Piani fattoriali in R

Mauro Gasparini

3 Dicembre 2024

Riferimento bibliografico

I seguenti esempi sono tratti da

McClave JT., Benson PG. e Sincich T. (2014).
Statistics for Business and Economics.
Pearson Education Limited.

Anova a una via

Si vogliono studiare dapprima gli effetti di 4 tipi diversi (A,B,C,D) di palla da golf sulla variabile risposta, cioé la distanza ottenuta con un tiro standard, cioé un tiro fatto da un robot (ANOVA a un fattore, o a una via). Se si assume che i tipi di palla vengano affidati ai tiri in maniera casuale, questo piano sperimentale ‘e chiamato **piano (ad un fattore) completamente randomizzato**.

Prima ripuliamo il nostro ambiente da possibili variabili con lo stesso nome

```
rm(list=ls())
```

Leggiamo i dati direttamente da linea.

```
golf1 <- read.table(header=T, text='
tipo distanza
   A    226.4
   A    232.6
   A    234.0
   A    220.7
   A    163.8
   A    179.4
   A    168.6
   A    173.4
   B    238.3
   B    231.7
   B    227.7
   B    237.2
   B    184.4
   B    180.6
   B    179.5
   B    186.2
   C    240.5
   C    246.9
   C    240.3
   C    244.7
   C    179.0
```

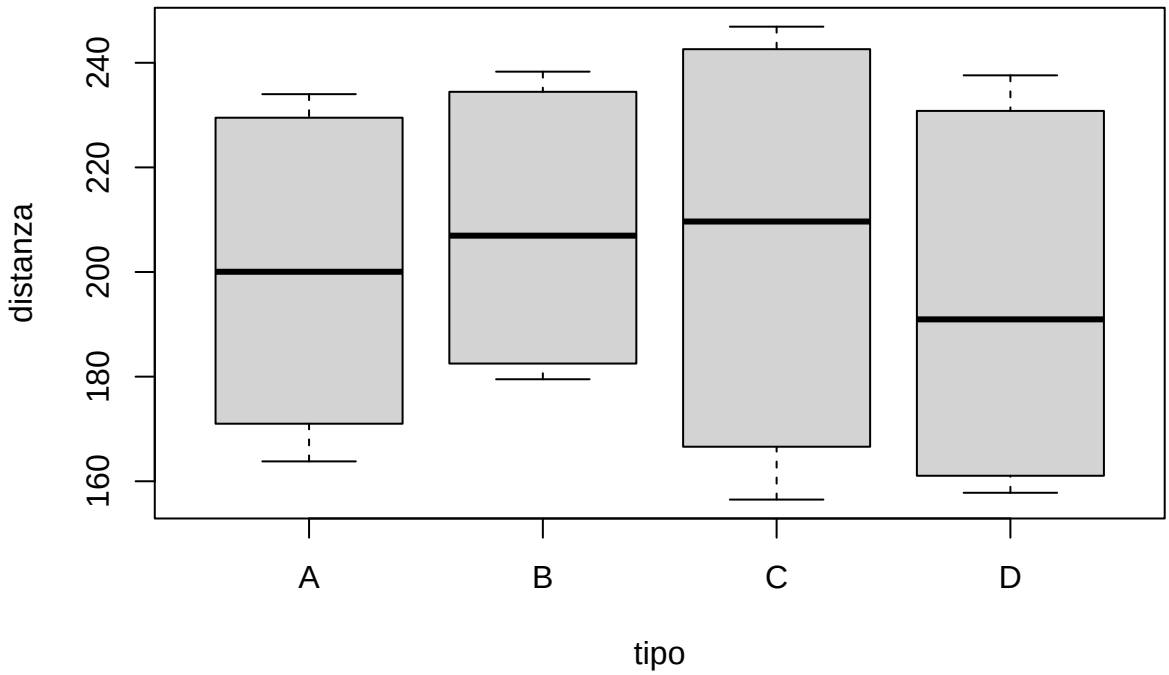
```
   C    168.0
   C    165.2
   C    156.5
   D    219.8
   D    228.7
   D    232.9
   D    237.6
   D    157.8
   D    161.8
   D    162.1
   D    160.3
')
attach(golf1)
summary(golf_lm <- lm(distanza ~ tipo))      ### con lm()
```

```
##
## Call:
## lm(formula = distanza ~ tipo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.638 -31.703  -0.481  32.947  42.475
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  199.862     12.262  16.299 8.04e-16 ***
## tipoB          8.338     17.341   0.481    0.634
## tipoC          5.275     17.341   0.304    0.763
## tipoD         -4.737     17.341  -0.273    0.787
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.68 on 28 degrees of freedom
## Multiple R-squared:  0.02322,    Adjusted R-squared:  -0.08143
## F-statistic: 0.2219 on 3 and 28 DF,  p-value: 0.8804
```

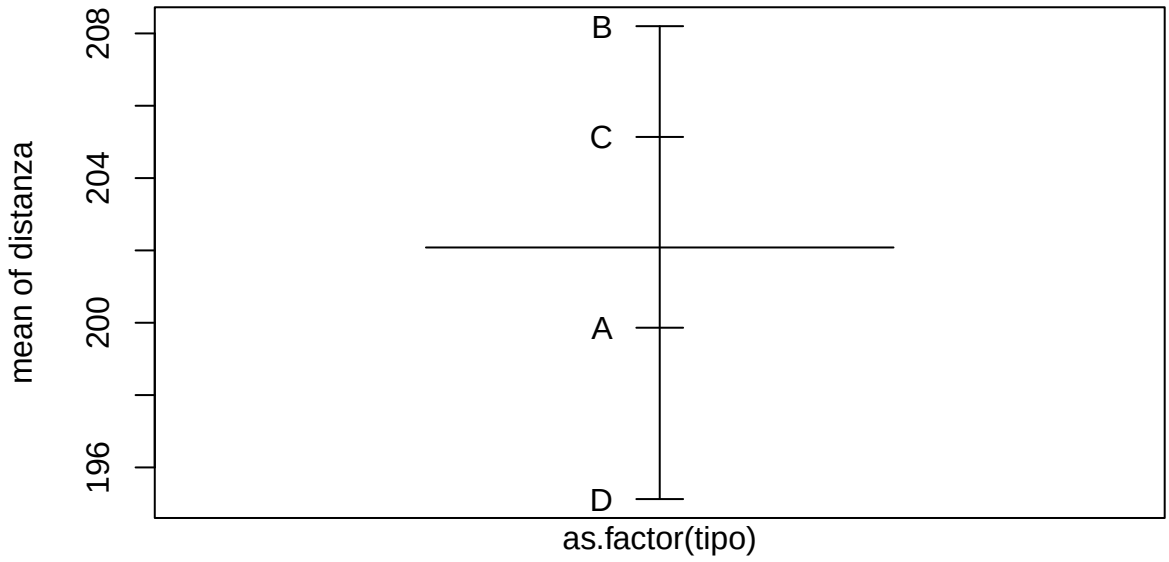
```
summary(golf_oneway <- aov(distanza ~ tipo))  ### con aov()
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## tipo         3    801   266.9   0.222   0.88
## Residuals   28  33681  1202.9
```

```
# questi due oggetti diversi danno informazioni diverse, quali?
boxplot(distanza ~ tipo)
```




```
plot.design(distanza ~ as.factor(tipo))
```

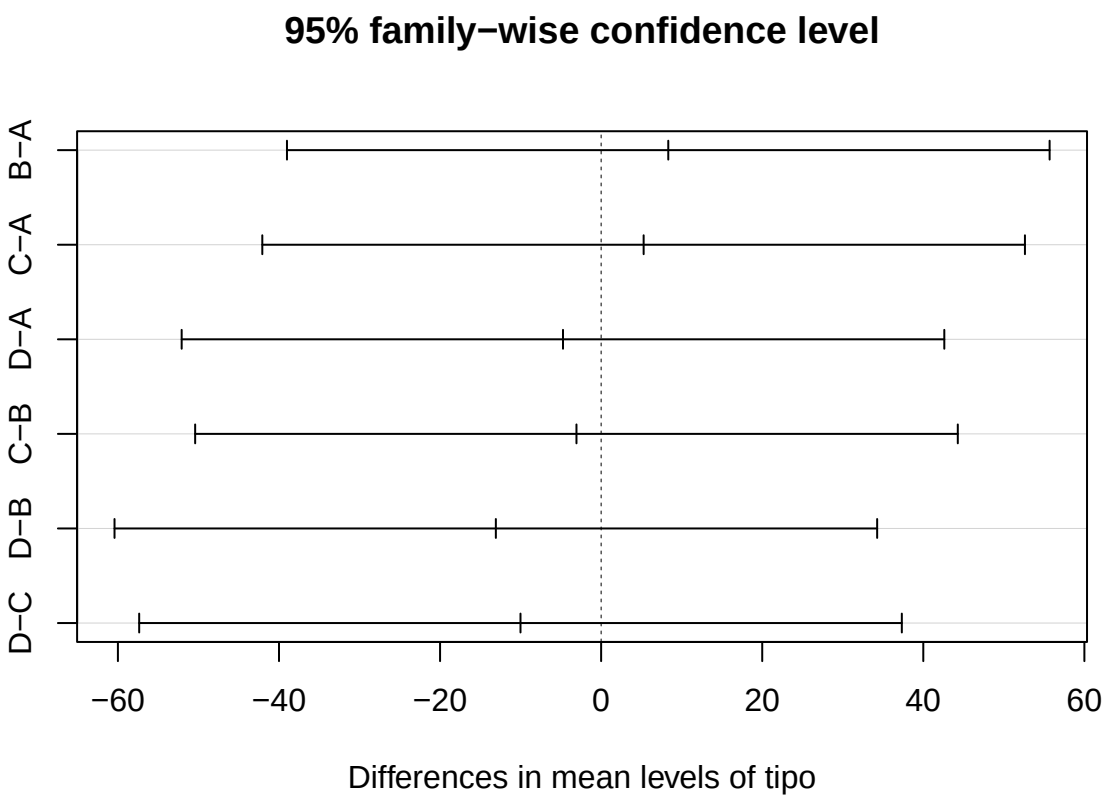



```
# gli intervalli di Tukey sono dei particolari metodi di inferenza multipla
TukeyHSD(golf_oneway)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = distanza ~ tipo)
##
## $tipo
##         diff       lwr      upr     p adj
```

```
## B-A   8.3375 -39.01007 55.68507 0.9627248
## C-A   5.2750 -42.07257 52.62257 0.9900118
## D-A  -4.7375 -52.08507 42.61007 0.9927111
## C-B  -3.0625 -50.41007 44.28507 0.9979954
## D-B -13.0750 -60.42257 34.27257 0.8741963
## D-C -10.0125 -57.36007 37.33507 0.9380369
```

```
plot(TukeyHSD(golf_oneway))
```




```
detach(golf1)
```

ANOVA a due vie in un piano fattoriale completo

In un secondo momento, si aggiunge il fattore mazza nei due livelli forniti da due mazze diverse (DRIVER e IRON). In questo piano sperimentale La combinazione tipo/mazza viene usata dal robot con 4 repliche per ciascuna combinazione, quindi il piano viene detto **piano fattoriale completo a due fattori** (o anche ANOVA a due fattori, o a due vie).

I dati in formato largo sono

```
golf2wide <- read.table(header=T, text='
mazza  A       B       C       D
DRIVER  226.4   238.3   240.5   219.8
DRIVER  232.6   231.7   246.9   228.7
DRIVER  234 227.7   240.3   232.9
DRIVER  220.7   237.2   244.7   237.6
IRON    163.8   184.4   179 157.8
IRON    179.4   180.6   168 161.8
IRON    168.6   179.5   165.2   162.1
IRON    173.4   186.2   156.5   160.3
```

```
')
```

E' preferibile un formato lungo (long) che si può ottenere in R con il pacchetto tidyr e l'istruzione gather() (ci sono altre possibilità, vedi laboratorio).

```
#install.packages("tidyr")
library(tidyr)
golf2 <- gather(golf2wide, tipo, distanza, A:D)
golf2
```

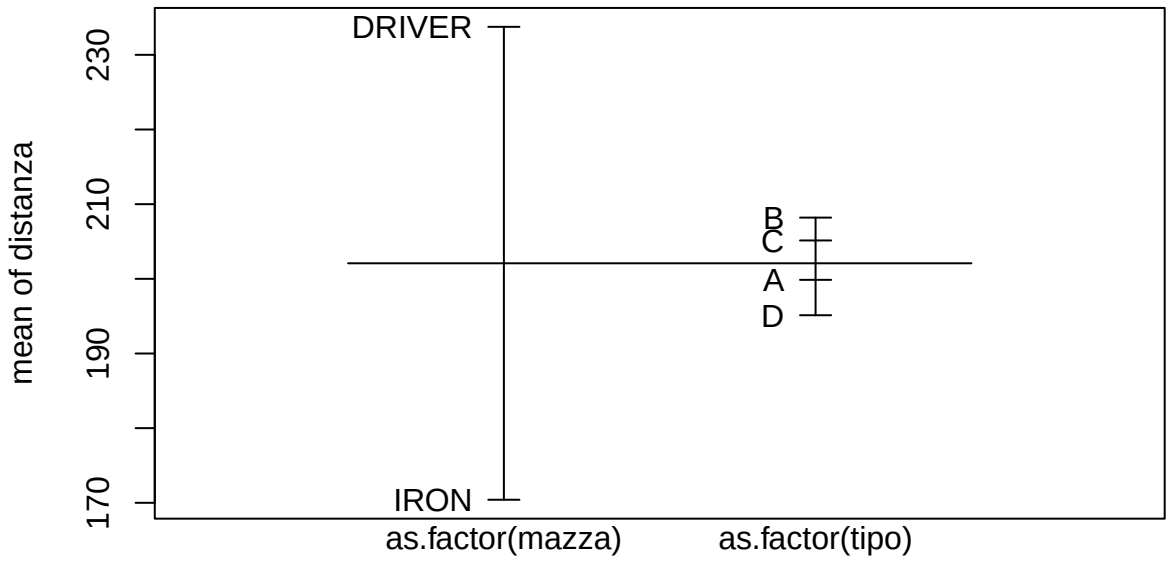
```
##     mazza tipo distanza
## 1  DRIVER    A    226.4
## 2  DRIVER    A    232.6
## 3  DRIVER    A    234.0
## 4  DRIVER    A    220.7
## 5    IRON    A    163.8
## 6    IRON    A    179.4
## 7    IRON    A    168.6
## 8    IRON    A    173.4
## 9  DRIVER    B    238.3
## 10 DRIVER    B    231.7
## 11 DRIVER    B    227.7
## 12 DRIVER    B    237.2
## 13   IRON    B    184.4
## 14   IRON    B    180.6
## 15   IRON    B    179.5
## 16   IRON    B    186.2
## 17 DRIVER    C    240.5
## 18 DRIVER    C    246.9
## 19 DRIVER    C    240.3
## 20 DRIVER    C    244.7
## 21   IRON    C    179.0
## 22   IRON    C    168.0
## 23   IRON    C    165.2
## 24   IRON    C    156.5
## 25 DRIVER    D    219.8
## 26 DRIVER    D    228.7
## 27 DRIVER    D    232.9
## 28 DRIVER    D    237.6
## 29   IRON    D    157.8
## 30   IRON    D    161.8
## 31   IRON    D    162.1
## 32   IRON    D    160.3
```

Ora facciamo una analisi ANOVA a due vie, perché abbiamo due fattori bilanciati.

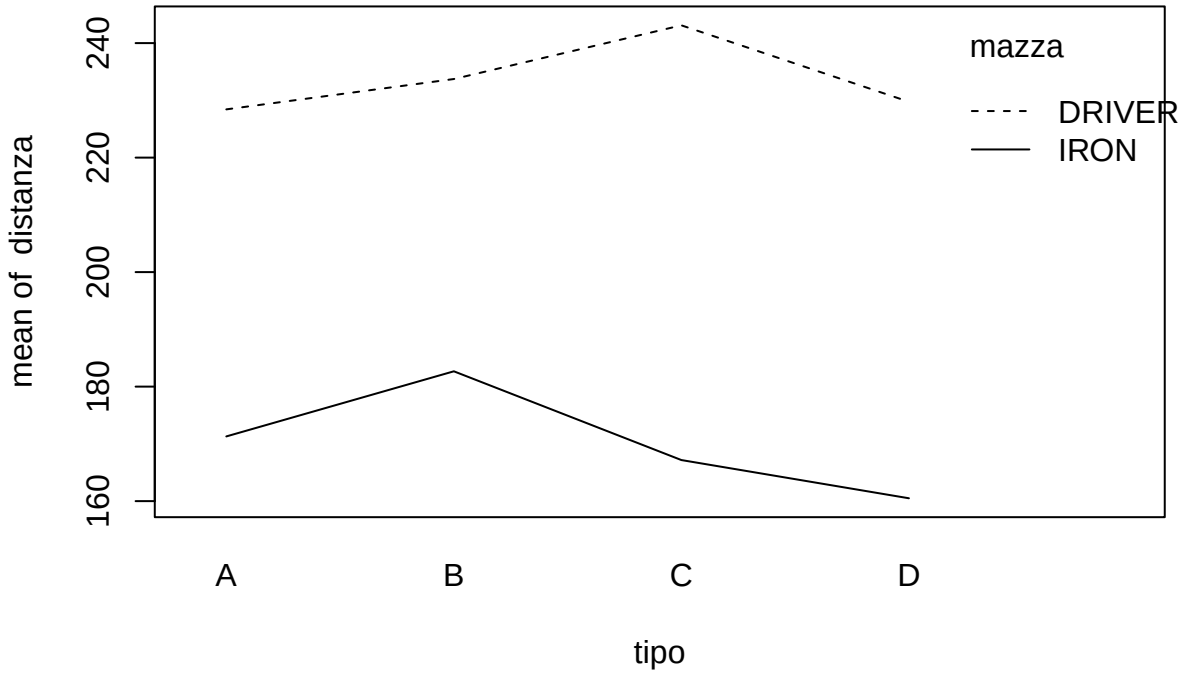
```
# Controlliamo che il piano sperimentale sia bilanciato:
attach(golf2)
table(mazza, tipo)
```

```
##         tipo
## mazza    A B C D
##   DRIVER 4 4 4 4
##   IRON   4 4 4 4
```

```r
#mazza <- as.factor(mazza)
#tipo <- as.factor(tipo)
# Una prima sommaria indagine:
plot.design(distanza ~ as.factor(mazza)*as.factor(tipo))
```




```r
# Disegnamo gli interaction plot:
interaction.plot(tipo, mazza, distanza)
```




```r
#Otteniamo la tabella ANOVA completa con i tre test:
model1 <- aov(distanza ~ mazza*tipo)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: distanza
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## mazza       1  32093   32093 936.7516 < 2.2e-16 ***
## tipo        3    801     267   7.7908 0.0008401 ***
## mazza:tipo  3    766     255   7.4524 0.0010789 **
## Residuals  24    822      34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Vediamo invece cosa ci dà il modello lineare:

summary(model2 <- lm(distanza ~ mazza*tipo))
```

```
##
## Call:
## lm(formula = distanza ~ mazza * tipo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6750  -2.7000   0.3125   3.4875  11.8250
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      228.425      2.927  78.051  < 2e-16 ***
## mazzaIRON        -57.125      4.139 -13.802 6.55e-13 ***
## tipoB              5.300      4.139   1.281  0.21259
## tipoC             14.675      4.139   3.546  0.00165 **
## tipoD              1.325      4.139   0.320  0.75163
## mazzaIRON:tipoB    6.075      5.853   1.038  0.30966
## mazzaIRON:tipoC  -18.800      5.853  -3.212  0.00373 **
## mazzaIRON:tipoD  -12.125      5.853  -2.072  0.04923 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.853 on 24 degrees of freedom
## Multiple R-squared:  0.9762, Adjusted R-squared:  0.9692
## F-statistic: 140.4 on 7 and 24 DF,  p-value: < 2.2e-16
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: distanza
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## mazza       1  32093   32093 936.7516 < 2.2e-16 ***
## tipo        3    801     267   7.7908 0.0008401 ***
## mazza:tipo  3    766     255   7.4524 0.0010789 **
## Residuals  24    822      34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
detach(golf2)
```

ANOVA a due vie in un piano fattoriale a blocchi randomizzati

A 10 giocatori di golf (GOLFER) vengono assegnate 10 sequenze casuali delle 4 marche di palle da golf (A,B,C,D). Viene poi misurata la distanza ottenuta da essi con un tiro standard sulle diverse palle. I dati in formato largo vengono trasformati in formato lungo.

```
golf3wide <- read.table(header=T, text='
GOLFER      A       B       C       D
1       202.4   203.2   223.7   203.6
2       242     248.7   259.8   240.7
3       220.4   227.3   240     207.4
4       230     243.1   247.7   226.9
5       191.6   211.4   218.7   200.1
6       247.7   253     268.1   244
7       214.8   214.8   233.9   195.8
8       245.4   243.6   257.8   227.9
9       224     231.5   238.2   215.7
10      252.2   255.2   265.4   245.2
')
golf3 <- gather(golf3wide, tipo, distanza, A:D)
golf3
```

```
##    GOLFER tipo distanza
## 1       1    A    202.4
## 2       2    A    242.0
## 3       3    A    220.4
## 4       4    A    230.0
## 5       5    A    191.6
## 6       6    A    247.7
## 7       7    A    214.8
## 8       8    A    245.4
## 9       9    A    224.0
## 10     10    A    252.2
## 11      1    B    203.2
## 12      2    B    248.7
## 13      3    B    227.3
## 14      4    B    243.1
## 15      5    B    211.4
## 16      6    B    253.0
## 17      7    B    214.8
## 18      8    B    243.6
## 19      9    B    231.5
## 20     10    B    255.2
## 21      1    C    223.7
## 22      2    C    259.8
## 23      3    C    240.0
## 24      4    C    247.7
## 25      5    C    218.7
## 26      6    C    268.1
## 27      7    C    233.9
## 28      8    C    257.8
```

```
## 29      9    C    238.2
## 30     10    C    265.4
## 31      1    D    203.6
## 32      2    D    240.7
## 33      3    D    207.4
## 34      4    D    226.9
## 35      5    D    200.1
## 36      6    D    244.0
## 37      7    D    195.8
## 38      8    D    227.9
## 39      9    D    215.7
## 40     10    D    245.2
```

Tale piano sperimentale 'e chiamato **piano a blocchi randomizzati**, in quanto ogni GOLFER fa da blocco di osservazioni omogenee di cui tenere conto, mentre il vero fattore di interesse è il tipo di palla.

Costruiamo un modello appropriato golfers.aov (senza interazione: perché?) e otteniamo la tabella ANOVA, da cui concludiamo che i GOLFER differiscono tra loro, come del resto anche i tipi di palla.

```
attach(golf3)

summary(golfers.aov <- aov(distanza ~ as.factor(GOLFER) +  as.factor(tipo)))
```

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(GOLFER)  9  12074  1341.5   66.27 4.50e-16 ***
## as.factor(tipo)    3   3299  1099.6   54.31 1.45e-11 ***
## Residuals         27    547    20.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

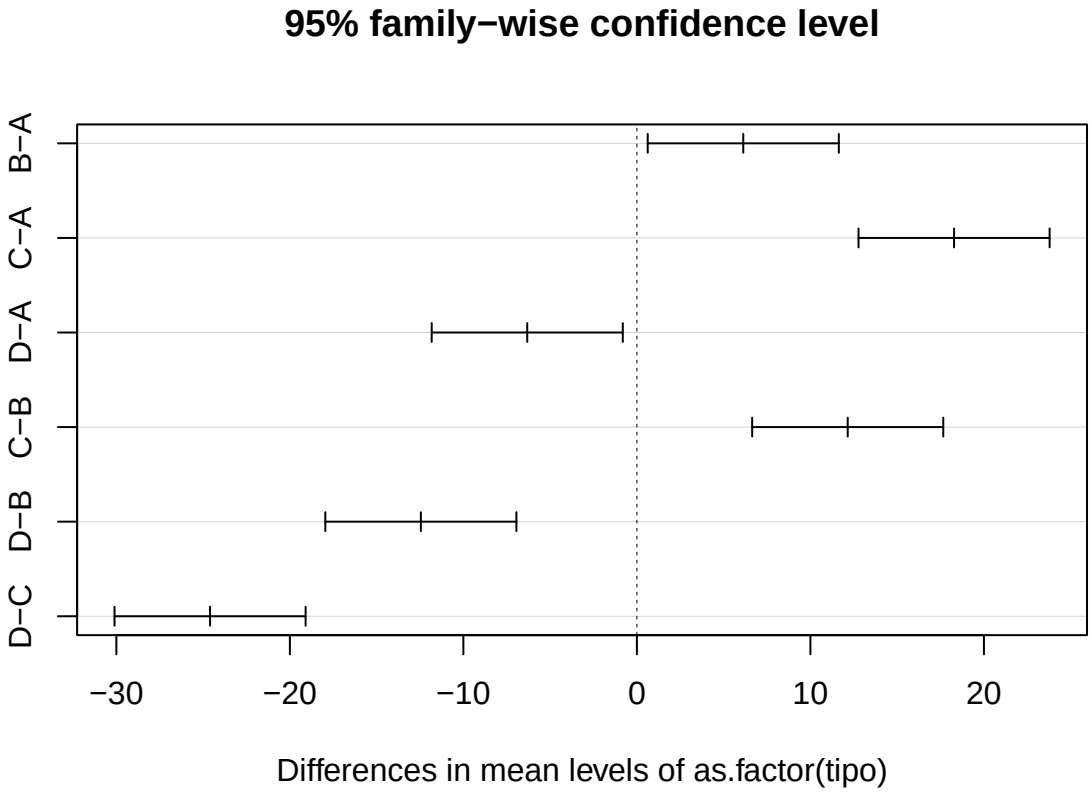
```
detach(golf3)
```

Ecco il risultato

```
                  Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(GOLFER)  9  12074  1341.5   66.27 4.50e-16 ***
tipo               3   3299  1099.6   54.31 1.45e-11 ***
Residuals         27    547    20.2
```

Per ottenere gli intervalli di Tukey per tutte le differenze tra tipi (non GOLFER), digitare

```
plot(TukeyHSD(golfers.aov,2))
```


95% family−wise confidence level
B−A
C−A
D−A
C−B
D−B
D−C
−30
−20
−10
0
10
20
Differences in mean levels of as.factor(tipo)