

- le varie proporzioni (**p_i**) di deceduti (o di sopravvissuti) a confronto,
- le dimensioni di ogni campione,
- la proporzione media complessiva ($\overline{p}$) delle **k** classi.

Per verificare se **le 4 proporzioni sono tra loro simili**, si può ricorrere alla formula

$$\chi^2_{(k-1)} = \sum_{i=1}^{k} \frac{(X_i - n_i \cdot p)^2}{n_i \cdot p \cdot q}$$

dove

- X_i è la frequenza osservata per ogni classe o campione,
- n_i è il numero di dati rilevati per ogni classe o dimensione di ogni campione,
- p può essere la **media generale** (nel caso dell'esempio è 0,317) **ponderata** oppure una **proporzione attesa qualsiasi** con la quale si vuole effettuare il confronto. Si ricorre alla media generale se essa rappresenta la stima migliore dell'effetto dei **k** gruppi, supposti tutti uguali; si utilizza una frequenza attesa per verificare se essa globalmente può essere il vero effetto medio dei k gruppi.

Per comprendere la formula precedente, è utile

- osservare come essa sia il quadrato di quella utilizzata per la distribuzione normale

$$Z = \frac{X - n \cdot p}{\sqrt{n \cdot p \cdot q}}$$

estesa a **k** proporzioni,

- ricordare che $\chi^2_{(1)} \cong Z^2$.

ESEMPIO 1. Verificare se le proporzioni (**p_i**) dei decessi nei 4 esperimenti sono uguali, utilizzando i dati dell'ultima tabella.

Risposta. Per verificare l'ipotesi nulla

$$\mathbf{H_0: \pi_A = \pi_B = \pi_C = \pi_D}$$

si calcola il valore del χ^2 con **3** gdl mediante

$$\chi^2_{(k-1)} = \sum_{i=1}^{k} \frac{(X_i - n_i \cdot p)^2}{n_i \cdot p \cdot q}$$

ottenendo

$$\chi^2_{(3)} = \frac{(32 - 88 \cdot 0{,}317)^2}{88 \cdot 0{,}317 \cdot 0{,}683} + \frac{(40 - 110 \cdot 0{,}317)^2}{110 \cdot 0{,}317 \cdot 0{,}683} + \frac{(17 - 83 \cdot 0{,}317)^2}{83 \cdot 0{,}317 \cdot 0{,}683} + \frac{(10 - 31 \cdot 0{,}317)^2}{31 \cdot 0{,}317 \cdot 0{,}683}$$

$$\chi^2_{(3)} = \frac{(32 - 27{,}9)^2}{19{,}1} + \frac{(40 - 34{,}9)^2}{23{,}8} + \frac{(17 - 26{,}3)^2}{18{,}0} + \frac{(10 - 9{,}8)^2}{6{,}7} = 0{,}88 + 1{,}09 + 4{,}81 + 0{,}01 = 6{,}79$$

χ^2 **= 6,79** con **3** gdl.

Poiché i valori critici con **3** gdl sono

- **7,815** per α **= 0.05**
- **6,231** per α **= 0.01**

al valore calcolato corrisponde una probabilità **P** compresa tra **0.10** e **0.05**.

In conclusione, con i dati raccolti non si può rifiutare l'ipotesi nulla, anche se è possibile affermare che l'esperimento è tendenzialmente significativo. Probabilmente, campioni con un numero maggiore di osservazioni avrebbero permesso di rifiutare l'ipotesi nulla.

Se l'informazione sui risultati è limitata alla **sola conoscenza delle proporzioni** (**p_i**), senza le dimensioni di ogni campione (**n_i**) o l'equivalente frequenza assoluta di casi postivi (**X_i**), **quando si disponga di misure ripetute** per ogni sostanza a confronto, è possibile utilizzare

- l'ANOVA dopo trasformazione angolare per omogeneizzare le varianze, se le proporzioni sono calcolate su campioni di dimensioni simili,
- l'analisi della varianza non parametrica (test di Kruskal-Wallis) come test più robusto, soprattutto se le dimensioni (**n_i**) di ogni campione sono diverse.

In entrambi i casi, dopo il rifiuto dell'ipotesi nulla, è possibile ricorrere ai confronti multipli, come già illustrato nei paragrafi dedicati all'ANOVA.

Anche nel caso del χ^2, **se e solo se il test risulta significativo**, è utile chiedersi **tra quali proporzioni la differenza sia significativa**.

Limitando la presentazione alle metodologie più diffuse e più semplici dei confronti a posteriori, è possibile ricorrere a

- la procedura **SNK** (Student-Newman-Keuls) o di **Tukey** per confronti singoli,
- il test di **Dunnett** per il confronto di ogni trattamento con il controllo.

I metodi sono simili a quelli già presentati nell'analisi della varianza. Tuttavia, trattandosi di proporzioni ($\mathbf{p_i}$), si presentano due problemi.

Il primo è come **rendere le varianze omogenee**, poiché esse dipendono strettamente dal valore delle proporzioni (σ^2 = npq) e sono indubbiamente differenti se il test sulle proporzioni è risultato significativo. Si ottengono varianze omogenee attraverso la **trasformazione angolare, cioè delle proporzioni in gradi**. Anche in questo caso, le proposte sono numerose.

La trasformazione più diffusa è quella classica in

$$p' = arcsin\sqrt{p'}$$

Nel 1948 F. J. **Anscombe** per risposte binarie come sono le proporzioni (vedi ***The trasformation of Poisson, binomial and negative binomial data*** pubblicato su **Biometrika** vol. 70, pp.: 227-234) ha proposto

$$p' = arcsin\sqrt{\frac{X+\frac{3}{8}}{n+\frac{3}{4}}}$$

Nel 1950 M. F. **Freeman** e J. W. **Tukey** (vedi ***Transformations related to the angular and the square root*** pubblicato su **Ann. Math. Statist.** vol. 21 pp.: 607-611) hanno proposto

$$p' = \frac{1}{2}\left(arcsin\sqrt{\frac{X}{n+1}} + arcsin\sqrt{\frac{X+1}{n+1}}\right)$$

Per calcolare la significatività del confronto tra la frequenza (trasformata in gradi) del generico gruppo A con quella del generico gruppo B si utilizza una formula analoga al test **t** per due campioni indipendenti, utilizzando la distribuzione del **q studentizzato**

$$Q_{\alpha/2,\infty,k} = \frac{'p_A - 'p_B}{es}$$

dove

- **α/2** è la probabilità prefissata in una distribuzione bilaterale (per il test di Tukey),
- **∞** è il numero di osservazioni in campioni grandi,
- **k** è il numero di gruppi a confronto (differente tra Tukey e Dunnett, dipendendo dal numero di confronti da effettuare).

Il secondo problema è come stimare l'**errore standard** (es), avendo un solo dato (la proporzione $\mathbf{p_i}$) per ogni campione. In questo caso, la relazione tra proporzione e varianza diventa un vantaggio poiché, dopo la trasformazione angolare o in gradi, la varianza è una costante e quindi l'errore standard si diversifica solo sulla base del numero **n** di osservazioni.

Per la stima dell'**errore standard**, espresso **in gradi**,

A) per il test di **Tukey**

- nel caso di **campioni bilanciati** di dimensioni **n** la formula è

$$es = \sqrt{\frac{820,7}{n + 0,5}}$$

- nel confronto tra due generici campioni A e B di dimensioni $\mathbf{n_A}$ e $\mathbf{n_B}$ è

$$es = \sqrt{\frac{410,35}{n_A + 0,5} + \frac{410,35}{n_B + 0,5}}$$

dove **820,7** è ricavato da $\mathbf{(180° / 2\pi)^2}$.

B) per il test di **Dunnett,** trattandosi di un test unilaterale

- nel caso di **campioni bilanciati** di dimensioni **n** la formula è

$$es = \sqrt{\frac{1641,4}{n + 0,5}}$$

- nel confronto tra due generici campioni A e B di dimensioni $\mathbf{n_A}$ e $\mathbf{n_B}$ è

$$es = \sqrt{\frac{820,7}{n_A + 0,5} + \frac{820,7}{n_B + 0,5}}$$

ESEMPIO 2 (i dati e lo sviluppo dell'esempio, applicati all'analisi di un ambientalista, sono tratti dal volume di J. **Zar** del 1999 (***Biostatistical Analysis***, 4th ed. Prentice Hall, New Jersey, a pag. 564). Rilevatori di inquinamento atmosferico per varie sostanze, collocati in quattro zone ritenute rappresentative della situazione generale di una città, hanno permesso di stimare quanto volte le misure effettuate hanno superato i livelli di attenzione

	Aree				
Risposte	A	B	C	D	TOTALE
Misure superiori ai livelli d'attenz.	**32**	**43**	**16**	**9**	100
Misure totali (n_i)	87	108	80	25	300
Proporzioni (p_i)	**0,368**	**0,398**	**0,200**	**0,360**	**0,333**

Valutare

- se tra le quattro aree esiste una differenza significativa nella proporzione delle misure che hanno superato il livelli di attenzione,
- tra quali aree esiste differenza significativa.

Risposta. Dapprima si deve stimare la significatività delle differenze tra le quattro proporzioni.
Con

$$\chi^2_{(3)} = \frac{(32 - 87 \cdot 0,333)^2}{87 \cdot 0,333 \cdot 0,667} + \frac{(43 - 108 \cdot 0,333)^2}{108 \cdot 0,333 \cdot 0,667} + \frac{(16 - 80 \cdot 0,333)^2}{80 \cdot 0,333 \cdot 0,667} + \frac{(9 - 25 \cdot 0,333)^2}{25 \cdot 0,333 \cdot 0,667}$$

$$\chi^2_{(3)} = \frac{(32 - 29,0)^2}{19,3} + \frac{(43 - 36,0)^2}{24,0} + \frac{(16 - 26,6)^2}{17,8} + \frac{(9 - 8,3)^2}{5,6} = 0,47 + 2,04 + 6,31 + 0,09 = 8,91$$

si ottiene un χ^2 = **8,91** con **3** gdl. Poiché i valori critici corrispondenti sono 7,815 per α = 0.05 e 9,348 per α = 0.025, al valore calcolato corrisponde una probabilità compresa tra 0.05 e 0.025.
Si rifiuta l'ipotesi nulla ed è possibile chiedersi tra quali proporzioni la differenza sia significativa.

A questo scopo,

1 - dapprima si ordinano le proporzioni in ordine crescente

Rango dei gruppi (**i**)	1	2	3	4
Campione	C	D	A	B
Dati $\mathbf{X_i}$ / $\mathbf{n_i}$	16 / 80	9 / 25	32 / 87	43 / 108
Proporzioni $\mathbf{p_i}$	0,200	0,360	0,368	0,398

2 - successivamente le proporzioni devono essere trasformate in gradi.

E' possibile utilizzare la formula che ricorre a **X** e **n**

$$p' = \frac{1}{2}\left(arcsin\sqrt{\frac{X}{n+1}} + arcsin\sqrt{\frac{X+1}{n+1}} \right)$$

ottenendo i risultati riportati nell'ultima riga della tabella successiva

Rango dei gruppi (**i**)	1	2	3	4
Campione	C	D	A	B
Proporzioni in gradi $\mathbf{'p_i}$	26,85	37,18	37,42	39,18

Ad esempio, applicata al campione C

$$'p_C = \frac{1}{2}\left(arcsin\sqrt{\frac{16}{80+1}} + arcsin\sqrt{\frac{16+1}{80+1}} \right)$$

$$'p_C = \frac{1}{2}\left(arcsin 0{,}444 + arcsin 0{,}458\right) = \frac{1}{2}\left(26{,}422 + 27{,}275\right) = 26{,}85$$

la proporzione **p = 0,200** trasformata in gradi diventa **'p = 26,85.**

3 - Per avere una visione generale delle analisi, è conveniente costruire un'altra tabella. In essa, si elencano tutti i **possibili confronti** (prima colonna della tabella successiva) in gradi, iniziando da quello che determina la differenza maggiore; per ogni confronto si riporta anche

- **la differenza** (in valore assoluto) **espressa in gradi** (seconda colonna),
- **l'errore standard della differenza espressa in gradi** (terza colonna),
- **il valore di Q calcolato** (quarta colonna),
- il **valore critico** di $Q_{\alpha/2,\, \nu,\, k}$ con la probabilità α bilaterale, i **gdl del test** (ν) e il **numero di gruppi** (k) (quinta colonna),
- la **significatività della differenza** (sesta colonna)

Confronto	Differenza	Err. St.	Q	$Q_{0.05,\, \infty, k}$	Conclusione
C vs. B (in ranghi 1 vs. 4)	39,18-26,85=12,33	2,98	4,137	3,633	Diff. Signif.
C vs. A (in ranghi 1 vs. 3)	37,42-26,85=10,57	3,13	3,378	3,633	Diff. Non Sig.
D vs. B (in ranghi 2 vs. 4)	39,18-37,18=2,00	4,46	0,448	3,633	Diff. Non Sig.
A vs. B (in ranghi 3 vs. 4)	Compresa in 2 vs. 4	---	---	---	Diff. Non Sig.
D vs. A (in ranghi 2 vs. 3)	Compresa in 2 vs. 4	---	---	---	Diff. Non Sig.
C vs. D (in ranghi 1 vs. 2)	Compresa in 1 vs. 3	---	---	---	Diff. Non Sig.

4 - Per i confronti che si effettuano, si stima l'errore standard con

$$es = \sqrt{\frac{410{,}35}{n_A + 0{,}5} + \frac{410{,}35}{n_B + 0{,}5}}$$

Ad esempio, nel caso del 1° confronto C vs. B

$$es = \sqrt{\frac{410{,}35}{80 + 0{,}5} + \frac{410{,}35}{108 + 0{,}5}} = \sqrt{5{,}098 + 3{,}782} = 2{,}98$$

si ottiene 2,98

5 - Con

$$Q = \frac{'p_A - 'p_B}{es}$$

si stima il valore di Q; nel caso del 1° confronto C vs. B

$$Q = \frac{39{,}18 - 26{,}85}{2{,}98} = \frac{12{,}33}{2{,}98} = 4{,}137$$

Q risulta uguale a 4,137.

6 - Il valore critico corrispondente della distribuzione Q alla probabilità $\alpha = 0.05$ bilaterale, con un numero di osservazioni molto alto (∞) e per 4 gruppi (fornito dalla tabella relativa), è uguale a 3,833. E' possibile rifiutare l'ipotesi nulla; la differenza tra la proporzioni minore e quella maggiore risulta significativo

7 - Si passa al confronto successivo, cioè tra due proporzioni che determinano una differenza minore

- la differenza tra la proporzione prima (C = 26,85) per rango rispetto alla terza (A = 37,42) determinano un valore di Q = 3,378: non risulta significativa;
- la differenza tra la proporzione seconda (D = 37,18) per rango rispetto alla quarta (B = 39,18) determinano un valore di Q = 0,448: non risulta significativa;

8 - Per il principio di cautela, tutte le differenze successive (minori di queste due, con i ranghi compresi nell'intervallo precedente) non potranno essere significative.
In conclusione il test tra le 4 proporzioni è risultato significativo. Tra tutti i possibili confronti singoli, solo quello tra la proporzione minore (C = 200) e quella maggiore (B = 0,398) è significativo.

15.8. IL TEST Q DI COCHRAN

Con un esperimento organizzato come i blocchi randomizzati già discussi nell'analisi della varianza a due criteri di classificazione, quando le **risposte sono di tipo binario** nella statistica non parametrica è possibile ricorrere al test **Q** di **W. G. Cochran** pubblicato nel 1950 (vedi ***The comparisons of percentages in matched samples***, su **Biometrika** vol. 37, pp. 256-266).
Il test serve per **verificare se il numero, la proporzione o la frequenza totale di successi o insuccessi di N prove ripetute** (da cui anche il nome di analisi della varianza a misure ripetute) **differiscono in modo significativo tra le varie situazioni a confronto**. In vari testi è presentato come **l'estensione a k campioni del test di McNemar**.

Il test **Q di Cochran** è utilizzato quando si devono valutare i risultati di sondaggi o giudizi qualitativi binari (sufficiente o insufficiente; accettabile o non-accettabile; migliorato o non migliorato) espressi dalle stesse persone su una serie di tre o più provvedimenti o situazioni.
Le valutazioni devono essere rappresentate mediante una votazione binaria, indicata solo mediante 0 e 1.

Per esempio, è possibile chiedere ad un gruppo di **N** persone

- se la situazione ambientale è ritenuta accettabile o no,
- se a loro parere nell'ultimo anno la qualità della vita sia migliorata oppure no,
- se la confezione di un prodotto industriale, oppure il colore o il sapore di un farmaco è gradito oppure no

confrontando simultaneamente K situazioni o prodotti

I risultati devono essere riportati in una tabella a due entrate come quella sottostante

INDIVIDUI	**QUARTIERI**			
	A	**B**	**---**	**K**
1	0	0	---	0
2	0	1	---	1
3	1	1	---	1
---	---	---	---	---
N	1	0	---	1

in cui, di norma,

- sulle righe sono indicati i valori attribuiti da ogni individuo e
- nelle colonne sono riportate le varie situazioni (in questo esempio i quartieri o leconfrzioni di prodotto).

Il test serve per verificare se le valutazioni assegnate alle varie colonne sono simili o statisticamente differenti: se le percentuali o proporzioni di 1 (oppure di 0) calcolate per colonna sono uguali (**H_0**) oppure significativamente differenti (**H_1**).

Con r righe e c colonne, si calcola il valore **Q** mediante

la formula di **Cochran**

$$Q = \frac{(k-1)\cdot(k\cdot\sum C_j^2 - (\sum C_j)^2)}{k\cdot\sum R_i - \sum R_i^2}$$

dove,

- **k** è il numero di colonne,

- $\mathbf{C_j}$ è il numero totale di successi nella colonna j-esima,
- $\mathbf{R_i}$ è il numero totale di successi nelle riga i-esima.

Quando il numero di righe non è troppo piccolo, **Q** segue la **distribuzione** χ^2 **con gdl k-1.**

Per l'uso dei valori critici riportati nella tabella del chi quadrato, **il campione è ritenuto di dimensioni accettabili quando il numero totale di osservazioni (N individui per k situazioni) è complessivamente uguale o maggiore di 24 e contemporaneamente il numero di righe (N) non è inferiore a 4**.

Per esperimenti di dimensioni minori, non presentati in questa trattazione, possono essere ottenute stime esatte dei valori critici mediante le permutazioni.

E' tuttavia diffusa la prassi di **utilizzare comunque la distribuzione chi quadrato**, poiché i calcoli con le permutazioni richiedono molto tempo e in conclusione determinano valori approssimativamente simili a quelli ottenuti con il chi quadrato.

Come potrà essere messo in evidenza con l'esempio, le righe composte solo da valori identici, siano essi tutti 0 oppure tutti 1 (come le risposte dell'individuo 1 e dell'individuo 3 nella tabella precedente) non influiscono sul valore dell'indice **Q**. Esso risente solo delle valutazioni differenti, che variano il numero di risposte positive tra colonne.

Di conseguenza, **per calcolare le dimensioni reali di una tabella si devono considerare solo le righe con valori che non siano tutti 1 oppure tutti 0**.

Anche in questo caso, **la metodologia è spiegata in modo semplice seguendo la soluzione di un problema**.

ESEMPIO. In una città con gravi problemi di traffico è stata profondamente modificata la circolazione del trasporto privato, cambiando sensi unici e divieti, sia di accesso**,** sia di sosta. Anche se la situazione in complesso appare migliorata, gli abitanti di alcuni quartieri non si ritengono soddisfatti, ritenendo che nella loro zona la nuova situazione non presenti gli stessi vantaggi evidenti in altri, riguardo all'inquinamento acustico e a quello dell'aria, al controllo del traffico e alla qualità della vita in genere.

A 15 esperti di vari settori della qualità della vita, è stato chiesto un giudizio sulla situazione di 4 quartieri (I, II, III, IV).

Essi devono rispondere se, rispetto all'anno precedente, in ogni quartiere la situazione è migliorata (1) oppure no (0) (cioè come prima o peggiorata).

.

La tabella riporta le 15 risposte, indicando con 1 quelle positive e con 0 quelle non positive (di parità oppure negative, poiché la risposta può essere solo binaria).

	Quartieri			
Esperti	**I**	**II**	**III**	**IV**
1	0	1	1	1
2	0	1	1	0
3	1	1	1	1
4	1	1	0	0
5	0	0	0	0
6	0	1	1	1
7	1	0	0	0
8	1	1	0	0
9	0	0	0	1
10	0	0	0	0
11	1	0	0	0
12	1	1	1	0
13	1	1	0	0
14	1	1	1	0
15	0	1	1	1
$\mathbf{C_j}$	**8**	**10**	**7**	**5**

Si può sostenere che i 4 quartieri hanno avuto variazioni differenti, per cui in essi la qualità della vita è migliorata in modo significativamente diverso?

Risposta. Per l'applicazione della formula di Cochran, occorre dapprima calcolare

- il totale di ogni colonna ($\mathbf{C_j}$),
- il totale di ogni riga ($\mathbf{R_i}$),
- e quello dei suoi quadrati ($\mathbf{R_i^2}$),

	Quartieri					
Esperti	I	II	III	IV	R_i	R_i^2
1	0	1	1	1	**3**	**9**
2	0	1	1	0	**2**	**4**
3	1	1	1	1	**4**	**16**
4	1	1	0	0	**2**	**4**
5	0	0	0	0	**0**	**0**
6	0	1	1	1	**3**	**9**
7	1	0	0	0	**1**	**1**
8	1	1	0	0	**2**	**4**
9	0	0	0	1	**1**	**1**
10	0	0	0	0	**0**	**0**
11	1	0	0	0	**1**	**1**
12	1	1	1	0	**3**	**9**
13	1	1	0	0	**2**	**4**
14	1	1	1	0	**3**	**9**
15	0	1	1	1	**3**	**9**
C_j	**8**	**10**	**7**	**5**	**30**	**80**

che risultano uguali a

$$\sum C_j = 30 \qquad \sum R_i = 30 \qquad \sum R_i^2 = 80$$

Applicando ai dati la formula di Cochran

$$Q = \frac{(4-1)\cdot(4\cdot(8^2+10^2+7^2+5^2)-30^2)}{4\cdot 30-80} = \frac{3\cdot(952-900)}{40} = \frac{156}{40} = 3{,}9$$

si ottiene un valore di **Q** uguale a 3,9.

La sua significatività deve essere verificata con i valori critici della distribuzione χ^2 con 3 gdl, che alla probabilità $\alpha = 0.05$ fornisce il valore 7,82.

Il valore calcolato (3,9) è inferiore non solo a quello critico della probabilità $\alpha = 0.05$ ma anche a quello della probabilità $\alpha = 0.10$.

Si deve concludere che non è dimostrato che nei 4 quartieri la situazione sia cambiata in modo significativamente diverso; anzi, l'alto valore di **P** (**P > 0.10**) permette di concludere che è probabilmente l'ipotesi nulla è vera (non solo non è rifiutata) e quindi che la qualità della vita nei 4 quartieri è migliorata in modo relativamente uniforme.

Quando il numero di colonne o risposte è uguale a 2, le N risposte in un test di Cochran, riportate nelle righe, possono essere riassunte in una tabella 2 x 2 e quindi analizzate come test di Mc Nemar, sempre attraverso la distribuzione chi quadrato.

15.9. ESTENSIONE DEL TEST DI McNEMAR O TEST DI BOWKER

Quando, **con due campioni dipendenti le risposte non sono binarie ma hanno tre o più modalità**, si costruisce una tabella quadrata N x N per analizzare la simmetria ai lati della diagonale. E' il test proposto da A. H. **Bowker** nel 1948 (con l'articolo ***A test for simmetry in contingency tables*** pubblicato su **J. Amer. Statist. Assoc.**, 43, pp. 572-4) per verificare, come permette il test di Mc Nemar con risposte binarie, l'ipotesi che nel confronto tra due situazioni sia avvenuto un cambiamento significativo nella frequenza delle varie modalità. Tra i testi di statistica non parametrica a maggior diffusione, questo metodo è presentato da **P. Sprent** nel volume ***Applied nonparametric statistical methods***, (second Edition, Chapman & Hall, London, 1993, pp. 239-240).

Nella ricerca ambientale, si ricorre al test di Bowker nel caso in cui si confrontano le risposte multiple (non binarie) delle stesse persone prima e dopo uno stimolo o i differenti effetti di due sostanze tossiche, somministrate allo stesso gruppo di cavie. E' un approccio diverso dal Q di Cochran, poiché in questo caso le risposte sono date non una volta sola ma due volte (prima e dopo l'evento) e esse sono di tipo ordinale, non binario.

Si supponga che un'amministrazione comunale abbia deciso la chiusura del centro storico al traffico automobilistico.
Prima della delibera è stato fatto un sondaggio nominativo su un campione di individui residenti nella zona storica, deducendo da una serie di risposte se erano favorevoli, incerti o contrari.
Dopo due mesi di applicazione del divieto, il test è stato ripetuto sugli stessi individui.

La sottostante tabella 3 x 3 riporta il numero di individui per ognuna delle 9 combinazioni di risposte.

		DOPO		
		Favorevoli	Incerti	Contrari
PRIMA	Favorevoli	85	**16**	**35**
	Incerti	**28**	34	**25**
	Contrari	**54**	**39**	74

La diagonale fornisce il numero di coloro che hanno mantenuto lo stesso parere, mentre le altre caselle riportano il numero di coloro che lo hanno cambiato in una direzione oppure nell'altra. Si vuole sapere se esistono differenze statisticamente significative tra prima e dopo.
Una indicazione preliminare, solo descrittiva, può essere fornita dai totali marginali

		DOPO			
		Favorevoli	Incerti	Contrari	TOTALI
PRIMA	Favorevoli	85	**16 (a)**	**35 (b)**	136
	Incerti	**28 (a)**	34	**25 (c)**	87
	Contrari	**54 (b)**	**39 (c)**	74	167
	TOTALI	167	89	134	390

dai quali risulta che i favorevoli sono aumentati da 136 (prima) a 167 (dopo) e quelli incerti sono aumentati da 87 (prima) a 89 (dopo) mentre i contrari sono diminuiti da 167 (prima) a 134 (dopo).
Ma, trattandosi di due campioni dipendenti, come per il test di Mc Nemar **la risposta per l'inferenza è fornita da quelli che hanno cambiato parere** (in grassetto nella tabella).
Il test presentato da **Bowker confronta le frequenze distribuite che sono in modo simmetrico rispetto alla diagonale**: se H_O è vera, dovrebbero essere approssimativamente uguali.

I concetti ed il metodo sono semplici e possono essere descritti in 4 passaggi logici.

1 - Se non è intervenuto nessun mutamento sistematico nelle opinioni espresse dagli intervistati (H_0 vera),

a) coloro che **prima** erano **Incerti** e **dopo** sono diventati **Favorevoli** (28) dovrebbero essere equivalenti a quelli che **prima** erano **Favorevoli** e **dopo Incerti** (16);
b) coloro che **prima** erano **Contrari** e **dopo Favorevoli** (54) dovrebbero essere equivalenti a quelli che **prima** erano **Favorevoli** e **dopo** sono **Contrari** (35);
c) coloro che **prima** erano **Contrari** e **dopo Incerti** (39) dovrebbero essere equivalenti a quelli che **prima** erano **Incerti** e **dopo Contrari** (25).

In modo più formale, l'ipotesi nulla è

$$\mathbf{H_0: \pi ij = \pi ji}$$

contro l'ipotesi alternativa

$$\mathbf{H_1: \pi ij \neq \pi ji}$$

con esclusione dei valori della diagonale.

2 – Il confronto è effettuato mediante un χ^2 che considera la differenza tra la frequenza assoluta di una casella non collocata sulla diagonale e quella ad essa simmetrica,
con una formula

$$\chi^2 = \sum \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

dove

- **la sommatoria è estesa a tutti gli *i* da *1* a *n-1* ma solo con *j > i*,**

Con i dati dell'esempio, si ottiene

$$\chi^2 = \frac{(28-16)^2}{28+16} + \frac{(54-35)^2}{54+35} + \frac{(39-25)^2}{39+25} = \frac{144}{44} + \frac{361}{89} + \frac{225}{64} = 3{,}27 + 4{,}06 + 3{,}52 = 10{,}85$$

un valore uguale a 10,85

3 – Sotto l'ipotesi nulla di simmetria dei dati, la distribuzione asintotica (teoricamente, per un numero **n** di dati tendente all'infinito; in pratica per **n** sufficientemente grande) è un chi quadrato con gdl

$$\text{gdl} = \frac{n \cdot (n-1)}{2}$$

Con i dati dell'esempio, i gdl

$$\text{gdl} = \frac{3 \cdot (3-1)}{2} = 3$$

sono 3, corrispondenti alle 3 somme fatte o 3 coppie di valori a confronto (indicate con a, b, c, nella tabella), derivanti dalla combinazione C_3^2

4 - Poiché il valore del chi quadrato con 3 gdl

- alla probabilità $\alpha = 0.05$ è uguale a 7,815,
- alla probabilità $\alpha = 0.01$ è uguale a 11,345,

si rifiuta l'ipotesi nulla, con una probabilità **P** inferiore a $\alpha = 0.05$.

ESEMPIO. Per valutare l'impatto di una discarica sulla popolazione residente in un comune, prendendo un campione nominativo è stato fatta un'indagine sulla percezione di odori sgradevoli, fornendo 4 livelli di risposta: mai, di rado, spesso, sempre.
Dopo un anno, l'indagine è stata ripetuta sugli stessi individui con i seguenti risultati complessivi:

		DOPO			
		Mai	Di rado	Spesso	Sempre
PRIMA	Mai	25	**86**	**97**	**25**
	Di rado	**65**	78	**54**	**33**
	Spesso	**58**	**47**	68	**28**
	Sempre	**12**	**39**	**17**	14

Si può sostenere che l'intervento ha modificato la situazione in modo statisticamente significativo?

Risposta. Indicando la posizione dei dati (esclusa la diagonale) con **Xij** come nella tabella

		DOPO			
		Mai	Di rado	Spesso	Sempre
PRIMA	Mai	25	**86** $X_{1,2}$	**97** $X_{1,3}$	**25** $X_{1,4}$
	Di rado	**65** $X_{2,1}$	78	**54** $X_{2,3}$	**33** $X_{2,4}$
	Spesso	**58** $X_{3,1}$	**47** $X_{3,2}$	68	**28** $X_{3,4}$
	Sempre	**12** $X_{4,1}$	**39** $X_{4,2}$	**17** $X_{4,3}$	14

Il metodo si fonda su quattro passaggi logici:

1- per verificare l'ipotesi nulla

$$\mathbf{H_0: \pi ij = \pi ji}$$

contro l'ipotesi alternativa

$$\mathbf{H_1: \pi ij \neq \pi ji}$$

per tutti gli *i* da *1* a *n-1* con *j > i*,

il confronto è tra i valori a destra della diagonale con quelli a sinistra, che occupano la posizione simmetrica (esempio $X_{1,2}$ contro $X_{2,1}$ e $X_{2,4}$ contro $X_{4,2}$…);

2 - con la formula

$$\chi^2 = \sum \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

in cui la sommatoria è estesa a tutti gli ***i*** da ***1*** a ***n-1*** con ***j > i***

si calcola un valore del chi quadrato

$$\chi^2 = \frac{(86-65)^2}{86+65} + \frac{(97-58)^2}{97+58} + \frac{(25-12)^2}{25+12} + \frac{(54-47)^2}{54+47} + \frac{(33-39)^2}{33+39} + \frac{(28-17)^2}{28+17} =$$

$$= \frac{441}{151} + \frac{1521}{155} + \frac{169}{37} + \frac{49}{101} + \frac{36}{72} + \frac{121}{45} = 2{,}92 + 9{,}81 + 4{,}57 + 0{,}49 + 0{,}50 + 2{,}69 = 20{,}98$$

che risulta uguale a 20,98.

3 – Con il rapporto

$$\text{gdl} = \frac{n \cdot (n-1)}{2}$$

applicato alla matrice 4 x 4

$$\text{gdl} = \frac{4 \cdot (4-1)}{2} = 6$$

si stimano 6 gdl, che corrispondono al numero di coppie di dati a confronto, cioè la combinazione

$$C_4^2 = \frac{4!}{(4-2)!2!} = \frac{1 \cdot 2 \cdot 3 \cdot 4}{1 \cdot 2 \cdot 1 \cdot 2} = 6$$

4 – Infine, poiché il valore critico del chi quadrato con 6 gdl

- alla probabilità α **= 0.05** è uguale a 12,592 e
- alla probabilità α **= 0.01** è uguale a 16,815

si rifiuta l'ipotesi nulla, con una probabilità P inferiore a 0.01.

15.10. TEST DI FRIEDMAN O ANALISI DELLA VARIANZA PER RANGHI A 2 CRITERI DI CLASSIFICAZIONE, CON UNA E CON K REPLICHE

Il test di Cochran si applica a dati dicotomizzati. Quando è possibile disporre di misure più precise su una **scala quantitativa continua**, in altri termini quando i valori raccolti sono

- **almeno di tipo ordinale**,
- senza o solo con **pochi valori identici** e
- disposti come nell'**ANOVA a due criteri** o a blocchi randomizzati,

per verificare l'ipotesi nulla sulla tendenza centrale

$$\mathbf{H_0:\ me_1 = me_2 = ... = me_K}$$

contro l'ipotesi alternativa

$$\mathbf{H_1}\text{: } \textbf{non tutte le k mediane sono uguali}.$$

si ricorre al test proposto da Milton **Friedman** nel 1937 (con l'articolo ***The use of ranks to avoid the assumptions of normality implicit in the analysis of variance*** pubblicato su **Journal of the American Statistical Association** Vol. 32, pp. 675-701). Come dice il titolo del suo articolo, l'uso dei ranghi da lui proposto permette di evitare le assunzioni di normalità, implicite nell'ANOVA.

E' l'alternativa non parametrica alll'**ANOVA a due criteri di classificazione o a blocchi randomizzati**, quando non sono rispettate le condizioni di validità richieste dai test parametrici. **E' uno dei test non parametrici più potenti e di uso più generale** . Come tale è riportato in quasi tutte le librerie statistiche informatiche, anche quelle che limitano la presentazione di quelli non parametrici a pochi test.

Per una presentazione chiara del problema, è utile che i dati siano riportati in una tabella, nella quale

- i valori delle varie righe fanno riferimento agli stessi soggetti e
- le colonne ai medesimi casi di studio,

come nell'analisi della varianza a due criteri di classificazione di cui è la versione non parametrica.
Ma, a differenza dell'ANOVA parametrica a due criteri,

- **l'ipotesi nulla è una sola e verte sulle k situazioni o casi o trattamenti,**

- **mentre i soggetti o individui sono considerati soltanto come repliche.**

<u>NEL CASO DI UN SOLO DATO PER CASELLA</u>, è utile presentare i dati raccolti come nella tabella

SOGGETTI O INDIVIDUI	**SITUAZIONI O CASI**			
	A	**B**	**---**	**K**
1	23	150	---	8
2	10	12	---	5
---	---	---	---	---
N	12	13	---	9
Mediane	$\mathbf{Me_A}$	$\mathbf{Me_B}$		$\mathbf{Me_K}$

Se l'ipotesi da verificare riguarda non il fattore riportato nella colonna ma quello nella riga, è sufficiente scambiare le righe con le colonne. Come risulterà evidente dalla illustrazione della metodologia, **la verifica dell'ipotesi sulla tendenza centrale dei due fattori presi in considerazione non possono essere condotte in modo simultaneo**, come è fatto nell'analisi della varianza. Quando interessano entrambi i fattori, la verifica dell'uguaglianza delle mediane per ognuno di essi deve essere condotta in tempi successivi, in modo indipendente.

Più sinteticamente, si ricorre al test di Friedman fondamentalmente in 2 situazioni:

1- quando sono state utilizzate misure continue in scale d'intervallo o di rapporto, di conseguenza con nessuna o pochissime misure ripetute, ma non sono rispettate le assunzioni di validità dell'ANOVA;

2- quando sono state usate misure discrete, semi-quantitative o di rango anche se rappresentate in forma simbolica.

<u>Il procedimento</u> è semplice e richiede pochi passaggi.

1 - Data una tabella a doppia entrata

SOGGETTI O INDIVIDUI	SITUAZIONI O CASI		
	A	**B**	**C**
1	23	50	18
2	12	10	5
3	23	28	19
4	12	13	9

2 - **trasformare i punteggi o le misure in ranghi entro la stessa riga**, assegnando **1** al punteggio minore e progressivamente valori maggiori fino a **k**, uguale al numero di colonne, al punteggio maggiore della medesima riga.

3 - Successivamente, **sommare per colonna** i valori dei ranghi (**T_i**)

SOGGETTI O INDIVIDUI	SITUAZIONI O CASI		
	A	**B**	**C**
1	2	3	1
2	3	2	1
3	2	3	1
4	2	3	1
Totali (T_i)	**9**	**11**	**4**

4 - Se **l'ipotesi nulla H_0 è vera**, nelle colonne a confronto i ranghi minori e quelli maggiori dovrebbero essere distribuiti casualmente;

pertanto, le somme dei ranghi nelle **k** colonne (**T_i** osservati) dovrebbero essere tra loro tutte equivalenti ed essere uguali ad un valore atteso, che dipende solo dal numero di osservazioni,

$$\mathbf{T_i}\ (\text{attesi}) = \frac{N \cdot (k+1)}{2}$$

dove **N** è il numero di righe.

Con i dati dell'esempio, **N** = 4 e **k** = 3

$$\mathbf{T_i}\ (\text{attesi}) = \frac{4 \cdot (3+1)}{2} = \mathbf{8}$$

la somma di ogni colonna (**T_i** attesi) dovrebbe essere uguale o almeno prossima a 8, considerando le variazioni casuali.

Se **l'ipotesi nulla H_0 è falsa** , in almeno una colonna si concentrano i ranghi minori o maggiori; di conseguenza, tale somma tende ad essere significativamente differente dal valore $\mathbf{T_i}$ atteso.

5 - Per decidere se queste somme ($\mathbf{T_i}$ osservati) sono significativamente differenti dell'atteso, si calcola la statistica **Fr**

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

che (come nella formula) è la sommatoria dei quadrati degli scarti tra i **k** totali osservati e i corrispondenti attesi.

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

- a **0** nel caso di accordo tra totali osservati e totali attesi ($\mathbf{H_0}$ vera),
- a un valore alto al crescere dello scarto tra essi ($\mathbf{H_0}$ falsa).

Con i dati dell'esempio, si ottiene un valore di **Fr**

$$Fr = (9 - 8)^2 + (11 - 8)^2 + (4 - 8)^2 = 1^2 + 3^2 + 4^2 = 1 + 9 + 16 = 26$$

uguale a 26.

Nel caso di **<u>piccoli campioni</u>**

(numero totale di dati N < 30-35 osservazioni che possono essere determinati, tra le situazioni più ricorrenti, da $k = 3$ e $N \leq 15$ oppure da $k = 4$ e $N \leq 9$)

la significatività della statistica **Fr** è fornita da tabelle specifiche (riportate nella pagina seguente).

Per **k = 3** e **N = 4, i valori critici** sono

- 24 alla probabilità $\alpha = 0.10$
- 26 alla probabilità $\alpha = 0.05$
- 32 alla probabilità $\alpha = 0.01$

Poiché il valore calcolato è stato 26, che coincide con la probabilità $\alpha = 0.05$, si può rifiutare l'ipotesi nulla.

VALORI CRITICI DI Fr DEL TEST DI FRIEDMAN
PER L'ANALISI DELLA VARIANZA NON PARAMETRICA A DUE CRITERI
IN PICCOLI CAMPIONI

k	N	0.10	0.05	0.01	0.001
3	**3**	18	18	---	---
	4	24	26	32	---
	5	26	32	42	50
	6	32	42	54	72
	7	38	50	62	86
	8	42	50	72	98
	9	50	56	78	114
	10	50	62	96	122
	11	54	72	104	146
	12	62	74	114	150
	13	62	78	122	168
	14	72	86	126	186
	15	74	96	134	194
4	**2**	20	20	---	---
	3	33	37	45	---
	4	42	52	64	74
	5	53	65	83	105
	6	64	76	102	128
	7	75	91	121	161
	8	84	102	138	184

Per **<u>campioni di dimensioni medie (k tra 3 e 5 con N tra 5 e 10 in modo che k·N sua tra 25 e 40)</u>** è proposto l'uso della distribuzione **F** di **Fisher**, con gdl uguali a **k-1** e **(k-1)·(N-1)**.

Si ricava il valore **Fr** attraverso

$$Fr = \frac{(N-1)\cdot\left[\dfrac{\sum_{i=1}^{k} T_i^2}{N} - \dfrac{N\cdot k\cdot(k+1)^2}{4}\right]}{\sum(r_{ij}^2) - \dfrac{\sum_{i=1}^{k} T_i^2}{N}}$$

dove

- $\mathbf{T_i}$ = Somma dei ranghi per colonna
- **N** = Numero di righe e **k** = numero di colonne o gruppi

Si rifiuta l'ipotesi nulla quando

- il valore di **Fr** calcolato è maggiore del valore critico di **F** di Fisher
- con gdl **(k-1)** e **(k-1)·(N-1)**.

ESEMPIO. Durante una settimana, in 5 zone di una città sono state calcolate le **quantità medie** delle polveri Pm 10, ottenendo la seguente serie di dati

Giorni	Zona				
	A	B	C	D	E
1	115	142	36	91	28
2	28	31	7	21	6
3	220	311	108	51	117
4	82	56	24	46	33
5	256	298	124	46	84
6	294	322	176	54	86
7	98	87	55	84	25

Verificare se esiste una differenza significativa nella presenza di polveri tra le 5 zone.

Risposta. La semplice lettura dei dati evidenzia, la presenza

- di grande variabilità tra giorni e tra zone,
- della non normalità della distribuzione,

- dell'uso di **valori medi**, al posto delle singole osservazioni (potrebbero anche essere le mediane , il terzo quartile o un indice qualsiasi di concentrazione).

Sono tutte condizioni che impongono il ricorso al test non parametrico di Friedman.

Quindi

- si trasformano i dati in ranghi, entro la stessa riga,
- e si calcolano i totali per colonna (**T_i**)

Giorni	Ranghi (r_{ij}) delle Zone per Giorno				
	A	B	C	D	E
1	4	5	2	3	1
2	4	5	2	3	1
3	4	5	2	1	3
4	5	4	2	3	1
5	5	4	1	3	2
6	4	5	3	1	2
7	5	4	2	3	1
T_i	31	32	14	17	11

Da questa tabella di ranghi, per usare la formula

$$Fr = \frac{(N-1)\cdot\left[\dfrac{\sum_{i=1}^{k} T_i^2}{N} - \dfrac{N\cdot k\cdot(k+1)^2}{4}\right]}{\sum(r_{ij}^2) - \dfrac{\sum_{i=1}^{k} T_i^2}{N}}$$

si devono prima ricavare

- $\sum_{i=1}^{k} T_i^2 = 31^2 + 32^2 + 14^2 + 17^2 + 11^2 = 961 + 1024 + 196 + 289 + 121 = 2591$
- $\sum\left(r_{ij}^2\right) = 4^2 + 5^2 + 2^2 + 3^2 + \ldots + 2^2 + 3^2 + 1^2 = 16 + 25 + 4 + 9 + \ldots + 4 + 9 + 1 = 385$
- N = 7 e k = 5

Da essi si ottiene

$$Fr = \frac{(7-1)\cdot\left[\frac{2591}{7} - \frac{7\cdot 5\cdot(5+1)^2}{4}\right]}{385 - \frac{2591}{7}} = \frac{6\cdot(370{,}143 - 315)}{385 - 370{,}143} = \frac{330{,}858}{14{,}857} = 22{,}270$$

un valore di **Fr = 22,27** con df **4** e **24** (ricavati da **(5 – 1)** e da (5-1)·(7-1)).

Poiché il valore critico di **F** con df 4 e 24

- alla probabilità $\alpha = 0.001$ è uguale a 7,39

si rifiuta l'ipotesi nulla di uguaglianza delle mediane dei valori giornalieri riportati, con probabilità di commettere un errore di primo tipo $P < 0.001$.

Per **<u>campioni di grandi dimensioni (k ≥ 5 e N abbastanza grande in modo che k·N > 40)</u>**, è stato proposto un indice χ^2_F che si distribuisce approssimativamente come il $\chi^2_{(k-1)}$ **con gdl k-1.**

Può essere stimato mediante la formula

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

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.

Per **semplificare i calcoli, sono state proposte <u>formule abbreviate</u>**. Una che ricorre con frequenza 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.

I risultati possono essere confrontati con **i valori critici di seguito riportati, tratti dalla distribuzione chi quadrato con df k-1,** o comunque presi direttamente da una distribuzione chi quadrato.

VALORI CRITICI DEL χ^2_F				
K	**α = 0.10**	**α = 0.05**	**α = 0.01**	**α = 0.001**
5	7,78	9,49	13,28	18,46
6	9,24	11,07	15,09	20,52
7	10,64	12,59	16,81	22,46
8	12,02	14,07	18,48	24,32
9	13,36	15,51	20,09	26,12
10	14,68	16,92	21,67	27,88

NEL CASO DI k DATI PER CASELLA, (ovviamente il campione diventa abbastanza grande e si può ricorrere alla distribuzione χ^2) come le misure raccolte

- in **k** stazioni o situazioni (riportate nei trattamenti),
- in **N** giorni o su **N** individui (riportati nei blocchi),
- con **r** repliche, uguali per ogni stazione o individuo (riportate all'incrocio tra trattamenti e blocchi) e quindi con **esperimenti bilanciati** come nella tabella sottostante,

per verificare la differenza nelle tendenze centrali d'inquinamento tra le **k** stazioni,
con ipotesi nulla

$$\mathbf{H_0: me_1 = me_2 = ... = me_K}$$

contro l'ipotesi alternativa

$$\mathbf{H_1: \text{non tutte le k mediane sono uguali}}$$

GIORNI	STAZIONI O LOCALITA'			
	A	**B**	**---**	**K**
1	23 28	150 123	--- ---	8 15
2	10 15	12 14	--- ---	5 4
---	--- ---	--- ---	--- ---	--- ---
N	12 10	13 15	--- ---	9 6
Mediane	$\mathbf{Me_A}$	$\mathbf{Me_B}$		$\mathbf{Me_K}$

si deve

1 - trasformare i dati in ranghi in modo indipendente per ogni blocco, considerando entro essi contemporaneamente i **kr** dati,

2 - calcolare i totali $\mathbf{T_i}$ di ogni colonna o trattamento degli **Nr** ranghi,

3 - stimare il valore χ_F^2

$$\chi_F^2 = \frac{12}{N \cdot k \cdot r^2 \cdot (k \cdot r + 1)} \cdot \sum \left[T_i - \frac{N \cdot r \cdot (k \cdot r + 1)}{2} \right]^2$$

4 - la cui significatività è data dalla distribuzione $\chi^2_{(k-1)}$.

Nel caso di un solo dato o replica per casella (quindi **r = 1**), la formula precedente coincide con

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

L'**<u>efficienza asintotica relativa</u>** del test **Friedman,** rispetto al test **F** di **Fisher-Snedecor** di cui è il corrispondente non parametrico,

- quando la distribuzione dei dati è una Normale, ha un valore uguale a 0,95k/(k+1) = (3/π)·k(k+1);
- quando la distribuzione dei dati è una Rettangolare, ha un valore uguale a 1k/(k+1);
- quando la distribuzione è Esponenziale Doppia, ha un valore uguale a 1,5k·(k+1) = (3/2) ·k·(k+1).

Quando la distribuzione dei dati è normale, il test non parametrico ha una potenza leggermente inferiore a quella dell'ANOVA, che è tanto più sensibile quanto maggiore è il numero di gruppi; ma quando ci si allontana dalla normalità mentre ha una potenza superiore. Resta il vantaggio, già ripetuto per i test non parametrici, che le conclusioni di questo test non sono confutabili come quelle del test F di Fisher, quando non sono rispettate le condizioni di normalità dei dati.

ESEMPIO 1 (**PICCOLI CAMPIONI E 1 DATO PER CASELLA**). In un depuratore sono stati posti 4 filtri (A, B, C, D) lungo il percorso di distribuzione dell'acqua; si vuole verificare se esiste differenza nella quantità di sostanze eliminate, pesata dopo una settimana di attività. La qualità dell'acqua e quindi la quantità del materiale trattenuto è molto variabile, derivando a volte da pozzi e altre da bacini di deposito:

	Filtri			
Prove	**A**	**B**	**C**	**D**
I	80	75	70	76
II	74	69	70	71
III	9	7	5	6
IV	7	8	7	6
V	8	6	5	6

Si vuole verificare se esistono differenze significative nella quantità di materiale in sospensione che viene trattenuto dai 4 filtri

Risposta. I dati raccolti sono estremamente variabili tra prove e non sono distribuiti in modo normale. Le quantità misurate (non importa il tipo di scala) sono quindi trasformate in ranghi entro ogni riga come nella tabella sottostante

	A	B	C	D
I	4	2	1	3
II	4	1	2	3
III	4	3	1	2
IV	2,5	4	2,5	1
V	4	2,5	1	2,5
Totali Ranghi T_i	**18,5**	**12,5**	**7,5**	**11,5**

Successivamente si calcolano i totali dei ranghi per colonna (**T_i**).
Applicando la formula per piccoli campioni

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

in cui

$$\left(\frac{N(k+1)}{2}\right) = \left(\frac{5(4+1)}{2}\right) = \frac{25}{2} = 12{,}5$$

si ottiene un valore di Fr

$$\text{Fr} = (18{,}5 - 12{,}5)^2 + (12{,}5-12{,}5)^2 + (7{,}5 - 12{,}5)^2 + (11{,}5 - 12{,}5)^2 = 36 + 0 + 25 + 1 = 62$$

uguale a 62.

Per **k = 4 e N = 5**
alla probabilità $\alpha = 0.05$ il valore critico riportato nella tabella precedente è 65.

Poiché il valore calcolato è 62, la probabilità **P** è maggiore di quella critica, fissata a 0.05. Quindi non è possibile rifiutare l'ipotesi nulla: non è dimostrata una differenza significative tra le mediane della quantità di materiale trattenuto dai 4 filtri.

ESEMPIO 2 (**GRANDI CAMPIONI E 1 DATO PER CASELLA**). Prodotto dalle combustioni contenenti zolfo, il biossido di zolfo SO_2 è uno dei fattori di inquinamento dell'aria più dannosi, a breve raggio. Secondo la legislazione italiana, il limite di accettabilità è fissato in 80 gm^{-3} (equivalenti a 0,03 ppm) come mediana delle concentrazioni medie giornaliere di un anno.

	Zone					
Giorno	**A**	**B**	**C**	**D**	**E**	**F**
1	55	75	32	60	36	48
2	57	70	69	65	41	52
3	60	24	43	68	98	54
4	52	85	51	58	33	64
5	57	69	72	56	28	53
6	59	45	107	64	38	55
7	58	73	33	35	66	54
8	58	67	75	71	35	56
9	57	48	51	67	55	36
10	59	79	48	81	52	63
11	88	70	53	64	43	65
12	61	40	42	71	81	55
13	57	45	38	65	39	53
14	59	76	40	67	38	51
15	55	73	57	69	42	56

Un'esposizione continua alla concentrazione di 0,03-0,05 ppm determina un peggioramento delle condizioni dei pazienti bronchitici; una esposizione di solo 20 secondi alla concentrazione 0,3-1 ppm può portare all'alterazione dell'attività cerebrale; più di 6 ore di esposizione ad oltre 20 ppm possono causare la saturazione delle vie e dei tessuti polmonari, con eventuale paralisi e/o morte.

Durante il periodo invernale, con una rilevazione continua, per 15 giorni sono state misurate le medie giornaliere presso 6 aree sia industriali che residenziali di una città.
Si vuole verificare se la tendenza centrale delle emissioni del periodo è significativamente diversa tra le 6 zone di monitoraggio.

Risposta. I dati raccolti formano un campione di grandi dimensioni e le misure riportate utilizzano una scala d'intervallo o di rapporto. Tuttavia, la distribuzione è fortemente asimmetrica, in alcuni gruppi: anche senza una misura oggettiva di valutazione delle simmetrie e di confronto della varianza, è possibile osservare che l'intervallo di variazione di alcuni gruppi è molto è differente.
Nel caso del problema, soprattutto **è l'uso di valori medi che vieta il ricorso all'ANOVA** parametrica.

E' vantaggioso ed appropriato utilizzare il test di Friedman. A tal fine,

- si trasformano i dati in ranghi entro la stessa riga;
- successivamente, si calcolano le somme per colonna, come nella tabella sottostante

Ricordando che **N = 15** e **k = 6** ,

dai dati della tabella successiva, in cui i valori sono stati trasformati in ranghi entro la stessa riga, si stima il valore χ_F^2

$$\chi_F^2 = \frac{12 \cdot (58^2 + 64^2 + 44^2 + 69^2 + 35^2 + 45^2)}{15 \cdot 6 \cdot 7} - 3 \cdot 15 \cdot 7$$

$$\chi_F^2 = \frac{12 \cdot (3364 + 4096 + 1936 + 4761 + 1255 + 2025)}{630} - 315$$

$$\chi_F^2 = \frac{12 \cdot 17407}{630} - 315 = \frac{208884}{630} - 315 = 331{,}57 - 315 = 16{,}57$$

che risulta uguale a 16,57.

	Zone					
Giorno	A	B	C	D	E	F
1	4	6	1	5	2	3
2	3	6	5	4	1	2
3	4	1	2	5	6	3
4	3	6	2	4	1	5
5	4	5	6	3	1	2
6	4	2	6	5	1	3
7	4	6	1	2	5	3
8	3	4	6	5	1	2
9	5	2	3	6	4	1
10	3	5	1	6	2	4
11	6	5	2	3	1	4
12	4	1	2	5	6	3
13	5	3	1	6	2	4
14	4	6	2	5	1	3
15	2	6	4	5	1	3
T_i	**58**	**64**	**44**	**69**	**35**	**45**

Il valore critico del χ^2 **con df 5 (k = 6)**

- alla probabilità α = 0.05 è uguale a 11,07
- mentre alla probabilità α = 0.01 è uguale a 15,09
- e alla probabilità α = 0.001 è uguale a 20,52.

Di conseguenza, con probabilità **P < 0.01** di commettere un errore di I° tipo, si rifiuta l'ipotesi nulla e si accetta l'ipotesi alternativa: le mediane dei 6 gruppi a confronto non sono tra loro tutte statisticamente uguali.

ESEMPIO 3 (**CON k REPLICHE O MISURE RIPETUTE PER CASELLA**). Per valutare la significatività delle differenze nei livelli d'inquinamento tra 4 stazioni collocate lungo un corso d'acqua, data la grande variabilità stagionale nelle portate d'acqua, i campioni sono stati raccolti alle stesse date, effettuando 2 prelievi in 5 giorni diversi

GIORNI	**STAZIONI O LOCALITA'**			
	A	**B**	**C**	**D**
I	20 28	33 31	34 39	40 41
II	55 58	61 54	69 63	73 70
III	18 22	24 19	15 23	28 26
IV	14 18	17 13	21 23	19 24
V	37 33	41 38	53 54	48 51

Esistono differenze significative nei livelli d'inquinamento delle 4 zone o stazioni?
Risposta.

GIORNI	STAZIONI O LOCALITA'			
	A	B	C	D
I	**1** **2**	**4** **3**	**5** **6**	**7** **8**
II	**2** **3**	**4** **1**	**6** **5**	**8** **7**
III	**2** **4**	**6** **3**	**1** **5**	**7** **8**
IV	**2** **4**	**3** **1**	**6** **7**	**5** **8**
V	**2** **1**	**4** **3**	**7** **8**	**5** **6**
Totali Ti	**23**	**32**	**56**	**69**

Dopo

- **aver trasformato i valori nei ranghi relativi entro lo stesso blocco**, con una impostazione grafica leggermente diversa da quella della tabella precedente, per meglio evidenziare che si tratta di repliche nelle stessa stazione e alla stessa data,

- e aver calcolato i totali dei ranghi per colonna,

il cui totale ΣT_i

$$(23 + 32 + 56 + 69) = 180$$

è uguale a 180 e che, come verifica di non aver commesso errori nella trasformazione in ranghi e nelle somme successive, **deve essere uguale** a

$$\mathbf{\Sigma T_i} = \frac{N \cdot k \cdot r \cdot (1 + k \cdot r)}{2}$$

(con i dati dell'esempio

$$\mathbf{\Sigma T_i} = \frac{5 \cdot 4 \cdot 2 \cdot (1 + 4 \cdot 2)}{2} = \frac{40 \cdot 9}{2} = \mathbf{180}$$

è infatti uguale a 180)

- con la formula generale

$$\chi_F^2 = \frac{12}{N \cdot k \cdot r^2 \cdot (kr + 1)} \cdot \sum \left[T_i - \frac{Nr(kr+1)}{2} \right]^2$$

dai dati dell'esempio, dopo aver stimato che

il totale atteso di ogni colonna è

$$\frac{N \cdot r \cdot (k \cdot r + 1)}{2} = \frac{5 \cdot 2 \cdot (4 \cdot 2 + 1)}{2} = \mathbf{45}$$

uguale a 45

- si calcola il valore di χ_F^2

$$\chi_F^2 = \frac{1}{5 \cdot 4 \cdot 2^2 \cdot (4 \cdot 2 + 1)} \cdot \left[(23-45)^2 + (32-45)^2 + (56-45)^2 + (69-45)^2 \right] =$$

$$= \frac{1}{80 \cdot 9} \cdot (484 + 169 + 121 + 576) = \frac{1}{720} \cdot 1350 = \mathbf{1{,}875}$$

che risulta uguale a 1,875 con 3 gdl.

Poiché alla probabilità $\alpha = 0.05$ il valore critico del chi quadrato con 3 gdl è uguale a 7,815 non è possibile rifiutare l'ipotesi nulla.

Correzione per valori identici o ranghi ripetuti (ties).

Con dati discreti e scale semiquantitative od ordinali, i valori identici possono essere numerosi. La varianza della distribuzione campionaria diventa minore; si rende necessario apportare una correzione al valore di χ_F^2, che ne aumenta il valore. La correzione non ha effetti quando non esistono ripetizioni; di conseguenza, alcuni testi la inseriscono nella formula generale (che ovviamente risulterà differente da quella precedentemente riportata), anche se in questo modo il calcolo manuale di χ_F^2 è più lungo e più frequentemente può determinare errori.

Si ottiene il **valore corretto di** χ_F^2 ponendo al denominatore non più **N·k·(k + 1)** ma tutta la formula seguente

$$N \cdot k \cdot (k+1) + \frac{N \cdot k - \sum_{i=1}^{N} \sum_{j=1}^{p} r_{ij}^3}{k-1}$$

dove:

- **N** è il numero di righe o osservazioni per gruppo,
- **k** è il numero di colonne o gruppi,
- **p** è il numero di dati con lo stesso valore, e quindi con lo stesso rango, nella medesima riga;
- $\mathbf{r_{ij}}$ è la dimensione dei ranghi ripetuti.

Nel calcolo della correzione, con $\mathbf{r_{ij} = 1}$ vengono inclusi anche i dati con valori che compaiono una sola volta nella stessa riga. Pertanto essi contribuiscono per un valore $r_{ij}^3 = \mathbf{1}$ ($1^3 = 1$), mentre

- i valori che compaiono 2 volte contribuiscono per un valore $r_{ij}^3 = 8$ ($2^3 = 8$),
- quelli che compaiono 3 volte per un valore $r_{ij}^3 = 27$ ($3^3 = 27$), ecc....

ESEMPIO 4 (**CON CORREZIONE PER TIES**). Al fine di ottenere un quadro possibilmente completo della qualità dell'aria, in 6 zone di una città (centro storico (A), 3 aree periferiche (B, C, D) e 2 zone industriali (E, F)) sono state collocate stazioni di rilevamento per misurare con continuità tutti i parametri indicati dalla legge.

Sono stati misurati gli inquinanti: anidride solforosa, ossidi di azoto, ossido di carbonio, fluoro, piombo, polveri, ozono, idrocarburi. Per ognuno di essi è stato valutato il grado di rispetto delle norme di legge. Inoltre sono stati rilevati alcuni parametri meteorologici che influiscono sulla qualità dell'aria: velocità e direzione del vento, temperatura, umidità, irraggiamento solare.

Per una più facile e diffusa comprensione della situazione da parte della popolazione e degli amministratori, i tecnici hanno divulgato i dati con cadenza settimanale ed hanno elaborato un criterio che definisce il "giudizio sintetico" di qualità dell'aria in 5 classi:

buona (++), discreta (+), mediocre (=), scadente (-), pessima (--).

I dati riportati in tabella sono i risultati di 10 settimane.

	ZONA					
Settimane	A	B	C	D	E	F
1	++	+	+	-	=	-
2	+	=	+	-	=	=
3	--	+	+	++	-	-
4	-	=	=	+	+	-
5	=	+	+	-	=	--
6	+	=	+	-	-	-
7	=	=	--	+	-	++
8	+	+	++	-	=	-
9	+	-	--	+	=	=
10	++	+	++	-	=	-

Si vuole verificare se tra le 6 zone esiste una differenza significativa nei giudizi mediani della qualità dell'aria.

Risposta. L'analisi statistica richiede preliminarmente la trasformazione dei giudizi sintetici in una scala di rango. E' un campione di grandi dimensioni, con misure ordinali o semiquantitative.
Un modo razionale di trasformazione dei simboli in ranghi potrebbe essere l'assegnazione del valore 1 alla zona che ha la migliore qualità dell'aria e progressivamente un valore maggiore al crescere dell'inquinamento, fino ad assegnare 6 alla zona la cui qualità dell'aria è stata giudicata la peggiore. Nulla cambia per l'analisi statistica se si agisse nel modo opposto, assegnando valori da 6 a 1.
Nel caso di valutazioni identiche entro la stessa riga, si deve assegnare lo stesso valore stimato come media aritmetica dei ranghi occupati.

Per facilitare il calcolo del**la correzione per misure ripetute i ties,** è utile riportare anche la somma dei valori r_{ij}^3

	ZONA						
Settimane	A	B	C	D	E	F	$\sum_j^p r_{ij}^3$
1	1	2,5	2,5	5,5	4	5,5	1 + 8 + 1 + 8 = 18
2	1,5	4	1,5	6	4	4	8 + 27 + 1 = 36
3	6	2,5	2,5	1	4,5	4,5	1 + 8 + 1 + 8 = 18
4	5,5	3,5	3,5	1,5	1,5	5,5	8 + 8 + 8 = 24
5	3,5	1,5	1,5	5	3,5	6	8 + 8 + 1 + 1 = 18
6	1,5	3	1,5	5	5	5	8 + 1 + 27 = 36
7	3,5	3,5	6	2	5	1	8 + 1 + 1 + 1 + 1 = 12
8	2,5	2,5	1	5,5	4	5,5	1 + 8 + 1 + 8 = 18
9	1,5	5	6	1,5	3,5	3,5	8 +8 + 1 + 1 + = 18
10	1,5	3	1,5	5,5	4	5,5	8 + 1 + 1 + 8 = 18
T_i	**29,0**	**31,0**	**27,5**	**38,5**	**39,0**	**46,0**	$\sum_{i=1}^{N} \sum_j^p r_{ij}^3 = 216$

La somma dei valori per i ranghi ripetuti ($\sum_{i=1}^{N} \sum_{j=1}^{p} r_{ij}^3$) è uguale a 216.

Nella formula generale

$$\chi^2_{\mathrm{F}} = \frac{12}{Nk(k+1)} \sum_{i=1}^{k} \left(T_i - \frac{N(k+1)}{2} \right)^2$$

al denominatore al posto di Nk(k+1)

si sostituisce la formula con i ties

$$N \cdot k \cdot (k+1) + \frac{N \cdot k - \sum_{i=1}^{N} \sum_{j=1}^{p} r_{ij}^3}{k-1}$$

dove:

- **N** è il numero di righe o osservazioni per gruppo,
- **k** è il numero di colonne o gruppi,
- **p** è il numero di dati con lo stesso valore, e quindi con lo stesso rango, nella medesima riga,
- $\mathbf{r_{ij}}$ è la dimensione dei ranghi ripetuti.

Con i dati dell'esempio, il valore χ^2_F corretto per i ties è

$$\chi_F^2 = \frac{12 \cdot (29^2 + 31^2 + 27{,}5^2 + 38{,}5^2 + 39^2 + 46^2)}{10 \cdot 6 \cdot 7 + (\frac{10 \cdot 6 - 216}{6 - 1})} - 3 \cdot 10 \cdot 7$$

e svolgendo i calcoli

$$\chi_F^2 = \frac{12 \cdot (841 + 961 + 756{,}25 + 1482{,}25 + 1521 + 1516)}{420 + (\frac{-156}{5})} - 210$$

si ottiene

$$\chi_F^2 = \frac{12 \cdot 7675{,}5}{420 - 31{,}2} - 210 = \frac{92130}{388{,}8} - 210 = 236{,}96 - 210 = 26{,}96$$

un risultato uguale a 26,96.

Il valore di χ_F^2 calcolato (26,96) è nettamente superiore al valore critico riportato nella tabella χ^2 per 5 df, anche alla probabilità α = 0.001 (20,52): si rifiuta l'ipotesi nulla, con probabilità P < 0.001.

Questo ultimo esempio dimostra come l'applicazione di **questo test possa essere esteso, nel pieno rispetto della sua validità, a indici complessi** determinati dalla somma o dalla media di una serie di parametri. A maggior ragione, **questo test può essere esteso a indici sintetici semplici come medie e mediane oppure quantili specifici**, come possono essere il primo e il terzo quartile. Ovviamente è richiesta una serie almeno una decina distribuzioni, raggruppate in **k** classi.

15.11. I CONFRONTI MULTIPLI TRA MEDIE DI RANGHI NELL'ANALISI DELLA VARIANZA NON PARAMETRICA, A DUE CRITERI DI CLASSIFICAZIONE

Anche nell'analisi della varianza non parametrica a due criteri di classificazione, **dopo aver rifiutato l'ipotesi nulla i confronti multipli permettono di individuare quali sono i trattamenti che risultano tra loro differenti**, alla probabilità α prefissata.

In altri termini, per le mediane di due generici gruppi A e B è possibile verificare l'ipotesi nulla

$$H_0: \; me_A = me_B$$

contro l'ipotesi alternativa **bilaterale**

$$H_1: \; me_A \neq me_B$$

Sulla base dei concetti del Bonferroni già illustrati nelle pagine dedicate ai confronti multipli a posteriori, quando si devono effettuare più confronti mantenendo costante la probabilità complessiva

(α_T) o experiment-wise, la singola probabilità (α) di ogni differenza o comparison-wise deve diminuire in rapporto al numero di confronti che si vogliono effettuare.

Con **k** gruppi, i confronti possibili, tra loro **non indipendenti**,
sono

$$\frac{k \cdot (k-1)}{2}$$

Se la probabilità scelta come experiment-wise complessivamente è $\boldsymbol{\alpha_T}$ = 0.05, per ogni singolo confronto la probabilità α diventa

$$\alpha = \frac{0.05}{\frac{k \cdot (k-1)}{2}}$$

Poiché si considerano le differenze in valore assoluto, **questi confronti sono bilaterali.** Di conseguenza, la probabilità α per ogni confronto deve essere dimezzata, diventando

$$\alpha = \frac{0.05}{k \cdot (k-1)}$$

Come ampiamente illustrato nel capitolo relativo ai confronti multipli, la stima della probabilità α comparison-wise ha avuto tante soluzioni differenti, che non portano sempre a risultati coincidenti, seppure spesso simili.

In questo paragrafo sono riportate le proposte divulgate da Sidney **Siegel** e N. John **Castellan** jr. Nel loro volume del 1988 ***Nonparametric Statistics for the Behavioral Sciences*** (edito da McGraw-Hill, Inc.) tradotto in italiano nel 1992 con ***Statistica non parametrica*** (2° ed., McGraw-Hill Libri Italia, Milano), essi propongono l'uso della distribuzione normale poiché, con i ranghi, sono sufficienti poche decine di osservazioni per ottenerla.
Quando il **numero di dati e di gruppi è sufficientemente grande**, si può ricorrere alla distribuzione normale applicata alla differenza

- tra le **<u>somme</u> dei ranghi** di due gruppi,
- tra le **<u>medie</u> dei ranghi** di due gruppi.

Sono significative **le differenze D tra le <u>somme dei ranghi</u>** del generico gruppo A (R_A) e del generico gruppo B (R_B)

$$D = |R_A - R_B|$$

quando

$$D \geq Z_{\frac{\alpha}{k \cdot (k-1)}} \cdot \sqrt{\frac{N \cdot k \cdot (k+1)}{6}}$$

dove:

- α è la probabilità complessiva prefissata,
- **k** è il numero di gruppi, tra i quali sono possibili $k \cdot (k-1)/2$ confronti,
- **N** è il numero di righe od osservazioni per ogni campione.

Sono significative **le differenze D tra le <u>medie dei ranghi</u>** del generico gruppo A ($\bar{r}_A$) e del generico gruppo B ($\bar{r}_B$)

$$D = |\bar{r}_A - \bar{r}_B|$$

quando

$$D \geq Z_{\frac{\alpha}{k \cdot (k-1)}} \cdot \sqrt{\frac{k \cdot (k+1)}{6N}}$$

All'aumentare del numero di confronti, il valore di α per ognuno di essi diventa sempre minore. Ne deriva che non è sempre facile trovare nella distribuzione normale l'esatto valore di **Z,** quando il numero di confronti è alto: con α molto piccolo, la stima diventa approssimata. In vari testi sono quindi proposte tabelle con stime precise, sulla base del valore della probabilità experiment-wise (α_T) e del numero di confronti (p) che si vogliono effettuare.

Nella tabella sono riporti solamente i valori di **Z** per α_T relativamente grandi.

VALORI DI **Z** PER CONFRONTI MULTIPLI

IN FUNZIONE DI α_T E DEL NUMERO **P** DI CONFRONTI

P		α_T	
	Test Bilaterale	0.10	0.5
	Test Unilaterale	0.05	0.025
1		1,645	1,960
2		1,960	2,241
3		2,128	2,394
4		2,241	2,498
5		2,326	2,576
6		2,394	2,638
7		2,450	2,690
8		2,498	2,724
9		2,539	2,773
10		2,576	2,807
11		2,608	2,838
12		2,638	2,886
15		2,713	2,935
21		2,823	3,038
28		2,913	3,125

Con 4 gruppi tutti i possibili confronti sono 6

Con 5 gruppi tutti i possibili confronti sono 10

Con 6 gruppi tutti i possibili confronti sono 15

Con 7 gruppi tutti i possibili confronti sono 21

Con 8 gruppi tutti i possibili confronti sono 28

ESEMPIO 1. In un esempio precedente, si è dimostrato che esiste una differenza significativa tra le mediane delle emissioni gassose giornaliere di 6 zone di una città, rilevate per 15 giorni.
Nella tabella sottostante, sono state riportate le medie dei ranghi:

	Emissioni giornaliere di 6 zone					
	A	B	C	D	E	F
Medie dei ranghi	**3,86**	**4,26**	**2,93**	**4,60**	**2,33**	**3,00**

Con confronti multipli a posteriori, si vuole verificare tra quali zone esista una differenza significativa, alla probabilità $\alpha_T = 0.05$.

Risposta. Tra le medie dei ranghi delle 6 zone è possibile calcolare 15 differenze ($\frac{6 \cdot 5}{2}$). Per una visone complessiva, esse sono utilmente riportate (in valore assoluto) in una matrice triangolare

		A	B	C	D	E	F
		3,86	4,26	2,93	4,60	2,33	3,00
B	4,26	**0,40**					
C	2,93	**0,93**	**1,33**				
D	4,60	**0,74**	**0,34**	**1,67**			
E	2,33	**1,53**	**1,93**	**0,60**	**2,27**		
F	3,00	**0,86**	**1,26**	**0,07**	**1,60**	**0,67**	

Per una probabilità complessiva prefissata $\alpha_T = 0.05$ in un test a due code, con 15 medie a confronto simultaneo la probabilità α per ogni media è **0.05/(15x2)** uguale a **0.00167**.
Alla probabilità $\alpha = 0.00167$ nella distribuzione normale corrisponde un valore di **Z** uguale a 2,935 (approssimativamente la metà fra Z = 2,93 della probabilità P = 0.0017 e Z = 2,94 della probabilità P = 0.0016).
La tabella dei valori Z riportata in precedenza semplifica la modalità per ottenere questa stima,

- fornendo 2,935 per α_T = 0.05 e P = 15.

Applicando la formula

$$D \geq Z_{\frac{\alpha}{k\cdot(k-1)}} \cdot \sqrt{\frac{k\cdot(k+1)}{6N}}$$

con **N = 15** e **k = 6**,

la differenza minima significativa **D** tra le due medie dei ranghi

$$\mathrm{D} \geq 2,935 \cdot \sqrt{\frac{6\cdot 7}{6\cdot 15}} = 2,935 \cdot \sqrt{0,467} = 2,935 \cdot 0,683 = 2,01$$

è uguale a 2,01.

Confrontando il valore calcolato con le 15 differenze riportate nella tabella triangolare, risulta significativa solamente la differenza (uguale a 2,27) tra la media dei ranghi del gruppo D (uguale a 4,60) e quella del gruppo E (uguale a 2,33).

Quando entro i **k** gruppi esiste **un trattamento usato come controllo**, per cui i risultati degli altri **k-1** trattamenti vengono confrontati solamente con il controllo, la significatività di ognuna delle **k-1** differenze possono essere verificate con la stessa metodologia.

Di conseguenza, la probabilità α del confronto di ogni singolo trattamento con il controllo risulta più alta di quella stimata per i confronti precedenti:

- il numero di confronti diventa **k-1**, per cui il valore di α_T complessivo deve essere diviso per **k-1**;
- inoltre, poiché sono confronti unilaterali, non è richiesto di dimezzare ulteriormente la probabilità;
- infine, occorre utilizzare la distribuzione Q di Dunnett, qui riportata, che tiene in considerazione anche gli effetti della **non indipendenza dei vari confronti**.

VALORI DEL **Q DI DUNNETT**

PER CONFRONTI MULTIPLI TRA **P TRATTAMENTI E UN CONTROLLO**

IN FUNZIONE DI α_T E DEL NUMERO **N** DI CONFRONTI

	Test unilaterale		Test bilaterale	
P	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
1	1,65	2,33	1,96	2,58
2	1,92	2,56	2,21	2,79
3	2,06	2,69	2,35	2,92
4	2,16	2,77	2,44	3,00
5	2,24	2,84	2,51	3,06
6	2,29	2,89	2,57	3,11
7	2,34	2,94	2,61	3,15
8	2,38	2,97	2,65	3,19
9	2,42	3,00	2,69	3,22
10	2,45	3,03	2,72	3,25
11	2,48	3,06	2,74	3,27
12	2,50	3,08	2,77	3,29
15	2,57	3,14	2,83	3,35
20	2,64	3,21	2,91	3,42

E' possibile valutare la significatività della differenza tra il controllo (C) e un generico trattamento (T) utilizzando la differenza

- tra le **<u>somme</u> dei ranghi**,
- tra le **<u>medie</u> dei ranghi**.

Sono significative **le differenze D tra le <u>somme dei ranghi</u>** del gruppo Controllo (R_C) e di un generico gruppo Trattamento (R_T)

$$D = \left| R_C - R_T \right|$$

quando

$$D \geq Q(\alpha_T, p) \cdot \sqrt{\frac{N \cdot k \cdot (k+1)}{6}}$$

Sono significative **le differenze D tra le <u>medie dei ranghi</u>** del generico gruppo A ($\bar{r}_A$) e del generico gruppo B ($\bar{r}_B$)

$$D = \left| \bar{r}_C - \bar{r}_T \right|$$

quando

$$D \geq Q(\alpha_T, p) \cdot \sqrt{\frac{k \cdot (k+1)}{6N}}$$

ESEMPIO 2. Riprendendo i dati dell'esempio precedente (N = 15)

	Emissioni giornaliere di 6 zone					
	A	B	C	D	E	P
Medie dei ranghi	**3,86**	**4,26**	**2,93**	**4,60**	**3,00**	**2,33**

si assuma che sia noto da tempo che la zona P, periferica e residenziale, caratterizzata da un traffico significativamente minore, abbia i livelli d'inquinamento più bassi della città e sia stata assunta come modello per una politica ambientale.

Verificare quali altre zone hanno un livello significativamente maggiore di essa, alla probabilità complessiva **α = 0.05** e alla probabilità **α = 0.01**.

Risposta. Le differenze tra la medie dei ranghi della zona di controllo **P** (con inquinamento minimo) e quelle delle altre 5 zone (**A, B, C, D, E**) sono riportate nella tabella

	A	B	C	D	E
Medie dei ranghi	3,86	4,26	2,93	4,60	3,00
Differenze da E (2,33)	**1,53***	**1,93****	**0,60**	**2,27****	**0,67**

Utilizzando la formula per le medie e ricavando dalla tabella che, in un **test unilaterale di 4 confronti** con **N = 15** e **k = 6,** poiché

- per α **= 0.05** il valore di **Q** è **2,16**
- per α **= 0.01** il valore di **Q** è **2,77**

il valore **D** per α **= 0.05**

è

$$D \geq 2{,}16 \cdot \sqrt{\frac{6 \cdot 7}{6 \cdot 15}} = 2{,}16 \cdot \sqrt{0{,}467} = 2{,}16 \cdot 0{,}683 = 1{,}475$$

uguale a 1,48 (arrotondato in eccesso, come richiesto dal principio di cautela);

il valore **D** per α **= 0.01**

è

$$D \geq 2{,}77 \cdot \sqrt{\frac{6 \cdot 7}{6 \cdot 15}} = 2{,}77 \cdot \sqrt{0{,}467} = 2{,}77 \cdot 0{,}683 = 1{,}892$$

uguale a 1,90 (arrotondato in eccesso, come richiesto dal principio di cautela).

In conclusione, alla probabilità complessiva

- α_T **= 0.05** rispetto all'inquinamento della zona di riferimento P (2,33) risulta maggiore quello rilevato nella zona **A** (3,86);
- per α_T **= 0.01** sono significativamente maggiori quello rilevato nella zona **B** (4,26) e quello della zona **D** (4,60).

<u>Questi test che utilizzano i ranghi sollevano obiezioni non banali sulla loro validità</u>. Come affermato nel 1966 da R. G. **Miller** jr. nel suo testo ***Simultaneous Statistical Inference*** (edito da **McGraw-Hill**, New York) e nel 1969 da K. R. **Gabriel** con l'articolo ***Simultaneous test procedures, some theory of multiple comparison*** (pubblicato **Annals of the Mathematical Statistics** Vol. 40,

pp.224-250), **le medie dei ranghi di due gruppi dipendono dai valori presenti in altri gruppi**, esclusi dal confronto specifico.

A maggior ragione, l'obiezione è valida per il test di **Kruskall-Wallis**.

15.12. TEST DI QUADE

La proposta di Friedman del 1937 è la metodologia che si è affermata nell'uso dei ricercatori prima e nei programmi informatici dopo. Ora è riportata in quasi tutti i testi e i programmi di statistica, dai più semplici ai più diffusi. In letteratura esistono test equivalenti, proposti negli anni immediatamente successivi, quali:

- **Wallis** W. A., 1939, ***The correlation ratio for ranked data***, pubblicato su **Journal of the American Statistical Association** Vol. 34, pp. 533-538.
- **Kendall** M. G., **Babington Smith** B., 1939, ***The problem of m-rankings***, pubblicato da **Annals of Mathematical Statistics** Vol. 10, pp. 275-287.
- **Brown** G. W. and **Mood** A. M., 1951, ***On Median Test for Linear Hypotheses,*** in **Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability**, Berkeley, University of California Press, pp. 159-166.

Esistono anche proposte più recenti. Tra esse possono essere ricordate soprattutto quelle che analizzano l'interazione o permettono di valutare l'effetto dei blocchi:

- **Lehmann** E. L., 1963, ***Asymptotically non parametric inference: an alternative approach to linear models***, su **Annals of Mathematical Statistics**, Vol. 34, pp. 1494-1506.
- **Mehra** K. L. **Sen** P. K., 1969, ***On a class of conditionally distribution-free tests for interaction in factorial experiments***, su **Annals of Mathematical Statistics**, Vol. 40, pp. 658-664.
- **Mehra** K. L. **Smith** G. E. J., 1970, ***On nonparametric estimation and testing for interactions in factorial experiments***, su **Journal of the American Statistical Association** Vol. 65, pp. 1283-1296.

Tra questi sembra emergere il test di Dana **Quade** del 1979 (vedi l'articolo ***Using weighted rankings in the analysis of complete blocks with additive block effects*** su **Journal of the American Statistical Association** Vol. 74, pp. 680-683), utile nel caso di piccoli campioni e quando le righe hanno una variabilità molto differente. Sviluppa idee già presenti in un articolo di J. W. **Tukey** del 1957, oltre 20 anni prima " ***Sums of Random Partitions of Ranks***" (su **Annals of Mathematical Statistics**, Vol. 28, pp. 987-992).

Nell'articolo di Dana **Quade** sono presi in considerazione

- le **tabelle con una sola osservazione** per trattamento x blocco,
- la **variabilità tra i blocchi**, in aggiunta a quella **entro i blocchi**.

Soprattutto nel caso di **pochi dati per trattamento**, quindi con pochi blocchi e complessivamente **pochi dati**, appare intuitivamente ragionevole assumere che **abbiano una credibilità maggiore i ranghi assegnati entro i blocchi con variabilità maggiore**. Infatti, per semplice ragionamento deduttivo, i blocchi con variabilità piccola,

- oltre ad avere con probabilità maggiore dei valori identici, che fanno perdere informazioni a tutto il sistema,
- hanno valori tra loro molto vicini che, per il solo effetto del caso, ne possono modificare l'ordine con probabilità maggiore.

Nell'artico citato, l'autrice scrive :"*Suppose the observations on different treatments are more distinct in some blocks than in the others; then it seems intuitively reasonable that the ordering of the treatments that these blocks suggest is more likely to reflect the underlying true ordering*".

La procedura

- utilizza la distribuzione **F** di **Fisher** con df **(k-1)** e **(k-1)·(N-1)**

e integra il metodo Friedman.

Per una spiegazione semplice della procedura e un confronto dei risultati, è stato impiegato lo stesso esempio sulle **quantità medie** delle polveri Pm 10, in 5 zone di una città:

Giorni	Zona				
	A	B	C	D	E
1	115	142	36	91	28
2	28	31	7	21	6
3	220	311	108	51	117
4	82	56	24	46	33
5	256	298	124	46	84
6	294	322	176	54	86
7	98	87	55	84	25

dove **k = 5** e **N = 7**.

Per verificare se esiste una differenza significativa tra le mediane dei valori riportati per le 5 zone, con ipotesi nulla

H_0: $me_A = me_B = me_C = me_D = me_E$ contro H_1: *le me delle 5 zone non sono tutte uguali*

dai dati dapprima si derivano

- sia la tabella dei **ranghi entro i blocchi**, con la metodologia di Friedman,

Giorni	Ranghi (r_{ij}) delle Zone entro Giorni				
	A	B	C	D	E
1	4	5	2	3	1
2	4	5	2	3	1
3	4	5	2	1	3
4	5	4	2	3	1
5	5	4	1	3	2
6	4	5	3	1	2
7	5	4	2	3	1

- sia **i ranghi dei blocchi** (righe), in funzione delle dimensioni del loro campo di variazione, assegnando **1 a quello minore** e **N a quello maggiore**:

Blocchi	Campo di variazione	Rango **Qj**
1	142 – 28 = 114	4
2	31 – 6 = 25	1
3	311 – 51 = 260	6
4	82 – 24 = 58	2
5	298 – 46 = 252	5
6	322 – 54 = 268	7
7	98 – 25 = 73	3

- Successivamente, mediante

$$T_{ij} = Q_j \cdot \left(r_{ij} - \frac{k+1}{2} \right)$$

dove

- Q_j è il rango del campo di variazione della riga (nel confronto delle N righe),
- r_{ij} è il rango di ogni valore entro riga (con k dati),

si modifica la tabella dei ranghi e si calcola la somma per colonna, ottenendo

Giorni	Zona				
	A	B	C	D	E
1	+4	+8	-4	0	-8
2	+1	+2	-1	0	-2
3	+6	+12	-6	-12	0
4	+4	+2	-2	0	-4
5	+10	+5	-10	0	-5
6	+7	+14	0	-14	-7
7	+6	+3	-3	0	-6
$\mathbf{T_i}$	+38	+46	-26	-26	-32

Ad esempio, in essa

- il rango di A1 (4 in quella di Friedman) diventa $4 \cdot \left(4 - \frac{5+1}{2}\right) = 4$
- il rango di A2 (5 in quella di Friedman) diventa $4 \cdot \left(5 - \frac{5+1}{2}\right) = 8$
- il rango di E7 (1 in quella di Friedman) diventa $3 \cdot \left(1 - \frac{5+1}{2}\right) = -6$

E' possibile verificare di non aver commesso errori in questa trasformazione dei ranghi di Friedman, in quanto

- la somma di ogni riga è uguale a zero
- la somma dei totali di colonna è uguale a zero.

Infine da questa ultima tabella mediante

$$W = \frac{(N-1) \cdot \left(\frac{\sum_{i=1}^{k} T_i^2}{N} \right)}{\sum (r_{ij}^2) - \frac{\sum_{i=1}^{k} T_i^2}{N}}$$

dove

- N = 7 e k = 5

- $\sum_{i=1}^{k} T_1^2$ = (38)^2 + (46)^2 + (-26)^2 +(-26)^2 + (-32)^2 = 1444 + 2116 + 676 + 676 + 1024 = 5936
- $\sum r_{ij}^2$ (dato dalla somma dei quadrati di tutti i ranghi modificati) = 1400 come nella tabella

Giorni	r_{ij}^2				
	A	B	C	D	E
1	16	64	16	0	64
2	1	4	1	0	4
3	36	144	36	144	0
4	16	4	4	0	16
5	100	25	100	0	25
6	49	196	0	196	49
7	36	9	9	0	36
$\sum r_{ij}^2 = 1400$					

si ottiene

$$W = \frac{(7-1)\cdot\frac{5936}{7}}{1400-\frac{5936}{7}} = \frac{5088}{552} = 9{,}217$$

un valore di **W = 9,217** che deve essere confrontato con

- il valore critico di F per gdl (5-1) e (5-1)·(7-1), cioè 4 e 24.

Poiché il valore critico di **F** con df 4 e 24

- alla probabilità α = 0.001 è uguale a 7,39

si rifiuta l'ipotesi nulla di uguaglianza delle mediane dei valori giornalieri riportati, con probabilità di commettere un errore di primo tipo P < 0.001.

Con il test di Friedman si era trovato un valore di **Fr = 22,27** ovviamente sempre con df **4** e **24**. Con il test di Quade si è ottenuto un valore inferiore, perché le righe con la variabilità maggiore (la 3 e la 6), quindi quelle che logicamente meglio permettono di definire i ranghi entro esse, hanno una distribuzione dei ranghi differente dalle altre 5.

In condizioni sperimentali differenti, può fornire un risultato più significativo del test di Friedman: è una opportunità in più per il ricercatore che vorrebbe dimostrare l'esistenza di differenze significative tra le mediane dei gruppi.

Nel suo articolo, Quade propone anche

- formule che utilizzano la distribuzione χ^2 e

- altre che utilizzano la distribuzione Z.

In un suo rapporto precedente del 1972, difficile da reperire nelle biblioteche e da lei citato, (***Analizing Randomized Blocks by Weighted Rankings***, Report SW 18/72 of the Mathematical Center, Amsterdam) **Dana Quade** ha anche fornito tabelle di valori critici per campioni di piccole dimensioni. Tra i testi italiani, il test di Quade è riportato da Mario **Castino** e Ezio **Roletto** in ***Statistica Applicata. Metodi di trattamento dei dati per studenti universitari, ricercatori e tecnici*** (Piccin, Padova, 1991, pp. 494). In esso è illustrato anche un metodo per i **confronti multipli a posteriori,** che utilizza il principio del **Bonferroni** e la distribuzione **t** di **Student** (qui è modificata solo la simbologia, per mantenerla uguale a quella utilizzata nel test).

Per **i confronti multipli del test di Friedman**, **Castino** e **Roletto** riportano che la differenza tra le **mediane** del generico **trattamento A** e del generico **trattamento B** sono significativamente differenti, quando

$$\left|T_A - T_B\right| > t_{(\alpha/2,\nu)} \cdot \sqrt{\frac{2N \cdot \left(\sum r_{ij}^2 - \frac{\sum T_i^2}{N}\right)}{(N-1)\cdot(k-1)}}$$

dove

- $\mathbf{T_A}$ e $\mathbf{T_B}$ sono i totali dei ranghi delle colonne (trattamenti) **A** e **B** a confronto,
- **t** è il di Student alla probabilità $\boldsymbol{\alpha/2}$ con gdl $\boldsymbol{\nu = (N-1)\cdot(k-1)}$.

Per **i confronti multipli del test di Quade**, riportano la stessa formula. Per i valori di $\mathbf{r_{ij}}$ e $\mathbf{T_i}$ devono essere utilizzati quelli calcolati nella tabella dei ranghi modificati secondo la metodologia di Quade.
Nella presentazione del metodo, anche in questo caso Castino e Roletto consigliano l'uso del t di Student senza tenere in considerazione i problemi di stima di α derivanti dal principio del Bonferroni, cioè senza valutare la relazione esistente tra comparison-wise e experiment-wise. Il loro metodo è quindi corretto solamente se a priori è stato programmato un solo confronto.

ESEMPIO. Poiché il risultato del test di Quade è stato significativo, confrontare tra loro le mediane delle 5 zone alla probabilità complessiva $\boldsymbol{\alpha_T = 0.05}$.

Risposta. La probabilità α di ogni confronto dipende dal numero di confronti che si vogliono effettuare. Se sono tutti quelli possibili a coppie tra le 5 mediane, essi sono 10; di conseguenza, per un probabilità totale o experiment-wise pari a $\boldsymbol{\alpha_T = 0.05}$ il valore di comparison-wise è $\boldsymbol{\alpha = 0.005}$.

La **differenza minima significativa** (metodo favorito dal fatto di avere necessariamente campioni con lo stesso numero di osservazioni) con

- N = 7 e k = 5
- $\sum_{i=1}^{k} T_1^2 = 5936$ e $\sum r_{ij}^2 = 1400$
- gdl = 24 e quindi t = 3,09 in una distribuzione bilaterale

è

$$|T_A - T_B| > 3{,}09 \cdot \sqrt{\frac{2 \cdot 7 \cdot \left(1400 - \frac{5936}{7}\right)}{(7-1)\cdot(5-1)}} = 3{,}09 \cdot \sqrt{332} = 55{,}45$$

uguale a 55,45

Dalla tabella delle somme dei ranghi di Quade effettuate per colonna (**T_i**)

Zone	A	B	C	D	E
T_i	+38	+46	-26	-26	-32

si ricava la matrice delle differenze in valore assoluto

		A	B	C	D	E
		+38	+46	-26	-26	-32
B	+46	**12**				
C	-26	**64***	**72***			
D	-26	**64***	**72***	**0**		
E	-32	**70***	**78***	**6**	**6**	

Alla probabilità complessiva **α_T = 0.05** sono significative tutte le differenze maggiori di 55,45 (segnate con asterisco).

Tale conclusioni devono essere trasferite alle mediane dei dati originali raccolti.

15.13. L'ESEMPIO DI KOCH: USO DI METODI NON PARAMETRICI, NELL'ANALISI STATISTISTICA DI UN ESPERIMENTO COMPLESSO CON K FATTORI

Tutti i test non parametrici presentati in questo capitolo e in quelli precedenti sono stati applicati a casi semplici. Anche nella verifica dell'uguaglianza tra k mediane o varianze, le analisi più sofisticate, è sempre stato preso in considerazione un solo fattore.
Per esperimenti più complessi, quali l'ANOVA a due o a più criteri con interazione, è prassi ricorrere alle tecniche parametriche. Tuttavia, è possibile giungere a soluzioni anche mediante l'uso congiunto di alcuni metodi parametrici già presentati, con il vantaggio rilevante di evitare le obiezioni collegate al tipo di scala e alla distribuzione dei dati.
Già nel 1970, Gary G. **Koch** nel riassunto del suo articolo ***The use of non-parametric methods in the statistical analysis of a complex split plot experiment*** (pubblicato su **Biometrics** Vol. 26, n. 1-4 pp. 105-128) scriveva: "*Da tempo gli statistici devono confrontarsi con dati che provengono da dati sperimentali complessi e che non soddisfano alle ipotesi tradizionali di validità dell'analisi della varianza, vale a dire alle ipotesi di indipendenza, d'omogeneità delle varianze e di normalità. Qualche volta lo statistico risolve queste situazioni aggiustando i dati con trasformazioni di variabili o arrivando a scartare certe osservazioni estreme. Benché queste procedure possano condurre spesso a analisi soddisfacenti, i risultati così ottenuti possono essere messi in discussione.*
L'oggetto di questo articolo è di mostrare che i metodi non parametrici sono in grado di fornire alternative realiste nell'analisi di questi dati. Le procedure presentate sono fondate sui ranghi. Benché siano valide in condizioni molto generali, esse hanno buone proprietà di potenza".

E' uno degli esempi didatticamente più completi, che ricorre all'uso congiunto di vari test e che può essere utilizzato come traccia per analisi simili in molti settori della statistica applicata. Questo articolo è riportato, con modifiche, da Giampiero **Landenna** e Donata **Marasini** nel loro volume del 1990 ***Metodi statistici non parametrici*** (Il Mulino, Bologna, pp. 511, a pp. 278-283). E' il testo italiano di statistica non parametrica più completo, valido anche in un confronto internazionale: ricchissimo di bibliografia, utilizza un linguaggio statisticamente sempre rigoroso, ma non semplice da comprendere, per "utenti" della statistica che non abbiano già acquisito conoscenze di buon livello.

L'esperimento considera tre fattori e le loro interazioni: la quantità di ferro assorbita dal fegato di topo albini, considerando

- l'effetto della temperatura (a due livelli),
- l'effetto della dieta (a due livelli),
- gli effetti del livello di acidità della soluzione (a tre livelli),
- con le loro interazioni di primo e di secondo livello.

		Etionina			Controllo		
Temp.	Coppia	Bassa	Media	Alta	Bassa	Media	Alta
37° C	1	2,23	2,59	4,50	1,34	1,40	3,87
	2	1,14	1,54	3,92	0,84	1,51	2,81
	3	2,63	3,68	10,33	0,68	2,49	8,42
	4	1,00	1,96	8,23	0,69	1,74	3,82
	5	1,35	2,94	2,07	2,08	1,59	2,42
	6	2,01	1,61	4,90	1,16	1,36	2,85
	7	1,64	1,23	6,84	0,96	3,00	4,15
	8	1,13	6,96	6,42	0,74	4,81	5,64
	Medie	***1,61***	***2,81***	***5,90***	***1,06***	***2,24***	***4,25***
25° C	9	3,00	6,77	2,95	1,56	4,71	2,64
	10	4,78	4,97	3,42	2,30	1,60	2,48
	11	0,71	1,46	6,85	1,01	0,67	3,66
	12	1,01	0,96	0,94	1,61	0,71	0,68
	13	1,70	5,59	3,72	1,06	5,21	3,20
	14	1,31	9,56	6,00	1,35	5,12	3,77
	15	2,13	1,08	3,13	1,40	0,95	3,94
	16	2,42	1,58	2,74	1,18	1,56	2,62
	17	1,32	8,09	3,91	1,55	1,68	2,40
	Medie	***2,04***	***4,45***	***3,74***	***1,45***	***2,47***	***2,82***
	Medie Totali	**1,85**	**3,68**	**4,76**	**1,27**	**2,36**	**3,48**

A questo scopo, sono stati presi 34 topi albini di sesso maschile, ognuno con un peso di circa 200 grammi.

Con questi 34 soggetti, sono stati costruiti **due campioni dipendenti, formando 17 coppie**. Un individuo per ogni coppia, scelto casualmente, ha avuto una dieta contenente etionina, l'altro una dieta di controllo.

Dopo 7 giorni, a questi 34 animali è stato estratto il fegato. Ogni fegato è stato suddiviso in tre parti e, per scelta casuale, posti in tre differenti soluzioni di ferro radioattivo con acidità bassa (pH 7,0 - 7,7), media (pH 4,5 – 5,5), alta (pH 2,0 – 3,0).

Infine sono stati fatti 2 blocchi: il primo, formato dalle tre parti del fegato di 8 coppie, scelte a caso, è stato posto in ambiente a 37° centigradi; il secondo blocco, formato dalle tre parti del fegato delle altre 9 coppie, è stato posto a 25° centigradi.

Conclusi i trattamenti, l'analisi chimica della quantità di ferro radioattivo assorbito dai 102 campioni ha dato i risultati riportati nella tabella precedente.

Nell'analisi statistica, che verrà descritta in modo dettagliato, l'autore dell'articolo considera gli individui delle 17 coppie come due **campioni dipendenti**. In realtà non lo sono, poiché non è stato seguito alcun criterio oggettivo. Infatti l'articolo riporta che **esse sono state formate per scelta puramente casuale**: "*The design was as follows: 34 male albino rats, each weighing approximately 200 grams,* ***were randomly divided into 17 pairs***".

Landenna e Marasini si accorgono di questo errore concettuale e rimediano aggiungendo, nella loro presentazione dell'esperimento, che ***le coppie sono state costruite in modo tale che la differenza di peso fra i due ratti fosse la minima possibile***.

E' un esperimento a struttura gerarchica, in cui

- le coppia sono nested nella temperatura,
- gli animali sono nested nelle coppie,
- le tre parti dello stesso fegato sono nested negli animali.

E' sufficiente la semplice lettura dei dati riportati nella tabella per rilevare la presenza di un numero non trascurabile di valori particolarmente alti, che allontanano la distribuzione dalla normalità facendole assumere forma bimodale. La loro eliminazione determina la perdita di informazioni, circa gli effetti dei fattori su alcuni animali.

L'assunzione di omogeneità della varianza non è corretta. Infatti le varianze (s^2) dei due blocchi, distinte per soluzione, sono

Varianza (s^2)						
Blocco	Etionina			Controllo		
Temperatura	Bassa	Media	Alta	Bassa	Media	Alta
37° C	0,35	3,46	6,81	0,22	1,42	3,88
25° C	1,56	10,89	3,10	0,15	3,80	1,00

Con il test più semplice e tra i meno potenti (il test di **Hartley**), il rapporto tra la varianza maggiore (10,89) e quella minore (0,15)

$$F_{(8,8)} = \frac{10,89}{0,15} = 72,6$$

è uguale a 72,6. Nella tabella dei valori critici di Hartley, il valore critico per k = 12 e df = 8 è

- 12,7 alla probabilità $\alpha = 0.05$
- 21 alla probabilità $\alpha = 0.01$.

Il test dimostra la non omogeneità delle varianze, con probabilità **P** nettamente inferiore a 0.01.

Sempre nell'articolo citato, **Koch** afferma che benché la eterogeneità della varianza possa essere eliminata con una trasformazione, certe inconsistenze potrebbero rimanere. Inoltre, nell'analisi dei risultati, molti ricercatori preferiscono utilizzare i dati non trasformati perché assumono significati più evidenti e semplici per l'interpretazione dei risultati.

Un altro aspetto che evidenza come manchino le condizioni di validità per applicare un test parametrico è la relazione significativa tra le serie dei valori entro i due blocchi, in particolare nelle combinazioni **temperatura x dieta**.

Correlazioni tra serie di valori							
BLOCCO		Etionina			Controllo		
37° C		Bassa	Media	Alta	Bassa	Media	Alta
Etionina	Bassa	---	-0,09	+0,32	+0,02	-0,20	+0,51
	Media		---	+0,18	-0,16	+0,79*	+0,52
	Alta			---	-0,78*	+0,37	+0,83*
Controllo	Bassa				---	-0,41	-0,56
	Media					---	+0,52
	Alta						---

Correlazioni tra serie di valori							
BLOCCO 25° C		Etionina			Controllo		
		Bassa	Media	Alta	Bassa	Media	Alta
Etionina	Bassa	---	+0,09	-0,29	+0,74*	+0,06	-0,05
	Media		---	+0,34	+0,18	+0,74*	+0,19
	Alta			---	-0,38	+0,20	+0,74*
Controllo	Bassa				---	-0,18	-0,42
	Media					---	+0,29
	Alta						---

La lettura delle due tabelle di correlazione parametrica evidenzia:

- le risposte delle differenti frazioni di fegato entro lo stesso animale tendono a non essere ugualmente correlate;
- le risposte degli animali entro la stessa coppia e associate allo stesso livello di acidità tendono ad essere più altamente correlate delle risposte con livelli di acidità differenti.

Per una interpretazione più completa, è utile ricordare che nella tabella per la significatività dei coefficienti di correlazione parametrica (vedi capitolo relativo) è riportato che

- con df = 6 (come per il Blocco 37° C) il valore critico alla probabilità $\alpha = 0.05$ è 0,7067 mentre alla probabilità $\alpha = 0.01$ è 0,8343
- con df = 7 (Blocco 25° C) il valore critico alla probabilità $\alpha = 0.05$ è 0,6664 mentre alla probabilità $\alpha = 0.01$ è 0,7977.

Nelle due tabelle precedenti, i valori di correlazione segnati con asterisco superano la soglia critica del 5%.

Nei capitoli dedicati all'ANOVA, è già stato illustrato come l'analisi statistica parametrica permetta di evidenziare l'effetto dei singoli fattori e delle loro interazioni, in accordo con il modello

$$X_{ijkp} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{oj} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ojk} + \varepsilon_{ijkp}$$

dove

- μ è la media generale,
- α_i è l'effetto della temperatura,
- β_j è l'effetto della dieta,
- γ_k è l'effetto del livello di acidità della soluzione,
- $\alpha\beta_{ij}$, $\alpha\gamma_{ik}$, $\beta\gamma_{jk}$ sono le rispettive interazione di primo livello,
- $\alpha\beta\gamma_{ijk}$ è l'interazione di secondo livello.

I risultati dell'ANOVA pubblicati nell'articolo di Koch sono i seguenti:

FONTE DI VARIAZIONE	Devianza	DF	Varianza	F	P
Totale	462,25	101	---	---	---
Temperatura	0,61	1	0,61	0,10	No
Coppie entro Temperatura	95,71	15	6,38	---	---
Dieta	28,55	1	28,55	28,27	< 0.001
Dieta x Temperatura	0,33	1	0,33	0,33	No
Dieta x Coppie entro Temperatura	15,19	15	1,01	---	---
Acidità	112,59	2	56,30	12,42	<0.001
Acidità x Temperatura	35,34	2	17,67	3,90	> 0.05
Acidità x Coppie entro Temperatura	136,03	30	4,53	---	---
Acidità x Dieta	2,82	2	1,41	1,41	≅ 0.50
Acidità x Dieta x Temperatura	5,01	2	2,50	2,50	≅ 0.20
Acidità x Dieta x Coppie entro Temp.	30,07	30	1,00	---	---

Trattandosi di un **esperimento nested complesso**, i diversi valori di F riportati nella tabella sono stati ottenuti da rapporti differenti, qui indicati nei punti fondamentali:

- il valore di F per la significatività delle differenze tra **le medie delle due temperature** è dato da

$$F_{(1,15)} = \frac{0,61}{6,38} = 0,10$$

e **non è significativo** in quanto addirittura inferiore a 1;

- il valore di F per la significatività delle differenze tra **le medie delle due diete** è dato da

$$F_{(1,15)} = \frac{28,55}{1,01} = 28,27$$

e risulta **altamente significativo** in quanto maggiore del valore critico, che per $F_{(1,15)}$ e $\alpha = 0.001$ è 19,5;

- il valore di F per la significatività delle differenze tra **le medie dei tre livelli di acidità**

$$F_{(2,30)} = \frac{56,30}{4,53} = 12,42$$

è altamente significativo, in quanto maggiore del valore critico di $F_{(2,30)}$ e $\alpha = 0.001$ che è 9,9.

Le varie interazioni (riportate nella tabella ANOVA) **non risultano significative**.

Con **test di statistica non parametrica** è possibile condurre le stesse verifiche, seppure separatamente per ogni fattore; ma alla fine si perviene ugualmente a una visione complessiva come la precedente

A) Per valutare l'**effetto della temperatura**, si aggregano i dati relativi ad ogni coppia di ratti come

(1)	(2)	(3)	(4)	(5)	(6)
Temp.	Coppia	Contr.	Farm.	Totale	r_i
37° C	1	6,61	9,32	15,93	8
	2	5,16	6,60	11,76	2
	3	11,59	16,64	28,23	17
	4	6,25	11,19	17,44	9
	5	6,09	6,36	12,45	4
	6	5,37	8,52	13,89	6
	7	8,11	9,71	17,82	10
	8	11,19	14,51	25,70	15
25° C	9	8,91	12,72	21,63	14
	10	6,38	13,17	19,55	12
	11	5,34	9,02	14,36	7
	12	3,00	2,91	5,91	1
	13	9,47	11,01	20,48	13
	14	10,24	16,87	27,11	16
	15	6,29	6,34	12,63	5
	16	5,36	6,74	12,10	3
	17	5,63	13,32	18,95	11

Per verificare l'ipotesi nulla **bilaterale**

$$H_0: \ me_{37} = me_{25} \qquad \text{contro} \qquad H_1: \ me_{37} \neq me_{25}$$

si ricorre al test **U di Mann-Whitney** per due campioni indipendenti. L'**ipotesi è bilaterale**, in quanto all'autore non è apparso logico non assumere a priori il vantaggio di una delle due condizioni sperimentali (temperatura a 35° C oppure a 25° C), nell'assorbimento della quantità di ferro (ricordare che l'ipotesi dipende solo dalla conoscenza a priori dell'effetto della temperatura).

In modo più dettagliato, in funzione della temperatura (colonna 1), si raggruppano i dati di tutte le coppie (colonna 2), separatamente per il Controllo (colonna 3) e il Farmaco (colonna 4), calcolandone la somma (colonna 5).

Il gruppo delle 8 coppie collocate a 37° C e quello delle 9 coppie a 25° C formano due campioni indipendenti.

Pure volendo utilizzare il test **U di Mann-Whitney**, in quanto più diffuso anche nei programmi informatici, la organizzazione dei dati nella tabella rende più semplice utilizzare il conteggio dei ranghi, secondo il metodo **T di Wilcoxon-Mann-Whitney**.

Si ottiene

$$T(37°) = 8 + 2 + 17 + 9 + 4 + 6 + 10 + 15 = 71$$

$$T(25°) = 14 + 12 + 7 + 1 + 13 + 16 + 5 + 3 + 11 = 82$$

Successivamente, sulla base della relazione

$$U_{37} = T_{37} - \frac{n_1 \cdot (n_1 + 1)}{2}$$

$$U_{25} = T_{25} - \frac{n_2 \cdot (n_2 + 1)}{2}$$

si ricavano i corrispondenti valori di U e U'

$$U = 71 - \frac{8 \cdot (8+1)}{2} = 71 - 36 = 35$$

$$U' = 82 - \frac{9 \cdot (9+1)}{2} = 82 - 45 = 37$$

La relazione

$$U + U' = n_1 \cdot n_2 = 35 + 37 = 8 \cdot 9 = 72$$

permette di verificare con facilità di non avere commesso errori nel calcolo dei ranghi e nella loro trasformazione in precedenze.

Poiché il valore critico (vedi tabella relativa al test U di Mann-Whitney nel capitolo dedicato ai test non parametrici per 2 campioni indipendenti) alla probabilità **α = 0.05** per un **test bilaterale** è **13** (molto minore del valore 35 calcolato), non si può rifiutare l'ipotesi nulla: le due differenti temperature non producono differenze nella quantità media di ferro assorbito.

Inoltre è utile osservare che il valore di U calcolato (35) non solo è maggiore di quello critico, ma è molto vicino alla media attesa (36)

$$\mu_U = \frac{n_1 \cdot n_2}{2} = \frac{8 \cdot 9}{2} = 36$$

E' quindi possibile concludere non solo che non è possibile rifiutare l'ipotesi nulla, ma che con altissima probabilità essa è vera.

B) Per valutare l'**<u>effetto della dieta</u> (α_i)** contenente etionina rispetto al controllo, sempre partendo dalla tabella dei dati presentati, si formano due campioni dipendenti, in cui la quantità per ogni ratto è dato dalla somma delle quantità misurate nelle tre soluzioni:

(1)	(2)	(3)	(4)	(5)	(6)
Coppia	Contr.	Farm.	d_i	$\lvert r_i \rvert$	r_i
1	6,61	9,32	+2,71	8	+8
2	5,16	6,60	+1,44	5	+5
3	11,59	16,64	+5,05	14	+14
4	6,25	11,19	+4,94	13	+13
5	6,09	6,36	+0,27	3	+3
6	5,37	8,52	+3,15	9	+9
7	8,11	9,71	+1,60	7	+7
8	11,19	14,51	+3,32	10	+10
9	8,91	12,72	+3,81	12	+12
10	6,38	13,17	+6,79	16	+16
11	5,34	9,02	+3,68	11	+11
12	3,00	2,91	-0,09	2	-2
13	9,47	11,01	+1,54	6	+6
14	10,24	16,87	+6,63	15	+15
15	6,29	6,34	+0,05	1	+1
16	5,36	6,74	+1,38	4	+4
17	5,63	13,32	+7,69	17	+17

Considerando le coppie di ratti, si confrontano le due mediane (quella del Farmaco e quella del Controllo) con **ipotesi unilaterale**

$$H_0: me_F \leq me_C \qquad \text{contro} \qquad H_1: me_F > me_C$$

attraverso il **test T di Wilcoxon per due campioni dipendenti**.

In questo caso, è stata scelta un'**ipotesi unilaterale,** in quanto ritenuto logico che il farmaco possa determinare un aumento reale della quantità di ferro assorbita dal fegato.

La descrizione dettagliata della procedura riportata nella tabella prevede: per ogni coppia (colonna 1), si sommano le quantità del Controllo (colonna 2) e del Farmaco (colonna 3), calcolandone l'incremento con il segno (colonna 4); con questi ultimi dati presi in valore assoluto, si stima il rango (colonna 5) e a ognuno si attribuisce il segno dell'incremento (colonna 6).
La somma minore dei valori con lo stesso segno (il negativo) risulta **T = 2**.
Poiché il valore critico (vedi tabella relativa al test **T di Wilcoxon** nel capitolo dedicato ai test non parametrici per 2 campioni dipendenti) alla probabilità **α = 0.0005** per un **test unilaterale** è **11**, si

rifiuta l'ipotesi nulla: il farmaco determina un aumento altamente significativo della quantità mediana di ferro.

C) Per valutare l'**effetto della acidità**, si aggregano i dati relativi ad ogni coppia di ratti come nella tabella seguente:

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	ACIDITA'					
Coppia	Somme			Ranghi		
	Bassa	Media	Alta	r_{bassa}	r_{media}	r_{alta}
1	3,57	3,99	8,37	1	2	3
2	1,98	3,05	6,73	1	2	3
3	3,31	6,17	18,75	1	2	3
4	1,69	3,70	12,05	1	2	3
5	3,43	4,53	4,49	1	3	2
6	3,17	2,97	7,75	2	1	3
7	2,60	4,23	10,99	1	2	3
8	1,87	11,77	12,06	1	2	3
9	4,56	11,48	5,59	1	3	2
10	7,08	6,57	5,90	3	2	1
11	1,72	2,13	10,51	1	2	3
12	2,62	1,67	1,62	3	2	1
13	2,76	10,80	6,92	1	3	2
14	2,66	14,68	9,77	1	3	2
15	3,53	2,03	7,07	2	1	3
16	3,60	3,14	5,36	2	1	3
17	2,87	9,77	6,31	1	3	2
Totali	---	---	---	24	36	42
Medie	3,12	6,04	8,25	1,41	2,12	2,47

Per verificare l'ipotesi nulla sulle mediane dei tre livelli di acidità con

H_0: $me_{Bassa} = me_{Media} = me_{Alta}$ contro H_1: *le me non sono tutte uguali*

si ricorre al test di **Friedman** per campioni dipendenti.

Per semplificare i calcoli, si può utilizzare la formula abbreviata che ricorre ai totali dei ranghi

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

dove:

- **N** è il numero di righe od osservazioni in ogni campione (17),
- **k** è il numero di colonne o campioni a confronto (3),
- T_i è il totale dei ranghi di ogni colonna (24, 36, 42)

ottenendo

$$\chi_F^2 = \frac{12 \cdot (24^2 + 36^2 + 42^2)}{17 \cdot 3 \cdot (3+1)} - 3 \cdot 17 \cdot (3+1) = \frac{12 \cdot 3636}{204} - 204 = 213{,}88 - 204 = 9{,}88$$

un valore uguale **9,88** da verificare nella distribuzione **chi-quadrato con 2 gdl**.

Poiché la tabella dei valori critici

- alla probabilità $\alpha = 0.01$ riporta 9,21
- alla probabilità $\alpha = 0.001$ riporta 13,82

si rifiuta l'ipotesi nulla con probabilità $P < 0.01$.

I tre livelli di acidità determino una differenza significativa nelle quantità mediane di ferro assorbito.

Per verificare se la quantità aumenta in modo significativo al crescere del livello di acidità, si dovrebbe applicare il test di Page, che è presentato nel capitolo successivo. Si tratta di un test unilaterale; applicato agli stessi dati, darebbe una probabilità dimezzata rispetto a quella ora calcolata, se effettivamente esiste un gradiente in crescita della quantità mediana di ferro assorbito nei tre livelli crescenti di acidità. (Ma Koch ha applicato il test di Friedman, forse perché test non parametrico analogo all'ANOVA a due criteri).

Nello stesso modo con il quale sono stati analizzati gli effetti dei tre fattori principali e possibile analizzare le loro interazioni. Si tratta di riorganizzare ogni volta la tabella dei dati originali in modo adeguato, applicando il metodo richiesto

- dall'ipotesi da verificare e
- dalla impostazione tabellare dei dati.

D) Per verificare se esiste **interazione Temperatura x Dieta**, i dati devono essere separati nei 4 gruppi:

- temperatura 37° C e farmaco (8 dati),
- temperatura 37° C e controllo (8 dati),
- temperatura 25° C e farmaco (9 dati),
- temperatura 25° C e controllo (9 dati),

(1)	(2)	(3)	(4)	(5)	(6)
Temp.	Coppia	Farmaco	Controllo	d_i	r_i
37° C	1	9,32	6,61	2,71	8
	2	6,60	5,16	1,44	5
	3	16,64	11,59	5,05	14
	4	11,19	6,25	4,94	13
	5	6,36	6,09	0,27	3
	6	8,52	5,37	3,15	9
	7	9,71	8,11	1,60	7
	8	14,51	11,19	3,32	10
25° C	9	12,72	8,91	3,81	12
	10	13,17	6,38	6,79	16
	11	9,02	5,34	3,68	11
	12	2,91	3,00	-0,09	1
	13	11,01	9,47	1,54	6
	14	16,87	10,24	6,63	15
	15	6,34	6,29	0,05	2
	16	6,74	5,36	1,38	4
	17	13,32	5,63	7,69	17

Separatamente per temperatura (colonna 1), per le 17 coppie (colonna 2) formate dal totale del Farmaco (colonna 3) e dal totale del Controllo (colonna 4) si determinano le differenze (colonna 5) e si attribuiscono ad esse i ranghi considerando tutta la colonna (colonna 6).

Infatti, il concetto di base è che, se non esiste interazione fra Temperatura e Dieta, la mediana delle differenze tra i totali delle coppie con temperatura di 37° C non differisce significativamente dalla mediana delle differenze tra i totali delle coppie con temperatura 25°C.

Diventa il confronto tra **le mediane di due campioni indipendenti**, che può essere verificata con il **test U di Mann-Whitney**. Anche in questo caso, a causa della disposizione tabellare dei dati, risulta più semplice utilizzare la somma dei ranghi per ricorrere al test T di Wilcoxon-Mann-Whitney o, mediante trasformazione, ricavare il valore di U.

$$T(37°) = 8 + 5 + 14 + 13 + 3 + 9 + 7 + 10 = 69$$
$$T(25°) = 12 + 16 + 11 + 1 + 6 + 15 + 2 + 4 + 17 = 84$$

Successivamente, sulla base della relazione utilizzata anche in precedenza si ricavano i corrispondenti valori di U e U' (preferito nei programmi informatici e nelle indicazioni di molti testi)

$$U = 69 - \frac{8 \cdot (8+1)}{2} = 69 - 36 = 33$$

$$U' = 84 - \frac{9 \cdot (9+1)}{2} = 84 - 45 = 39$$

La relazione

$$U + U' = n_1 \cdot n_2 = 33 + 39 = 8 \cdot 9 = 72$$

permette di verificare con facilità di non avere commesso errori nel calcolo dei ranghi e nella loro trasformazione in precedenze.

Poiché il valore critico (vedi tabella relativa al test U di Mann-Whitney nel capitolo dedicato ai test non parametrici per 2 campioni indipendenti) alla probabilità **α = 0.05** per un **test bilaterale** è **13** (molto minore del valore 33 calcolato), non è possibile rifiutare l'ipotesi nulla: **non esiste una interazione significativa** tra Temperatura e Dieta.

Ancora una volta si può osservare, che il valore di U calcolato (39) non solo è maggiore di quello critico, ma è molto vicino alla media attesa (36)

$$\mu_U = \frac{n_1 \cdot n_2}{2} = \frac{8 \cdot 9}{2} = 36$$

E' quindi possibile concludere non solo che non è possibile rifiutare l'ipotesi nulla, ma che probabilmente essa è vera: **l'interazione** tra Temperatura e Dieta **è nulla**.

E) Per verificare la significatività delle interazioni nelle quali si prende in considerazione l'Acidità, l'analisi diventa più complessa, poiché essa è a tre livelli. Di conseguenza, il modello di analisi applicato in precedenza alla dieta deve considerare le differenze tra tutte le possibili coppie di acidità, che per tre livelli sono tre: Bassa – Media, Bassa – Alta, Media – Alta.

Ne deriva una matrice di ranghi delle 17 coppie di ratti, distinti nei due livelli di temperatura, con i tre ranghi delle differenze tra le coppie dei livelli di acidità.

E' un'analisi con ricorre con poca frequenza e che pertanto non viene illustrata. È riportata solamente la tabella dei dati che dovrebbero essere analizzati:

(1)	(2)	(3)	(4)	(5)
Temp.	Coppia	Somme delle Acidità		
		Bassa	Media	Alta
37° C	1	3,57	3,99	8,37
	2	1,98	3,05	6,73
	3	3,31	6,17	18,75
	4	1,69	3,70	12,05
	5	3,43	4,53	4,49
	6	3,17	2,97	7,75
	7	2,60	4,23	10,99
	8	1,87	11,77	12,06
25° C	9	4,56	11,48	5,59
	10	7,08	6,57	5,90
	11	1,72	2,13	10,51
	12	2,62	1,67	1,62
	13	2,76	10,80	6,92
	14	2,66	14,68	9,77
	15	3,53	2,03	7,07
	16	3,60	3,14	5,36
	17	2,87	9,77	6,31

L'articolo di Gary G. **Koch** chiarisce che anche le tecniche non parametriche dimostrano la non significatività sia di tutte le interazioni di primo livello, sia di quella di secondo livello.

Nelle conclusioni presenta una tabella riassuntiva di tutte le analisi (leggermente modificata).

Per rendere i confronti omogenei e di facile lettura, i risultati di tutti i test non parametrici effettuati sono stati trasformati nei valori χ^2 corrispondenti,

Fonte di variazione	DF	χ^2	P
Temperatura	1	0.01	Non sign.
Dieta	1	13,24	< 0.001
Acidità	2	7,54	< 0.01
Temperatura x Dieta	1	0,08	Non sign.
Temperatura x Acidità	2	5,84	> 0.05
Acidità x Dieta	2	0,62	Non sign.
Temperatura x Dieta x Acidità	2	3,30	> 0.10

Il confronto con i valori critici mostra che si è giunti globalmente alle stesse conclusioni permesse dall'ANOVA parametrica a due fattori con interazione.

Questi test non parametrici non solamente sono applicabili in condizioni molto più generali, ma hanno esattamente la stessa potenza nell'evidenziare effetti significativi.

DF	Valori critici del χ^2			
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
1	2,706	3,841	6,635	10,826
2	4,605	5,991	9,210	13.816

L'articolo di Koch evidenzia anche un ultimo aspetto, che sottolinea concetti già presentati in questo corso, nel paragrafo introduttivo ai test non parametrici. Per la sua importanza viene riportata in modo integrale (modificando solo la grafica, per mettere in risalto alcuni concetti):

Finally, some remarks should be made about a paradoxical aspect of the analysis given here. Namely, we claim that because its underlying assumptions do not hold, the analysis of variance is not appropriate. However, it is performed anyway and then is used as a standard to which the results of the non-parametric tests are compared. This has been done for two reasons:

1. *The analysis of variance plays a sacred role in statistical analysis because of its robustness properties.*
2. *For this set of data, the results of the parametric and non-parametric analysis are similar.*

To avoid the confusions created, we recommend that three cases be delineated as follows:

I . The assumptions of the analysis of variance ***definitely hold****.*

II. The assumptions of the analysis of variance are ***not drastically violated****.*

III. The assumptions of the analysis of variance are ***drastically violated***

If Case I applies, ***only the analysis of variance need be performed****.*

If Case III applies, ***only the non-parametric analysis should be performed****.*

If Case II applies, ***both the analysis of variance and the non-parametric analysis should be performed****, and hopefully they will lead to the* ***same conclusions****.*

The data considered in this paper represent a Case II situation.