ANalysis **O**f **VA**riance

Analisi della Varianza - ANOVA
Aprile, 2005

Nel linguaggio delle variabili le operazioni fondamentali sono tre

- Descrizione
- Spiegazione
- Interpretazione

• Descrizione

La relazione tra variabili viene sintetizzata per meglio cogliere gli aspetti più importanti . Le variabili non vengono distinte in manifeste e latenti o in indipendenti e dipendenti

Es: Chi^2, coefficienti di correlazione ….

• Spiegazione

Si studiano relazioni di dipendenza tra le variabili manifeste.

Variabili indipendenti ➔ variabili dipendenti

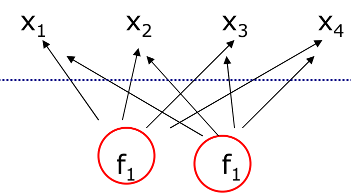
X Y

Es: ANOVA Regressione ….

•Interpretazione

Si studiano relazioni tra variabili manifeste interpretandole come azione di variabili latenti (fattori).

Variabili manifeste



Variabili manifeste

Es: Analisi fattoriale - Rasch

•ANOVA ad una via (one-way)

Una sola variabile dipendente ed una indipendente

•ANOVA fattoriale

Una sola variabile dipendente più indipendenti

•MANOVA

Più variabili dipendenti e diverse indipendenti

Logica

La variabile dipendente (Y) misurata a livello di scala di intervalli in relazione di causa-effetto con la variabile indipendente (X) misurata a livello di scala categoriale (o ordinale, ma con un mumero limitato di modalità).

$$Y=f(X)$$

La variabile indipendente (X) con k modalità, suddivide il campione in k gruppi ognuno dei quali caratterizzato da una certa media e da una certa varianza della variabile dipendente (Y).

Logica

L'ipotesi nulla Ho consiste nel sostenere che i parametri μ

delle popolazioni dalle quali sono stati estratti i K gruppi siano tra loro eguali (i k gruppi sono stati estratti dalla medesima popolazione).

L'ipotesi alternativa che almeno uno dei k parametri sia diverso dagli altri (almeno un gruppo è estratto da una differente popolazione)

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

Logica

Variabile indipendente X

Gruppo 1 $\bar{y}_1, s_1^2 \rightarrow \mu_1, \sigma_1^2$

Gruppo 2 $\bar{y}_2, s_2^2 \rightarrow \mu_2, \sigma_2^2$

Gruppo... $\bar{y}_{...}, s_{...}^2 \rightarrow \mu_{...}, \sigma_{...}^2$

Gruppo K $\bar{y}_k, s_k^2 \rightarrow \mu_k, \sigma_k^2$

Analisi della Varianza - ANOVA
Aprile, 2005

Assunti

•LIVELLO DI SCALA

X a livello categoriale Y a livello cardinale

•OMOSCHEDASTICITA'

La variabile Y deve avere la stessa varianza nelle K popolazioni

•NORMALITA'

La variabile Y deve distribuirsi normalmente nelle K popolazioni

•INDIPENDENZA

I k gruppi devono essere indipendenti tra di loro

Logica

Se gli assunti sono veri (omoschedaticità,normalità) e l'ipotesi nulla anche, i k gruppi sono estratti da una popolazione nella quale la variabile Y è distribuita normalmente (μ,σ).

Partendo dai K gruppi possiamo stimare la varianza della popolazione in due modi:

a) Ricostruiamo con le k medie la distribuzione della V.a. media campionaria. Sapendo che la varianza della v.a. media campionaria è σ^2 /n ne otteniamo una stima

b) Calcoliamo la media delle varianze di ciascun gruppo(valore atteso della v.a. varianza campionaria) e correggiamo il valore ottenuto

Logica

La prima stima (a) è corretta solo se l'ipotesi nulla è vera e quindi i gruppi provengono dalla medesima popolazione. Se l'ipotesi nulla è falsa il metodo (a) ci conduce ad una stima distorta.

La seconda stima è invece indipendente dalla veridicità dell'ipotesi nulla, in quanto è garantita dal rispetto dell'assunto di omoschedasticità.

Dunque per verificare se l'ipotesi nulla è vera o falsa dobbiamo confrontare le due stime.

Se le due stime sono identiche il loro rapporto sarà =1

Ingredienti

$\bar{y}_k$ = media del k-esimo gruppo

$\bar{y}_{..}$ = media generale

SS_T (Dev_T) = devianza generale

SS_M (Dev_M) = devianza tra le medie

SS_P (Dev_P) = devianza parziale tra le osservazioni

media del k-esimo gruppo	$\bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} y_{i,k}$
media generale (se bilanciato)	$\bar{y}_{..} = \frac{1}{K}\sum_{k=1}^{K} \bar{y}_k$
Devianza totale	$SS_T = \sum_{i=1}^{n}\sum_{k=1}^{K}\left(y_{ik} - \bar{y}_{..}\right)^2$
Devianza tra le medie (tra gruppi)	$SS_M = \sum_{k=1}^{K}\left(\bar{y}_{.k} - \bar{y}_{..}\right)^2$
Devianza parziale (nei gruppi)	$SS_{p_k} = \sum_{i=1}^{n}\left(y_{ik} - \bar{y}_{.k}\right)^2$

Metodo a

Se l'ipotesi nulla è vera, i k gruppi sono campioni estratti da un'unica popolazione. Le k medie costituiscono quindi osservazioni della variabile aleatoria media campionaria che sappiamo si distribuisce con varianza σ^2/n.

Possiamo quindi usare la varianza calcolata tra le medie campionarie per calcolare σ^2

Calcoliamo la Devianza tra le medie

$$Dev_M = \sum_{k=1}^{K} \left(\bar{y}_k - \bar{y}_{..}\right)^2$$

Trasformiamo la Devianza tra le medie in Varianza dividendo per i gradi di libertà, ottenendo una stima della varianza delle medie campionarie

$$\frac{\sigma^2}{n} = \frac{\sum_{k=1}^{K} \left(\bar{y}_k - \bar{y}_{..}\right)^2}{K-1}$$

Moltiplicando per n otteniamo una prima stima della varianza della variabile Y che chiamiamo Varianza Between.
<u>Questa stima è corretta solamente se l'ipotesi nulla è vera !</u>

$$\frac{\sigma^2}{n} = \frac{\sum_{k=1}^{K}\left(\bar{y}_k - \bar{y}_{..}\right)^2}{K-1}$$

$$\sigma^2 = VAR_B = n \cdot \frac{\sum_{k=1}^{K}\left(\bar{y}_k - \bar{y}_{..}\right)^2}{K-1}$$

n.b. esempio di disegno bilanciato

Analisi della Varianza - ANOVA
Aprile, 2005

Metodo b

Per ognuno dei k gruppi possiamo calcolare la varianza (ogni osservazione rispetto alla media del proprio gruppo).

Ognuna delle varianze così calcolate rappresenta una oservazione della variabile aleatoria varianza campionaria.

Questo è vero sempre, purché valga l'assunto di omoschedasticità.

Sappiamo che il valore atteso della v.a. varianza campionaria coincide, a parte un fattore di correzione con quello della varianza della variabile stessa

$$\mathrm{E}\left(s^2\right) = \frac{n-1}{n}\sigma^2$$

Metodo b

Calcoliamo per ciascun gruppo la Devianza parziale

$$Dev_{p_k} = \sum_{i=1}^{n} \left(y_{ik} - \bar{y}_{.k} \right)^2$$

Trasformiamo la devianza parziale in varianza dividendo per i gradi di libertà (n-1), facciamo la media delle K stime della varianza così ottenuta ed otteniamo una stima che chiamiamo Varianza Within

$$VAR_W = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \left(y_{ik} - \bar{y}_{.k} \right)^2}{n-1} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n} \left(y_{ik} - \bar{y}_{.k} \right)^2}{N-K}$$

Verifica dell'ipotesi

Se l'ipotesi nulla è vera le due stime hanno lo stesso valore atteso. Se l'ipotesi nulla è falsa il valore atteso della varianza tra i gruppi sarà la somma tra la varianza della popolazione e la varianza dei gruppi

$$F = \frac{VAR_B}{VAR_W}$$

La distribuzione campionaria è quella della v.a. F di Snedecor con (k-1) e (n-K) gradi di libertà

Teorema della scomposizione della devianza

La devianza totale di un insieme di osservazioni può essere sempre riscritta come la somma di due devianze

$$SS_T = SS_B = SS_W$$

Partendo da questa osservazione possiamo costruire un indice di determinazione che ci informa sulla quota di varianza riprodotta

$$\eta^2 = 1 - \frac{VAR_W}{VAR_T}$$

Analisi della Varianza - ANOVA
Aprile, 2005

esercitazione

ident	ore lettura settimana	genere lettura	ORE TV	TITOLO DI STUDIO
2	20	1	6	2
5	8	1	14	3
6	5	1	22	2
8	3	1	36	2
10	13	1	11	3
15	1	1	41	3
20	6	1	23	3
1	15	2	7	1
9	2	2	53	1
11	9	2	15	1
13	3	2	38	3
14	1	2	29	1
16	5	2	25	1
17	2	2	17	1
3	5	3	21	1
4	4	3	26	3
7	9	3	13	2
12	15	3	11	2
18	2	3	56	2
19	11	3	15	2
21	5	3	27	3
		1=giallo		1=obbligo
		2=classico		2=m.sup
		3=fant.		3=laurea