Labbrapport i Statistik

Laboration 11

732G46

Mattias Hällgren, Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-11-22

Innehåll

Introduktion

I denna laboration kommer ett dataset analyseras.

I datasetet som kommer från en studie där effekten av respondenternas ögonkontakt (Faktor A) mäts och respondentens kön (Faktor B) för att jämföra sannolikheten att få jobbet. 10 män och 10 kvinnor fick betygsätta en personens chans att få jobbet beroende på huruvida närvarande de var på en skala 0 (Totalt underkänt) till 20 (utomordentlig prestation).

Målet med denna laboration är att lära sig formulera tvåvägs ANOVA-modell på två olika sätt och förkalra hur modellens parameterar kan tolkas. Förstå begreppet interaktion. Använda data för att göra inferens med modellparametrarna, samt använda dator för att lösa balanserade tvåvägs-ANOVA-problem.

Databehandling

```
Eye_contact <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/Eye_contact") #Ladda
colnames(Eye_contact) <-c("Rating","Presence","Gender","Groups")#Namnger kolumnerna
Eye_contact$`Presence`<-as.factor(Eye_contact$`Presence`)
Eye_contact$`Gender`<-as.factor(Eye_contact$`Gender`)
```

Uppgifter

19.12 Eye contact effect

a) Obtain the fitted values for ANOVA model (19.23)

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
##  9.2  9.2  9.2  9.2  9.2 13.6 13.6 13.6 13.6 13.6 13.0 13.0 13.0 13.0 13.0 16.4
##   17   18   19   20
## 16.4 16.4 16.4 16.4
```

Ovan syns de anpassade värdena för Aov modellen, för de första 5 observationerna där det är Män som närvarande så är medelvärdet 9.2. I den andra grupppen som är närvarande kvinnor är medelvärdet 13.6. För den tredje gruppen som är frånvarande män är medelvärdet 13.0. Slutligen är medelvärdet för frånvarande kvinnor 16.4.

b) Obtain the residulas. Do they sum to zero for each treatment?

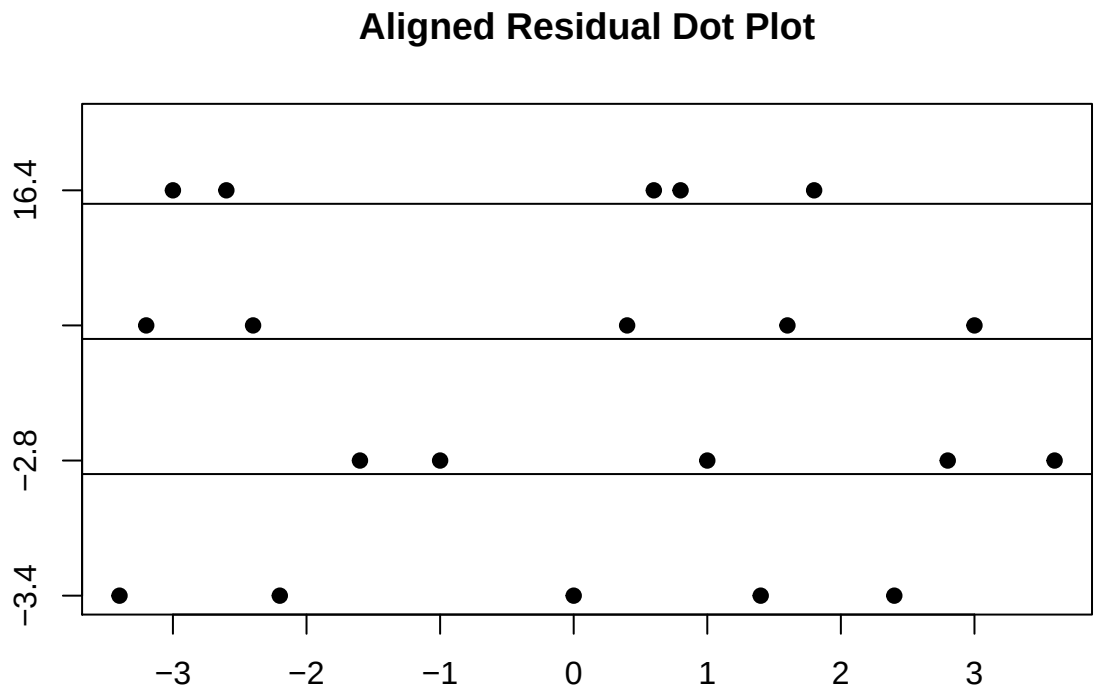
```
residuals(Eye_c)
```

```
##           1           2           3           4           5           6
##  1.8000e+00 -2.2000e+00  2.8000e+00 -3.2000e+00  8.0000e-01  1.4000e+00
##           7           8           9          10          11          12
## -1.6000e+00  4.0000e-01 -2.6000e+00  2.4000e+00 -1.0000e+00  3.0000e+00
##          13          14          15          16          17          18
## -3.0000e+00 -1.6653e-16  1.0000e+00 -2.4000e+00  6.0000e-01 -3.4000e+00
##          19          20
##  3.6000e+00  1.6000e+00
```

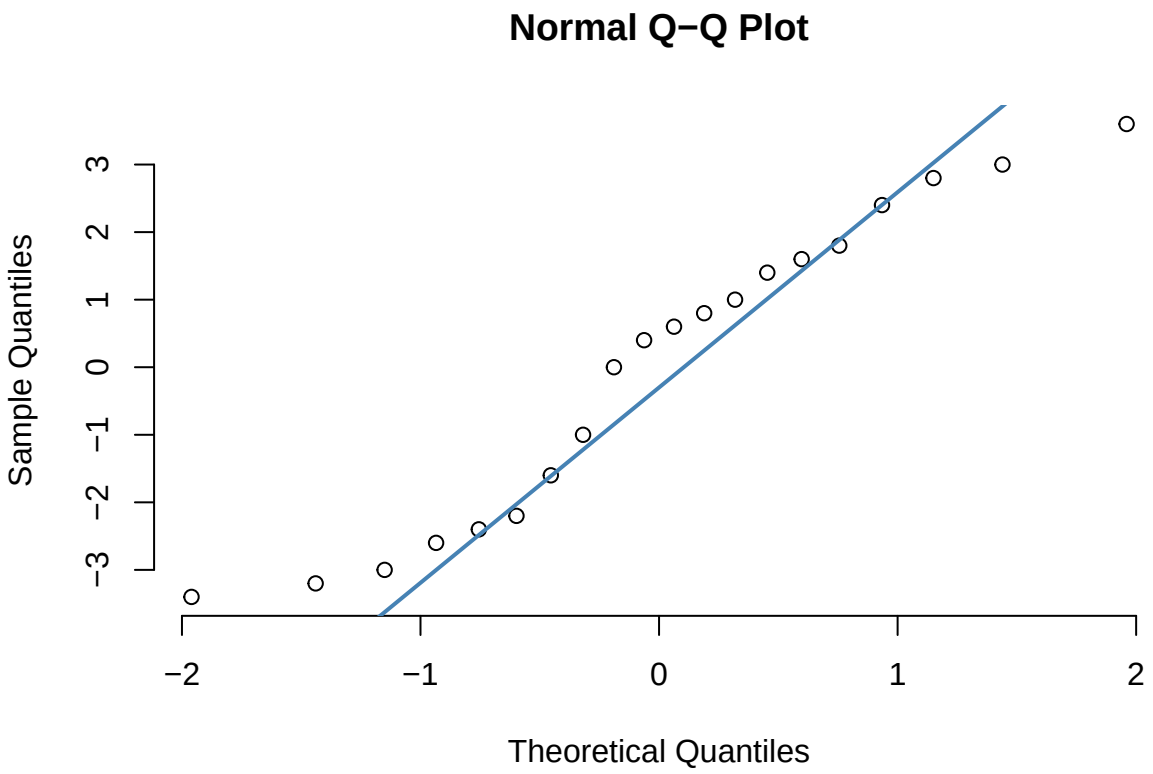
Ja, för nästintill varenda observation syns det hur extremt små residualvärdena som summeras till noll returneras.

c) Prepare aligned fresidual dot plots for the treatments. What depratues from ANOVA model (19.23) can be studied from these plots? What are your findings?

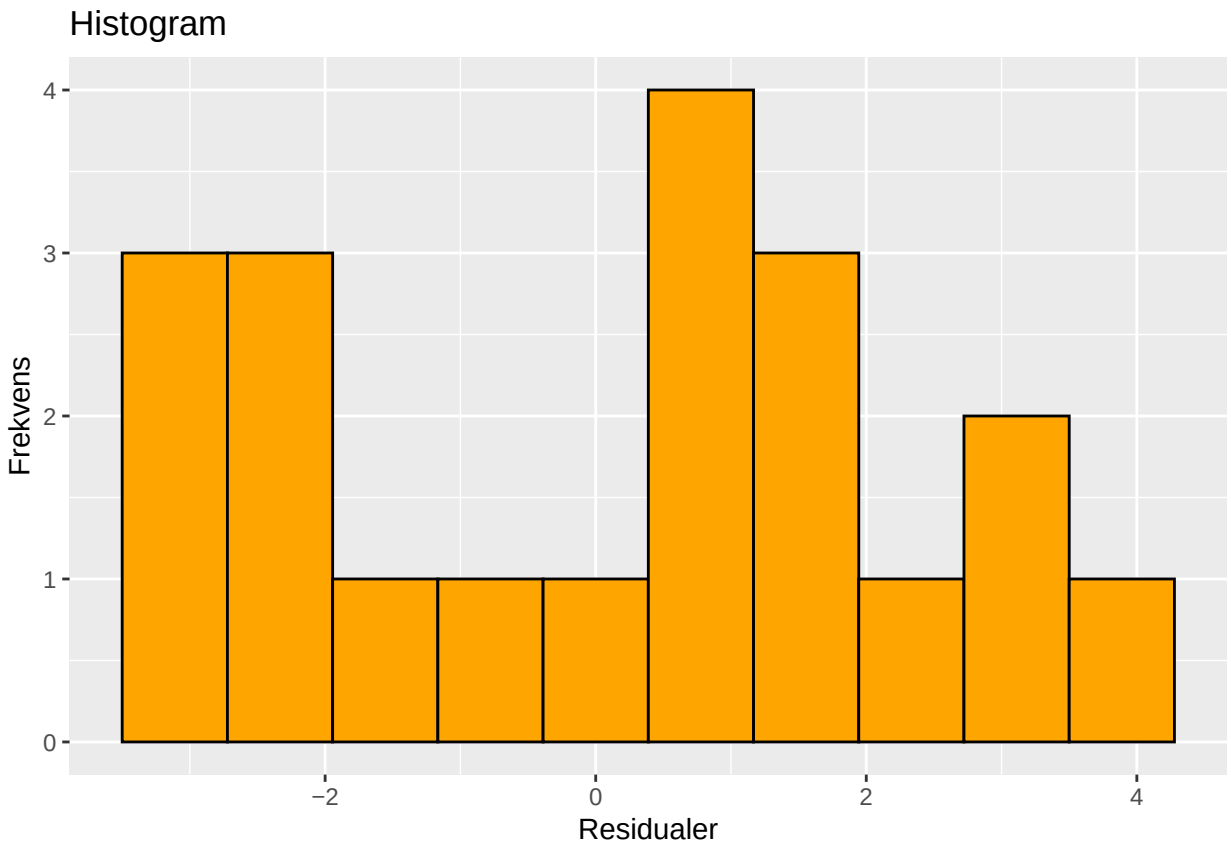
```
stripchart(split(resid(Eye_c), Eye_c$coefficients), method = "stack",  pch = 19)
abline(h = seq(2, 4)-0.1)
title("Aligned Residual Dot Plot")
```



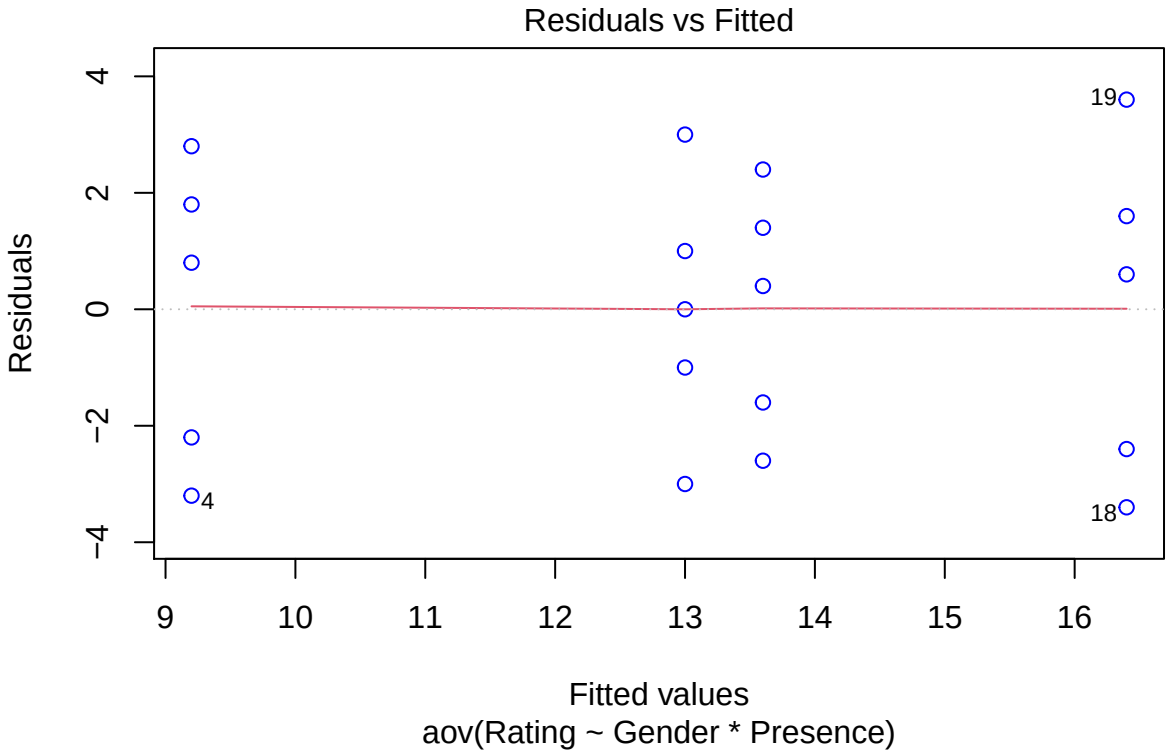
I den linjära residualplotten syns spridning av residualerna baserat på faktorgrupp, residualerna är jämnt fördelade i diagrammet. Däremot är grupper såsom grupp 1 på den första raden positivt skeva eller såsom grupp 3 på tredje raden negativt skeva.



Ovanför är en Q-Q Plot som mäter normalfördelning här kan man se att värdena varierar lite och ploten visar att datan inte är perfekt normalfördelad men att den ändå följer linjen. Man ser även några extremvärden, främst i början samt slutet.



På histogrammet ovanför så visar denna plot att det är tydligt att datan inte är normalfördelad eftersom detta histogram går mer som en våg istället för att vara en fin båge över staplarna. Histogrammet kan även vara en otydlig plot att göra när man ej har så många värden att arbeta med.

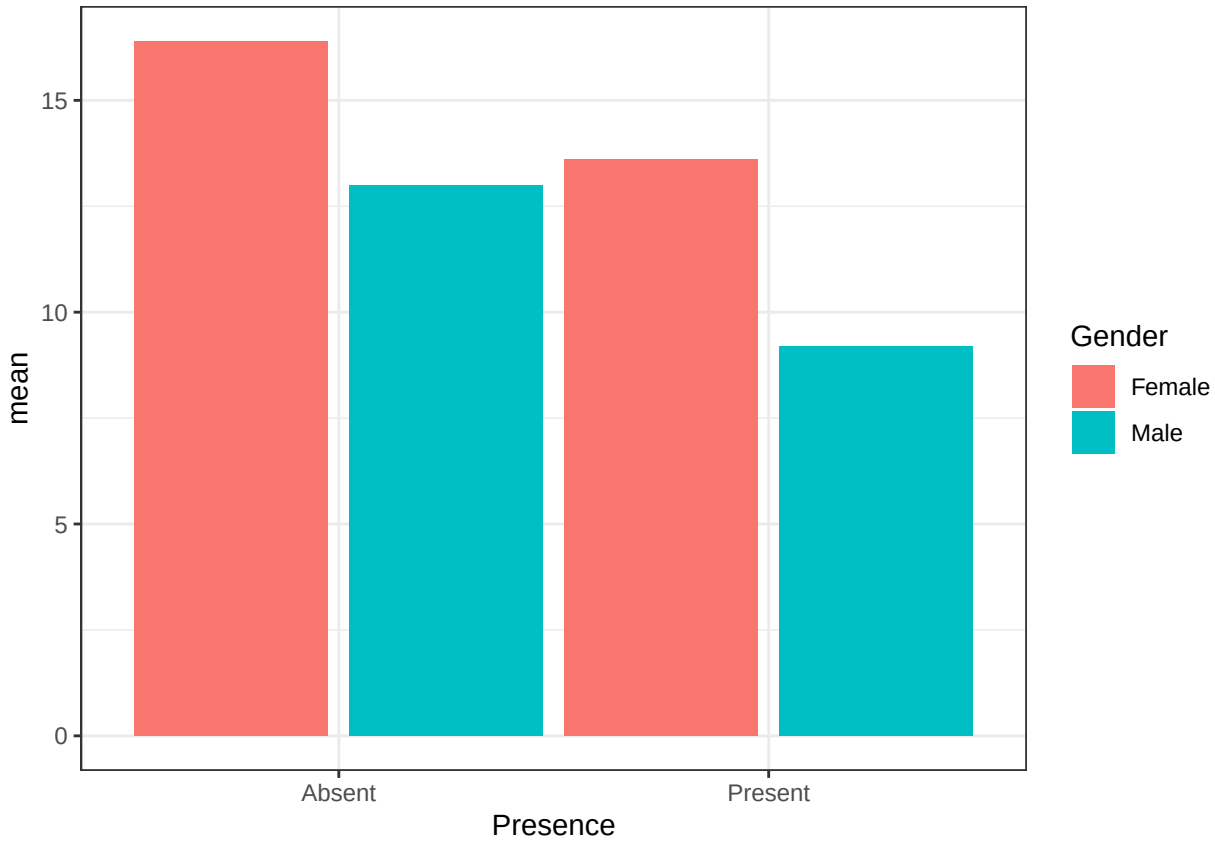


Ovan visas en jämförelse av anpassade värden och residualer, fördelningen är utspridd, däremot inte jämnt utspridd och tre extremvärden uppmärksammas (Observation 4,18 och 19).

19.13 Refer to Eye contact effect problem 19.12

a) Prepare an estimated treatment means plot. Does it appear that any factor effects are present? Explain

```
library(FSA)
Sum = Summarize(Rating ~ Presence*Gender,
                data=Eye_contact)
pd = position_dodge(1)
ggplot(Sum,                  ### The data frame to use.
       aes(x     = Presence,
           y     = mean,
           fill  = Gender)) +
  geom_bar(stat     = "identity",
           position = pd)+
  theme_bw()
```



I stapeldiagrammet ovan syns medelvärden för de olika faktorgrupperna, gruppen med högst medelvärde är gruppen med frånvarande kvinnor, följt av gruppen med närvarande kvinnor. På tredje plats gruppen med frånvarade män följt av gruppen med närvarande män. Slutsatserna som går att dra utifrån att kolla på

diagrammet är att icke närvarande respondenter i snitt får högre betyg i studien och att kvinnor i sin helhet får högre rating än män.

b) Set up the ANOVA table. Does any one source account for most of the total

variability in the success ratings in the study? Explain

```
anova(Eye_c)
```

```
## Analysis of Variance Table
##
## Response: Rating
##                 Df Sum Sq Mean Sq F value Pr(>F)
## Gender           1   76.0    76.0   12.52 0.0027 **
## Presence         1   54.4    54.4    8.96 0.0086 **
## Gender:Presence  1    1.2     1.2    0.21 0.6562
## Residuals       16   97.2     6.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ja! Det syns hur kön är den variabeln som är mest signifikant i detta fall med 4 gånger så lågt p-värde samt ett högre värde från F-testet gentemot variabeln närvaro. Trots detta är är fortfarande variabeln närvaro en signfikant variabel på 1 % signifikansnivå.

c) Test whether or not interaction effects are present; use a =.01 State the alternatives, decision rule, and conclusion. What is the P-value of the test?

$$H_0 : \sigma^2_{\alpha\beta} = 0$$
$$H_1 : \sigma^2_{\alpha\beta} > 0$$
$$F(.99, 1, 16) = 8.531$$

$$F^* = \frac{MSAB}{MSE} \Rightarrow \frac{1.25}{6.08} = 0.2056$$

I fall av att teststatistikan överstiger det kritiska värdet ($F > 8.531$) förkastas H_0 om att det inte råder interaktionseffekt mellan variablerna.

I detta fall då ($0.2056 < 8.531$) samtidigt som att vi får ett p-värde som summeras till (0.6562) kan vi med 95 % konfidens anta H_0 om att det inte råder en interaktionseffekt mellan variablerna. Alltså kan vi med 95 % konfidens säga att det inte råder någon interaktionseffekt mellan variablerna.

d) Test whether or not eye contact and gender main effects are present. In each,

use a=.01 and state the alternatives, decision rule, and conclusion. What is the

P-value of each test? Is it meaningful here to test for main factor effects? Explain

Gender

$$H_0 : \sigma_\alpha^2 = 0$$
$$H_1 : \sigma_\alpha^2 > 0$$

$$F(.99, 1, 16) = 8.531$$

$$F^* = \frac{MSA}{MSE} \Rightarrow \frac{76.05}{6.08} = 12.519$$

I fall av att teststatistikan överstiger det kritiska värdet ($F > 8.531$) förkastas H_0 om att det inte råder interaktionseffekt mellan variablerna.

I detta fall då ($12.519 > 8.531$) samtidigt som att vi får ett p-värde som summeras till ($0.00273 $) kan vi med 95 % konfidens förkasta H_0 om att kön inte har någon effekt på resultatet. Alltså kan vi med 95 % konfidens säga att att kön har effekt på resultatet.

Eye contact

$$H_0 : \sigma_\beta^2 = 0$$
$$H_1 : \sigma_\beta^2 > 0$$

$$F(.99, 1, 16) = 8.531$$

$$F^* = \frac{MSB}{MSE} \Rightarrow \frac{54.45}{6.08} = 8.96$$

I fall av att teststatistikan överstiger det kritiska värdet ($F > 8.531$) förkastas H_0 om att ögonkontakt inte har någon effekt på resulatet.

I detta fall då ($8.96 > 8.531$) samtidigt som att vi får ett p-värde som summeras till (0.00273) kan vi med 95 % konfidens förkasta H_0 om att ögonkontakt inte har någon effekt på resulatet. Alltså kan vi med 95 % konfidens säga att ögonkontakt har effekt på resultatet.

f) Do the results in parts(c) and (d) confirm your graphic analysis in part(a)?

```
options(digits=6)
anova(Eye_c)
```

```
## Analysis of Variance Table
##
## Response: Rating
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## Gender           1  76.05   76.05  12.519 0.00273 **
## Presence         1  54.45   54.45   8.963 0.00859 **
## Gender:Presence  1   1.25    1.25   0.206 0.65620
## Residuals       16  97.20    6.08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ja med tanke på att kvinnor generellt sett får ett högre medelvärde vad gäller rating och oberoende av närvaro får en kvinna som är närvarande högre rating än en män som är frånvarande, detta eftersom att kön är en starkare variabel sett till p-värde och teststatistiska på f-test.

19.31 Refer to Eye contact effect Problem 19.12 and 19.13

a) Estimate Estiamte u21 with a 99 percent confidence interval. Interpret

your interval estimate.

```
Mean_Absent_male <-13
n <-5
sqrt(s_y21 <- 2*6.08/10)
```

```
## [1] 1.10272
```

```
qt(0.995,16)
```

```
## [1] 2.92078
```

$$\bar{Y} \pm qt\left(1 - \frac{alfa}{2}, df\right)\sqrt{\frac{2 \cdot MSE}{nb}} = 13 \pm (2.9208)\sqrt{\frac{2 \cdot 6.08}{2 \cdot 5}} = 9.7789 : 16.221$$

I ett 99 % konfidensintervall kommer medelvärdet för en frånvarande män att ligga mellan 9.7789 : 16.221 Sett till intervallet på medelvärdet är det väldigt stort vilket kan förklaras med den höga konfidensgraden.

b) Estimate Estiamte u1 with a 99 percent confidence interval. Interpret

your interval estimate.

```
mean <-filter(Eye_contact, Presence == "Present")
mean(mean$Rating)
```

```
## [1] 11.4
```

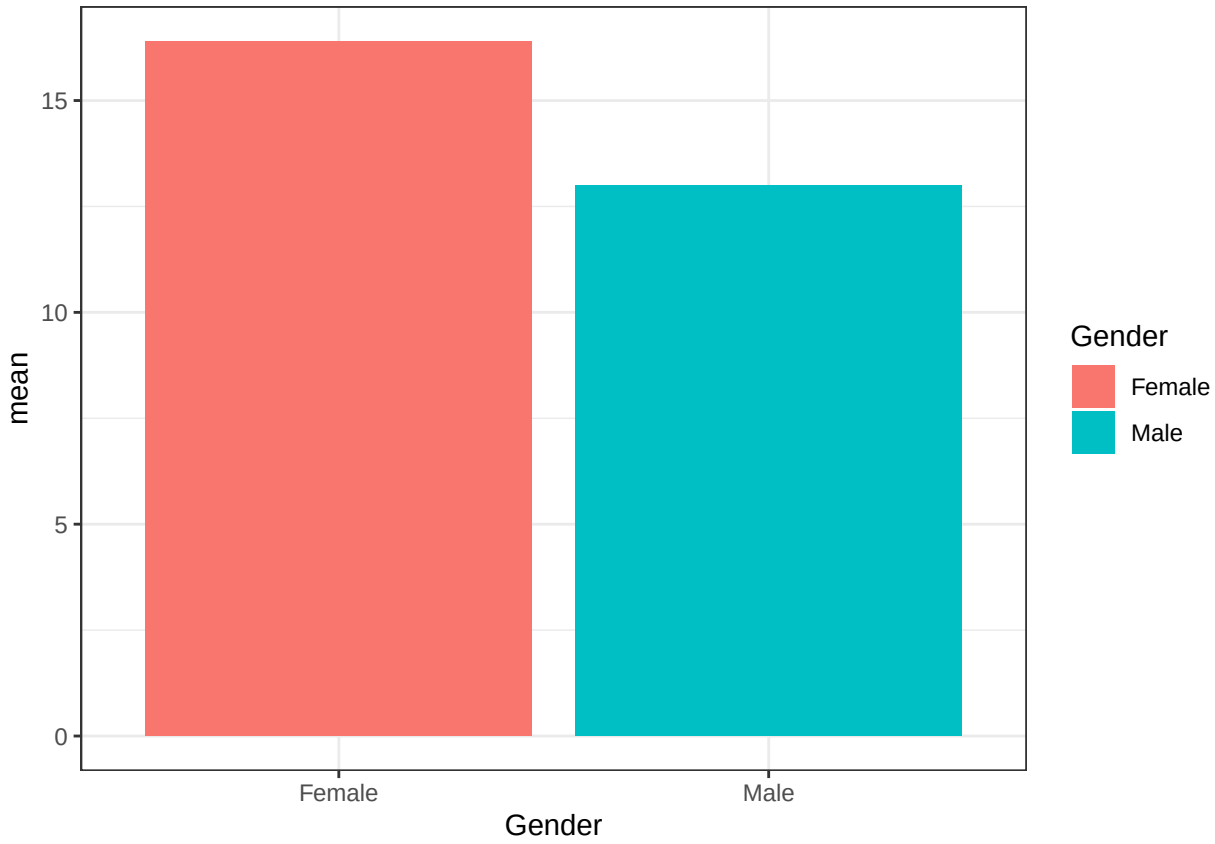
$$\bar{Y} \pm qt\left(1 - \frac{alfa}{2}, df\right)\sqrt{\frac{MSE}{bn}} = 11.4 \pm (2.9208)\sqrt{\frac{6.08}{10}} = 9.122 : 13.677$$

Medelvärdet för en respondent som är närvarande kommer medelvärdet på rating att ligga mellan 9.122 : 13.677 vilket är ett mycket mindre interval gentemot den ovanstående jämförelsen som genomfördes. Detta trots samma konfidensgrad.

c) Prepare a bar graph of the estimated Factor gender level means. What does this plot suggest about the Factor gender main effects?

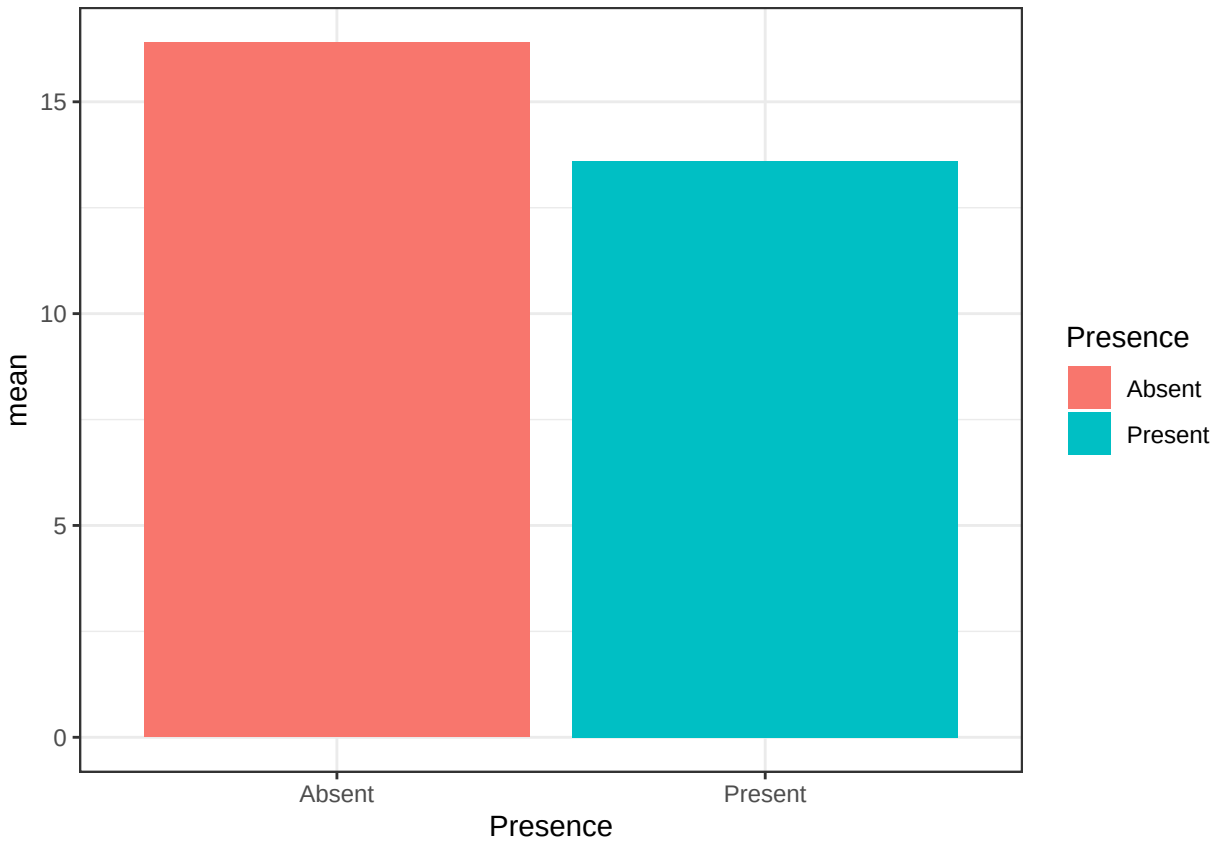
```
library(FSA)
SumT = Summarize(Rating ~ Presence*Gender,
                 data=Eye_contact)

pd = position_dodge(1)
ggplot(SumT,                  ### The data frame to use.
       aes(x     = Gender,
           y     = mean,
           fill  = Gender)) +
  geom_bar(stat     = "identity",
           position = pd)+
  theme_bw()
```



Stapeldiagrammet visar tydligt hur kvinnor generellt sett får högre Rating i testerna sett till medelvärde.

e) Prepare a bar graph of the estimated factor Presence level means. What does this plot suggest about the Factor gender main effects?



Stapeldiagrammet ovan visar hur stor skillnad det är mellan faktorvariabeln närvarande och frånvarande. Kopplar man detta till hur skillnaderna mellan män och kvinnor ser ut, kan slutsatsen kring att trots den stora skillnaden mellan närvaronivåerna, kan kvinnornas medelvärde trots att de tillhör gruppen med lägst medelvärde vad gäller närvarande, fortfarande få ett högre medelvärde än män.

f) Obtain confidence intervals for D1 =u2-u1 and D2=U2-u; use the Bonferroni

procedure and a 95 percent family confidence coefficent. Summarize your findings.

Are your findings consisent with those parts (c) and (e)

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = Rating ~ Gender * Presence, data = Eye_contact)
##
## $`Gender:Presence`
##                               diff        lwr      upr    p adj
## Male:Absent-Male:Present       3.8 -0.6598885  8.25989 0.109642
## Female:Present-Male:Present    4.4 -0.0598885  8.85989 0.053812
## Female:Absent-Male:Present     7.2  2.7401115 11.65989 0.001460
## Female:Present-Male:Absent     0.6 -3.8598885  5.05989 0.979929
## Female:Absent-Male:Absent      3.4 -1.0598885  7.85989 0.170744
## Female:Absent-Female:Present   2.8 -1.6598885  7.25989 0.310829
```

I samband med simultana jämförelser används två kriterier för att klassa en differens som signifikant eller inte, den första är att se ifall intervaller täcker 0 eller inte, den andra är att titta på p-värdet från tabellen och se ifall värdet över eller understiger signifikansnivån. I detta fall är signifikansnivån 5 %. De jämförelser som är signifikanta på denna nivå är enbart den tredje jämförelsen, varför den andra jämförelsen inte är signifikant på 5 % signifikansnivån beror på att intervallet täcker 0. Alltså är det bara jämförelsen mellan en närvarande kvinna och en närvarande man som är signifikant skild på 5 % signifikansnivå.

$s\left\{\bar{D}_{ij}\cdot\right\} = \sqrt{\frac{2*MSE}{n}} = \sqrt{\frac{2*6.08}{10}} = 1.1027$

$\hat{D}_1 \pm B \cdot s\left\{\hat{D}_{ij,i'j'}\right\} = 0.574 : 6.026$

För variabeln närvaro så kommer skillnaderna i medelvärde mellan frånvarande och närvarande i ett 95 % konfidens intervall att ligga mellan 0.574 och 6.026.

$\hat{D}_3 \pm B \cdot s\left\{\hat{D}_{ij,i'j'}\right\} = 1.174 : 6.626$

För variabeln Kön så kommer skillnaderna i medelvärde mellan Kvinna och Man att i ett 95 % konfidens intervall ligga mellan 1.174 och 6.626.

Lärdomar

Målet med denna laboration var att bekanta sig med två-vägsanovamodeller, i efterhand upplever laborationsgruppen att vi lyckats med detta och att det har varit intressant, i detta fall var interaktionen mellan variablerna intressant då Gender*Presence inte var en signifikant variabel, medan man kunde se signifikanta skillnader mellan de olika könen och deras närvaronivå. Kul att inferensen förekommer och man får möjlighet att repetera den!