* Inferenza statistica

*Test sulla differenza tra le medie di due popolazioni con varianza nota**

Supponiamo di dover decidere se due popolazioni hanno medie significativamente diverse sulla base di due campioni, rispettivamente di ampiezza n_1>29 ed n_2>29, estratti da esse. Considerando H_0: *"non c'è differenza fra le medie delle popolazioni"*, cioè $\mu_1=\mu_2=\mu$ e ricordando che la distribuzione campionarie delle differenze delle medie è approssimativamente normale con media e deviazione standard rispettivamente:

$$\mu_{\bar{x}_1-\bar{x}_2}=\mu_1-\mu_2 \qquad \sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\sigma^2_{\bar{x}_1}+\sigma^2_{\bar{x}_2}}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

possiamo decidere di rifiutare o meno l'ipotesi nulla prendendo in considerazione il valore standard z della differenza delle medie quando $\mu_1=\mu_2=\mu$ ***(H_0: "μ-μ_2=0";H_1: "μ_1-$\mu_2\neq0$").***

Se z cade in una delle due regioni critiche della distribuzione normale standardizzata rifiuteremo H_0, altrimenti non la rifiuteremo. Il valore standardizzato z è dato da:

$$z=\frac{\bar{x}_1-\bar{x}_2-\mu_{\bar{x}_1-\bar{x}_2}}{\sigma_{\bar{x}_1-\bar{x}_2}}=\frac{\bar{x}_1-\bar{x}_2-0}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$$

NOTA: Come si può osservare, anche in questo caso, viene rispettata la formula generale per le statistiche dei test.

Nel caso di varianze ignote e/o di campioni poco numerosi (n<30), si applicherà il test *t di Student* per campioni indipendenti .

diap. 8.1.4

* *Esercizi 8.1.4.1* *

Esercizio tratto da: "Statistica Medica" di G.Fabbrocini e M.Quarto. Gruppo Editoriale Esselibri - Simone. Napoli.

- Due campioni di pazienti di ampiezza 70 e 110 sono sottoposti a due trattamenti diversi che producono tempi medi di guarigione rispettivamente di 60 e 57 giorni con deviazioni standard di 8 e 12 giorni. Se i tempi di risposta sono distribuiti normalmente e le due popolazioni hanno le stesse varianze dei campioni, verificare se i due trattamenti hanno pari efficacia (α=0,02).

SOLUZIONE

H_0: μ_1-μ_2 = 0; H_1: μ_1-$\mu_2\neq 0$

Calcoliamo il valore standard z della differenza delle medie quando $\mu_1=\mu_2$

$$z=\frac{\bar{x}_1-\bar{x}_2-0}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}=\frac{60-57}{\sqrt{64/70+144/110}}=\frac{3}{\sqrt{0,9+1,3}}=2,02$$

Poiché -2,33<2,02<2,33 non si può rifiutare l'ipotesi H_0.

diap. 8.1.4.1

* Inferenza statistica

Il test t di Student

• Il *test z* sulla differenza di due medie è applicabile quando sono note le varianze delle popolazioni a confronto ed i campioni sono sufficientemente ampi *(n_1 ed n_2 maggiori di 30).*

• Se queste condizioni non sono soddisfatte è necessario utilizzare ***il test t di Student**** per campioni indipendenti che utilizza le varianze campionarie e che risulta valido sia per piccoli che per grandi campioni.

• Anche in questo caso, supponendo di avere due campioni con medie diverse, si tratta di decidere se la differenza tra le due medie sia significativa, ovvero se esiste una differenza reale fra le medie delle popolazioni da cui sono estratti i campioni.

• Analogamente al *test z*, il test ***t di Student*** è determinato dal rapporto:

$$t = \frac{differenza\_medie\_campionarie}{errore\_standard\_della\_differenza\_delle\_medie\_campionarie}$$

* Le condizioni di applicabilità del test riguardano le distribuzioni (approssimativamente normali) e le varianze delle popolazioni (pressochè uguali fra loro).

diap. 9.1

* Inferenza statistica

Il test t di Student: calcolo di t e della stima combinata della varianza

• In formule, per campioni di diversa numerosità n_1 e n_2, abbiamo:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

dove: $\bar{x}_1$ e $\bar{x}_1$ sono le due medie campionarie

$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ è la stima combinata della varianza della popolazione

• Per campioni di uguale numerosità $n_1 = n_2 = n$, abbiamo:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2\left(\frac{s^2}{n}\right)}}$$

dove: $\bar{x}_1$ e $\bar{x}_1$ sono le due medie campionarie

$s^2 = \frac{1}{2}(s_1^2 + s_2^2)$ è la stima combinata della varianza della popolazione

diap. 9.2

* Inferenza statistica

Il test t di Student: valori critici

Valori critici di t (due code)

ν	0.20	0.10	0.05	0.02	0.01
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.97	9.93
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.42	1.90	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.06
13	1.35	1.77	2.16	2.65	3.01
14	1.34	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.32	1.72	2.09	2.53	2.84
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.48	2.79
.	.	.	.	.	.
40	1.30	1.68	2.02	2.42	2.70
.	.	.	.	.	.
50	1.30	1.68	2.01	2.40	2.68
.	.	.	.		
∞	1.282	1.645	1.960	2.326	2.580

- **La tabella a fianco mostra i valori critici t_c corrispondenti a diversi *livelli di significatività* α e diversi gradi di libertà ν.**

Per campioni di uguale numerosità abbiamo:

$$\nu = 2(n-1)$$

- **Per campioni di diversa numerosità abbiamo:**

$$\nu = n_1 + n_2 - 2$$

- **NOTA: il valore di t necessario per rifiutare l'ipotesi H_0 decresce al crescere di ν e cresce al diminuire di α.**

diap. 9.4

* *Esercizi 9.4.1* *

In una fattoria si vogliono provare gli effetti di un certo fertilizzante sulla produzione di grano. Per effettuare l'esperimento sono stati scelti 12 appezzamenti di terreno di uguale area: una metà è stata trattata con il fertilizzante, mentre l'altra metà no (gruppo di controllo). In media la quantità di grano prodotta sugli appezzamenti non trattati è stata di 4,8 staia, con scarto quadratico medio di 0,40 staia, mentre la quantità di grano prodotta dagli appezzamenti trattati è stata in media di 5,1 staia, con scarto quadratico medio di 0,36 staia. Si può concludere che ci sia stata una differenza significativa nella produzione di grano fra i due tipi di appezzamenti, usando il livello di significatività del 5% ?

SOLUZIONE

H_0: "L'uso del fertilizzante non ha effetto, ovvero $\mu_1=\mu_2$"; H_1: "$\mu_1\neq\mu_2$".

$$s^2 = \frac{1}{2}\left(s_1^2 + s_2^2\right) = \frac{0{,}40^2 + 0{,}36^2}{2} = 0{,}145 \qquad t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{2\frac{s^2}{n}}} = \frac{5{,}1 - 4{,}8}{\sqrt{2\frac{0{,}145}{6}}} = 1{,}370 \qquad \nu = 2(n-1) = 10$$

per *10 gradi di libertà* il valore critico per *α=5%* è *2,23*. Pertanto NON si rifiuta H_0 (la differenza osservata non è significativa)

diap. 9.4.1

* Esercizi 9.4.2 *

Si vuole verificare se tra le fasce di età 20-30 anni e 40-50 anni nei maschi di una determinata area geografica esistono differenze (α=0,01) per una data concentrazione ormonale analizzando le statistiche descrittive di due campioni estratti casualmente (vedi tabella).

Età	N	Media	varianza
20-30	20	66,10	73,79
40-50	22	75,76	115,99

$$H_0 : "\mu_{40-50} - \mu_{20-30} = 0" ; H_1 : "\mu_{40-50} - \mu_{20-30} \neq 0"$$

SOLUZIONE

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{21 \times 115{,}99 + 19 \times 73{,}79}{40} = 95{,}945$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{75{,}76 - 66{,}10}{\sqrt{\frac{95{,}945}{22} + \frac{95{,}945}{20}}} = 2{,}963 \qquad \nu = 22 + 20 - 2 = 40$$

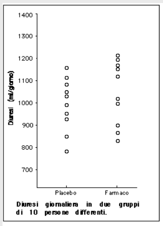
La regione di rifiuto per *α=0,01 e 40 gdl* è quella esterna all'intervallo *[-2,70;+2,70]*. Pertanto rifiutiamo l'ipotesi nulla di indipendenza fra le fasce d'età e la concentrazione ormonale.

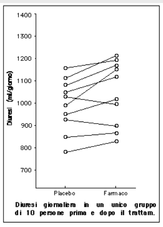
* Inferenza statistica

Il test t per dati appaiati

• Il test t di Student per dati appaiati si applica negli esperimenti nei quali ciascun soggetto è osservato prima e dopo un determinato trattamento o una determinata terapia *. Si tratta cioè di verificare l'ipotesi sulla variazione media prodotta dal trattamento, anziché quella sulla differenza nelle risposte medie in assenza o in presenza del trattamento stesso.

• Questo approccio riduce quella parte della variabilità delle osservazioni che è legata alla differenza tra soggetti diversi e dà luogo, pertanto, ad un test più sensibile.





Campioni indipende Dati appaiati

Esempio di verifica dell'efficacia di un diuretico tramite due disegni sperimentali a campioni indipendenti e per dati appaiati:

Nella primo caso il diuretico non risulta efficace perché la differenza fra le medie in relazione alle deviazioni standard è piccola *(t=1,33 con 18 gdl).*

Nel secondo caso il farmaco mostra un aumento della diuresi in 8 su dieci soggetti del campione (i segmenti uniscono i valori associati ai soggetti). Questo risultato suggerisce che il farmaco potrebbe essere efficace.

* Si parla di dati appaiati anche nel caso di due campioni simili per tutte le caratteristiche eccetto quella in studio (campioni non indipendenti)

* Inferenza statistica

Il test t per dati appaiati: calcolo di t

• Il ***test t per dati appaiati*** può essere utilizzato per verificare l'ipotesi che, mediamente, non ci sia nelle unità statistiche del campione alcuna variazione fra la prima e la seconda misurazione. L'interesse è cioè focalizzato sulle differenze rilevate in ogni unità statistica del campione.

• L'indice ***t*** è, in questo caso, determinato dal rapporto:

$$t = \frac{\bar{d}}{s_{\bar{d}}}$$

dove $\bar{d}$ è la differenza media dovuta al trattamento
e $S_{\bar{d}}$ l'errore standard della differenza media

• Il valore di ***t*** che ne risulta è confrontato col valore critico di ***t*** per ***n-1*** gradi di libertà.

• Le condizioni di applicabilità del test prevedono che le variazioni associate al trattamento debbano essere distribuite in modo approssimativamente normale.

diap. 9.6

* Inferenza statistica

Il test t per dati appaiati: riepilogo del procedimento

Ricapitolando, quando analizziamo dati tratti da un esperimento nel quale è possibile osservare ciascun individuo prima e dopo la somministrazione di un solo trattamento:

• Calcoliamo la differenza della risposta di ciascun individuo in concomitanza col trattamento.

• Calcoliamo la differenza media e l'errore standard della differenza media.

• Calcoliamo la statistica ***t*** come rapporto fra la differenza media e l'errore standard della differenza media.

• Confrontiamo questo valore di ***t*** col valore critico per *$\nu = n - 1$ gradi di libertà*, dove *n* è il numero degli individui coinvolti nell'esperimento.

diap. 9.7

*   *Esercizi 9.8*   *

* Esempio tratto da: "Statistica in Medicina" di T. Colton. Edizioni Piccin & Vallardi. Padova

Paziente	Placebo	Idroclorot.	differenze
1	211	181	30
2	210	172	38
3	210	196	14
4	203	191	12
5	196	167	29
6	190	161	29
7	191	178	13
8	177	160	17
9	173	149	24
10	170	119	51
11	163	156	7

La tabella mostra la pressione sistolica misurata ad 11 pazienti ipertesi dopo la somministrazione di placebo e di idroclorotiazide.

In base ai risultati osservati, vi è una qualsiasi evidenza di una differenza nella pressione media sistolica sanguigna a seguito di questi due trattamenti?

*   *Soluzione Esercizio 9.9*   *

H_0: "la media delle differenze Δ_0 è 0"

$H1$: "la media delle differenze Δ_0 è diversa da 0"

$$\bar{d} = \frac{\sum d_i}{n} = \frac{264}{11} = 24{,}0$$

$$S_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{(30-24)^2 + (38-24)^2 + \ldots + (7-24)^2}{10}} = 13{,}09$$

$$t = \frac{\bar{d}}{s_{\bar{d}}} = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} = \frac{24}{\frac{13{,}09}{\sqrt{11}}} = 6{,}08$$

$$\nu = n - 1 = 10$$

In base al t-test di Student per dati appaiati possiamo notare un abbassamento significativo della pressione sistolica in corrispondenza della somministrazione di idroclorotiazide ($p<0,01$)

Misure ripetute

I test appena visti non si possono applicare a misure ripetute sugli stessi soggetti:

in questi ogni soggetto deve avere una sola misura per ogni condizione

Quindi nel caso di più misure su ogni soggetto, se si vogliono confrontare più soggetti tra loro o con un valore di riferimento, è necessario

- *Mediare le misure o comunque estrarre un valore caratteristico*
- *Usare un test statistico che possa tenere conto di misure ripetute (uno di questi è il test ANOVA per misure ripetute)*

diap. 9.9

* Inferenza statistica

Il test del CHI-QUADRATO

• Esiste una procedura più flessibile del test z che consente anche di confrontare più di due campioni o più di due esiti: il ***test Chi-Quadrato*** (χ^2).

• Esso rientra nei test non parametrici, in quei metodi cioè che non presuppongono alcuna condizione circa la natura o i parametri della popolazione dalla quale i campioni provengono.

• Per descrivere la procedura analizziamo le tabelle di contingenza derivate da uno studio clinico controllato in ***doppio cieco*** per stabilire l'efficacia di dosi leggere di aspirina nella prevenzione di trombi in pazienti dializzati (naturalmente i gruppi risultavano omogenei per quanto riguarda il tempo in dialisi ed altre presunte variabili di confondimento).

diap. 10.1

* Inferenza statistica

Il test del CHI-QUADRATO: tabella delle frequenze osservate

La seguente ***tabella a doppia entrata*** *mostra i risultati ottenuti nell'esperimento con l'aspirina e con il placebo, ovvero il numero di pazienti in ogni gruppo (trattati e non trattati) che sviluppò o non sviluppò i emorragia:*

Gruppo campionario	**Sviluppo** *emorragia*	**Non sviluppo** *emorragia*	**Totale righe**
Placebo	18	7	25
Aspirina	6	13	19
Totale colonne	24	20	44

• La differenza riscontrata nei due gruppi è più grande di quella che ci aspetteremmo se l'aspirina agisse come un placebo (ipotesi nulla) ?

• Poiché le maggiori frequenze appaiono sulla diagonale della tabella (18 pazienti non trattati hanno sviluppato *emorragia*, 13 pazienti trattati con aspirina non li hanno sviluppati), tenderemmo a rifiutare l'ipotesi nulla e concludere che esiste un'associazione fra presenza di *emorragia* e assenza del trattamento con aspirina.

Esempio tratto da "Statistica per discipline biomediche". Stanton A. Glantz. McGraw-Hill

diap. 10.2

* Inferenza statistica

Il test del CHI-QUADRATO:tabella delle frequenze attese

La seguente ***tabella a doppia entrata*** *mostra come potrebbero apparire i dati se fosse vera l'ipotesi nulla, ovvero se l'aspirina non avesse alcun effetto:*

Gruppo campionario	**Sviluppo** *emorragia*	**Non sviluppo** *emorragia*	**Totale righe**
Placebo	13,64	11,36	25
Aspirina	10,36	8,64	19
Totale colonne	24	20	44

• Essa è ricavata considerando che, a prescindere dai gruppi, una proporzione di 24/44 dei pazienti ha sviluppato *emorragia* e 20/44 non li ha sviluppati. Così, sotto l'ipotesi che il trattamento non ha alcun effetto, ci aspetteremmo che 24/44 dei 25 soggetti non trattati e 24/44 dei 19 soggetti trattati sviluppassero *emorragia* *.

• Come possiamo quantificare la differenza fra la tabella delle frequenze attese e quella delle frequenze osservate?

* Notare come, nel caso di una tabella 2X2, basti calcolare la frequenza attesa in una cella, per determinare per differenza dalle frequenze totali marginali tutte le altre.

diap. 10.3

* Inferenza statistica

Il calcolo del CHI-QUADRATO

• Per valutare la disomogeneità esistente fra le frequenze osservate e quelle attese sotto l'ipotesi che non ci fosse associazione fra trattamenti ed esiti, definiamo il test statistico χ^2 come:

$$\chi^2 = \sum \frac{(O-A)^2}{A}$$

dove O e A rappresentano rispettivamente le frequenze osservate e attese in una certa cella. La somma è calcolata per tutte le celle della tabella di contingenza..

• Il valore di χ^2 è tanto più elevato quanto più le frequenze osservate differiscono da quelle attese.

• Come tutti i test statistici, per effetto del campionamento casuale, χ^2 può assumere uno spettro di valori anche quando non c'è associazione fra trattamenti ed esiti e la sua distribuzione dipende dal numero di gradi di libertà che è dato da:

$$\nu = (r-1)(c-1)$$

dove r è il numero di righe e c il numero di colonne della tabella.

* Inferenza statistica

I valori critici del CHI-QUADRATO

Analogamente a quanto visto nei test precedenti, dopo aver calcolato χ^2 dai dati rifiuteremo l'ipotesi nulla con un rischio di errore $p<\alpha$ se il valore ottenuto supererà il corrispondente valore critico. Poiché, nel caso dell'esperimento sulla formazione di *emorragia* abbiamo:

$$\chi^2 = \sum \frac{(O-A)^2}{A} = \frac{(18-13{,}64)^2}{13{,}64} + \frac{(7-11{,}36)^2}{11{,}36} + \frac{(6-10{,}36)^2}{10{,}36} + \frac{(13-8{,}64)^2}{8{,}64} = 7{,}10$$

Valori critici del χ^2

	α			
ν	**0,10**	**0,05**	**0,025**	**0,01**
1	2,706	3,841	5,024	6,635
2	4,605	5,991	7,387	9,210
3	6,251	7,815	3,348	11,345
4	7,779	9,488	11,143	13,277
5	9,236	11,070	12,833	15,086
6	10,645	12,592	14,449	16,812
7	12,017	14,067	16,013	18,475
8	13,362	15,507	17,535	20,090
9	14,684	16,019	19,023	21,666
10	15,987	18,307	20,483	23,209

Probabilità di valori maggiori, p

e il valore ottenuto supera il valore critico 6,635 associato ad $\alpha=0{,}01$, possiamo affermare che l'aspirina è associata ad un tasso più basso di emorragia e possiamo rifiutare l'ipotesi di assenza di associazione con un rischio di errore $p<0{,}01$.

• Nelle tabelle 2X2 tutte le frequenze attese devono essere >5 affinché il test sia accurato*. Inoltre, per una migliore compatibilità con la distribuzione teorica, sarebbe opportuno applicare la correzione di Yates per la continuità:

$$\chi^2 = \sum \frac{(|O-A|-0{,}5)^2}{A}$$

** in tabelle più grandi di 2X2 le frequenze attese in ogni casella non dovrebbero mai essere inferiori a 1 e non più del 20% inferiori a 5.*

* Inferenza statistica

Applicazione del CHI-QUADRATO a tabelle con dimensioni > di 2x2

• La tabella mostra i risultati di un'indagine sulle comorbidità in un campione di 98 pazienti ospedalizzati con diagnosi di Schizofrenia, Disturbi Schizoaffettivi e Disturbi Bipolari. Fra parentesi sono indicate le frequenze attese sotto l'ipotesi nulla che non ci sia una prevalenza di comorbidità significativamente più alta in qualcuno dei tre gruppi.

Comorbidità	Schizofrenia	Dist. Schizoaff.	Dist. Bipolari	Totale righe
SI	17 (10,79)	6 (8,91)	23 (26,30)	46
NO	6 (12,21)	13 (10,09)	33 (29,70)	52
Totale colonne	23	19	56	98

$$\chi^2 = \frac{(17-10{,}79)^2}{10{,}79} + \frac{(6-8{,}91)^2}{8{,}91} + \frac{(23-26{,}30)^2}{26{,}30} + \frac{(6-12{,}21)^2}{12{,}21} + \frac{(13-10{,}09)^2}{10{,}09} + \frac{(33-29{,}70)^2}{29{,}70} = 9{,}293 \qquad \nu = (3-1)(2-1) = 2$$

• La tabella dei valori critici mostra che, se la differenza fra le frequenze attese e osservate è dovuta solo al caso, meno del 1% delle volte il valore del χ^2 supera 9,21. Dunque concludiamo che c'è una relazione fra la classificazione diagnostica e la presenza di comorbidità ($p<0{,}01$). Tuttavia non sappiamo quale o quali gruppi si differenziano maggiormente.

diap. 10.6

* Test non parametrici basati sui ranghi

Quando vanno applicati

- I test parametrici (z-test, t-test e analisi di varianza) sono basati sul presupposto che le osservazioni siano tratte da popolazioni di valori normalmente distribuiti, nelle quali le varianze sono approssimativamente uguali.
- Questi metodi sono detti parametrici perché si basano sulla stima dei due parametri della popolazione, la media e la varianza, che definiscono completamente una distribuzione normale.
- Se i dati sperimentali non sono compatibili con queste condizioni preliminari o se addirittura, le osservazioni sono misurate su una scala qualitativa ordinale, i metodi parametrici diventano poco attendibili poiché la media e la varianza non sono sufficienti per una completa descrizione della popolazione.
- In questi casi è possibile utilizzare, invece delle osservazioni, i ranghi, cioè il numero d'ordine delle osservazioni stesse, al fine di calcolare test non parametrici (ovvero test liberi da distribuzione) nella verifica delle ipotesi.

diap. 13.1

* Test non parametrici basati sui ranghi

Il test per la somma dei ranghi di Mann-Whitney

▪Analogamente al T-test di Student per dati indipendenti, questo test è utilizzato in esperimenti in cui si confrontano due campioni indipendenti, ma la variabile in studio non rispetta le condizioni di applicabilità dei test parametrici.

Illustriamo il procedimento utilizzando un piccolo esperimento effettuato su 7 pazienti con depressione maggiore che mira a valutare l'efficacia di un farmaco antidepressivo . Il livello di gravità è valutato calcolando il punteggio totale della Scala HAM-D dopo un ciclo di un mese in cui a 3 soggetti è stato somministrato un placebo e a 4 il farmaco.

I ranghi sono attribuiti indipendentemente dal gruppo di appartenenza: al punteggio più basso è attribuito rango 1, al più elevato rango 7.

Osservazioni relative ad un esperimento su un farmaco antidepressivo

Placebo (controllo)		**Farmaco (trattamento)**	
Punteggio Scala HAM-D	Rango	Punteggio Scala HAM-D	Rango
19	3	21	4
24	5	27	7
16	1	18	2
	T=9	25	6

diap. 13.2

* Test non parametrici basati sui ranghi

Il test di Mann-Whitney, considerazioni.

▪Se il farmaco avesse diminuito i punteggi, dovremmo aspettarci i ranghi più bassi nel gruppo in trattamento; viceversa se gli avesse aumentati, dovremmo aspettarci i ranghi più bassi nel gruppo di controllo (placebo).

▪Come test statistico possiamo utilizzare la somma T dei ranghi del gruppo più piccolo (nell'esempio il gruppo di controllo).

▪Nel nostro esempio, il valore T=9 è sufficientemente "estremo" da giustificare il rifiuto dell'ipotesi nulla che il farmaco non abbia effetto? Ovvero, prendendo in considerazione la distribuzione di tutti i possibili valori di T, quanto è probabile osservare una somma T altrettanto "estrema" di quella osservata?

▪Per stimare questa probabilità potremmo elencare tutti i possibili ranghi attribuibili a ciascuna osservazione e calcolando per ogni combinazione la somma dei ranghi del gruppo di controllo.

diap. 13.3

* Test non parametrici basati sui ranghi

Il test di Mann-Whitney. Possibili attribuzioni di ranghi

Possibili ranghi e somme di ranghi per 3 individui su 7

rango 1	2	3	4	5	6	7	*Somma dei ranghi T*
x	x	x					6
x	x		x				7
x	x			x			8
x	x				x		9
x	x					x	10
x		x	x				8
x		x		x			9
x		x			x		10
x		x				x	11
x			x	x			10
x			x		x		11
x			x			x	12
x				x	x		12
x				x		x	13
x					x	x	14
	x	x	x				9
	x	x		x			10
	x	x			x		11
	x	x				x	12
	x		x	x			11
	x		x		x		12
	x		x			x	13
	x			x	x		13
	x			x		x	14
	x				x	x	15
		x	x	x			12
		x	x		x		13
		x	x			x	14
		x		x	x		14
		x		x		x	15
		x			x	x	16
			x	x	x		15
			x	x		x	16
			x		x	x	17
				x	x	x	18

Nel nostro esempio, il numero totale di modi diversi con cui i ranghi possono essere attribuiti al gruppo placebo corrisponde alle combinazioni di 7 oggetti a tre a tre, ovvero:

$$_7C_3 = \frac{7!}{3! \times 4!} = 35$$

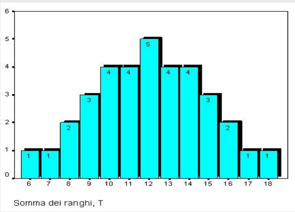
La tabella a lato mostra i 35 modi; le crocette identificano i ranghi associabili ai tre soggetti del gruppo placebo.

La colonna sulla destra indica la somma dei ranghi attribuibile al gruppo di controllo per ogni possibile combinazione di ranghi.

diap. 13.3.1

* Test non parametrici basati sui ranghi

Il test di Mann-Whitney. Distribuzione di T.



- **La figura a lato mostra la distribuzione dei possibili valori di T per ampiezze campionarie pari a 3 e 4.**
- **L'insieme dei due valori più estremi (T=6 e T=18) costituisce il 5,7% (p=0,057) di tutti i possibili modi di combinare i ranghi.**
- **Possiamo quindi definire questi valori come valori soglia (valori critici) per individuare i valori "estremi" di rifiuto dell'ipotesi nulla di inefficacia del trattamento (α=0,057)*.**
- **Dal momento che nel nostro esempio abbiamo ottenuto T=9, possiamo concludere che il valore non è sufficientemente "estremo" da giustificare il rifiuto dell'ipotesi che il farmaco non abbia effetto sulla depressione.**

* A differenza dei test parametrici in cui la distribuzione dei possibili valori della statistica è una distribuzione continua, la distribuzione dei valori T è discreta e quindi non è possibile considerare valori soglia corrispondenti esattamente al 5% dei valori più estremi (livello di significatività α=0,05), ma solo quelli che più gli si avvicinano.

diap. 13.4

* Test non parametrici basati sui ranghi

Il test di Mann-Whitney. Valori critici.

n_1, n_2	Livello di significatività a due code 0.05	0.01	0.001
2, 8	3, 19		
2, 9	3, 21		
2, 10	3, 23		
2, 11	4, 24		
2, 12	4, 26		
2, 13	4, 28		
2, 14	4, 30		
2, 15	4, 32		
2, 16	4, 34		
2, 17	5, 35		
2, 18	5, 37		
2, 19	5, 39	3, 41	
2, 20	5, 41	3, 43	
2, 21	6, 42	3, 45	
2, 22	6, 44	3, 47	
2, 23	6, 46	3, 49	
2, 24	6, 48	3, 51	
2, 25	6, 50	3, 53	
3, 5	6, 21		
3, 6	7, 23		
3, 7	7, 26		
3, 8	8, 28		
3, 9	8, 31	6, 33	
3, 10	9, 33	6, 36	
3, 11	9, 36	6, 39	
3, 12	10, 38	7, 41	
3, 13	10, 41	7, 44	
3, 14	11, 43	7, 47	
3, 15	11, 46	8, 49	
3, 16	12, 48	8, 52	
3, 17	12, 51	8, 55	
3, 18	13, 53	8, 58	
3, 19	13, 56	9, 60	
3, 20	14, 58	9, 63	
3, 21	14, 61	9, 66	6, 69
3, 22	15, 63	10, 68	6, 72
3, 23	15, 66	10, 71	6, 75
3, 24	16, 68	10, 74	6, 78
3, 25	19, 71	11, 76	6, 81
4, 4	10, 26		
4, 5	11, 29		
4, 6	12, 32	10, 34	
4, 7	13, 35	10, 38	
4, 8	14, 38	11, 41	
4, 9	15, 41	11, 45	
4, 10	15, 45	12, 48	
4, 11	16, 48	12, 52	
4, 12	17, 51	13, 55	
4, 13	18, 54	14, 58	10, 62
4, 14	19, 57	14, 62	10, 66
4, 15	20, 60	15, 65	10, 70
4, 16	21, 63	15, 69	11, 73
4, 17	21, 67	16, 72	11, 77
4, 18	22, 70	16, 76	11, 81
4, 19	23, 73	17, 79	12, 84
4, 20	24, 76	18, 82	12, 88
4, 21	25, 79	18, 86	12, 92
4, 22	26, 82	19, 89	13, 95
4, 23	27, 85	19, 93	13, 99
4, 24	28, 88	20, 96	13, 103
4, 25	28, 92	20, 100	14, 106
5, 5	17, 38	15, 40	
5, 6	18, 42	16, 44	
5, 7	20, 45	17, 48	
5, 8	21, 49	17, 53	
5, 9	22, 53	18, 57	15, 60
5, 10	23, 57	19, 61	15, 65
5, 11	24, 61	20, 65	16, 69
5, 12	26, 64	21, 69	16, 74
5, 13	27, 68	22, 73	17, 78
5, 14	28, 72	22, 78	17, 83
5, 15	29, 76	23, 82	18, 87
5, 16	31, 79	24, 86	18, 92
5, 17	32, 83	25, 90	19, 96
5, 18	33, 87	26, 94	19, 101
5, 19	34, 91	27, 98	20, 105
5, 20	35, 95	28, 102	20, 110
5, 21	37, 98	29, 106	21, 114
5, 22	38, 102	29, 111	21, 119
5, 23	39, 106	30, 115	22, 123
5, 24	40, 110	31, 119	23, 127
5, 25	42, 113	32, 123	23, 132
6, 6	26, 52	23, 55	
6, 7	27, 57	24, 60	
6, 8	29, 61	25, 65	21, 69
6, 9	31, 65	26, 70	22, 74
6, 10	32, 70	27, 75	23, 79
6, 11	34, 74	28, 80	23, 85

Adattata da White [114].

La tabella a fianco mostra i valori critici del test T di Mann-Whitney per la somma dei ranghi.

Essi dipendono dalla numerosità del campione più piccolo (n_1) e dalla numerosità del campione più grande (n_2).

diap. 13.4.1

* Test non parametrici basati sui ranghi

Procedimento di calcolo del test di Mann-Whitney.

Ricapitolando, il procedimento per il calcolo del test di Mann-Whitney per la verifica dell'ipotesi che un trattamento non abbia effetto, consiste nei seguenti passi:

- Attribuire un rango a tutte le osservazioni sulla base della loro grandezza, assegnando rango 1 all'osservazione più piccola*.
- Calcolare T, la somma dei ranghi del campione più piccolo**.
- Confrontare il valore ottenuto di T con la distribuzione di tutte le possibili somme dei ranghi, per vedere se il valore ottenuto sia compatibile con l'ipotesi che le differenze fra i due gruppi a confronto siano dovute solo alla casualità della loro composizione.

* A osservazioni con lo stesso valore deve essere assegnato lo stesso rango, pari alla media dei ranghi che sarebbero stati attribuiti alle singole osservazioni se fosse stato possibile distinguerle.

** Se entrambi i campioni hanno la stessa ampiezza, calcoliamo T da uno qualunque di essi.

diap. 13.5

* Test non parametrici basati sui ranghi
Il Test con segno di Wilcoxon

• Analogamente al test t di Student per dati appaiati, il test con segno di Wilcoxon è utilizzato negli esperimenti in cui i dati risultano appaiati (ad es., nel caso di esperimenti che prevedono un unico gruppo di soggetti osservati prima e dopo un certo trattamento) ed in cui non siano rispettate le condizioni di applicabilità dei test parametrici.

• Si calcolano le differenze causate dal trattamento in ciascun soggetto in studio, e si assegna un rango a ciascuna differenza in relazione al valore assoluto; successivamente attribuiamo al rango il segno della differenza e, infine, calcoliamo la somma algebrica dei ranghi in modo da ottenere il test statistico W.

• Confrontiamo poi il valore di W con la distribuzione di tutti i possibili valori di W per campioni di ampiezza uguale a quella del nostro studio, per verificare se le osservazioni sono compatibili con l'ipotesi che il trattamento non abbia avuto effetto.

diap. 13.7

* Test non parametrici basati sui ranghi
Il Test con segno di Wilcoxon: procedimento di calcolo

Illustriamo il procedimento di calcolo utilizzando un esperimento ipotetico sull'efficacia di un diuretico somministrato a 6 persone. La tabella mostra i risultati di questo esperimento, insieme alla variazione nella diuresi in ciascun soggetto dovuta al trattamento.

	Diuresi giornaliera, ml/giorno			Rango delle	Rango con Segno della
Soggetto	Prima della somministrazione	Dopo la somministrazione	Differenza	differenze	differenza
1	1490	1600	110	5	5
2	1300	1850	550	6	6
3	1380	1420	40	2	2
4	1410	1500	90	4	4
5	1350	1400	50	3	3
6	1010	1000	-10	1	-1
					W=19

La diuresi giornaliera è aumentata in 5 soggetti su 6. Questo è sufficiente a giustificare l'asserzione che il farmaco è efficace?

diap. 13.8

* Test non parametrici basati sui ranghi

Il Test con segno di Wilcoxon: considerazioni

• Se il trattamento non avesse effetto, i ranghi associati a variazioni positive dovrebbero essere pressoché pari ai ranghi associati a variazioni negative e la somma dei ranghi con segno W dovrebbe assumere un valore prossimo a 0.

• Viceversa, quanto più il trattamento modifica i valori della variabile oggetto di studio, tanto più le variazioni tenderanno ad assumere lo stesso segno. In questo caso la somma dei ranghi con segno W tenderà ad assumere un valore positivo o negativo "grande".

Per delimitare il confine fra "piccolo" e "grande" dobbiamo considerare tutte le possibili combinazioni di ranghi e calcolare per ognuna di esse W, in modo da identificare quali sono i suoi valori più "estremi" di rifiuto dell'ipotesi nulla di inefficacia del trattamento (per un dato il livello di significatività α).

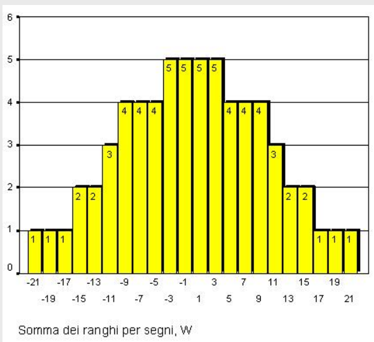
diap. 13.9

* Test non parametrici basati sui ranghi

Il Test con segno di Wilcoxon: distribuzione di W

La tabella mostra una parte dei possibili modi (n.tot=2^6=64) con cui i segni (qui codificati con 0=positivo e con 1=negativo) sono attribuiti ai ranghi associati ai 6 soggetti in studio ed i corrispondenti valori W.

RANGO						Somma W
1	2	3	4	5	6	
0	0	0	0	0	0	21
0	0	0	0	0	1	9
0	0	0	0	1	0	11
0	0	0	0	1	1	-1
0	0	0	1	0	0	13
0	0	0	1	0	1	1
0	0	0	1	1	0	3
0	0	0	1	1	1	-9
0	0	1	0	0	0	15
0	0	1	0	0	1	3
0	0	1	0	1	0	5
0	0	1	0	1	1	-7
0	0	1	1	0	0	7
0	0	1	1	0	1	-5
0	0	1	1	1	0	-3
0	0	1	1	1	1	-15
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1	1	1	1	1	0	-9
1	1	1	1	1	1	-21



Il diagramma a barre mostra la distribuzione completa dei valori di W.

Come si può osservare i valori –19 e 19 delimitano il 6,25% (pari a 4/64) dei valori più estremi della distribuzione di W.

diap. 13.10

* Test non parametrici basati sui ranghi

Il Test di Wilcoxon: quando W è sufficientemente estremo da rifiutare H_0

• Va osservato che, come nel test di Mann-Whitney per la somma dei ranghi, la natura discreta della distribuzione dei possibili valori di W implica l'impossibilità di determinare sempre i valori critici associati esattamente al livello di significatività voluto (generalmente α=0,05).

• Come si può notare nel nostro esempio, il livello α più prossimo al 5% è 6,25% *(p=4/64)* a cui corrispondono i valori critici –19 e +19.

• Pertanto il valore W=19 ottenuto ci autorizza a rifiutare l'ipotesi che il trattamento non abbia avuto effetto (cioè che il diuretico non sia efficace). Così facendo ci accolliamo un rischio di errore del 6,25%.

diap. 13.11

* Test non parametrici basati sui ranghi

Il Test di Wilcoxon: valori critici.

Valori critici del test W di Wilcoxon (test a due code)

n	*Valore critico*	*p*	*n*	*Valore critico*	*p*
5	15	0.062	13	65	0.022
6	21	0.032		57	0.048
	19	0.062	14	73	0.020
7	28	0.016		63	0.050
	24	0.046	15	80	0.022
8	32	0.024		70	0.048
	28	0.054	16	88	0.022
9	39	0.020		76	0.050
	33	0.054	17	97	0.020
10	45	0.020		83	0.050
	39	0.048	18	105	0.020
11	52	0.018		91	0.048
	44	0.054	19	114	0.020
12	58	0.020		98	0.050
	50	0.052	20	124	0.020
				106	0.048

Fonte: adattata da F.Mosteller e R.Rourke, *Study, Statistics, Nonparametrics and Order Statistics*, Addison-Wesley, Reading, Mass, 1973, Tab. A-11.

La tabella a fianco mostra i valori di W per esperimenti che coinvolgono fino a 20 unità statistiche.

diap. 13.12

* Test non parametrici basati sui ranghi

Riepilogo del procedimento di calcolo del test di Wilcoxon.

Ricapitolando, il procedimento per il calcolo del test di Wilcoxon per confrontare gli effetti di un trattamento effettuato su un unico gruppo di soggetti, consiste nel:

- Calcolare per ciascun soggetto la variazione verificatasi nella variabile oggetto di studio.
- Attribuire un rango a tutte le differenze in relazione al loro valore assoluto.
- Assegnare a ciascun rango il segno della differenza corrispondente.
- Sommare i ranghi con segno in modo da ottenere il test statistico W.
- Confrontare il valore di W ottenuto con la distribuzione dei possibili valori di W per decidere se è abbastanza "elevato" da far rifiutare l'ipotesi che il trattamento non ha effetto.

NOTA: ad una eventuale differenza nulla non è attribuito rango e la corrispondente unità statistica non viene presa in considerazione (l'ampiezza campionaria viene diminuita di tante unità quante sono le differenze nulle).

diap. 13.13