ANOVA per misure ripetute

(Le pagine che seguono sono tratte dal capitolo 4 del libro di R. Micciolo, G. Espa, L. Canal *Ricerca con* R. Apogeo, Milano, 2013.)

Esempio

Esempio 0.1 Per un nuovo prodotto sono stati proposti tre diversi metodi di assemblaggio e si vuole determinare quale dei tre metodi consente di produrre, in un'ora, un maggior numero di componenti. Sono stati selezionati 10 lavoratori ciascuno dei quali ha impiegato, in ordine casuale, tutti e tre i metodi proposti. Di seguito è riportato il numero di unità prodotte in un'ora da ciascun lavoratore.

	Lavoratore									
Metodo	A	B	C	D	E	F	G	H	I	J
A	97	73	93	100	73	91	100	86	92	95
B	93	100	93	55	77	91	85	73	90	83
C	99	94	87	66	59	75	84	72	88	86

Verificare, ad un livello di significatività del 5%, l'ipotesi nulla che il numero medio orario di componenti prodotte è lo stesso per i tre metodi. (Anderson, Sweeney, Williams, 2011).

ANOVA

L'esempio proposto richiede di verificare l'ipotesi nulla $H_0 : \mu_A = \mu_B = \mu_C$ nel caso in cui i tre campioni su cui sono state rilevati i dati siano dipendenti. Si potrebbero eseguire tre test t di Student per dati appaiati adottando la correzione di Bonferroni (si veda il capitolo 3), ma, come già fatto notare in precedenza, questo approccio risulta piuttosto penalizzante. Dobbiamo quindi estendere l'analisi della varianza per un disegno completamente randomizzato (che abbiamo descritto nel terzo capitolo) al caso in cui si abbiano dei campioni dipendenti (ovvero delle *misure ripetute*). Chiamiamo questa estensione "analisi della varianza a blocchi randomizzati", dove, nell'esempio in parola, un "blocco" è rappresentato da un soggetto, mentre la randomizzazione consiste nell'attribuzione a ciascun soggetto di una sequenza delle tre procedure A, B e C in ordine casuale.

Illustreremo la procedura di calcolo e la logica dell'analisi impiegando i dati dell'esempio. Iniziamo leggendo i dati e memorizzandoli sotto forma di una "matrice" in cui le righe rappresentano i soggetti e le colonne rappresentano i tre diversi metodi posti a confronto.

```
> A <- c(97, 73, 93, 100, 73, 91, 100, 86, 92, 95)
> B <- c(93, 100, 93, 55, 77, 91, 85, 73, 90, 83)
> C <- c(99, 94, 87, 66, 59, 75, 84, 72, 88, 86)
> Y <- cbind(A,B,C); colnames(Y) <- c("A","B","C")
> Y
        A   B  C
 [1,]  97  93 99
 [2,]  73 100 94
 [3,]  93  93 87
 [4,] 100  55 66
 [5,]  73  77 59
 [6,]  91  91 75
 [7,] 100  85 84
 [8,]  86  73 72
 [9,]  92  90 88
[10,]  95  83 86
```

Possiamo calcolare sia la media generale, sia il numero medio di componenti prodotti in un'ora con ciascuno dei tre metodi:

```
> mean(Y)
[1] 85
> colMeans(Y)
 A  B  C
90 84 81
```

Argomentando in modo analogo a quanto fatto nel capitolo 3, possiamo calcolare una devianza totale

```
> (devt <- sum((Y - mean(Y))^2))
[1] 4256
```

e una devianza fra gruppi (ovvero fra i metodi a confronto)

```
> (devf <- 10*sum((colMeans(Y) - mean(Y))^2))
[1] 420
```

Se non esistesse l'"appaiamento" (ovvero se i tre metodi fossero stati "testati" da tre gruppi diversi di lavoratori) a questo punto non ci rimarrebbe che calcolare (per differenza) la devianza residua ed eseguire il test F. Ma nell'esempio in parola è possibile anche calcolare il numero medio di componenti assemblate da ciascun lavoratore

```
> round(rowMeans(Y),1)
 [1] 96.3 89.0 91.0 73.7 69.7 85.7 89.7 77.0 90.0 88.0
```

medie che, del tutto ragionevolmente, variano da lavoratore a lavoratore. È quindi importante tenere conto delle diverse "abilità" dei lavoratori, calcolando una devianza fra lavoratori (ovvero, in generale, fra i "blocchi")

```
> (devb <- 3*sum((rowMeans(Y) - mean(Y))^2))
[1] 1992.667
```

È questo punto che possiamo calcolare la devianza residua per differenza:

```
> (devr <- devt - (devf+devb))
[1] 1843.333
```

La verifica dell'ipotesi nulla di uguaglianza fra le tre procedure si calcola mediante un test F dato dal rapporto fra la varianza fra i tre gruppi (al numeratore) e la varianza residua (al denominatore). Come sempre, passiamo dalle devianze alle varianze dividendo per i corrispondenti gradi di libertà; nell'esempio che stiamo affrontando i gradi di libertà sono 2 per il numeratore (il numero dei gruppi meno uno) e 18 per il denominatore. Questo valore è dato dalla differenza fra i gradi di libertà totali (che sono 29, ovvero 30, il numero delle osservazioni complessive, meno uno) e la somma fra 2 (i gradi di libertà della varianza fra i gruppi) e 9 (i gradi di libertà della varianza fra i blocchi, cioè fra i soggetti: $10 - 1$).

Calcoliamo quindi la varianza fra i tre gruppi e la varianza residua e facciamone il rapporto (cioè calcoliamo il valore del test F):

```
> (varf <- devf/2)
[1] 210
> (varr <- devr/18)
[1] 102.4074
> (F <- varf/varr)
[1] 2.050633
```

Il *p-value* associato al valore del test è l'area a destra di 2.051 in una distribuzione F con 2 gradi di libertà al numeratore e 18 gradi di libertà al denominatore (ricordiamo che il test F è un test a una coda):

```
> pf(F,2,18,lower.tail=FALSE)
[1] 0.1576515
```

Il *p-value* è superiore al livello di significatività del 5%, per cui non possiamo rifiutare l'ipotesi nulla.

Eseguire con R un'ANOVA per un disegno a blocchi randomizzati prevede che i dati siano preparati in un formato particolare e non del tutto "naturale" per il ricercatore. Infatti l'esistenza dell'appaiamento porterebbe a memorizzare i dati come abbiamo fatto poco sopra, ovvero in una matrice (o in un *data-frame*) le cui righe rappresentano i blocchi e le colonne i gruppi (trattamenti) a confronto. Purtroppo questo tipo di archiviazione (detto anche in forma *wide*) non va bene per R (per lo meno non va bene se vogliamo sfruttare la funzione `aov` come abbiamo fatto nel capitolo 3). I dati vanno invece archiviati in forma *long*. Vediamo come si fa.

In un disegno a blocchi come quello dell'esempio proposto ciascuna delle n osservazioni complessive è identificata da due "coordinate": la prima identifica il blocco (il soggetto), la seconda identifica il trattamento. Ecco quindi che dovremo preparare un *data-frame* con 3 colonne: una colonna conterrà le informazioni che identificano il blocco, una colonna conterrà le informazioni che identificano il trattamento e una colonna conterrà le informazioni relative a ciò che il ricercatore vuole misurare (il numero di componenti prodotte in un'ora, nel caso del nostro esempio). Incominciamo da quest'ultima.

```
> (y <- c(A,B,C))
 [1]  97  73  93 100  73  91 100  86  92  95
[11]  93 100  93  55  77  91  85  73  90  83
[21]  99  94  87  66  59  75  84  72  88  86
```

Passiamo ora ad identificare i trattamenti (i gruppi); per come abbiamo inserito le osservazioni nel vettore `y` dovremo avere

```
> (x <- factor(c(rep("A",10),rep("B",10),rep("C",10))))
 [1] A A A A A A A A A A
[11] B B B B B B B B B B
[21] C C C C C C C C C C
Levels: A B C
```

(notate che il vettore `x` è un *fattore*). Infine il vettore che identifica i blocchi (i soggetti)

```
> (z <- factor(rep(c(1:10),3)))
 [1] 1  2  3  4  5  6  7  8  9  10
[11] 1  2  3  4  5  6  7  8  9  10
[21] 1  2  3  4  5  6  7  8  9  10
10 Levels: 1 2 3 4 5 6 7 8 ... 10
```

(anche questo è un fattore). Il *data-frame* è quindi

```
> df <- data.frame(z,x,y)
```

e invitiamo il lettore a visualizzarne il contenuto; noi non lo facciamo per ragioni di spazio.

La sintassi per eseguire l'ANOVA nel caso di un disegno a blocchi è la seguente:

```
> fit <- aov(y ~ x + Error(z),data=df)
```

La sintassi è del tutto analoga a quella che avremmo impiegato nel caso di un disegno completamente randomizzato (`aov(y ~ x)`) se non fosse per l'impiego del termine `Error` che permette di specificare quale variabile contiene le informazioni relative ai blocchi. A dire il vero la sintassi più corretta sarebbe `Error(z/x)` per specificare in modo esplicito che le osservazioni all'interno dei gruppi (`x`) sono ripetute su ogni soggetto (`z`). Tuttavia in una situazione come quella dell'esempio in parola, le due sintassi (`Error(z/x)` e `Error(z)`) sono equivalenti. Per vedere il risultato dell'ANOVA impieghiamo, come sempre, la funzione `summary`

```
> summary(fit)

Error: z
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9   1993   221.4

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
x          2    420   210.0   2.051  0.158
Residuals 18   1843   102.4
```

che ci presenta le devianze, i gradi di libertà, le varianze e il valore del test F con il *p-value* associato.

Se desideriamo eseguire dei confronti multipli possiamo impiegare le procedure viste nel terzo capitolo.

La "sfericità"

Nel disegno a blocchi (ovvero per misure ripetute) gli assunti alla base dell'ANOVA sono sostanzialmente gli stessi richiesti nel caso di una ANOVA per un disegno completamente randomizzato: normalità degli "errori" ed omogeneità delle varianze. A questi se ne aggiunge tuttavia un terzo, particolarmente critico quando i blocchi sono rappresentati, ad esempio, da soggetti e i "trattamenti" da tempi distinti (che, ovviamente, in quanto tali non possono certo essere randomizzati!): la *sfericità*. Questa condizione implica che anche le varianze delle differenze tra i diversi trattamenti a confronto siano uguali fra loro. Se questa condizione non è soddisfatta, il *p-value* associato al test F è sottostimato (ovvero è più piccolo di quanto non sia in realtà). Questo implica che un test F che non risulti statisticamente significativo non ha bisogno di alcuna "correzione". Diversa è la situazione in cui il *p-value* sia inferiore

a quello critico: in questo caso il *p-value* va aumentato. Sono state proposte delle "correzioni" ai gradi di libertà impiegando un fattore ϵ che vale 1 quando l'assunto di sfericità è soddisfatto e raggiunge un limite inferiore pari a $1/(k-1)$ (dove k è il numero di trattamenti (tempi) a confronto) nel caso di massima violazione dell'assunto di sfericità. Il fattore di correzione $\epsilon = 1/(k-1)$, proposto da Box (1954), è quello più penalizzante. Infatti, dal momento che i gradi di libertà per la varianza fra i trattamenti sono $k-1$, adottando questa correzione il test farà riferimento ad una distribuzione F con un solo grado di libertà al numeratore, mentre i gradi di libertà al denominatore saranno ridotti di $k-1$ volte (quindi dimezzati, nel caso "migliore", in cui i trattamenti a confronto sono tre e $\epsilon = 1/(k-1) = 1/2$). Facciamo notare a questo proposito che, nel caso di un confronto fra due campioni appaiati $(k = 2)$, non è necessaria alcune "correzione", dal momento che $\epsilon = 1$. Per ottenere dei *p-value* meno "conservativi" sono state proposte delle opportune statistiche per stimare il valore di ϵ; le più usate sono quelle di Greenhouse e Geisser (ϵ_{GG}) e di Huynh e Feldt (ϵ_{HF}).

L'applicazione di queste correzioni ai risultati dell'esempio che abbiamo precedentemente affrontato non modificherebbe il risultato dell'analisi in quanto il test F era già risultato non significativo. Per illustare il procedimento di "correzione per la (non) sfericità" faremo ricorso ai dati contenuti nel *package* `rmf` relativi ad una indagine eseguita in Italia fra il 2005 e il 2006 volta a valutare il mutamento dei modelli di "civismo". A 946 soggetti è stato chiesto di specificare, impiegando una scala da 1 a 10, quanto erano d'accordo con le tre seguenti affermazioni:

- La scienza e la tecnologia stanno rendendo la nostra vita più sana, facile e più confortevole.
- Grazie alla scienza ed alla tecnologia ci saranno maggiori opportunità per la prossima generazione.
- Dipendiamo troppo dalla scienza e non abbastanza dalla fede.

Le risposte individuali a ciascuna delle tre domande (identificate rispettivamente con le etichette A, B e C) sono contenute in forma *long* nel *data-frame* `wvs.long` e in forma *wide* nel *data-frame* `wvs.wide`. In questo ultimo caso le tre colonne del *data-frame* corrispondono alle tre domande

```
> str(wvs.wide)
'data.frame':   946 obs. of  3 variables:
 $ A: num  6 8 3 7 4 8 5 5 5 6 ...
 $ B: num  5 9 4 7 8 10 5 6 5 6 ...
 $ C: num  10 10 8 2 10 3 8 8 5 5 ...
```

mentre nel primo caso le colonne del *data-frame* contengono, rispettivamente, l'identificativo del soggetto, l'etichetta della domanda e il punteggio associato all'affermazione (il valore minimo, 1, indica un completo disaccordo, mentre quello massimo, 10, un completo accordo):

```
> str(wvs.long)
'data.frame':   2838 obs. of  3 variables:
 $ id   : Factor w/ 946 levels "1","2","3","4",..: 1 2 3
 $ aff  : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1
 $ score: num  6 8 3 7 4 8 5 5 5 6 ...
```

Calcoliamo innanzitutto i punteggi medi associati a ciascuna affermazione

```
> tapply(wvs.long$score,wvs.long$aff,mean)
       A        B        C
6.326638 6.419662 6.120507
```

chiedendoci se possiamo considerare significative (al 5%) le differenze fra le tre medie.

Per sottoporre a verifica l'ipotesi nulla di uguaglianza fra le medie dei tre punteggi facciamo ricorso ad una ANOVA per misure ripetute, analogamente a quanto fatto nell'esempio trattato in precedenza; la sintassi è del tutto equivalente:

```
> fit <- aov(score ~ aff + Error(id),data=wvs.long)
> summary(fit)

Error: id
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 945   6329   6.697

Error: Within
            Df Sum Sq Mean Sq F value  Pr(>F)
aff          2     44  22.174   4.653 0.00965 **
Residuals 1890   9008   4.766
```

Il risultato ottenuto (ovvero un test F pari a 4.653 con 2 e 1890 gradi di libertà) sarebbe significativo anche adottando un livello di significatività dell'1%. Dobbiamo però chiederci se rimarrebbe tale anche tenendo conto di una eventuale violazione dell'assunto di sfericità.

Il fattore di correzione di Box nel caso in esame $(k = 3)$ è $\epsilon = 1/2$; moltiplichiamo quindi per tale valore i gradi di libertà associati al test F e calcoliamo il *p-value* "corretto"

```
> eps <- 1/2
> pf(4.653,2*eps,1890*eps,lower.tail=FALSE)
[1] 0.03125088
```

osservando che, pur essendo notevolmente aumentato, risulta ancora inferiore al 5%. Possiamo quindi concludere rifiutando l'ipotesi nulla. In questo caso risulta inutile

procedere oltre calcolando i *p-value* associati alle correzioni ϵ_{GG} e ϵ_{HF}. Tuttavia noi lo faremo allo scopo di illustrare questa procedura.

Il procedimento da seguire, non del tutto trasparente per chi non ha solide conoscenze sui legami fra ANOVA e modello lineare, si basa su una ANOVA multivariata e sulla funzione `anova.mlm`. In questo testo non ci sembra opportuno affrontare in dettaglio questo argomento, per il quale rimandiamo il lettore interessato all'articolo di Peter Dalgaard (2007) pubblicato su *R News*. La funzione `aov.mr` di `rmf` si occupa di passare a `aov.mlm` la sintassi appropriata per analizzare il disegno in parola. A differenza di `aov`, la funzione `aov.mr` vuole i dati in forma *wide* e impiega una sintassi davvero essenziale: l'unico argomento della funzione è il *data-frame* con i dati da analizzare.

```
> aov.mr(wvs.wide)
Analysis of Variance Table

Greenhouse-Geisser epsilon: 0.6786
Huynh-Feldt epsilon:        0.6792

  Res.Df Df Gen.var.      F num Df den Df   Pr(>F)   G-G Pr   H-F Pr
1    945      1.9964
2    946  1   2.0021 4.6525      2   1890 0.009647 0.020574 0.020544
```

I due fattori di correzione ϵ_{GG} e ϵ_{HF} sono molto simili fra loro (coincidono se arrotondati al terzo decimale) e anche molto vicini al limite inferiore di $1/2$, segno di una violazione piuttosto pesante dell'assunto di sfericità. L'*output* della funzione propone, oltre alle due stime già segnalate, anche il test F con i gradi di libertà e il *p-value* "originali" (senza apportare cioè alcuna correzione) e, a fianco, i due *p-value* ottenuti impiegando la correzione sui gradi di libertà proposta da Greenhouse e Geisser e quella proposta da Huynh e Feldt. Verifichiamo che, analogamente a quanto fatto applicando la correzione di Box, il *p-value* corretto secondo Greenhouse e Geisser si ottiene moltiplicando i gradi di libertà (al numeratore e al denominatore della distribuzione F) per ϵ_{GG}

```
> eps <- 0.6786
> pf(4.6525,2*eps,1890*eps,lower.tail=FALSE)
[1] 0.02057492
```

In modo del tutto analogo, si ottiene il *p-value* corretto secondo Huynh e Feldt

```
> eps <- 0.6792
> pf(4.6525,2*eps,1890*eps,lower.tail=FALSE)
[1] 0.02054585
```

Come anticipato il *p-value* calcolato impiegando ϵ_{HF} è più piccolo di quello calcolato impiegando ϵ_{GG}. Il lettore si sarà anche accorto che i gradi di libertà "corretti" sono, in generale, frazionari; questo tuttavia non è un problema, dal momento che

esistono distribuzioni F con gradi di libertà frazionari, anche se le tavole riportano, del tutto ragionevolmente, soltanto gradi di libertà "interi".

Prima di concludere questo argomento ci piace segnalare quanto scrive Peter Dalgaard nell'articolo già citato. Infatti, la formula originale riportata nell'articolo di Huynh e Feldt conteneva un errore al numeratore nel caso di un disegno *split-plot*. Diamo ora la parola a Peter Dalgaard: *"A correction was published 15 years later (Lecoutre, 1991), but another 15 years later, this error is still present in SAS and SPSS."* E, aggiungiamo noi, anche in STATA. Sembra davvero incredibile che dei programmi commerciali non siano "attenti" alla pubblicazione di correzioni di metodiche implementate quando queste sono segnalate a distanza di (molto) tempo, mentre un programma *free* come R lo sia. Ma forse questo non è poi così "incredibile". R è sviluppato da una comunità di ricercatori; un ricercatore deve tenersi sempre aggiornato sulle tecniche che impiega e non può farsi "scappare" una correzione di un errore, anche se la correzione è fatta a distanza di molto tempo. Nella ricerca gli errori sono errori e non hanno una "data di scadenza".