NOTE DALLE LEZIONI DI STATISTICA MEDICA ED ESERCIZI

CONFRONTO DI PIU' MEDIE

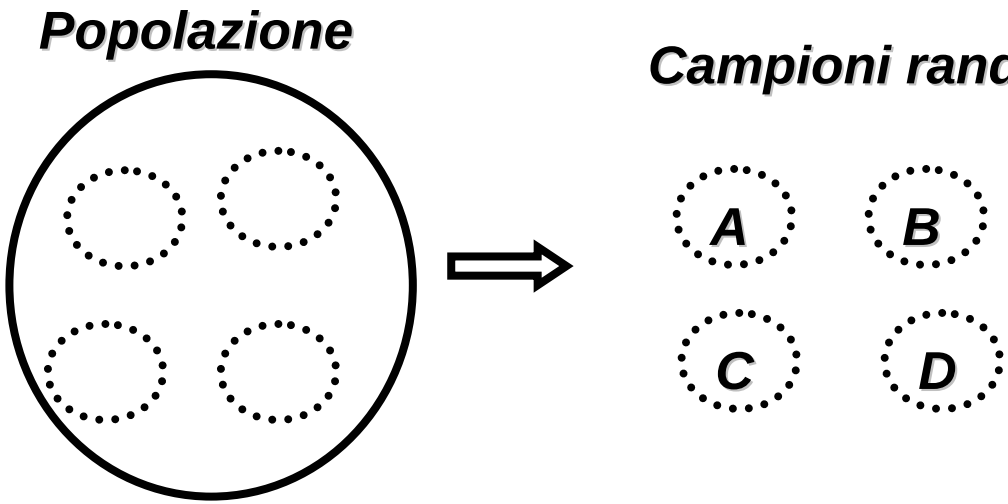
IL METODO DI

ANALISI DELLA VARIANZA

IL PROBLEMA

Supponiamo di voler studiare l'effetto di 4 diverse diete su un campione casuale di 24 cavie rilevando per ciascuna cavia il peso in gr..
Suddividiamo le 24 cavie in 4 gruppi, quanti sono i trattamenti, in modo casuale (random)

	Trattamenti			
Repliche	**A**	**B**	**C**	**D**
1	716	634	668	728
2	761	676	723	786
3	684	678	710	754
4	792	682	762	789
5	838	726	744	836
6	774	694	732	842



Dati

Si dispone del peso in grammi delle 24 cavie che sono assegnate casualmente ai 4 diversi trattamenti.

le osservazioni relative a ciascun trattamento, le possiamo indicare genericamente y_{ij}.

Con **i=1,2,3,…,k gruppi**
j=1,2,3,4,…,n_i osservazioni

Assunzioni

Distribuzione della variabile : Gauss.
Campioni indipendenti.
Varianze omogenee.

Ipotesi

$$\begin{cases} H_0: & \mu_a=\mu_b=\mu_c=\mu_d=\ldots=\mu_k=\mu \\ H_1: & \mu_r\neq\mu_s \end{cases}$$

L'ipotesi da saggiare è che tutti i trattamenti siano uguali contro un ipotesi alternativa che almeno due siano diversi tra loro.

Repliche	1	2	…	i	…	k	
1	y_{11}	y_{12}	…	y_{1i}	…	y_{1k}	
2	y_{21}	y_{22}	…	y_{2i}	…	y_{1k}	
…	…	…	…		…	…	
j	y_{j1}	y_{j2}	…	y_{ij}	…	y_{jk}	
…	…	…	…	…	…	…	
N_i	y_{ni1}	y_{ni2}	…	y_{nii}	…	y_{nik}	
$\Sigma y_{ij}=T_i$	T_1	T_2	…	T_i	…	T_k	T
$\bar{y}_i$	$\bar{y}_1$	$\bar{y}_2$	…	$\bar{y}_i$	…	$\bar{y}_k$	$\bar{y}$
T^2_i/n_i	T^2_1/n_1	T^2_2/n_2	…	T^2_i/n_i	…	T^2_k/n_k	$\Sigma (T^2_i/n_i)$
$\Sigma y^2_{ij}=S_i$	S_1	S_2	…	S_i	…	S_k	S

RIEMPITE LA TABELLA PRECEDENTE CON I DATI DEL PROBLEMA E CALCOLATE LE DIFFERENTI QUANTITA'

Nel nostro esempio la tabella precedente diventa:

	Trattamento				
Repliche	A	B	C	D	
1	716	634	668	728	
2	761	676	723	786	
3	684	678	710	754	
4	792	682	762	789	
5	838	726	744	836	
6	774	694	732	842	
$\Sigma y_{ij}=T_i$	4565	4090	4339	4735	T=17729
$\bar{y}_i$	760,8	681,7	723,2	789,2	$\bar{y}$ =738,7
T^2_i/n_i	3475204	2788017	3137820	3736704	$\Sigma(T^2_i/n_i)$ =13137745
$S_i=\Sigma y^2_{ij}$	3488217	2792452	3143057	3746677	S = 13170403

Costruzione della Statistica test

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y} \right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i \right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{y}_i - \bar{y} \right)^2$$

devianza totale = devianza entro gruppi + dev. tra gruppi.

G.I. N-1 N-k k-1

Con i=1,2,3,…,k gruppi
j=1,2,3,4,…,ni osservazioni

La statistica test consiste nel valutare quanta parte della variabilità totale è attribuibile alla differenza tra i trattamenti.

☞**La statistica test sarà :**

$$F = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2 \Big/ k-1}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \Big/ N-k} = \frac{\text{varianza tra gruppi (g.l. k - 1)}}{\text{varianza entro gruppi (g.l. N - k)}}$$

Distribuzione della statistica test

La distribuzione della statistica test è F-Fisher, dipende dai gradi di libertà del numeratore (k-1) e del denominatore (n-K)

Regola di decisione

Fisso α (livello di significatività, errore di I tipo) accettabilmente basso (0,05).

e in corrispondenza dei gradi di libertà del numeratore e del denominatore si determina un valore tabulato che delimita la zona di rifiuto.

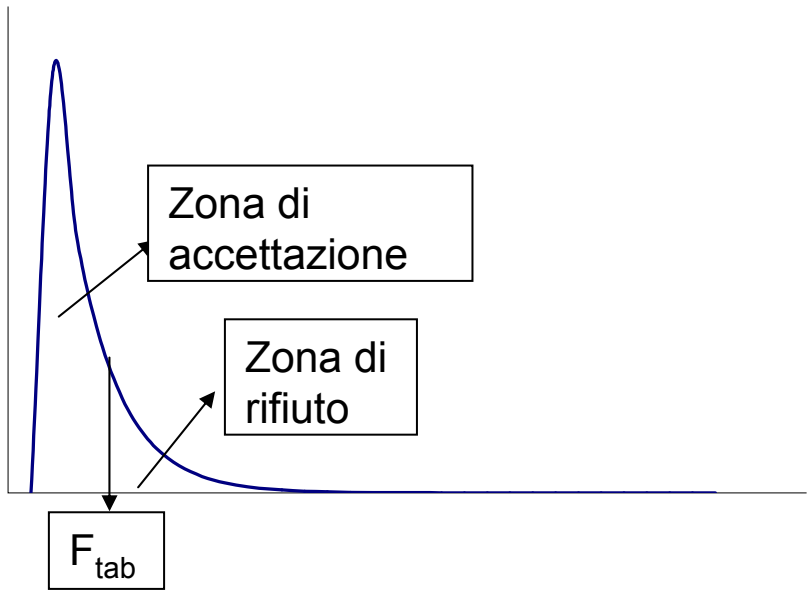


Tavola per il calcolo dell'analisi della varianza (ANOVA)

Sorgenti di variazione	Devianze	Gradi di libertà	Varianza	Fcalcolato
Trattamento	**Dev. tra gruppi**	**k-1**	**S_f=Dev. tra gruppi/k-1**	**$F=S_f / S_e$**
Errore	**Dev. entro gruppi**	**N-k**	**S_e=Dev. entro gruppi/N-k**	
Totale	**Dev. totale**	**N-1**	**Dev. tot/N-1**	

$$SSqTOT = \sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j y_{ij}^2 - \frac{\left(\sum_i \sum_j y_{ij}\right)^2}{N} = S - \frac{T^2}{N}$$

$$SSqENTRO = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 = \sum_i \sum_j y_{ij}^2 - \sum_i \frac{\left(\sum_j y_{ij}\right)^2}{n_i} = S - \sum_i \frac{T_i^2}{n_i}$$

$$SSqTRA = \sum_i \sum_j (\bar{y}_i - \bar{y})^2 = \sum_i \frac{\left(\sum_j y_{ij}\right)^2}{n_i} - \frac{\left(\sum_i \sum_j y_{ij}\right)^2}{N} = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

$S = \Sigma\Sigma y_{ij}^2$ $T = \Sigma\Sigma\, y_{ij}$ $T_i = \Sigma y_{ij}$ N = num. delle osservazioni

UTILIZZATE
LE FORMULE PRECEDENTI
E CALCOLATE
LE DIFFERENTI DEVIANZE
NONCHE'
LA TAVOLA DI
ANALISI DELLA VARIANZA

Applicando le formule precedenti si ottiene:

devianza totale = $S - (T^2 / N)$=13170403-13096560= 73843

devianza entro gruppi = $S - \Sigma(T^2_i / n_i)$=13170403-13135745=34658

devianza tra gruppi = $\Sigma (T^2_i / n_i) - (T^2 / N)$=13135745-13096560=39185

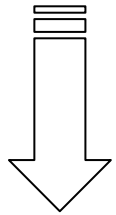
Da cui:

TAVOLA DI ANALISI DELLA VARIANZA

Sorgenti	Devianze	Gradi di libertà	Varianze	F
Tra gruppi	39185	3	13061,67	7,54
Entro gruppi	34658	20	1732,90	
Totale	73843	23		

Conclusioni

Dato che il valore calcolato F = 7,54
è maggiore del valore tabulato
per α=0,05 $F_{3,20} = 3,10$
allora posso rifiutare H_0



Almeno due diete sono differenti tra loro.

TEST DI BARTLETT
PER L'OMOGENEITA' DELLE VARIANZE

Dati k gruppi di osservazioni, le corrispondenti varianze devono tutte risultare stime della stessa varianza incognita della popolazione, e in base a questo presupposto la varianza di errore può essere ottenuta come media ponderata delle k varianze stimate in ciascun gruppo

$$H_0: \quad \sigma_1^2 = \sigma_2^2 = \ldots\ldots = \sigma_k^2 = \sigma^2$$

contro l'alternativa che almeno uno dei valori σ_i^2 risulti diverso dai rimanenti k-1

La statistica test

$$\frac{A}{B} = \frac{2.3026 \bullet \left[\left(\sum n_i - k \right) \bullet \lg_{10} S^2 - \sum_i (n_i - 1) \bullet \lg_{10} S_i^2 \right]}{1 + \frac{\sum_i \frac{1}{n_i - 1} - \frac{1}{\sum_i n_i - k}}{3(k-1)}}$$

La distribuzione della statistica test è $\chi 2$ con k-1 gradi di libertà
S^2 è la varianza residua e S^2_i è la varianza del gruppo i-esimo

Si rifiuta l'ipotesi se il valore calcolato di A/B risulta maggiore del valore tabulato $\chi 2$ con k-1 gradi di libertà

$$\frac{A}{B}=\frac{2.3026\bullet\left[\left(\sum n_i-k\right)\bullet\lg_{10}S^2-\sum(n_i-1)\bullet\lg_{10}S_i^2\right]}{1+\dfrac{\sum_i\dfrac{1}{n_i-1}-\dfrac{1}{\sum_i n_i-k}}{3(k-1)}}$$

$$\left(\sum n_i-k\right)=20 \qquad \lg_{10}S^2=\lg_{10}1632{,}9=3{,}21$$

$$\sum(n_i-1)\bullet\lg_{10}S_i^2=17{,}07+14{,}73+15{,}10+16{,}49=63{,}4$$

$$1+\frac{\sum_i\dfrac{1}{n_i-1}-\dfrac{1}{\sum_i n_i-k}}{3(k-1)}=1+\frac{\sum_i\dfrac{1}{6-1}-\dfrac{1}{24-4}}{3(4-1)}=1{,}049$$

$$\frac{A}{B}=\frac{2{,}3026\bullet\left[20\bullet3{,}21-63{,}4\right]}{1{,}049}=1{,}79$$

$\chi 2$ con 3 gradi di libertà = 7,81>A/B=1,79 $\Rightarrow$Varianze omogenee

CONFRONTI MULTIPLI

Vanno effettuati quando l'ANOVA ha portato al rifiuto dell'ipotesi nulla

Nella valutazione della significatività bisogna stare attenti poichè questi tests possono comportare una distorsione dell'errore di I tipo così come della potenza del test

Procedura LSD di Fisher

esegue tutti i possibili test t

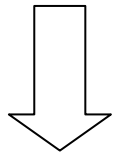
(Least Significant Difference = Minima differenza significativa)

Se eseguita solo nel caso in cui l'ipotesi nulla su tutte le medie viene rifiutata risulta piuttosto efficiente nel mantenere un ragionevole grado di controllo sui falsi errori.

$$LSD = t_{\alpha/2} \sqrt{(2\ Var\ Residua/n)}$$

$t_{\alpha/2}$ Valore della t-Student con gradi di libertà N-k

- Si ordinano le medie dalla più piccola alla più grande
- Si confronta la media più grande con la più piccola
  - calcolando la differenza
  - confrontando la differenza ottenuta con il valore LSD



Se la differenza supera il valore LSD
si conclude che le due medie sono diverse

Si procede confrontando la più grande con la seconda più piccola e via di seguito

Nessuna coppia di medie può essere dichiarata significativamente diversa, se si trova all'interno di un'altra coppia già dichiarata non differente

UTILIZZATE
LE FORMULE PRECEDENTI
E CALCOLATE
LA MINIMA DIFFERENZA SIGNIFICATIVA (LSD)
DETERMINATE LA DIFFERENZA TRA LE MEDIE DEI GRUPPI
CONFRONTATE IL VALORE DI LSD CON LA DIFFERENZA TRA LE MEDIE

$$\mathbf{LSD = t_{\alpha/2}\sqrt{(2\ Var\ Residua/n)} = 2.086\sqrt{(2x1732.90/6)} = 50.14}$$

$\bar{y}_i$	681.7	723.2	760.8	789.2

Dif1 = 789.2 – 681.7 = 107.5 > 50,14 ⇒ medie differenti

Dif2 = 789.2 – 723.2 = 66.0 > 50,14 ⇒ medie differenti

Dif3 = 789.2 – 760.8 = 28.4 < 50,14 ⇒ medie uguali

Dif4 = 760.8 – 681.7 = 79.1 > **50.14** ⇒ medie differenti

Dif5 = 760.8 – 723.2 = 37.6 < **50.14** ⇒ medie uguali

Dif6 = 723.2 – 681.7 = 41.5 < **50.14** ⇒ medie uguali

Test t multiplo di Bonferroni

E' basato sulla costruzione degli intervalli di confidenza per la differenza delle medie poste a confronto

$$\left[\left(\bar{y}_r - \bar{y}_s\right) - t_{1-\alpha/2m,g.l.}\sqrt{\frac{2S^2_{residua}}{n}} \le \mu_r - \mu_s \le \left(\bar{y}_r - \bar{y}_s\right) + t_{1-\alpha/2m,g.l.}\sqrt{\frac{2S^2_{residua}}{n}}\right]$$

m rappresenta il numero di intervalli predeterminati per il confronto
g.l sono i gradi di libertà della varianza residua ottenuti nell'ANOVA
α è il livello di significatività stabilito per l'ANOVA

Se l'intervallo ottenuto non contiene lo zero le medie poste a confronto possono ritenersi diverse

UTILIZZATE
LE FORMULE PRECEDENTI
E CALCOLATE
L'INTERVALLO DI CONFIDENZA SECONDO
LA CORREZIONE DI BONFERRONI

$$\left[(\bar{y}_r - \bar{y}_s) - t_{1-\alpha/2m,g.l.}\sqrt{\frac{2S^2_{residua}}{n}} \le \mu_r - \mu_s \le (\bar{y}_r - \bar{y}_s) + t_{1-\alpha/2m,g.l.}\sqrt{\frac{2S^2_{residua}}{n}}\right]$$

$\bar{y}_i$	681.7	723.2	760.8	789.2	$\alpha/12 = 0.004$

Int 1-4 $107.5 \pm 2.84 \times 24.03 = 107.5 \pm 68.25$ ⇒ 38.75 --- 175.25

Int 2-4 $66.0 \pm 2.84 \times 24.03 = 66.0 \pm 68.25$ ⇒ - 2.25 --- 134.25

Int 3-4 $28.4 \pm 2.84 \times 24.03 = 28.4 \pm 68.25$ ⇒ - 39.85 --- 96.65

Int 1-3 $79.1 \pm 2.84 \times 24.03 = 79.1 \pm 68.25$ ⇒ 10.85 --- 147.35

Int 2-3 $37.6 \pm 2.84 \times 24.03 = 37.6 \pm 68.25$ ⇒ - 30.65 --- 105.85

Int 1-2 $41.5 \pm 2.84 \times 24.03 = 41.5 \pm 68.25$ ⇒ - 26.75 --- 109.75

QUANDO NON E' POSSIBILE FARE L'ASSUNZIONE DELLA DISTRIBUZIONE GAUSSIANA SULLA VARIABILE, IL CONFRONTO DI PIU' GRUPPI SI EFFETTUA CON IL METODO DELL'ANALISI DELLA VARIANZA NON PARAMETRICA

IL PROBLEMA

Si vuole studiare l'effetto di due farmaci sul tempo di reazione ad un certo stimolo su animali da laboratorio; il terzo gruppo è quello di controllo.
Possiamo affermare che i tre campioni si differenziano rispetto ai tempi di reazione (misurati in secondi)?

GRUPPO 1	GRUPPO 2	GRUPPO 3
17	8	2
20	7	5
40	9	4
31	8	3
35		

ANALISI DELLA VARIANZA NON PARAMETRICA

TEST DI KRUSKAL-WALLIS

Le n_1, n_2, …, n_k osservazioni provenienti da k campioni sono aggregate in un'unica serie di dati di dimensione n e messe in ordine crescente

Alle n osservazioni vengono assegnati i ranghi. Quando due o più osservazioni hanno lo stesso valore, ad ogni osservazione viene assegnata la media dei ranghi di tutte le osservazioni con lo stesso valore

I ranghi assegnati alle osservazioni in ognuno dei k gruppi vengono sommati tra loro ottenendo k somme dei ranghi

SEGUENDO I PASSI DESCRITTI NEL PRECEDENTE ALGORITMO EFFETTUATE I CALCOLI

Sui nostri dati

Gr I	**R1**	Gr. II	**R2**	Gr III	**R3**
17	**9**	8	**6,5**	2	**1**
20	**10**	7	**5**	5	**4**
40	**13**	9	**8**	4	**3**
31	**11**	8	**6,5**	3	**2**
35	**12**		**26**		**10**
	55				

I valori in giallo sono i dati originari

I valori in bianco rappresentano i ranghi

I valori in verde sono le somme dei ranghi

La statistica test

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

Dove:
K= numero dei campioni
nj = numero di osservazioni nel j-esimo campioni
n = numero totale delle osservazioni
Rj =somma dei ranghi nel j-esimo campioni

ADESSO CALCOLATE

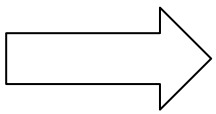
LA STATISTICA TEST

E PRENDETE LA DECISIONE

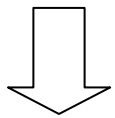
Per il nostro insieme di dati:

$$H = \frac{12}{13(13+1)}\left[\frac{55^2}{5} + \frac{26^2}{4} + \frac{10^2}{4}\right] - 3(13+1) = 10.68$$

Per α=0,009
Htab=7.76



Rifiuto l'ipotesi nulla



I tempi di reazione sono diversi

RICORDATE

Quando ci sono tre campioni e 5 o meno osservazioni in ogni campione, la significatività di H viene determinata usando le tavole di Kruskal-Wallis.

Quando ci sono più di 5 osservazioni in uno o più campioni H viene confrontato con i valori tabulati del χ^2 con k-1 gradi di libertà.

Se ci sono osservazioni con il medesimo valore (ties) bisogna correggere la statistica H

$$1-\sum T \Big/ n^3-n$$

$$T=t^3-t$$

$$H_{corr}=\frac{H}{1-\left(\frac{\sum T}{n^3-n}\right)}$$

Nel nostro caso

Correzione $1-\sum T \Big/ n^3-n$

$T=t^3-t$

$T=t^3-t=2^3-2=6$

Correzione = 1 - (6 / 13^3-13) = 0,9973

$$H_{corr}=\frac{H}{1-\frac{\sum T}{n^3-n}}=\frac{10.68}{0.9973}=10.71$$