Laboratorio 8

Analisi della varianza

8.1 Analisi del dataset STURDY.DAT

I dati riportati nel file `sturdy.dat` si riferiscono ad un esperimento effettuato per studiare la resistenza allo strappo di diverse marche di impermeabili. Capi di cinque diverse marche sono stati sottoposti alla stessa sollecitazione. La resistenza allo strappo è stata misurata con il tempo (in minuti e frazioni decimali di minuti) intercorso tra la sollecitazione e lo strappo.

1. Esistono differenze nella resistenza tra le varie marche?
2. È possibile dire se le marche A, B, C sono più resistenti, in media, delle marche D, E?

```
Marca A: 2,34  2,46  2,83  2,04  2,69
Marca B: 2,64  3,00  3,19  3,83
Marca C: 2,61  2,07  2,80  2,58  2,98  2,30
Marca D: 1,32  1,62  1,92  0,88  1,50  1,30
Marca E: 0,41  0,83  0,58  0,32  1,62
```

Acquisiamo i dati.

```
> sturdy <- read.table("I:/modelli/sturdy.dat")
> sturdy
     V1
 1 2.34
 2 2.46
 ......
23 0.83
24 0.58
25 0.32
26 1.62

> names(sturdy) <- "time"
```

Dobbiamo costruire il vettore indicatore delle cinque marche ed unirlo al dataframe.

```
> group <- c(rep("A",5), rep("B",4), rep("C",6), rep("D",6), rep("E",5))
> group <- factor(group)
> group
> sturdy <- data.frame(sturdy, group)
> sturdy
```

Possiamo provare a fare un po' di analisi esplorativa, anche se abbiamo pochissime osservazioni per gruppo.

```
> attach(sturdy)
> plot(time~group)
```

Il grafico mostra una evidente differenza tra i gruppi in termini di mediane. (Il gruppo B ha la mediana più elevata).
Anche se a livello assolutamente indicativo, la variabilità entro i gruppi parrebbe comparabile.
Considerato il numero ridotto di osservazioni per gruppo, le distribuzioni dentro i gruppi appaiono sufficientemente simmetriche (con qualche eccezione per il gruppo E). La sola simmetria non garantisce comunque la normalità dei dati. Tuttavia, in questo caso abbiamo solo 5 osservazioni per ogni gruppo. Non è molto sensato verificare la normalità attraverso `qqnorm` quando le numerosità sono cosìesigue.
Quindi, in questo caso assumeremo la normalità delle osservazioni valutando graficamente solo i boxplot.
Assumendo che la distribuzione della variabile risposta nei gruppi sia normale (con medie e varianze non necessariamente uguali) possiamo costruire il test del rapporto di verosimiglianza per verificare l'ipotesi nulla di omoschedasticità, ossia di uguaglianza delle varianze. In `R` questo test è effettuato con il comando `bartlett.test`.
La sintassi può essere indifferentemente in forma di formula (`bartlett.test(time~group)`) oppure mettendo prima il nome della variabile con i valori della variabile continua e poi il nome del fattore di gruppo, separati da virgola:

```
> bartlett.test(time,group)

        Bartlett test for homogeneity of variances

data:  time and group
Bartlett's K-squared = 1.8016, df = 4, p-value = 0.7722
```

Il test porta all'accettazione dell'ipotesi nulla, ossia le varianze dei gruppi si possono considerare uguali.
L'osservazione circa la diversità delle mediane, unita all'ipotesi di normalità della variabile risposta, ci rende inclini a pensare che esista una differenza anche in termini di medie.
Verifichiamo se c'è diversità nella medie utilizzando la funzione `lm`:

```
> sturdy.lm<-lm(time~group)
```

La variabile `group` è un fattore e viene interpretata all'interno della funzione `lm` come una sequenza di variabile *dummy*, omettendo la variabile dummy relativa alla prima modalità. Il risulato indica una significativa differenza tra le medie dei gruppi, poiché vi è almeno un coefficiente significativamente diverso da zero:

```
> summary(sturdy.lm)

Call:
lm(formula = time ~ group)

Residuals:
      Min        1Q    Median        3Q       Max
-0.543333 -0.235500  0.005667  0.212667  0.868000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.47200    0.17929  13.788 5.40e-12 ***
groupB       0.69300    0.26893   2.577 0.017583 *
groupC       0.08467    0.24276   0.349 0.730733
groupD      -1.04867    0.24276  -4.320 0.000302 ***
groupE      -1.72000    0.25355  -6.784 1.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4009 on 21 degrees of freedom
Multiple R-Squared: 0.8433,     Adjusted R-squared: 0.8135
F-statistic: 28.26 on 4 and 21 DF,  p-value: 3.475e-08
```

L'intercetta rappresenta la stima della media nel gruppo A. I restanti coefficienti si interpretano come scostamenti dalla media del gruppo A. Ad esempio, la stima del coefficiente del gruppo B è pari a 0.693. Ciò vuol dire che la stima della media nel gruppo B è pari a 2.472+0.693=3.165.
Se volevamo verificare la diversità delle medie nei gruppi, potevamo riccorrere direttamente all'analisi della varianza, con il comando:

```
> sturdy.aov <- aov(time~group)
```

Vediamo che cosa contiene l'oggetto risultante dall'analisi.

```
> sturdy.aov
Call:
   aov(formula = time ~ group)

Terms:
                    group Residuals
Sum of Squares  18.168119  3.375127
Deg. of Freedom         4        21

Residual standard error: 0.4008994
Estimated effects may be unbalanced
```

Per avere informazioni circa l'esito del test F utilizziamo la funzione `summary`:

```
> summary(sturdy.aov)
            Df  Sum Sq Mean Sq F value    Pr(>F)
group        4 18.1681  4.5420  28.261 3.475e-08 ***
Residuals   21  3.3751  0.1607
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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). Questo è ottenibile come:

```
>  1-pf(28.261, 4, 21)
[1] 3.474262e-08
```

e coincide con il `p-value` del test F che si trova in `summary(sturdy.lm)` .
Per avere invece la soglia della regione di rifiuto, calcoliamo il quantile di livello 0.95 per una distribuzione $F_{4,21}$.

```
> qf(0.95, 4, 21)
[1] 2.8401
```

Conclusioni: I risultati delle analisi effettuate portano ad un rifiuto dell'ipotesi nulla. Per rispondere ora al quesito 2, possiamo accorpare i gruppi A, B, C ed i gruppi D, E per verificare poi la significatività della differenza tra le medie delle due nuove popolazioni (quella composta dalle marche A, B, C e quella composta dalle marche D, E).
Per fare ciò possiamo creare un nuovo vettore indicatore dei due nuovi gruppi ed aggiungerlo al dataframe.

```
> detach()
> gnew <- c(rep("G1", 15), rep("G2", 11))
> gnew <- factor(gnew)
> sturdy1 <- data.frame(sturdy, gnew)
> attach(sturdy1)
```

Per valutare la significatività della differenza tra le due medie, possiamo applicare il test t di Student. Non è possibile in questo caso applicare ancora l'ANOVA invece del test t perché l'ipotesi che vogliamo verificare ha una alternativa chiaramente unilaterale.

Esercizio: Prima di effettuare il test verificare le assunzioni.

Quindi:

```
> t.test(time[gnew=="G1"], time[gnew=="G2"], alternative="greater",
+         var.equal=TRUE)

        Two Sample t-test

data:  time[gnew == "G1"] and time[gnew == "G2"]
t = 8.0229, df = 24, p-value = 1.500e-08
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.237152      Inf
sample estimates:
mean of x mean of y
 2.690667  1.118182
```

Conclusioni: L'ipotesi nulla è decisamente rifiutata a favore dell'alternativa, che prevede che la resistenza media del gruppo di marche G1 sia maggiore della resistenza media del gruppo G2.

```
> detach()
```

8.2 Analisi del dataset RATS.DAT

I dati contenuti nel file `rats.dat`, si riferiscono ad uno studio 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 scelte a caso. Quindi 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

> topi$veleno <- factor(topi$veleno)
> topi$trattamento <- factor(topi$trattamento)
> attach(topi)
```

Per il momento consideriamo solamente i dati relativi al veleno II, creando il nuovo data frame

```
> topiII<-data.frame(tempo=tempo[veleno=='II'],
+ trattamento=trattamento[veleno=='II'])
> attach(topiII)
```

Vogliamo verificare se c'è differenza in media tra i diversi trattamenti, mediante un'analisi della varianza.
Proviamo a controllare le ipotesi di normalità e di omoschedasticità (tenendo conto che abbiamo solo 4 osservazioni per tipo di trattamento!):

```
> plot(trattamento,tempo)
```

Sembra esserci una diversità nella varianza dei vari gruppi. Proviamo a verificare l'ipotesi di omoschedasticità attraverso il test di Bartlett:

```
> bartlett.test(tempo,trattamento)

        Bartlett test for homogeneity of variances

data:  tempo and trattamento
Bartlett's K-squared = 9.5432, df = 3, p-value = 0.02288
```

Il livello di significatività osservato porta ad un rifiuto dell'ipotesi nulla di uguaglianza delle varianze.
Possiamo provare a trasformare la variabile risposta, considerando il reciproco del tempo di sopravvivenza

```
> plot(trattamento,1/tempo)
```

La situazione è migliorata. Questo è confermato anche dal test di Bartlett che porta all'accettazione dell'ipotesi nulla di uguaglianza delle varianze

```
> bartlett.test(1/tempo,trattamento)

        Bartlett test for homogeneity of variances

data:  1/tempo and trattamento
Bartlett's K-squared = 1.2807, df = 3, p-value = 0.7337
```

Quindi possiamo pensare di verificare l'uguaglianza delle medie dei vari trattamenti lavorando con i reciproci dei tempi di sopravvivenza

```
> topiII.aov<-aov(1/tempo~trattamento)
> summary(topiII.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
trattamento  3 9.1424  3.0475  7.3913 0.004594 **
Residuals   12 4.9477  0.4123
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I risultati evidenziano una diversità nelle medie dei reciproci dei tempi di sopravvivenza per i diversi trattamenti.

8.3 Analisi del dataset MORLEY.DAT

I dati riportati nel file `Morley.dat` rappresentano misurazioni della velocità della luce nell'aria fatte tra il 5/6/1879 e il 2/7/1879. I dati sono stati raccolti in 5 esperimenti (numerati da 1 a 5), ognuno dei quali consisteva di 20 misurazioni (dette `Run`) della velocità della luce nell'aria (in km/s). Si vuole valutare se esistono differenze significative nelle medie delle 5 popolazioni da cui provengono i dati.
Leggiamo i dati.

```
> mor <- read.table("I:/modelli/Morley.dat", header=T)
> mor
    Expt Run Speed
 1     1   1   850
 2     1   2   740
 ..............
 98    5  18   800
 99    5  19   810
100    5  20   870
```

Comunichiamo a `R` che `Expt` è una etichetta che individua i 5 esperimenti:

```
> mor$Expt <- factor(mor$Expt)
> attach(mor)
```

Effettuiamo un'analisi preliminare dei dati. Volendo effettuare un'analisi della varianza vogliamo esplorare l'ipotesi di normalità ed omoschedasticità della distribuzione della velocità all'interno dei gruppi.

```
> par(mfrow=c(1,1))
> plot(Speed~Expt)
```

Circa l'omoschedasticità, i quattro gruppi mostrano varianza campionaria abbastanza diversa. Basandoci solo sulla valutazione grafica, forse l'ipotesi di omoschedasticità pare un po' azzardata.
Circa la normalità, possiamo dire che i gruppi 2 e 4 presentano una distribuzione sufficientemente simmetrica, mentre gli altri gruppi, in particolare il 3 ed il 5, mostrano asimmetria. Per saggiare graficamente l'ipotesi di normalità possiamo comunque usare i `qqnorm`.

```
> par(mfrow=c(2,3), pty="s")
> qqnorm(Speed[Expt==1])
> qqline(Speed[Expt==1])
> qqnorm(Speed[Expt==2])
> qqline(Speed[Expt==2])
> qqnorm(Speed[Expt==3])
> qqline(Speed[Expt==3])
> qqnorm(Speed[Expt==4])
> qqline(Speed[Expt==4])
> qqnorm(Speed[Expt==5])
> qqline(Speed[Expt==5])
```

Anche i `qqnorm` indicano che l'assunzione di normalità è ragionevole per i gruppi 2 e 4, mentre ci sono delle anomalie negli altri 3 gruppi.
Il terzo gruppo in particolare è anomalo. Cerchiamo di capire a che cosa è dovuta la strana forma della distribuzione.

```
> summary(Speed[Expt==3])
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   620     840     855     845     880     970
> sort(Speed[Expt==3])
 [1] 620 720 720 840 840 840 840 840 850 850
[11] 860 860 870 880 880 880 880 910 950 970

> par(mfrow=c(1,1))
> hist(Speed[Expt==3])
```

In pratica, le prime tre osservazioni risultano essere *outliers* rispetto alle altre (vedi `boxplot` e `qqnorm`). Questo ha l'effetto di alterare tutta l'immagine della distribuzione.
Per il momento ignoriamo il problema della non perfetta normalità, della possibile eteroschedasticità e procediamo come se le ipotesi dell'ANOVA fossero rispettate.

```
> a <- aov(Speed~Expt)
> a
Call:
   aov(formula = Speed ~ Expt)

Terms:
                   Expt Residuals
Sum of Squares    94514    523510
Deg. of Freedom       4        95

Residual standard error: 74.23363
Estimated effects may be unbalanced
```

Per avere informazioni circa l'esito del test F utilizziamo la funzione `summary`:

```
> summary(a)
            Df Sum Sq Mean Sq F value   Pr(>F)
Expt         4  94514   23629  4.2878 0.003114 **
Residuals   95 523510    5511
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La funzione ci presenta la classica tabella di scomposizione della varianza. La quantità indicata con `Pr(>F)` indica il p-value. Questo è ottenibile come:

```
> 1-pf(4.2878, 4, 95)
[1] 0.003114458
```

Per avere invece la soglia della regione di rifiuto, calcoliamo il quantile di livello 0.95 per una distribuzione $F_{4,95}$.

```
> qf(0.95, 4, 95)
[1] 2.467494
```

Conclusioni: Assumendo normalità e omoschedasticità delle popolazioni, l'evidenza empirica porta al rifiuto dell'ipotesi di uguaglianza delle medie delle 5 popolazioni.

Per provare a risolvere il problema del gruppo 3, possiamo ripetere l'analisi dopo aver eliminato gli outliers nel gruppo 3

```
> Speed[Expt==3]
 [1] 880 880 880 860 720 720 620 860 970 950
[11] 880 910 850 870 840 840 850 840 840 840
```

Le tre osservazioni di interesse occupano la posizione 45, 46, 47 nel dataframe. Possiamo allora crearci un nuovo dataframe in cui non siano presenti le tre righe di interesse.

```
> mor1 <- mor[-(45:47),]
> detach()
> attach(mor1)

> par(mfrow=c(1,2))
> plot(Speed~Expt)
> hist(Speed[Expt==3])
```

La situazione rimane anomala. Proviamo a vedere alcune statistiche di sintesi sulla velocità nel terzo gruppo.

```
> summary(Speed[Expt==3])
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  840.0   840.0   860.0   872.9   880.0   970.0
> sort(Speed[Expt==3])
[1] 840 840 840 840 840 850 850 860 860 870 880 880 880 880 910 950 970
```

Il primo quartile coincide con il minimo. Questo è dovuto al fatto che il valore minimo è stato osservato per ben 5 volte, cioè più del 25% dei casi. La distribuzione appare asimmetrica e l'ipotesi di normalità pare comunque forzata.
A questo punto possiamo provare a ripetere l'ANOVA, anche se rimangono le perplessità sulle assunzioni di base che avevamo sollevato nel caso precedente.

```
> a2 <- aov(Speed~Expt)
> a2
Call:
   aov(formula = Speed ~ Expt)

Terms:
```

```
                    Expt Residuals
Sum of Squares   98043.2  428362.9
Deg. of Freedom        4        92

Residual standard error: 68.23576
Estimated effects may be unbalanced

> summary(a2)
            Df Sum Sq Mean Sq F value    Pr(>F)
Expt         4  98043   24511  5.2642 0.0007301 ***
Residuals   92 428363    4656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusioni: Non cambiano rispetto a prima.
Quando le assunzioni non sono soddisfatte, è possibile ricorrere a diversi test. In questo caso si potrebbe utilizzare il test non parametrico di Kruskal-Wallis, attraverso il comando:

```
> kruskal.test(Speed,Expt)

        Kruskal-Wallis rank sum test

data:  Speed and Expt
Kruskal-Wallis chi-squared = 15.0221, df = 4, p-value = 0.004656
```

L'ipotesi nulla di uguaglianza dei parametri di posizione delle distribuzioni viene rifiutata.

```
> detach()
```