



TECNICHE DI ANALISI DEI DATI

AA 2019/2020

PROF. V.P. SENESE

Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:

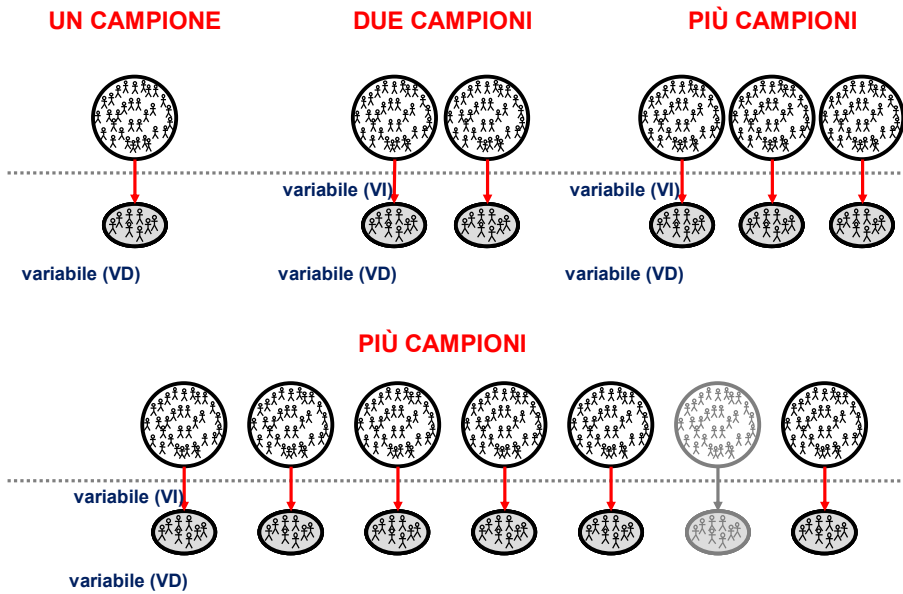
http://psiclab.altervista.org/TecnAnDat2020/2019_2020.html

Università della Campania «Luigi Vanvitelli» – Dipartimento di Psicologia – TECNICHE DI ANALISI DEI DATI – © Prof. V.P. Senese

ANALISI UNIVARIATE

- VERIFICA DELLE IPOTESI SU DI UN CAMPIONE
- IL CONFRONTO TRA DUE CAMPIONI
- REGRESSIONE
- ANOVA

ANALISI UNIVARIATE



TECNICHE DI ANALISI DEI DATI

AA 2018/2019

PROF. V.P. SENESE

Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:

<https://goo.gl/xY15fR>

UN CAMPIONE

Quando abbiamo a disposizione un **unico campione** possiamo essere interessati a indagare se:

- 1 c'è differenza negli indici di posizione (o parametri) relativi alla tendenza centrale tra campione e popolazione;
- 2 c'è differenza tra frequenze osservate e frequenze attese in base ad un modello teorico;
- 3 è ragionevole ritenere che il campione derivi da una popolazione avente una forma o una distribuzione specifica (normale, uniforme, ecc.).

Z TEST E T TEST

Quando vogliamo confrontare la distribuzione campionaria di **un parametro** relativo a una variabile misurata su **scala quantitativa** possiamo utilizzare la distribuzione *normale* o *t di Student*, scegliendo in base all'ampiezza del campione e alla conoscenza dei parametri della popolazione.

n > 30

$$z_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

n < 30

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n-1}$$

Popolazione finita
o camp. senza reinserimento

$$\sigma_{DCM} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

σ ignota

$$\hat{s} = \sigma = \sqrt{\frac{s^2}{n-1}}$$

\bar{x} = media del campione
 μ = media della popolazione
 σ = ds della popolazione
 n = ampiezza del campione

$$gdl = n - 1$$

TEST CHI-QUADRATO

Diversamente, quando consideriamo una variabile **qualitativa**, per confrontare i dati campionari con quelli della popolazione possiamo utilizzare la distribuzione delle frequenze e confrontare i valori osservati con quelli attesi nella popolazione (teorici).

La statistica che misura la discrepanza tra le frequenze osservate e quelle attese è:

$$\chi^2 = \text{Chi - quadrato}$$

TEST CHI-QUADRATO

$$\chi^2 = \sum_{j=1}^k \frac{(f_o - f_a)^2}{f_a}$$

k = numero di celle

f_o = frequenza osservata

f_a = frequenza attesa

$$gdl_{\chi^2} = k - 1$$

Assunzioni:

(1) osservazioni indipendenti;

(2) nessuna freq. osservata = 0;

(3a) se dicotomica, nessuna freq. teorica < 5;

(3b) se politomica, nessuna freq. teorica < 1 e meno del 20% < 5.

TEST CHI-QUADRATO

Questo test consente di verificare se:

- 1 se ci sono **differenze nelle frequenze** tra diverse categorie di una stessa variabile qualitativa;
- 2 se la distribuzione delle frequenze tra le diverse categorie di una variabile qualitativa **rispecchia una determinata (teorica) distribuzione** di frequenze;
- 3 se **due variabili** qualitative sono associate tra loro.

TEST CHI-QUADRATO

Si applica su **variabili qualitative** (originali o trasformate) e in due casi:

- una sola variabile (dicotomica o politomica);
- due variabili ($A \times B \Rightarrow 2 \times 2, 2 \times 3, 3 \times 4, \text{ecc.}$).

Quando ci sono **più di due variabili qualitative** si utilizza il *chi quadro del rapporto di verosimiglianza*, che si usa con la tecnica dei *modelli log-lineari*.

TEST CHI-QUADRATO

Si procede in questo modo:

- (1) **raccolta** e codifica dei dati: **frequenze osservate**;
- (2) inserimento dei dati in una tabella di frequenze;
- (3) definizione **ipotesi nulla** e **ipotesi alternativa**;
- (4) calcolo delle **frequenze teoriche** (in base a H_0);
- (5) calcolo del **chi-quadrato** e dei **gdl**;
- (6) si verifica l'ipotesi in base alla **distribuzione teorica del chi-quadrato**;
- (7) si interpretano i risultati.

ESEMPIO #1

Uno psicologo è interessato a verificare se il **tipo di patologie (VD, O)** che vengono diagnosticate al reparto ospedaliero dove lavora siano tutte ugualmente frequenti, o se invece le diagnosi mantengono la stessa distribuzione descritta in ambito nazionale.

A tal scopo registra il **tipo e numero (frequenza)** di diagnosi che vengono effettuate nel reparto durante un mese.

ESEMPIO #1

Diagnosi effettuate
in un mese.

DIAGNOSI			
Normalità ← → Patologia			
Non patologici (1)	Nevrotici (2)	Psicotici (3)	<i>N</i>
356	213	170	739

f_o

L'ipotesi generale è che tra i soggetti che prendono contatto con il reparto, la maggior parte dei pazienti sia non patologico poi, in ordine decrescente, nevrotici e psicotici.

$$H_0 \Rightarrow f_1 = f_2 = f_3$$

$$\alpha = .05$$

$$H_1 \Rightarrow f_1 \neq f_2 \neq f_3$$

ESEMPIO #1

Diagnosi effettuate
in un mese.

DIAGNOSI			
Normalità ← → Patologia			
Non patologici (1)	Nevrotici (2)	Psicotici (3)	<i>N</i>
356	213	170	739

f_o

Calcolo le frequenze
attese in base a H_0
(ipotesi nulla) →

$$H_0 \Rightarrow f_1 = f_2 = f_3 \Rightarrow f_1 = f_2 = f_3 = \frac{1}{3}$$

$$\Rightarrow f_1 = f_2 = f_3 = \frac{739}{3} = 246.33$$

ESEMPIO #1

Diagnosi effettuate
in un mese.

		DIAGNOSI			
		Normalità ←	→ Patologia		
		Non patologici (1)	Nevrotici (2)	Psicotici (3)	<i>N</i>
f_o		356	213	170	739
f_a		246.33	246.33	246.33	

$$\chi^2 = \frac{(356 - 246.33)^2}{246.33} + \frac{(213 - 246.33)^2}{246.33} + \frac{(170 - 246.33)^2}{246.33} =$$

ESEMPIO #1

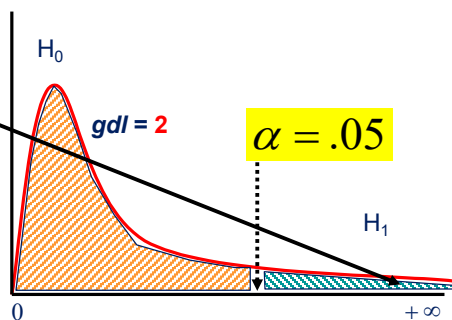
$$\chi^2 = \frac{(356 - 246.33)^2}{246.33} + \frac{(213 - 246.33)^2}{246.33} + \frac{(170 - 246.33)^2}{246.33} =$$

$$= 48.83 + 4.51 + 23.65 =$$

$$\chi^2 = 76.99$$

$$gdl = 3 - 1 = 2$$

```
> pchisq(76.99, 2, lower.tail=0)
[1] 1.913524e-17
p = 0.00000000000000001913524
```



ESEMPIO #1

	Observed N	Expected N	Residual
1.00	356	246.3	109.7
2.00	213	246.3	-33.3
3.00	170	246.3	-76.3
Total	739		

Test Statistics

	VAR00001
Chi-Square ^a	76.988
df	2
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 246.3.

ESEMPIO #1

Questo risultato ci porta a **respingere l'ipotesi nulla** e a **supportare l'ipotesi alternativa**. Tuttavia sappiamo che esistono **almeno due frequenze diverse** non sappiamo quali e soprattutto come differiscono.

$$H_0 \Rightarrow f_1 = f_2 = f_3$$

$$H_1 \Rightarrow f_1 \neq f_2 \neq f_3$$

Per sapere quali sono le specifiche categorie che differiscono e in che modo, dopo aver verificato l'ipotesi è necessario stimare per ciascuna cella la **distanza** tra f_o e f_a (in base all'ipotesi nulla H_0): **residui standardizzati (R)**.

$$R = \frac{f_o - f_a}{\sqrt{f_a}}$$

Si interpretano come dei **punti z** e utilizzando la **distribuzione normale standard**.

$$|R| > 1.96, p < .05$$

$$|R| > 2.58, p < .01$$

ESEMPIO #1

Diagnosi effettuate
in un mese.

DIAGNOSI			
Normalità		Patologia	
Non patologici (1)	Nevrotici (2)	Psicotici (3)	<i>N</i>
f_o 356	213	170	739
f_a 246.33	246.33	246.33	
$R = \frac{356 - 246.33}{\sqrt{246.33}}$ $R = +6.98$	$R = \frac{213 - 246.33}{\sqrt{246.33}}$ $R = -0.55$	$R = \frac{170 - 246.33}{\sqrt{246.33}}$ $R = -4.86$	
+	=	-	

ESEMPIO #1

In base ai risultati raccolti e alle analisi effettuate possiamo dire che:

le differenti tipologie di diagnosi non sono tutte ugualmente probabili, $\chi^2(2) = 76.99$, $p < .05$, $N = 739$. La maggior parte degli individui (48%) è diagnosticato come normale, $R = 6.98$; $p < .05$, mentre significativamente inferiore è il numero degli individui diagnosticati come psicotici (23%), $R = -4.86$, $p < .05$.

ESEMPIO #2

Diagnosi effettuate
in un mese.

DIAGNOSI			
Normalità ← → Patologia			
Non patologici (1)	Nevrotici (2)	Psicotici (3)	<i>N</i>
356	213	170	739

f_o

Immaginiamo ora di avere una ipotesi di ricerca molto definita (**Modello**):

$$H_0 \Rightarrow f_1 = 50\%; f_2 = 30\%; f_3 = 20\%$$

$$\alpha = .05$$

$$H_1 \Rightarrow f_1 \neq 50\%; f_2 \neq 30\%; f_3 \neq 20\%$$

ESEMPIO #2

Diagnosi effettuate
in un mese.

DIAGNOSI			
Normalità ← → Patologia			
Non patologici (1)	Nevrotici (2)	Psicotici (3)	<i>N</i>
356	213	170	739

f_o

f_a

369.5	221.7	147.8
-------	-------	-------

Calcolo le frequenze
attese in base a H_0
(ipotesi nulla) →

$$f_1 = 50\%; f_2 = 30\%; f_3 = 20\%$$

$$739(.5) = 369.5$$

$$739(.3) = 221.7$$

$$739(.2) = 147.8$$

ESEMPIO #2

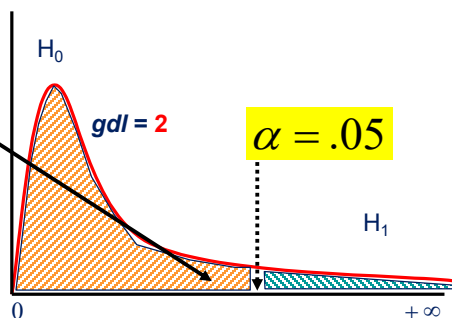
$$\chi^2 = \frac{(356 - 369.5)^2}{369.5} + \frac{(213 - 221.7)^2}{221.7} + \frac{(170 - 147.8)^2}{147.8} =$$

$$= .49 + .34 + 3.34 =$$

$$\chi^2 = 4.17$$

$$gdl = 3 - 1 = 2$$

```
> pchisq(4.17, 2, lower.tail=0)
[1] 0.1243071
```



ESEMPIO #2

Questo risultato ci porta a supportare l'ipotesi nulla (**Modello**).

$$H_0 \Rightarrow f_1 = 50\%; f_2 = 30\%; f_3 = 20\%$$

$$H_1 \Rightarrow f_1 \neq 50\%; f_2 \neq 30\%; f_3 \neq 20\%$$

Possiamo dire che:

i risultati confermano quanto previsto in base ai dati riferiti alla popolazione, $\chi^2(2) = 4.17$, $p = .124$, $N = 739$. Il **50%** degli individui è diagnosticato come normale, il **30%** degli individui è diagnosticato come nevrotico, mentre il restante **20%** sono diagnosticati come psicotici.



TECNICHE DI ANALISI DEI DATI

AA 2019/2020

PROF. V.P. SENESE

Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:

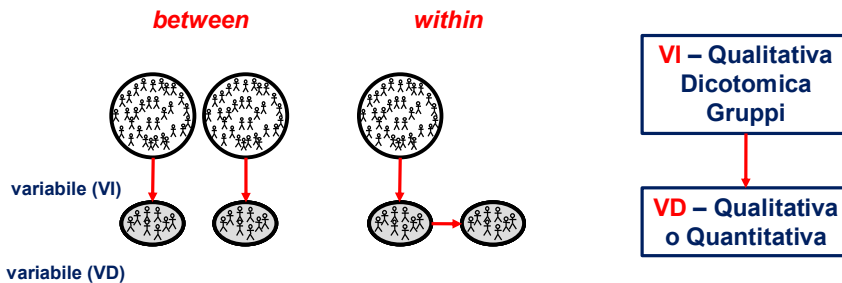
http://psiclab.altervista.org/TecnAnDat2020/2019_2020.html

Università della Campania «Luigi Vanvitelli» – Dipartimento di Psicologia – TECNICHE DI ANALISI DEI DATI – © Prof. V.P. Senese

DUE CAMPIONI

In alcuni casi l'obiettivo del ricercatore è quello di confrontare la distribuzione di una variabile all'interno di due campioni di misurazioni.

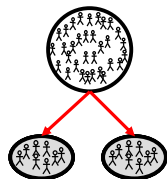
Quando le misure sono relative alle stesse unità osservative parliamo di **misure ripetute** (*within subject*), mentre quando le unità sono diverse parliamo di **misure indipendenti** (*between subject*).



T TEST - *WITHIN*

Quando le misure sono **ripetute** o **within**, se la variabile **dipendente** è **quantitativa** (intervalli o rapporti), la statistica più adatta a verificare se ci sono delle variazioni nella distribuzione della variabile tra le due misurazioni è il **t test** per misure dipendenti.

H_0 within

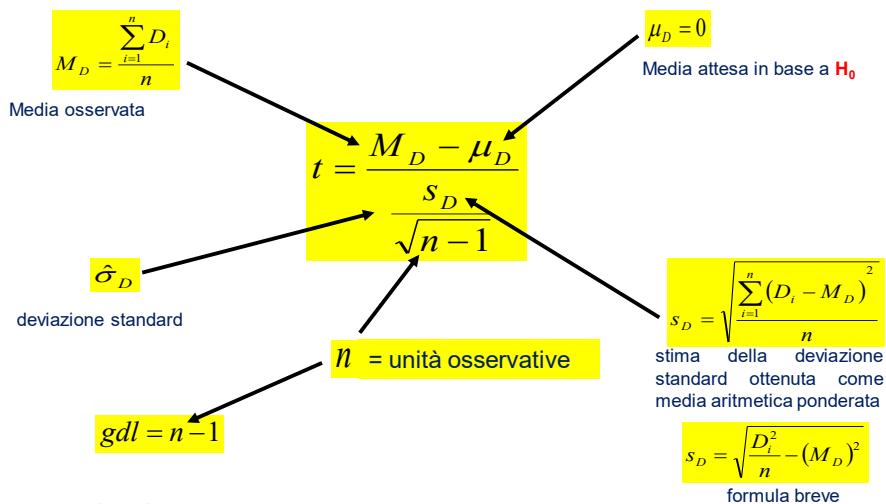


$$D_i = x_i - y_i$$

$$M_D = \frac{\sum_{i=1}^n D_i}{n}$$

In base alla **distribuzione campionaria delle medie**, possiamo considerare questo caso come un confronto tra la **media di un campione** (differenza) e la **media di una popolazione** (H_0).

T TEST - *WITHIN*



Assunzioni:

- (1) la variabile dipendente deve essere distribuita normalmente
- (2) le varianze devono essere omogenee.

T TEST - WITHIN

Le formule per il calcolo della forza dell'effetto **d** o **r**:

$$d = \frac{M_D - \mu_D}{s_D \sqrt{\frac{n}{n-1}}}$$

Effect size	d
small	.20
medium	.50
large	.80

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

Effect size	r
small	.10
medium	.30
large	.50

T TEST - WITHIN

- (1) raccolta e codifica dei dati (**distr. osservate**);
- (2) inserimento dei dati in una matrice;
- (3) definizione **ipotesi nulla** e **ipotesi alternativa**;
- (4) calcolo del *t* Test e dei *gdl*;
- (5) verifica dell'ipotesi in base alla distribuzione teorica del *t di Student*;
- (6) se significativo il test, calcolo dell'*effect size*;
- (7) interpretazione dei risultati.

ESEMPIO #3

Su 8 pazienti con attacchi di panico viene rilevata la frequenza degli attacchi **prima** e **dopo** (VD, R) una psicoterapia breve (VI, manipolata, un solo gruppo).

PRE TEST (x_i)	<i>trattamento</i> ⊙	POST TEST (y_i)
------------------------------	-------------------------	-------------------------------

Disegno di ricerca con un solo gruppo a due misure ripetute (**within**): pre-test e post-test.

ESEMPIO #3

Prima (x_i)	5	8	9	6	8	4	4	8
Dopo (y_i)	4	5	6	4	9	5	2	7

L'ipotesi generale è che ci sia una **riduzione** significativa del numero di sintomi manifestati dopo il trattamento.

$$H_0 \Rightarrow \mu_D = 0$$

$$H_1 \Rightarrow \mu_D > 0$$

$$\alpha = .05$$

**COMPOSTA
MONODIREZIONALE**

ESEMPIO #3

Si procede con il calcolo di M_D e s_D (utilizzando la formula abbreviata):

Sogg.	x_i	y_i	D_i	D_i^2
1	5	4	1	1
2	8	5	3	9
3	9	6	3	9
4	6	4	2	4
5	8	9	-1	1
6	4	5	-1	1
7	4	2	2	4
8	8	7	1	1
		Σ	10	30

$$M_D = \frac{10}{8} = 1.25$$

$$s_D = \sqrt{\frac{30}{8} - (1.25)^2} = 1.48$$

ESEMPIO #3

$$M_D = \frac{10}{8} = 1.25$$

$$s_D = \sqrt{\frac{30}{8} - (1.25)^2} = 1.48$$

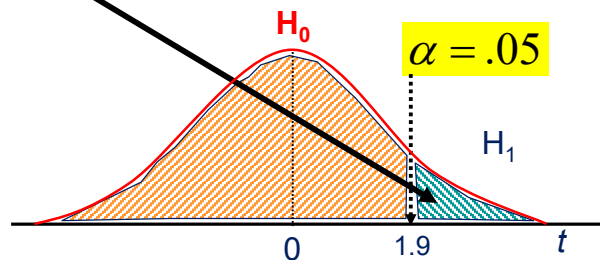
$$t = \frac{1.25}{\frac{1.48}{\sqrt{8-1}}} = 2.23$$

$$d = \frac{1.25}{1.48 \sqrt{\frac{8}{8-1}}} = .790$$

$$r = .367$$

$$gdl = 8 - 1 = 7$$

```
> pt(2.22,7,lower.tail=0)
[1] 0.03093863
```



ESEMPIO #3

Paired Samples Test

		Paired Differences				t	df	Sig. (2-tailed)	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower				Upper
Pair 1	Prima - Dopo	1.25000	1.58114	.55902	-.07187	2.57187	2.236	7	.060

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Prima	6.50000	8	2.00000	.70711
	Dopo	5.25000	8	2.12132	.75000

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Prima & Dopo	8	.707	.050

ESEMPIO #3

Questo risultato ci porta a respingere l'ipotesi nulla e a supportare l'ipotesi alternativa.

$$H_0 \Rightarrow \mu_D = 0$$

$$H_1 \Rightarrow \mu_D > 0$$

I risultati evidenziano che il trattamento riduce significativamente il numero di sintomi, $t(7) = 2.23$, $p = .031$ (una coda). In particolare, i dati evidenziano che dopo il trattamento, in media, si osserva una riduzione di 1.2 attacchi di panico, $d = .790$.



TECNICHE DI ANALISI DEI DATI

AA 2019/2020

PROF. V.P. SENESE

Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:

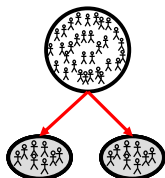
http://psiclab.altervista.org/TecnAnDat2020/2019_2020.html

Università della Campania «Luigi Vanvitelli» – Dipartimento di Psicologia – TECNICHE DI ANALISI DEI DATI – © Prof. V.P. Senese

TEST WILCOXON - *WITHIN*

Quando le misure sono **ripetute** o ***within***, se la variabile **dipendente** è **qualitativa** (ordinale), la statistica più adatta a verificare se ci sono delle variazioni nella distribuzione della variabile tra le due misurazioni è il **test dei segni per ranghi di Wilcoxon** per misure dipendenti.

H₀ within



$$d_i = x_i - y_i \longrightarrow \text{rango}(d_i)$$

La logica è la stessa del t test *within* solo che si basa sui **ranghi**.

TEST WILCOXON - WITHIN

$$T^+ = \sum \text{ranghi}(d_i^+)$$

$$T^- = \sum \text{ranghi}(d_i^-)$$

$$\sum \text{ranghi}(d_i) = \frac{N_c(N_c + 1)}{2}$$

d_i = differenza ($x_i - y_i$)
 T^+ = totale ranghi positivi
 T^- = totale ranghi negativi
 N_c = ampiezza campione corretta
 (esclusi i valori 0)

d_i	rango
-11	-1
12	2.5
-12	-2.5
13	4
0	
18	5

Quando due o più d_i hanno lo stesso valore si assegna il valore corrispondente alla **media dei ranghi**; ad esempio:

$$\text{rango medio} = \frac{(2+3)}{2} = 2.5$$

TEST WILCOXON - WITHIN

Se il campione è grande ($N > 15$) la somma dei ranghi T^+ (ΣT^+) è distribuita in modo approssimativamente normale:

$$\mu_T = \frac{N_c(N_c + 1)}{4}$$

$$\sigma^2_T = \frac{N_c(N_c + 1)(2N_c + 1)}{24}$$

$$z_{T^+} = \frac{T^+ - \mu_T}{\sigma_T} = \frac{T^+ - \frac{N_c(N_c + 1)}{4}}{\sqrt{\frac{N_c(N_c + 1)(2N_c + 1)}{24}}}$$

TEST WILCOXON - WITHIN

La formula per il calcolo della forza dell'effetto r_c (coefficiente di correlazione biseriale tra ranghi appaiati) e r .

$$r_{c^+} = \frac{4(T^+ - \mu_T)}{N(N+1)}$$

$$r = \frac{z}{\sqrt{2 \cdot N_c}}$$

Effect size	r
small	.10
medium	.30
large	.50

TEST WILCOXON - WITHIN

1 coda 2 code	.05	.025	.01	.005	.001
N_c					
5	0				
6	2	0			
7	3	2	0		
8	5	3	1	0	
9	8	5	3	1	
10	10	8	5	3	0
11	13	10	7	5	1
12	17	13	9	7	2
13	21	17	12	9	4
14	25	21	15	12	6
15	30	25	19	15	8

Per la verifica delle ipotesi:

- se $N_c < 15$ (campione piccolo) si utilizza la tavola;
- se $N_c > 15$ (campione grande) si può utilizzare la distribuzione normale standardizzata.

H_1	H_0	$H_0 \rightarrow \min(T^+, T^-) > T_{critico}$	$H_0 _{N_c=11, \alpha=.05} \rightarrow T_{critico} = 13$
-------	-------	--	---

TEST WILCOXON - WITHIN

- (1) raccolta e codifica dei dati (**distr. osservate**);
- (2) inserimento dei dati in una matrice;
- (3) definizione **ipotesi nulla** e **ipotesi alternativa**;
- (4) calcolo della differenza d_i ;
- (5) assegnazione dei ranghi e del segno ai valori d_i ;
- (6) verifica dell'ipotesi in base alla tabella dei valori critici oppure alla distribuzione teorica normale standardizzata;
- (7) se significativo il test, calcolo dell'*effect size*;
- (8) interpretazione dei risultati.

ESEMPIO #4

Su 12 pazienti con sintomi depressivi viene rilevata il livello di depressione **prima** e **dopo** (**VD**, **O**) una psicoterapia breve (**VI**, manipolata, un solo gruppo).



$$H_0 \Rightarrow M\varepsilon_{pre} = M\varepsilon_{post}$$

$$H_1 \Rightarrow M\varepsilon_{pre} > M\varepsilon_{post}$$

$$\alpha = .05$$

**COMPOSTA
MONODIREZIONALE**

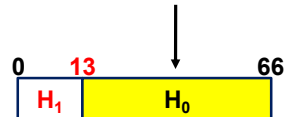
ESEMPIO #4

SS	X_i	Y_i	$d_i (X_i - Y_i)$	$R d_i $	R(+)	R(-)
1	33	38	-5	7		7
2	45	43	+2	2	2	
3	50	42	+8	10	10	
4	45	44	+1	1	1	
5	46	49	-3	3.5		3.5
6	45	41	+4	5.5	5.5	
7	28	22	+6	8	8	
8	43	46	-3	3.5		3.5
9	32	32	0			
10	40	31	+9	11	11	
11	34	27	+7	9	9	
12	40	44	-4	5.5		5.5
TOT				66	46.5	19.5

$$H_0 \rightarrow \min(T^+, T^-) > T_{critico}$$

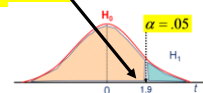
$$T^-_{critico} = 13$$

$$\min(T^+, T^-) = 19.5$$



$$r_{c^+} = +.41 \quad r = +.10$$

$$z_{T^+} = 1.2$$



ESEMPIO #4

Questo risultato ci porta ad accettare l'ipotesi nulla.

$$H_0 \Rightarrow M\varepsilon_{pre} = M\varepsilon_{post}$$

$$H_1 \Rightarrow M\varepsilon_{pre} > M\varepsilon_{post}$$

I risultati non evidenziano una variazione significativa nel livello di depressione dopo il trattamento, $T = 19.5$, $N_c = 11$, $z = 1.2$, $p = .12$, $r = +.1$.



TECNICHE DI ANALISI DEI DATI

AA 2019/2020

PROF. V.P. SENESE

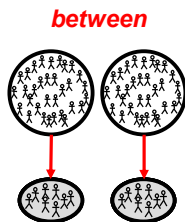
Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:

http://psiclab.altervista.org/TecnAnDat2020/2019_2020.html

Università della Campania «Luigi Vanvitelli» – Dipartimento di Psicologia – TECNICHE DI ANALISI DEI DATI – © Prof. V.P. Senese

T TEST - *BETWEEN*

Quando le misure sono **indipendenti** o **between**, se la variabile **dipendente** è **quantitativa** (intervalli o rapporti), la statistica più adatta a verificare se ci sono delle variazioni nella distribuzione della variabile tra le due misurazioni è il **t test** per misure indipendenti.



Assunzioni:

- (1) la variabile dipendente deve essere distribuita normalmente
- (2) le varianze devono essere omogenee.

T TEST - BETWEEN

Media gruppo 1 e 2

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{1/N_1 + 1/N_2}}$$

$$\sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}$$

stima della varianza ottenuta come
media aritmetica ponderata

Ampiezza gruppo 1 e 2

$$gdl = (N_1 + N_2 - 2)$$

T TEST - BETWEEN

H_0

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2 n_1 + s_2^2 n_2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$gdl = (N_1 + N_2 - 2)$$

T TEST - *BETWEEN*

Le formule per il calcolo della forza dell'effetto **d** o **r**:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

Effect size	d
small	.20
medium	.50
large	.80

$$r = \frac{|d|}{\sqrt{d^2 + \frac{(n_1 + n_2)^2}{n_1 \cdot n_2}}}$$

Effect size	r
small	.10
medium	.30
large	.50

T TEST - *BETWEEN*

- (1) raccolta e codifica dei dati (**distr. osservate**);
- (2) inserimento dei dati in una matrice;
- (3) definizione **ipotesi nulla** e **ipotesi alternativa**;
- (4) calcolo del *t* Test e dei *gdl*;
- (5) verifica dell'ipotesi in base alla distribuzione teorica del *t di Student*;
- (6) se significativo il test, calcolo dell'*effect size*;
- (7) interpretazione dei risultati.

ESEMPIO #5

Uno studente di Psicologia, ha letto che esistono due tipologie di persone in funzione del **locus of control**: i cosiddetti **esterni** e i cosiddetti **interni**; e che le **donne sono generalmente più esterne degli uomini**. Decide allora di verificare se questo fenomeno si manifesta anche tra i suoi amici. Somministra il questionario di Rotter sul LOC a 20 persone, 10 maschi e 10 femmine.

GRUPPO A	<i>M</i>	Test 1 (x_i)
GRUPPO B	<i>F</i>	Test 1 (y_i)

Disegno di ricerca correlazionale con due gruppi a misure indipendenti (*between*).

ESEMPIO #5

GRUPPO 1	<i>M</i>	Test 1 (x_i)
GRUPPO 2	<i>F</i>	Test 1 (y_i)

L'ipotesi generale è che il sesso (**VI** non manipolata, N) influisca sul grado di esternalità (**VD**, I).

$$H_0 \Rightarrow \mu_1 = \mu_2 = \mu$$

$$H_1 \Rightarrow \mu_2 > \mu_1$$

$$\alpha = .05$$

COMPOSTA
MONODIREZIONALE

ESEMPIO #5

Maschi

SS	Sesso	LOC
10	1	13
01	1	12
15	1	12
19	1	12
02	1	11
18	1	11
03	1	10
06	1	10
12	1	10
08	1	09

$N_1 = 10$

$$\bar{x}_1 = 11$$

$$\bar{x}_2 = 12.5$$

$$s_1^2 = 1.7$$

$$s_2^2 = 2.6$$

$$s_1 = 1.3$$

$$s_2 = 1.6$$

Femmine

SS	Sesso	LOC
05	2	15
11	2	14
20	2	14
09	2	13
17	2	13
04	2	12
16	2	12
13	2	11
14	2	11
07	2	10

$N_2 = 10$

ESEMPIO #5

$$\bar{x}_1 = 11$$

$$\bar{x}_2 = 12.5$$

$$s_1^2 = 1.7$$

$$s_2^2 = 2.6$$

$$s_1 = 1.3$$

$$s_2 = 1.6$$

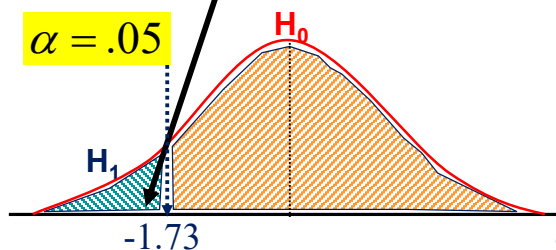
$$gdl = 10 + 10 - 2 = 18$$

$$\sigma = \sqrt{\frac{10(1.7) + 10(2.6)}{10 + 10 - 2}} = 1.5456$$

$$d = -1.05$$

$$r = .465$$

$$t = \frac{11 - 12.5}{1.5456 \sqrt{\frac{1}{10} + \frac{1}{10}}} = -2.1701$$



ESEMPIO #5

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
LOC	Equal variances assumed	.844	.370	-2.355	18	.030	-1.50000	.63683	-2.83794	-.16206
	Equal variances not assumed			-2.355	17.074	.031	-1.50000	.63683	-2.84316	-.15684

Group Statistics

		Sesso	N	Mean	Std. Deviation	Std. Error Mean
LOC	Maschi		10	11.0000	1.24722	.39441
	Femmine		10	12.5000	1.58114	.50000

ESEMPIO #5

Questo risultato ci porta a respingere l'ipotesi nulla e a supportare l'ipotesi alternativa.

$$H_0 \Rightarrow \mu_1 = \mu_2 = \mu$$

$$H_1 \Rightarrow \mu_2 > \mu_1$$

I risultati evidenziano una differenza significativa nel *Locus of control* in funzione del genere, $t(18) = -2.355$, $p = .015$ (una coda), $d = 1.05$. In particolare, i dati evidenziano che le donne ($M = 12.5$) hanno un grado maggiore di esternalità rispetto agli uomini ($M = 11.0$).



TECNICHE DI ANALISI DEI DATI

AA 2019/2020

PROF. V.P. SENESE

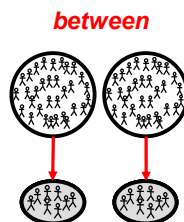
Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:

http://psiclab.altervista.org/TecnAnDat2020/2019_2020.html

Università della Campania «Luigi Vanvitelli» – Dipartimento di Psicologia – TECNICHE DI ANALISI DEI DATI – © Prof. V.P. Senese

TEST U DI MANN-WHITNEY

Se la variabile è misurata su scala Ordinale la statistica più adatta per confrontare due gruppi (due misurazioni indipendenti) è l'applicazione del **test U Mann-Whitney**.



TEST U DI MANN-WHITNEY

Mediante questo *test*, usando come parametro il **rango** (o la **mediana**), è possibile verificare se due campioni provengono dalla medesima popolazione.

$$H_0 \Rightarrow M\varepsilon_{G1} = M\varepsilon_{G2}$$

$$H_1 \Rightarrow M\varepsilon_{G1} \neq M\varepsilon_{G2}$$

$$H_1 \Rightarrow M\varepsilon_{G1} > M\varepsilon_{G2}$$

$$H_1 \Rightarrow M\varepsilon_{G1} < M\varepsilon_{G2}$$

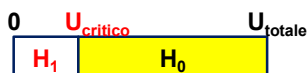
TEST U DI MANN-WHITNEY

$$U_{G1} = \sum ranghi_{G1} - \frac{n_{G1} \cdot (n_{G1} + 1)}{2}$$

$$U_{G2} = \sum ranghi_{G2} - \frac{n_{G2} \cdot (n_{G2} + 1)}{2}$$

Se n_1 o $n_2 < 8$

$$\min(U_{G1}, U_{G2}) \Leftrightarrow U_{critico}$$



Se n_1 e $n_2 > 8$

$$|z| = \frac{U_{Gx} - \frac{n_{G1} \cdot n_{G2}}{2}}{\sqrt{\frac{n_{G1} \cdot n_{G2} \cdot (n_{G1} + n_{G2} + 1)}{12}}}$$

TEST U DI MANN-WHITNEY

La formula per il calcolo della forza dell'effetto r_g (*correlazione rango biseriale*):

$$r_g = \frac{2 \cdot |\bar{R}_{G1} - \bar{R}_{G2}|}{n_{G1} + n_{G2}}$$

Effect size	r
small	.10
medium	.30
large	.50

$$r = \frac{|z|}{\sqrt{n_{G1} + n_{G2}}}$$

TEST U DI MANN-WHITNEY

- (1) raccolta e codifica dei dati (valori osservati);
- (2) inserimento dei dati in una matrice;
- (3) definizione ipotesi nulla e ipotesi alternativa;
- (4) calcolo dei ranghi;
- (5) calcolo del *valore* U_{min} ;
- (6) verifica dell'ipotesi: se n_1 o $n_2 < 8$ si utilizzando le tabelle;
se n_1 o $n_2 > 8$ si utilizza la distribuzione teorica *normale*;
- (7) si interpretano i risultati.

ESEMPIO #6

GRUPPO 1	<i>M</i>	Test 1 (x_j)
GRUPPO 2	<i>F</i>	Test 1 (y_i)

L'ipotesi generale è che il sesso (**VI** non manipolata, N) influisca sul grado di esternalità (**VD**, I → **O**).

$$H_0 \Rightarrow M\varepsilon_M = M\varepsilon_F$$

$$H_1 \Rightarrow M\varepsilon_M < M\varepsilon_F$$

$$\alpha = .05$$

COMPOSTA
MONODIREZIONALE

ESEMPIO #6

$n_1 = 10$

SS	Sesso	LOC	Rango
10	1	13	16
01	1	12	12
15	1	12	12
19	1	12	12
02	1	11	7.5
18	1	11	7.5
03	1	10	3.5
06	1	10	3.5
12	1	10	3.5
08	1	09	1
Tot			78.5

$n_2 = 10$

SS	Sesso	LOC	Rango
05	2	15	20
11	2	14	18.5
20	2	14	18.5
09	2	13	16
17	2	13	16
04	2	12	12
16	2	12	12
13	2	11	7.5
14	2	11	7.5
07	2	10	3.5
Tot			131.5

$$\text{Rango medio}_{10} = \frac{2^\circ + 3^\circ + 4^\circ + 5^\circ}{4} = 3.5$$

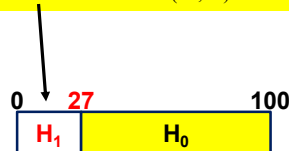
ESEMPIO #6

$$U_{G1} = \sum ranghi_{G1} - \frac{n_{G1} \cdot (n_{G1} + 1)}{2}$$

$$U_M = 78.5 - \frac{10 \cdot 11}{2} = 23.5$$

$$U_F = 131.5 - \frac{10 \cdot 11}{2} = 76.5$$

$$23.5 < U_{critico(10,10)} = 27$$



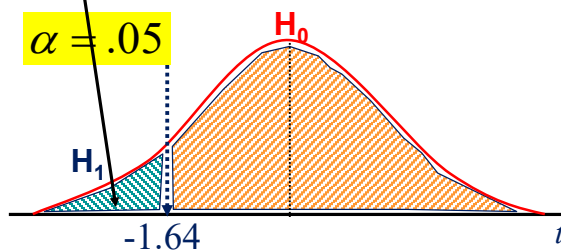
ESEMPIO #6

$$U_M = 78.5 - \frac{10 \cdot 11}{2} = 23.5$$

$$|z| = \frac{23.5 - \frac{10 \cdot 10}{2}}{\sqrt{\frac{10 \cdot 10 \cdot (10 + 10 + 1)}{12}}} = -2.003$$

$$r_g = \frac{2 \cdot |7.85 - 13.15|}{10 + 10} = .53$$

$$r = \frac{|2.003|}{\sqrt{10 + 10}} = .45$$



ESEMPIO #6

Group Statistics

Sesso	N	Mean	Std. Deviation	Std. Error Mean
LOC Maschi	10	11.0000	1.24722	.39441
Femmine	10	12.5000	1.58114	.50000

Ranks

Sesso	N	Mean Rank	Sum of Ranks
LOC Maschi	10	7.85	78.50
Femmine	10	13.15	131.50
Total	20		

Test Statistics^b

	LOC
Mann-Whitney U	23.500
Wilcoxon W	78.500
Z	-2.038
Asymp. Sig. (2-tailed)	.042
Exact Sig. [2*(1-tailed Sig.)]	.043 ^a

a. Not corrected for ties.

b. Grouping Variable: Sesso

ESEMPIO #6

Questo risultato ci porta a respingere l'ipotesi nulla e a supportare l'ipotesi alternativa.

$$H_0 \Rightarrow M\varepsilon_M = M\varepsilon_F$$

$$H_1 \Rightarrow M\varepsilon_M < M\varepsilon_F$$

I risultati evidenziano una differenza significativa e forte nel *Locus of control* in funzione del genere, $z_U = -2.03$, $p = .021$ (una coda), $r_g = .53$. In particolare, i dati evidenziano che le donne ($M = 12.5$) hanno un grado maggiore di esternalità rispetto agli uomini ($M = 11.0$).



TECNICHE DI ANALISI DEI DATI

AA 2019/2020

PROF. V.P. SENESE

Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:

http://psiclab.altervista.org/TecnAnDat2020/2019_2020.html

Università della Campania «Luigi Vanvitelli» – Dipartimento di Psicologia – TECNICHE DI ANALISI DEI DATI – © Prof. V.P. Senese

CHI-QUADRATO

Se la variabile dipendente (**VD**) è misurata su scala **Ordinale** o **Nominale** il test adatto è il **test del Chi-quadrato**, applicato alle frequenze dei due gruppi (due misurazioni indipendenti).

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{o_{ij}} - f_{a_{ij}})^2}{f_{a_{ij}}}$$

r = numero di righe, 1, 2, ..., i
 c = numero di colonne, 1, 2, ..., j
 f_o = frequenza osservata
 f_a = frequenza attesa

$$gdl_{\chi^2} = (r - 1)(c - 1)$$

Assunzioni:

- (1) osservazioni indipendenti;
- (2) nessuna freq. osservata = 0;
- (3) nessuna freq. teorica < 1 e meno del 20% < 5.

ESEMPIO #7

f_o	RICADUTA		
CONSUMO ALCOOL	Sì	No	TOT
Sì	20	13	33
No	48	96	144
TOT	68	109	177

L'ipotesi generale è che tra coloro che consumano alcool hanno una maggiore frequenza di ricaduta nel fumo.

$$H_0 \Rightarrow f_{11} = f_{12} \text{ e } f_{21} = f_{22}$$

$$\alpha = .05$$

$$H_1 \Rightarrow f_{11} \neq f_{12} \text{ e } f_{21} \neq f_{22}$$

ESEMPIO #7

f_o	RICADUTA		
CONSUMO ALCOOL	Sì	No	TOT
Sì	20	13	33
No	48	96	144
TOT	68 (38%)	109 (62%)	177

$$H_0 \Rightarrow f_{11} = f_{12} \text{ e } f_{21} = f_{22}$$

$$p_{11} = \frac{68}{177} = .384$$

In base alle
proporzioni
marginali..

..oppure, in alternativa..

$$f_{a_{11}} = p_{11} \cdot f_{1.} = .384 \cdot 33 = 12.7$$

$$68 : 177 = f_a : 33 \Rightarrow f_a = \frac{68 \cdot 33}{177} = 12.7$$

ESEMPIO #7

f_o	RICADUTA		
CONSUMO ALCOOL	Sì	No	TOT
Sì	20	13	33
No	48	96	144
TOT	68	109	177

$$H_0 \Rightarrow f_{11} = f_{12} \text{ e } f_{21} = f_{22}$$

$$f_{a_{11}} = \frac{68 \cdot 33}{177} = 12.7$$

$$f_{a_{12}} = \frac{109 \cdot 33}{177} = 20.3$$

$$f_{a_{21}} = \frac{68 \cdot 144}{177} = 55.3$$

$$f_{a_{22}} = \frac{109 \cdot 144}{177} = 88.7$$

ESEMPIO #7

	RICADUTA		
CONSUMO ALCOOL	Sì	No	TOT
Sì	20	13	33
No	48	96	144
TOT	68	109	177

	Ricaduta Sì	Ricaduta No
Alcool Sì	12.7	20.3
Alcool No	55.3	88.7

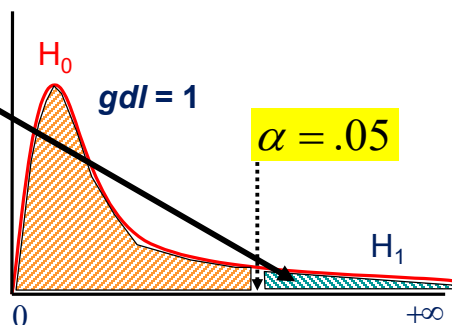
ESEMPIO #7

$$\chi^2 = \frac{(20-12.7)^2}{12.7} + \frac{(13-20.3)^2}{20.3} + \frac{(48-55.3)^2}{55.3} + \frac{(96-88.7)^2}{88.7} =$$

$$= 4.21 + 2.64 + 0.96 + 0.61 =$$

$$\chi^2 = 8.42$$

$$gdl = (2-1)(2-1) = 1$$



ESEMPIO #7

Chi-quadrato

	Valore	df	Sig. asint. (2 vie)	Sig. esatta (2 vie)	Sig. esatta (1 via)
Chi-quadrato di Pearson	8,441 ^a	1	,004		
Correzione di continuit�	7,327	1	,007		
Rapporto di verosimiglianza	8,223	1	,004		
Test esatto di Fisher				,005	,004
Associazione lineare-lineare	8,393	1	,004		
N. di casi validi	177				

a. 0 celle (.0%) hanno un conteggio atteso inferiore a 5. Il conteggio atteso minimo   12.68.

b. Calcolato solo per una tabella 2x2

Tabella di contingenza Alcool * Ricaduta

		Riaduta		Totale	
		1.00 Si	2.00 No		
Alcool	1.00 Si	Conteggio	20	13	33
		Conteggio atteso	12,7	20,3	33,0
		% entro Alcool	60,6%	39,4%	100,0%
		Residui stand.	2,1	-1,6	
2.00 No		Conteggio	48	96	144
		Conteggio atteso	55,3	88,7	144,0
		% entro Alcool	33,3%	66,7%	100,0%
		Residui stand.	-1,0	,8	
Totale		Conteggio	68	109	177
		Conteggio atteso	68,0	109,0	177,0
		% entro Alcool	38,4%	61,6%	100,0%

	Valore	Sig. appross.
V di Cramer	,218	,004
N. di casi validi	177	

ESEMPIO #7

In base ai risultati raccolti e alle analisi effettuate possiamo dire che:

il consumo di alcool è associato alla frequenza di ricaduta nel fumo, $\chi^2(1) = 8.44$, $p = .004$ (due code), $N = 177$, $V = .218$. Infatti, tra coloro che consumano alcool la percentuale di ricadute nel fumo è del **61%**, $R = +2.01$; $p < .05$, mentre tra coloro che non consumano alcool la percentuale delle recidive scende al **33%**.