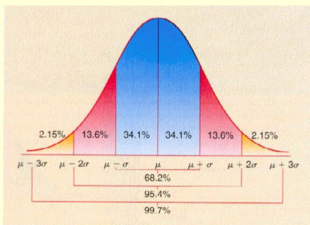
Test statistici non-parametrici

Il test t di Student e l'ANOVA sono basati su alcune assunzioni…

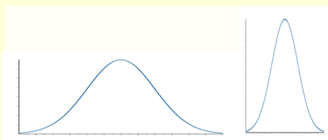
1. Variabili continue o almeno misurate in un intervallo (es. non conosco il valore assoluto, ma posso quantificare le differenze fra due valori)

2. Indipendenza fra media e varianza (l'errore di misura deve essere indipendente dal valore misurato)

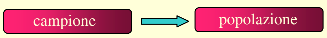
3. Variabili distribuite in modo (approssimativamente) normale



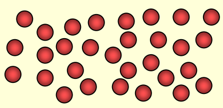
4. Omogeneità delle varianze



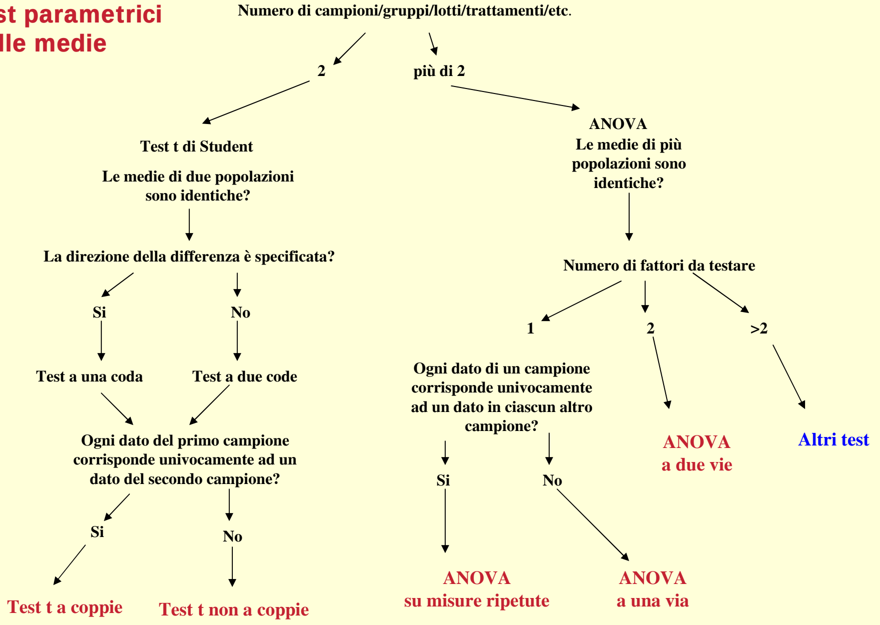
5. I risultati ottenuti con l'analisi di campioni si applicano alle popolazioni



6. Dimensione campione > 10 (meglio se ≥ 30)



Una chiave per i test parametrici sulle medie



Se queste assunzioni (una o più sono violate)…

Assunzione	Altri test?	Rimedi?
1. Variabile non continua	Si	
2. Indipendenza media-varianza	No	Migliori metodi di misura
3. Distribuzione non normale	Si	Trasformazione dei dati
4. Varianze disomogenee	Si	
5. Campione $\neq$ popolazione	Si	
6. n<10	Si	Raccogliere più dati

Test non-parametrici

- Questi test si impiegano quando almeno una delle assunzioni alla base del test t di Student o dell'ANOVA è violata.
- Sono chiamati "non-parametrici" perchè essi non implicano la stima di parametri statistici (media, deviazione standard, varianza, etc.).

Ne esistono almeno due grandi categorie:

1) Test di conformità (confronto fra valori osservati e valori attesi opportunamente calcolati)
2) Test equivalenti di test parametrici

Un esempio di test di conformità

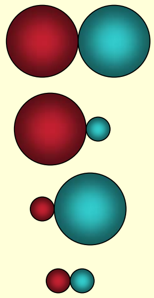
Frequenza dei fenotipi di *Bipalla rotunda*

A_B_

A_bb

aaB_

aabb



rapporto 9:3:3:1 (secondo la 3.a legge di Mendel)

- 320 individui campionati

	A_B_	A_bb	aaB_	aabb
osservato (o)	194	53	67	6

- 320 individui campionati

	A_B_	A_bb	aaB_	aabb
osservato (o)	194	53	67	6
atteso (e)	180	60	60	20

- 320 individui campionati

	A_B_	A_bb	aaB_	aabb
osservato (o)	194	53	67	6
atteso (e)	180	60	60	20
o - e	14	-7	7	-14

- 320 individui campionati

	A_B_	A_bb	aaB_	aabb
osservato (o)	194	53	67	6
atteso (e)	180	60	60	20
o - e	14	-7	7	-14
$(o - e)^2$	196	49	49	196

- 320 individui campionati

	A_B_	A_bb	aaB_	aabb
osservato (o)	194	53	67	6
atteso (e)	180	60	60	20
o - e	14	-7	7	-14
$(o - e)^2$	196	49	49	196
$\frac{(o - e)^2}{e}$	1.08	.82	.82	9.8

- 320 individui campionati

	A_B_	A_bb	aaB_	aabb
osservato (o)	194	53	67	6
atteso (e)	180	60	60	20
o - e	14	-7	7	-14
$(o - e)^2$	196	49	49	196
$\frac{(o - e)^2}{e}$	1.08	.82	.82	9.8

$$X^2 = \sum \frac{(o - e)^2}{e} = 1.08 + .82 + .82 + 9.8 = 12.52$$

GdL = numero di fenotipi -1 = 3

$X^2 = 12.52$ Il valore critico per 3 gradi di libertà al livello .05 è 7.82

Tavola di X^2

	Significance level					
df	0.20	0.10	0.05	0.025	0.01	0.001
1	1.64	2.71	3.84	5.02	6.64	10.83
2	3.22	4.61	5.99	7.38	9.21	13.82
3	4.64	6.25	7.82	9.35	11.34	16.27
4	5.99	7.78	9.49	11.14	13.28	18.47
5	7.29	9.24	11.07	12.83	15.09	20.52
6	8.56	10.64	12.59	14.45	16.81	22.46
7	9.80	12.02	14.07	16.01	18.48	24.32
8	11.03	13.36	15.51	17.53	20.09	26.12
9	12.24	14.68	16.92	19.02	21.67	27.88
10	13.44	15.99	18.31	20.48	23.21	29.59

La vera probabilità di $X^2 = 12.52$ e GdL = 3 è $.01 > p > .001$

Lo scarto fra frequenze osservate dei fenotipi e frequenze previste in base alla 3.a legge di Mendel è tale che la probabilità di osservare scarti ancora maggiori è molto piccola (<0.01). Quindi verosimilmente essi non sono stati estratti da una popolazione mendeliana.

Un caso particolare: la correzione di Yates

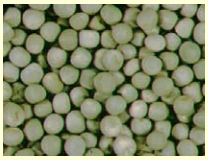
La formula per il calcolo del X^2 è:

$$X^2 = \sum \frac{(o - e)^2}{e}$$

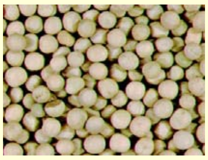
Però, nel caso in cui GdL= 1 (cioè se le categorie di dati sono solo due) la formula diventa:

$$X^2 = \sum \frac{(|o - e| - 0.5)^2}{e}$$

Mendel e i piselli: un test di eterogeneità



Verdi (g)



Gialli (Y)

Da un incrocio Yg x Yg il rapporto atteso di Y_ rispetto a gg è di 3:1

H_0: il rapporto è 3:1
H_1: il rapporto <u>non</u> è 3:1

L'esperimento viene ripetuto 10 volte…

Esperimento	Gialli	Verdi	n	X^2	GdL
1	25	11	36	0.33	1
2	32	7	39	.69	1
3	14	5	19	.02	1
4	70	27	97	.28	1
5	24	13	37	1.52	1
6	20	6	26	0	1
7	32	13	45	.46	1
8	44	9	53	1.42	1
9	50	14	64	.19	1
10	44	18	62	.34	1
Totali	355	117	472	5.25	10

Problema: i dati dei singoli esperimenti possono essere combinati fra loro?

Per essere combinati, essi devono essere stati estratti dalla medesima popolazione. Quindi:

H_0: i dati sono stati estratti dalla medesima popolazione.

(dalla slide precedente)

Esperimenti	**Gialli**	**Verdi**	**n**	**X^2**	**GdL**
Tutti	**355**	**117**	**472**	**0.003**	**1**

Totale dei X^2 = 5.25, GdL = 10 X^2 dei totali = .003, GdL = 1

X^2 = Totale dei X^2 - X^2 dei totali = 5.25 - .003 = 5.247 e GdL = 9

Per $X^2 = 5.247$ e GdL=9, $p \approx 0.81$

Conclusioni: i dati sono estratti dalla stessa popolazione e quindi possono essere combinati.

Test di McNemar

- Confronto di due campioni non indipendenti
- Si usa per variabili nominali rilevate più volte sugli stessi individui

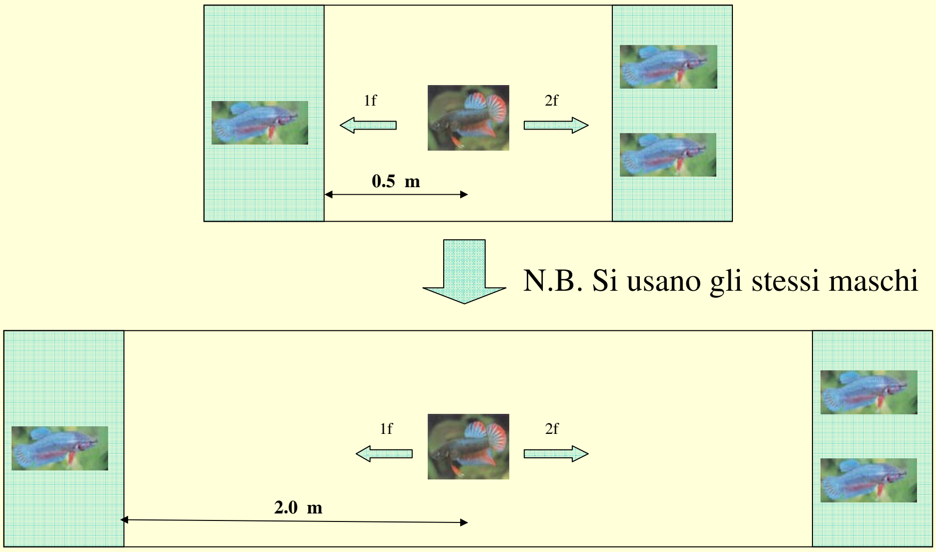
Esempio: la scelta di una o due femmine in *Betta splendens* (pesce combattente del Siam)

♂

Problema: la distanza a cui un maschio si sposta varia con il numero delle femmine disponibili?

♀

Disegno sperimentale



Risultati

Vasca 0.5 m

Vasca 2 m

	1f	2f	
1f	8	5	13
2f	9	8	17
	17	13	30

Vasca 0.5 m

Vasca 2 m

	1f	2f	
1f	a	b	a+b
2f	c	d	c+d
	a+c	b+d	30

Risultati

Vasca 0.5 m

Vasca 2 m		**1f**	**2f**	
	1f	**8**	**5**	**13**
	2f	**9**	**8**	**17**
		17	**13**	**30**

Vasca 0.5 m

Vasca 2 m		**1f**	**2f**	
	1f	**a**	**b**	**a+b**
	2f	**c**	**d**	**c+d**
		a+c	**b+d**	**30**

Cosa significano questi dati?

5 animali scelgono 2f (vanno verso 2 femmine) nella vasca piccola, ma preferiscono 1f (andare verso 1 femmina) nella vasca grande

H_0: la frequenza degli animali che cambiano risposta da 1f a 2f è la stessa di quelli che cambiano da 2f a 1f

ovvero H_0: $b - c = 0$

Test di McNemar: calcoli

$$X^2 = \frac{(b-c)^2}{b+c}$$

Se b+c < 200, si introduce un fattore di correzione:

$$X^2 = \frac{(|b-c| - 1)^2}{b+c} = \frac{(|5-9| - 1)^2}{5+9} = .6429$$

$X^2_{(.05,\ df = 1)} = 3.84$

Poiché $.6429 << 3.84$, si accetta H_0

Test esatto di Fisher

Si usa su dati nominali, con due campioni indipendenti

Esempio: una serie di misure del numero di Emitteri e di Coleotteri sulle facce superiori ed inferiori di foglie

L'Emittero *Lygus lineolaris*

Il Coleottero *Altica sylvia*

H_0: la frequenza di Emitteri e Coleotteri è indipendente dalla faccia delle foglie.

Dati sperimentali

	Emitteri	Coleotteri	Totali
Faccia superiore	**12(a)**	**7(b)**	**19 (a+b)**
Faccia inferiore	**2(c)**	**8(d)**	**10(c+d)**
Totali	**14(a+c)**	**15(b+d)**	**N=29**

$$p = \frac{\dfrac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{n!}}{a!b!c!d!} = \frac{\dfrac{(19)!\,(10)!\,(14)!\,(18)!}{29!}}{12!7!2!8!} = .02923$$

p = .02923 (cioè < .05), quindi si rigetta H_0

N.B. Questo test consente il calcolo diretto di p.

Test Q di Cochran

Si usa con dati nominali e con più di due campioni non indipendenti (misure ripetute)

Effetto dell'abbigliamento sulle punture di zanzara (pungono=1, non pungono=0)

	Leggero, comodo	Leggero, aderente	Scuro, lungo	Scuro, corto	Non punto	Totale punture
Soggetto						
1	0	0	0	1	0	1
2	*1*	*1*	*1*	*1*	*1*	*5*
3	0	0	0	1	1	2
4	1	0	0	1	1	3
5	0	1	1	1	1	4
6	0	0	0	1	1	2
7	0	1	1	0	1	3
8	0	1	1	1	0	2
G_i	1	3	3	6	4	17

a = 5 (numero di trattamenti)
b = 7 (numero di soggetti - N.B. il soggetto 2 è stato escluso perché la risposta è sempre positiva)

Test Q di Cochran: calcoli

H_0: le punture non dipendono dall'abbigliamento.

$$Q = \frac{a-1\left[\Sigma G_i^2 - \frac{(\Sigma G_i)^2}{a}\right]}{\Sigma b^2 - \frac{(\Sigma b)^2}{a}} = 6.947 = X^2$$

$X^2 = 6.947 < X^2_{\text{crit (.05, 4 GdL)}} = 9.488$, quindi $p > .05$ (p=0.14)

Quindi si accetta l'ipotesi nulla H_0

In conclusione…

Tipo di dati	Numero di campioni	Dati indipendenti?	Test da usare
Nominali	2	No	McNemar
Nominali	2	Si	Esatto di Fisher
Nominali	>2	No	Q di Cochran

Test di Tukey-Duckworth

- Il test di Tukey-Duckworth è uno dei test statistici più semplici da applicare
- E’ cosi’ semplice che praticamente non richiede calcoli
- Ovviamente ha dei limiti operativi e non è altrettanto “potente” quanto altri test non-parametrici o parametrici

Test di Tukey-Duckworth

- Il numero dei dati nei due campioni deve essere: $4 <= n_1 <= n_2 <= 30$
- H_o: i campioni sono identici
- H_a: i campioni differiscono fra loro
- La statistica da calcolare è C
- Il test esiste solo nella forma a due code
- Ci sono solo due valori critici:
  $C_{0.05} = 7$
  $C_{0.01} = 10$

Test di Tukey-Duckworth

1. Si determinano il valore massimo e quello minimo assoluto riferiti ad entrambi i campioni.
2. Per il campione che contiene il valore massimo assoluto si contano i dati il cui valore è maggiore del massimo (relativo) dell’altro campione.
3. Per il campione che non contiene il massimo assoluto, si contano i valori che sono più piccoli del minimo valore dell’altro campione.
4. La statistica C è la somma delle due conte.

Test di Tukey-Duckworth

Altezze in piedi delle palme in due quadrati campione

sito

A	B
80	86
82	87
83	90
84	91
85	91
86	92
87	93
89	93
92	95
93	96
94	98
96	99
	101
	103

$C = \boxed{5} + \boxed{4} = 9$

$C_{0.05} = 7$
$C_{0.01} = 10$

$C > C_{0.05}$

Si rigetta H_0

Test di Komolgorov-Smirnov (KS)

Un campione, dati ordinali (ranghi)

Esperimento:
preferenza per l'umidità di porcellini di terra (Isopoda, Porcellionidae)

Si dà una scelta fra vari livelli di umidità (da 1 a 5)

1 umido ↔ 5 secco

H_0: nessuna preferenza per un particolare livello di umidità
H_1: preferenza per un particolare livello di umidità

I dati…

Classe di umidità	1	2	3	4	5	
fi	2	18	10	4	1	Frequenza osservata
$\hat{f}_i$	7	7	7	7	7	Frequenza attesa
F_i	2	20	30	34	35	Freq. oss. cumulativa
$\hat{F}_i$	7	14	21	28	35	Freq. att. cumulativa
$\lvert d_i \rvert$	5	6	9	6	0	Valore assoluto della differenza

Statistica di Kolmogorov-Smirnov: $d_{max} = 9$

Valore critico: $d_{max(5, 35)} = 7$ **Quindi, si rigetta** H_0

Test di Wilcoxon

Due campioni non indipendenti, dati ordinali

Il test di Wilcoxon dovrebbe essere usato come alternativa non-parametrica al t di Student per campioni non indipendenti se una qualsiasi delle assunzioni necessarie per quest’ultimo è violata.

Test di Wilcoxon

Esperimento
Misura del tempo per cui si nutrono degli uccelli, come numero di minuti di attività nella mattina e nel pomeriggio

Uccello	Mattina	Pomeriggio	Differenza	Rango |differenza|	Rango con segno
1	23	46	17	4	4
2	28	51	23	7	7
3	37	29	-8	2	-2
4	24	49	25	8	8
6	27	46	19	5	5
6	27	39	22	6	6
7	31	30	-1	1	-1
8	28	41	13	3	3

H_0: non c'è differenza fra mattina e pomeriggio
H_1: esiste una differenza fra mattina e pomeriggio

Test di Wilcoxon: calcoli

Uccello	Mattina	Pomeriggio	Differenza	Rango |differenza|	Rango con segno
1	23	46	17	4	4
2	28	51	23	7	7
3	37	29	-8	2	-2
4	24	49	25	8	8
6	27	46	19	5	5
6	27	39	22	6	6
7	31	30	-1	1	-1
8	28	41	13	3	3

Somma dei ranghi positivi: $T_+ = 4+6+8+7+5+3 = 33$
Somma dei ranghi negativi: $T_- = 2+1=3$

Si rigetta H_0 se T_+ o $T_- \leq$ valore critico tabulare
In questo caso, poichè $T_{(.05,\ n=8)} = 3$, si rigetta H_0

Test U di Mann-Whitney

Due campioni indipendenti, dati ordinali

Il test U di Mann-Whitney dovrebbe essere usato come alternativa non-parametrica ad un test t di Student su campioni indipendenti, se una qualsiasi delle assunzioni necessarie è violata.

Test U di Mann-Whitney

Esperimento
Distanze al vicino più prossimo fra Nudibranchi in due quadrati campione

Quadrato 1	Quadrato 2
193	175
188	173
185	168
183	165
180	163
178	
170	

H_0: <u>non</u> c'è differenza fra i quadrati nella distanza al vicino più prossimo
H_1: c'è differenza fra i quadrati nella distanza al vicino più prossimo

Test U di Mann-Whitney: calcoli

Dati ordinati
193
188
185
183
180
178
175
173
170
168
165
163

Quadrato 1	Quadrato 2	Ranghi quadrato 1	Ranghi quadrato 2
193	175	1	7
188	173	2	8
185	168	3	10
183	165	4	11
180	163	5	12
178		6	
170		9	
$n_1 = 7$	$n_2 = 5$	$\Sigma R_1 = 30$	$\Sigma R_2 = 48$

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \Sigma R_1 = 7 \cdot 5 + \frac{7 \cdot 8}{2} - 30 = 33$$

$$U' = n_1 n_2 - U = 7 \cdot 5 - 33 = 2$$

Se U o U' $\geq U_{crit(.05,\ 7,\ 5)}$, si rigetta H_0

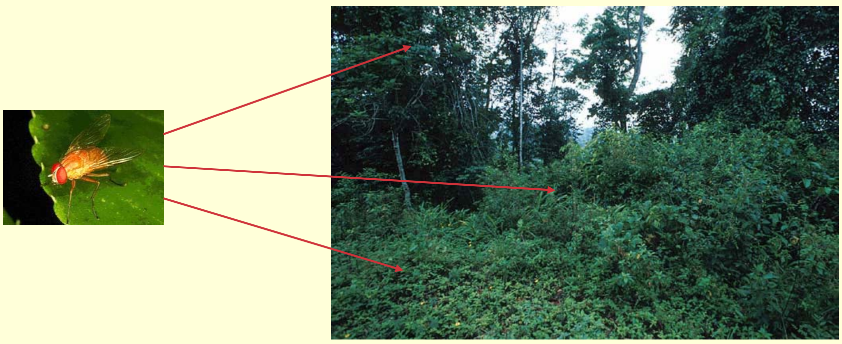
Poichè $U_{crit(.05,\ 7,\ 5)} = 30$ e U=33> 30, si rigetta H_0

Test di Kruskal-Wallis

Un analogo dell'ANOVA a una via da usare quando le assunzioni necessarie per quest'ultima sono violate.

Esperimento

Si studia la distribuzione verticale delle mosche nella vegetazione.



Test di Kruskal-Wallis

Numero di mosche /m²

Vegetazione erbacea	Arbusti	Alberi
14	8.4	6.9
12.1	5.1	7.3
9.6	5.5	5.8
8.2	6.6	4.1
10.2	6.3	5.4

H_0: la distribuzione delle mosche è omogenea fra strati
H_1: la distribuzione delle mosche non è omogenea fra strati

Test di Kruskal-Wallis: calcoli

Si calcolano i ranghi dei dati

Vegetazione erbacea	Arbusti	Alberi
14 (15)	8.4 (11)	6.9 (8)
12.1 (14)	5.1 (2)	7.3 (9)
9.6 (12)	5.5 (4)	5.8 (5)
8.2 (10)	6.6 (7)	4.1 (1)
10.2 (13)	6.3 (6)	5.4 (3)

$n_1 = 5$ $R_1 = 64$ **$n_2 = 5$ $R_2 = 30$** **$n_3 = 5$ $R_3 = 26$**

N = 15

Test di Kruskal-Wallis: calcoli

Si calcolano i ranghi dei dati

Vegetazione erbacea	Arbusti	Alberi
14 (15)	8.4 (11)	6.9 (8)
12.1 (14)	5.1 (2)	7.3 (9)
9.6 (12)	5.5 (4)	5.8 (5)
8.2 (10)	6.6 (7)	4.1 (1)
10.2 (13)	6.3 (6)	5.4 (3)

$n_1 = 5$ $R_1 = 64$ **$n_2 = 5$ $R_2 = 30$** **$n_3 = 5$ $R_3 = 26$**

N = 15

$$H = \frac{12}{N(N+1)}\left[\sum \frac{R_i^2}{n_i}\right] - 3(N+1) = \frac{12}{15 \cdot 16}\left[\frac{64^2}{5} + \frac{30^2}{5} + \frac{26^2}{5}\right] - 3 \cdot 16 = 8.72$$

per n =5, 5, 5 si ha $H_{crit(0.05)} = 5.78$ Poichè $H=8.72>H_{crit(0.05)}=5.78$, si rigetta H_0

Test di Friedman

Analogo non-parametrico dell'ANOVA a due vie

Esperimento
Accrescimento di cavie in funzione della dieta

	Diete 1	2	3	4
Lotti 1	1.5	2.7	2.1	1.3
2	1.4	2.9	2.2	1.0
3	1.4	2.1	2.4	1.1
4	1.2	3.0	2.0	1.3
5	1.4	3.3	2.5	1.5

Test di Friedman

Analogo non-parametrico dell'ANOVA a due vie

Esperimento
Accrescimento di cavie in funzione della dieta

	Diete 1	2	3	4
Lotti				
1	1.5	2.7	2.1	1.3
	2	4	3	1
2	1.4	2.9	2.2	1.0
	2	4	3	1
3	1.4	2.1	2.4	1.1
	2	4	3	1
4	1.2	3.0	2.0	1.3
	1	4	3	2
5	1.4	3.3	2.5	1.5
	1	4	3	2
R_i (somma dei ranghi)	8	19	16	7

Ranghi calcolati all'interno di ciascun lotto

Test di Friedman: calcoli

$$X^2 = \frac{12}{ba(a+1)} \Sigma R_i^2 - 3b(a+1) = 12.6$$

Dove: a = numero delle diete
b = numero dei lotti

H_0: l'accrescimento è omogeneo fra diete e fra lotti
H_1: l'accrescimento <u>non</u> è omogeneo fra diete e fra lotti

Il valore critico per p=0.05 è $X^2_{(0.05,\ GdL=3)} = 7.815$

Poichè $X^2 > X^2_{(0.05,3)}$, si rigetta H_0

Sommario dei test presentati

Tipo di dati	Numero di campioni	Campioni indipendenti?	Test da usare
Nominali	2	No	McNemar
Nominali	2	Si	Fisher's Exact
Nominali	>2	No	Cochran's Q
Ordinali	1	-	Komolgorov- Smirnov
Ordinali	2	Si	Tukey-Duckworth
Ordinali	2	No	Wilcoxon (analogo del t-test per campioni non indipendenti)
Ordinali	2	Si	Mann-Whitney U (analogo del t-test per campioni indipendenti)
Ordinali	>2	Si	Kruskal-Wallis (analogo dell'ANOVA a una via)
Ordinali	>2	No	Friedman (ANOVA a due vie)

- I test statistici si imparano con la pratica. Quando serve applicarli, ci aiutano i libri ed il software dedicato.
- Preparare i dati e fare calcoli di base è invece un'abilità primaria, la dovete avere sempre con voi.
- Quindi, saper usare Excel o un altro foglio di calcolo è fondamentale.

Excel in un pillola:

http://www.salvatorepagano.brianzaest.it/corsi/moduli_fortic/Excel%20Fundamentals%20(Antonio%20Potenza).htm

Una lista di tutorial di base ed avanzati:

http://www.salvatorepagano.brianzaest.it/corsi/sommario.htm

N.B. Ben fatti quelli del Ministero del Tesoro in formato PDF