Laboratorio 9

Analisi della varianza a due fattori

9.1 Analisi del dataset PENICILLIN.DAT

I dati contenuti nel file `penicillin.dat`, si riferiscono ad un esperimento di produzione di penicillina tendente a valutare gli effetti di 4 modi differenti di produzione (A, B, C, D). Si è osservato precedentemente come la miscela adottata nella produzione sia piuttosto variabile e questo possa in qualche maniera influire sulla produzione. Quindi si è deciso di controllare anche l'effetto della miscela considerando 5 miscele (I, II, II, IV, V) e impiegando ognuna di queste nei quattro processi produttivi. Si noti come in questo caso l'interesse è rivolto a verificare se esiste una diversità d'effetto sulle quantità di penicillina prodotte tra modi di produzione, indipendentemente dal tipo di miscela impiegata.

```
> pen <- read.table("I:\\modelli\\penicillin.dat", header=TRUE)
> pen
   miscela modo penicillina
1        I    A          89
2        I    B          88
3        I    C          97
4        I    D          94
5       II    A          84
6       II    B          77
7       II    C          92
8       II    D          79
9      III    A          81
10     III    B          87
11     III    C          87
12     III    D          85
13      IV    A          87
14      IV    B          92
15      IV    C          89
16      IV    D          84
17       V    A          79
18       V    B          81
19       V    C          80
```

```
20       V    D          88
```

Si noti che si ha una sola osservazione sulla variabile risposta in ciascuno dei 20 sottogruppi.
Procediamo dapprima con un'analisi esplorativa.

```
> attach(pen)
> par(mfrow=c(1,2))
> plot(penicillina~modo)
```

Il grafico mostra una evidente differenza tra modi in termini di mediane (il modo C ha la mediana più elevata), anche se dobbiamo tener conto della ridotta numerosità campionaria. La variabilità entro i gruppi parrebbe comparabile. Notiamo una certa asimmetria nelle distribuzioni.
Consideriamo in primo luogo la dipendenza dal `modo`, senza tener conto del tipo di miscela, ossia considerando l'analisi della varianza ad un fattore (`modo`)

```
> pen.aov <- aov(penicillina ~ modo)
> summary(pen.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
modo         3  70.00   23.33  0.7619 0.5318
Residuals   16 490.00   30.62
```

La funzione ci presenta la classica tabella di scomposizione della varianza. La quantità indicata con `Pr(>F)` indica il p-value (livello di significatività osservato) del test per verificare l'ipotesi nulla di uguale effetto dei 4 modi di produzione. I risultati delle analisi effettuate portano, contrariamente a quanto suggeriva l'analisi esplorativa, ad accettare l'ipotesi nulla. Questo potrebbe essere dovuto alla presenza di diverse miscele. Proviamo a tenere conto di questo. Dapprima consideriamo il grafico

```
> plot(penicillina~miscela)
> pen.aov <- aov(penicillina~miscela)
> summary(pen.aov)
```

Possiamo osservare come la miscela sembra avere un effetto sulla produzione.
Consideriamo ora il modello di analisi della varianza a due fattori:

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij}.$$

```
> pen2.aov <- aov(penicillina ~ modo + miscela)
> summary(pen2.aov)
            Df  Sum Sq Mean Sq F value  Pr(>F)
modo         3  70.000  23.333  1.2389 0.33866
miscela      4 264.000  66.000  3.5044 0.04075 *
Residuals   12 226.000  18.833
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Si osservi come la miscela abbia un effetto significativo sulla produzione, mentre il fattore che più ci interessa, il `modo`, sembra non dare differenze di produzione significative.

```
> detach()
```

9.2 Analisi del dataset RATS.DAT

I dati contenuti nel file `rats.dat`, si riferiscono ad un disegno fattoriale 3×4 completamente randomizzato sull'effetto di un agente tossico. Si considerano 3 tipi di veleno (I, II, III) e 4 trattamenti (A, B, C, D). Ogni combinazione veleno-trattamento viene somministrata a 4 cavie. Per ogni cavia viene osservato il tempo di sopravvivenza espresso in decine di ore.

```
> topi <- read.table("I:\\modelli\\rats.dat", header=TRUE)
> topi
  tempo veleno trattamento
1  0.31      I     A
2  0.82      I     B
3  0.43      I     C
4  0.45      I     D
5  0.45      I     A
...
44 0.31    III     D
45 0.23    III     A
46 0.29    III     B
47 0.22    III     C
48 0.33    III     D

> attach(topi)
```

Cominciamo ad analizzare le distribuzioni del tempo di sopravvivenza rispetto ad ognuno dei fattori.

```
> par(mfrow=c(1,2))
> plot(tempo ~ veleno + trattamento)
```

Da questi grafici possiamo vedere che le distribuzioni non appaiono simmetriche e che, in generale, i diversi livelli dei fattori hanno un effetto sulla risposta. Con i comandi seguenti possiamo valutare graficamente l'effetto di interazione.

```
> par(mfrow=c(1,1))
> interaction.plot(veleno, trattamento, tempo)
> interaction.plot(trattamento, veleno, tempo)
```

Se le spezzate appaiono parallele vi è un'indicazione che non vi sia interazione. Nel caso in esame possiamo sottoporre a verifica questa ipotesi, visto che abbiamo più di una osservazione ($n_{jk} = 4$) per ogni combinazione dei fattori. L'effetto di interazione è stimabile tramite il modello

$$Y_{ijk} = \mu + \alpha_j + \gamma_k + \delta_{jk} + \varepsilon_{ijk}.$$

Si faccia attenzione alla sintassi del comando `lm`

```
> g <- lm(tempo ~ veleno * trattamento)
> anova(g)
Analysis of Variance Table

Response: tempo
                   Df  Sum Sq Mean Sq F value    Pr(>F)
veleno              2 1.03301 0.51651 23.2217 3.331e-07 ***
trattamento         3 0.92121 0.30707 13.8056 3.777e-06 ***
veleno:trattamento  6 0.25014 0.04169  1.8743    0.1123
Residuals          36 0.80072 0.02224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Utilizzare `anova()` su un oggetto derivante dalla funzione `lm()` è equivalente a utilizzare `aov()`.

```
> g.aov <- aov(tempo ~ veleno * trattamento)
> summary(g.aov)
                   Df  Sum Sq Mean Sq F value    Pr(>F)
veleno              2 1.03301 0.51651 23.2217 3.331e-07 ***
trattamento         3 0.92121 0.30707 13.8056 3.777e-06 ***
veleno:trattamento  6 0.25014 0.04169  1.8743    0.1123
Residuals          36 0.80072 0.02224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Come si può osservare l'effetto di interazione (`veleno:trattamento`) non è significativo. La funzione `anova` calcola la statistica test

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/(p - p_0)}{\hat{\sigma}^2/(n - p)}$$

dove sotto H_0 il modello è senza interazione

$$Y_{ijk} = \mu + \alpha_j + \gamma_k + \varepsilon_{ijk}.$$

Si noti che, $n = 48$, $p = 12$, $p_0 = 6$.

```
> g.h0 <- lm(tempo ~ veleno + trattamento)
> F.test <- (( sum(g.h0$res^2) -sum(g$res^2) ) /
+             (g.h0$df.res-g$df.res) )/(sum(g$res^2)/g$df.res)
> F.test
[1] 1.874333
> 1-pf(F.test, g.h0$df.res-g$df.res, g$df.res)
[1] 0.1122506
```

Consideriamo ora i residui.

```
> par(pty='s')
> res <- rstandard(g)
> qqnorm(res)
> qqline(res)
> plot(fitted(g), res)
```

L'ipotesi di omoschedasticità non sembra suffragata. Proviamo a trasformare i dati.

```
> g <- lm(log(tempo) ~ veleno * trattamento)
> plot(fitted(g), rstandard(g))
```

Questa trasformazione non sembra aver portato qualche miglioramento. Proviamo allora a considerare il reciproco del tempo di sovravvivenza.

```
> g <- lm(1/tempo ~ veleno * trattamento)
> plot(fitted(g), rstandard(g))
```

Il modello sembra catturare meglio la variabilità della risposta ed anche il grafico quantile-quantile è migliorato. Si ricordi anche che la trasformazione adottata era suggerita anche nel Laboratorio 8 (vedi Esempio 8.2).

```
> qqnorm(rstandard(g))
> qqline(rstandard(g))
```

L'analisi della varianza non ci mostra nessuna indicazione della presenza di interazione.

```
> anova(g)
Analysis of Variance Table

Response: 1/tempo
                   Df Sum Sq Mean Sq F value    Pr(>F)
veleno              2 34.877  17.439 72.6347 2.310e-13 ***
trattamento         3 20.414   6.805 28.3431 1.376e-09 ***
veleno:trattamento  6  1.571   0.262  1.0904    0.3867
Residuals          36  8.643   0.240
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A questo punto possiamo specificare un modello senza interazione.

```
> g <- lm(1/tempo ~ veleno + trattamento)

> anova(g)
Analysis of Variance Table

Response: 1/tempo
            Df Sum Sq Mean Sq F value    Pr(>F)
veleno       2 34.877  17.439  71.708 2.864e-14 ***
trattamento  3 20.414   6.805  27.982 4.192e-10 ***
```

```
Residuals   42 10.214   0.243
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> detach()
```

Esercizio: I dati contenuti nel file `nails.dat`, si riferiscono ad un esperimento per mezzo del quale si vuole valutare l'efficacia di uno smacchiatore nello sciogliere macchie di smalto per unghie dai tessuti. Vengono impiegati due tipi di solventi e 3 tipi di smalto. L'esperimento consisteva nell'immergere in una bacinella con un certo solvente 5 tessuti macchiati da un certo tipo di smalto, misurarando il tempo (in minuti) affinchè la macchia si dissolvesse. Si chiede di valutare se esistono diversità di azione dei solventi anche rispetto al tipo di macchia.