ST437/537 – HW #05 - Solution

Due date: February 25, 2019

Instructions

Please follow the instructions below when you prepare and submit your assignment.

- **Include a cover-page** with your homework. It should contain

i. Full name,
ii. Course#: ST 437/537 and
iii. HW-#
iv. Submission date

- Assignments should be submitted in class on the date specified ("due date").
- Neatly typed or hand-written solution on standard letter-size papers (stapled on the top-left corner) should be submitted. **All R code/output should be well commented, with relevant outputs highlighted.**
- **Always staple (upper left corner) your homework before coming to class. Ten percent points will be deducted otherwise.**
- When you solve a particular problem, do not only give the final answer. Instead **show all your work** and the steps you used (with proper explanation) to arrive at your answer to get full credit.
- **DO NOT** give printouts of whole dataset or matrices. Present only the relevant output when answering a question.

Problems

Solve the following problems. You may use `R` for these problems unless I specifically instruct otherwise.

DO NOT give printouts of whole dataset or matrices. Present only the relevant output/graphs when answering a question.

Problem 1: (10 points) A researcher measured three indices (concerning severity of hreat attacks), X_1, X_2 and X_3, for each of $n = 40$ heart attack patients, and produced summary statistics:

$$\bar{x} = \begin{bmatrix} 46.1 \\ 57.3 \\ 50.4 \end{bmatrix} \qquad \boldsymbol{S} = \begin{bmatrix} 101.3 & 63.0 & 71.0 \\ 63.0 & 80.2 & 55.6 \\ 71.0 & 55.6 & 97.4 \end{bmatrix}.$$

Test for the equality of mean indices at $\alpha = 0.05$. [Hint: write a proper contrast matrix, and write H_0 first].

Define the vector of mean indices $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^T$. We want to test

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$

We can define a contrast matrix

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix},$$

and thus an equivalent hypothesis is

$$H_0 : \boldsymbol{C\mu} = \boldsymbol{0}.$$

We can test H_0 using the Hotelling's $T2$ test.

```
p <- 3
n <- 40

# xbar and S
xbar <- c(46.1, 57.3, 50.4)
S <- matrix( c(101.3, 63.0, 71.0,
63.0, 80.2, 55.6,
71.0, 55.6, 97.4), ncol = 3, byrow = T)

# contrast matrix
C <- cbind(c(1,1), -diag(1, 2))
q <- nrow(C)

# test
invCSC <- solve( C %*% S %*% (t(C)) )
Cxbar <- C %*% xbar
T2 <- n*(n-q)/((n-1)*q) * (t(Cxbar)) %*% invCSC %*% (Cxbar)

critical_value_F = qf(p = 0.05, df1 = q, df2 = n-q, lower.tail = F)

# p-value
pv <- pf(T2, df1 = q, df2 = n-q, lower.tail = F)

# results
results <- data.frame(T2 = T2, critical = critical_value_F,
                      df1 = q, df2 = n-q, pvalue = pv)
results
```

```
##        T2 critical df1 df2       pvalue
## 1 44.0871 3.244818   2  38 1.251548e-10
```

Since $T2$ is larger than the critical value (equivalently p-value is very small) we reject H_0 as $\alpha = 0.05$.

Problem 2: (20 points) Consider the anesthasia data discussed in class.

a. Write another contrast matrix corresponding to $H_0 : \mu_1 = \ldots = \mu_4$ different from the ones presented in class, that is, **a different contrast than the two matrices below** (do not just multiply a constant and call it different contrast):

$$\boldsymbol{C}_1 = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix} \qquad \boldsymbol{C}_2 = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Another contrast matrix is

$$\boldsymbol{C}_4 = \begin{bmatrix} -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

b. For your contrast matrix in part (a), test H_0 and compare the results to those in lecture notes.

We use the same function defined in class:

```
T2.contrast <- function(data.matrix, contrast.matrix, alpha = 0.05){
  # Input args
  #   data.matrix: n x p matrix, each row is one subject, each col is one treatment
  #   contrast.matrix: q x p matrix C, each row is one contrast
  #   alpha: significance level, default 0.05

  dat <- data.matrix
  C <- contrast.matrix

  # sample mean vector
  xbar <- colMeans(dat)
  # sample covariance matrix
  S <- cov(dat)

  # parameters
  n <- nrow(dat)
  q <- nrow(C)

  # Intermediate quantities
  invCSC <- solve(C%*%S %*% (t(C)))
  Cxbar <- C %*% xbar
  # test statistic
  T2 <-  n*(n-q)/((n-1)*q) * (t(Cxbar)) %*% invCSC %*% (Cxbar)
  # critical value
  critical_value_F = qf(p = 0.05, df1 = q, df2 = n-q, lower.tail = F)
  # p-value
  pv <- pf(T2, df1 = q, df2 = n-q, lower.tail = F)

  # display the results
  results <- data.frame(T2 = T2, critical = critical_value_F,
                        df1 = q, df2 = n-q, pvalue = pv)

  return(results)
}
```

Now load the data and test using the contrast $\boldsymbol{C}_4$.

```
dat <- as.matrix( read.table("../data/T6-2.dat") )
colnames(dat)  = c("trt 1", "trt 2","trt 3", "trt 4")
# contrast
C <- cbind(-diag(1, 3), c(1,1,1))
T2.contrast(dat, C)
```

```
##         T2 critical df1 df2       pvalue
## 1 34.37521 3.238872   3  16 3.317767e-07
```

We obtain identical results as for the other contrast matrix shown in class.

c. Separately, test for interaction effect, the main effect of halothane, and the main efffect of CO2, and interprete the results.

Test for interaction:

```
# contrast for no interaction
C <- matrix(c(1, -1, -1, 1), nrow = 1)

# test
T2.contrast(dat, C)
```

```
##          T2 critical df1 df2    pvalue
## 1 0.4112318 4.413873   1  18 0.5294265
```

Test for Halothane main effect:

```
# contrast for no Halothane main effect
C <- matrix(c(1, 1, -1, -1), nrow = 1)

# test
T2.contrast(dat, C)
```

```
##         T2 critical df1 df2       pvalue
## 1 88.25581 4.413873   1  18 2.314867e-08
```

Test for CO2 main effect:

```
# contrast for no CO2 main effect
C <- matrix(c(1, -1, 1, -1), nrow = 1)

# test
T2.contrast(dat, C)
```

```
##         T2 critical df1 df2      pvalue
## 1 13.18751 4.413873   1  18 0.001908795
```

It is evedent that there is no interaction, but both the main effects are significant at 5% level.

d. Suppose we want to test whether `CO2 effect (High - Low) when Halothane is present` is twice `CO2 effect (High - Low) when Halothane is absent`. Write the null hypothesis and the corresponding contrast matrix. Test this hypothesis and interprete the results.

The null hypothesis is $H_0 : \mu_1 - \mu_2 = 2(\mu_3 - \mu_4)$.

```
# contrast
C <- matrix(c(1, -1, -2, 2), nrow = 1)

# test
T2.contrast(dat, C)
```

```
##          T2 critical df1 df2    pvalue
## 1 0.1319143 4.413873   1  18 0.7206858
```

Based on the large p-value, we can not reject H_0, and conclude that indeed `CO2 effect (High - Low) when Halothane is present` is twice `CO2 effect (High - Low) when Halothane is absent`.

Problem 3: (20 points) The dataset [here] (../data/T6-9.dat) gives measurements on the carapaces on 24 male and 24 female turtles.

```
dat <- read.table("../data/T6-9.dat", header = F)
colnames(dat) <- c("Length", "Width", "Height", "Gender")
head(dat)
```

```
##   Length Width Height Gender
## 1     98    81     38 female
## 2    103    84     38 female
## 3    103    86     42 female
## 4    105    86     42 female
## 5    109    88     44 female
## 6    123    92     50 female
```

```
tail(dat)
```

```
##    Length Width Height Gender
## 43    121    95     42   male
## 44    125    93     45   male
## 45    127    96     45   male
## 46    128    95     45   male
## 47    131    95     46   male
## 48    135   106     47   male
```

a. Test for equality of mean measurements between the two genders.

```
library(car)
```

```
## Loading required package: carData
```

```
gender <- dat[, 4]
Y <- as.matrix(dat[, 1:3])
lmres <- lm(Y ~ gender)
out <- manova( lmres )
summary( out, test = "Wilks" )
```

```
##           Df   Wilks approx F num Df den Df    Pr(>F)
## gender     1 0.38857   23.078      3     44 3.967e-09 ***
## Residuals 46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The small p-value of the Wilks lambda statistics suggests that the mean of the two groups are different.

b. Create Bonferroni intervals for each component of the difference of the mean vector.

```
library(emmeans)
```

```
## Warning: package 'emmeans' was built under R version 3.5.2
```

```
# number of variables
p <- ncol(Y)

# Create a list to store the results
pair.lst <- vector("list", p)

# name the list according to variabls (for convenience)
names(pair.lst) <- colnames(Y)

# run emmeans for each variable to estimate the group means etc
for(j in 1:p){
  wts <- rep(0, p)
  wts[j] <- 1
  pair.lst[[j]] <- emmeans(out, "gender", weights=wts)
}

# number of groups
g <- 2 # (male  vs  female)

# old significance level
alpha <- 0.05

# number of comparison
nc <- p * g * (g-1) / 2

# new significance level
alphanew <- 0.05 / nc
```

Contrast for `Length`:

```
# obtain the contrasts first
cont <- contrast(pair.lst$Length, "pairwise")

# pair-wise differences for `Length`
bb <- confint(cont, level=1-alphanew, adj="none")
bb
```

```
##  contrast      estimate   SE df lower.CL upper.CL
##  female - male     22.7 4.96 46     10.3       35
##
## Results are averaged over the levels of: rep.meas
## Confidence level used: 0.983333333333333
```

Contrast for `Width`:

```
# obtain the contrasts first
cont <- contrast(pair.lst$Width, "pairwise")

# pair-wise differences for `Width`
bb <- confint(cont, level=1-alphanew, adj="none")
bb
```

```
##  contrast      estimate   SE df lower.CL upper.CL
##  female - male     14.3 3.04 46     6.74     21.8
##
## Results are averaged over the levels of: rep.meas
## Confidence level used: 0.983333333333333
```

Contrast for `Height`:

```
# obtain the contrasts first
cont <- contrast(pair.lst$Height, "pairwise")

# pair-wise differences for `Height`
bb <- confint(cont, level=1-alphanew, adj="none")
bb
```

```
##  contrast      estimate   SE df lower.CL upper.CL
##  female - male     11.3 1.78 46     6.91     15.8
##
## Results are averaged over the levels of: rep.meas
## Confidence level used: 0.983333333333333
```

Problem 4: (20 points) Consider the `Pottery` data in the `car` library; see the help page for `Pottery` for details.

```
library(car)
head(Pottery)
```

```
##        Site   Al   Fe   Mg   Ca   Na
## 1 Llanedyrn 14.4 7.00 4.30 0.15 0.51
## 2 Llanedyrn 13.8 7.08 3.43 0.12 0.17
## 3 Llanedyrn 14.6 7.09 3.88 0.13 0.20
## 4 Llanedyrn 11.5 6.37 5.64 0.16 0.14
## 5 Llanedyrn 13.8 7.06 5.34 0.20 0.20
## 6 Llanedyrn 10.9 6.26 3.47 0.17 0.22
```

```
dat <- as.matrix(Pottery[,-1])
site <- Pottery[,1]
```

The first column defines the groups.

a. Estimate mean vector of each site (group), and the overall mean vector.

```
aggregate(dat, by = list(site), mean)
```

```
##       Group.1       Al       Fe       Mg        Ca        Na
## 1 AshleyRails 17.32000 1.512000 0.606000 0.0520000 0.0480000
## 2    Caldicot 11.70000 5.415000 3.855000 0.2950000 0.0500000
## 3  IsleThorns 18.18000 1.712000 0.674000 0.0260000 0.0540000
## 4   Llanedyrn 12.56429 6.372143 4.826429 0.2021429 0.2507143
```

b. Perform a MANOVA to determine wheather the group means are equal or not. Give the sum of squares and cross product matrices (B and E) as defined in class.

```
# manova
out <- manova(dat ~ site)
summary(out, test = "Wilks")
```

```
##           Df    Wilks approx F num Df den Df   Pr(>F)
## site       3 0.012301   13.088     15 50.091 1.84e-12 ***
## Residuals 22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject H_0 due to the p-value being smaller than 0.05.

c. If you reject the hypothesis of equality of means in part (a) [hint: your results in part (a) should reject H_0], investigate which components are different using pair-wise comparisons.

```
library(emmeans)

# number of variables
p <- ncol(dat)

# Create a list to store the results
pair.lst <- vector("list", p)

# name the list according to variabls (for convenience)
names(pair.lst) <- colnames(dat)

# run emmeans for each variable to estimate the group means etc
for(j in 1:p){
  wts <- rep(0, p)
  wts[j] <- 1
  pair.lst[[j]] <- emmeans(out, "site", weights=wts)
}

# number of groups
g <- 4 # four sites

# old significance level
alpha <- 0.05

# number of comparison
nc <- p * g * (g-1) / 2

# new significance level
alphanew <- 0.05 / nc

# pair-wise contrats
for(ii in 1:p){
  # obtain the contrasts first
  cont <- contrast(pair.lst[[ii]], "pairwise")
  # pair-wise differences for `mb`
  bb <- confint(cont, level=1-alphanew, adj="none")
  print(paste("Variable: ", colnames(dat)[ii]))
  print(bb)
  writeLines("\n\n")
}
```

```
## [1] "Variable:  Al"
##  contrast                estimate    SE df lower.CL upper.CL
##  AshleyRails - Caldicot      5.620 1.240 22     1.18    10.06
##  AshleyRails - IsleThorns   -0.860 0.937 22    -4.22     2.50
##  AshleyRails - Llanedyrn     4.756 0.772 22     1.99     7.52
##  Caldicot - IsleThorns      -6.480 1.240 22   -10.92    -2.04
##  Caldicot - Llanedyrn       -0.864 1.120 22    -4.87     3.15
##  IsleThorns - Llanedyrn      5.616 0.772 22     2.85     8.38
##
## Results are averaged over the levels of: rep.meas
## Confidence level used: 0.998333333333333
##
##
##
## [1] "Variable:  Fe"
##  contrast                estimate    SE df lower.CL upper.CL
##  AshleyRails - Caldicot     -3.903 0.590 22    -6.02   -1.789
##  AshleyRails - IsleThorns   -0.200 0.446 22    -1.80    1.398
##  AshleyRails - Llanedyrn    -4.860 0.368 22    -6.18   -3.544
##  Caldicot - IsleThorns       3.703 0.590 22     1.59    5.817
##  Caldicot - Llanedyrn       -0.957 0.533 22    -2.87    0.953
##  IsleThorns - Llanedyrn     -4.660 0.368 22    -5.98   -3.344
##
## Results are averaged over the levels of: rep.meas
## Confidence level used: 0.998333333333333
##
##
##
## [1] "Variable:  Mg"
##  contrast                estimate    SE df lower.CL upper.CL
##  AshleyRails - Caldicot     -3.249 0.701 22   -5.758    -0.74
##  AshleyRails - IsleThorns   -0.068 0.530 22   -1.965     1.83
##  AshleyRails - Llanedyrn    -4.220 0.436 22   -5.783    -2.66
##  Caldicot - IsleThorns       3.181 0.701 22    0.672     5.69
##  Caldicot - Llanedyrn       -0.971 0.633 22   -3.238     1.30
##  IsleThorns - Llanedyrn     -4.152 0.436 22   -5.715    -2.59
##
## Results are averaged over the levels of: rep.meas
## Confidence level used: 0.998333333333333
##
##
##
## [1] "Variable:  Ca"
##  contrast                estimate     SE df lower.CL upper.CL
##  AshleyRails - Caldicot   -0.2430 0.0405 22  -0.3879  -0.0981
##  AshleyRails - IsleThorns  0.0260 0.0306 22  -0.0836   0.1356
##  AshleyRails - Llanedyrn  -0.1501 0.0252 22  -0.2404  -0.0599
##  Caldicot - IsleThorns     0.2690 0.0405 22   0.1241   0.4139
##  Caldicot - Llanedyrn      0.0929 0.0366 22  -0.0381   0.2238
##  IsleThorns - Llanedyrn   -0.1761 0.0252 22  -0.2664  -0.0859
##
## Results are averaged over the levels of: rep.meas
## Confidence level used: 0.998333333333333
```

```
##
##
##
## [1] "Variable:  Na"
##  contrast                estimate     SE df lower.CL upper.CL
##  AshleyRails - Caldicot    -0.002 0.0796 22   -0.287   0.2831
##  AshleyRails - IsleThorns  -0.006 0.0602 22   -0.222   0.2096
##  AshleyRails - Llanedyrn   -0.203 0.0496 22   -0.380  -0.0252
##  Caldicot - IsleThorns     -0.004 0.0796 22   -0.289   0.2811
##  Caldicot - Llanedyrn      -0.201 0.0719 22   -0.458   0.0569
##  IsleThorns - Llanedyrn    -0.197 0.0496 22   -0.374  -0.0192
##
## Results are averaged over the levels of: rep.meas
## Confidence level used: 0.998333333333333
```

d. What assumptions on the population/sample are you making in this situation?

We are assuming

- Each group/population is normal but with equal covariance
- The populations are independent

Problem 5: (20 points) The dataset [here] (../data/T6-17.dat) gives measurements on `Yield` (X_1), `Sound mature kernels` (X_2) and `Seed size` (X_3) on peanuts from different `Location` and `Variety` (two factors).

```
dat <- read.table("../data/T6-17.dat", header = F)
colnames(dat) <- c("Location", "Variety", "Yield", "SdMatKer", "Size")
dat
```

```
##    Location Variety Yield SdMatKer Size
## 1         1       5 195.3    153.1 51.4
## 2         1       5 194.3    167.7 53.7
## 3         2       5 189.7    139.5 55.5
## 4         2       5 180.4    121.1 44.4
## 5         1       6 203.0    156.8 49.8
## 6         1       6 195.9    166.0 45.8
## 7         2       6 202.7    166.1 60.4
## 8         2       6 197.6    161.8 54.1
## 9         1       8 193.5    164.5 57.8
## 10        1       8 187.0    165.1 58.6
## 11        2       8 201.5    166.8 65.0
## 12        2       8 200.0    173.8 67.2
```

a. Perform a MANOVA on this dataset. Test for a location-variety interaction effect, location effect and variety effect.

```
lmres <- lm(cbind(Yield, SdMatKer, Size) ~ Location*Variety, data = dat)
summary( manova(lmres) )
```

```
##                  Df  Pillai approx F num Df den Df  Pr(>F)
## Location          1 0.58764   2.8502      3      6 0.12727
## Variety           1 0.66085   3.8971      3      6 0.07362 .
## Location:Variety  1 0.45620   1.6778      3      6 0.26968
## Residuals         8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b. Using results in part (a), can we conclude that the effects of `Location` and `Variety` are additive?

Indeed, the interaction effect is not significant, and thus the effect of `Location` and `Variety` are additive (no interaction).

c. Investigate whether location-variety interaction show up for some variables but not others by running three univariate ANOVA models. [Hint: don't worry about multiple comparison here.]

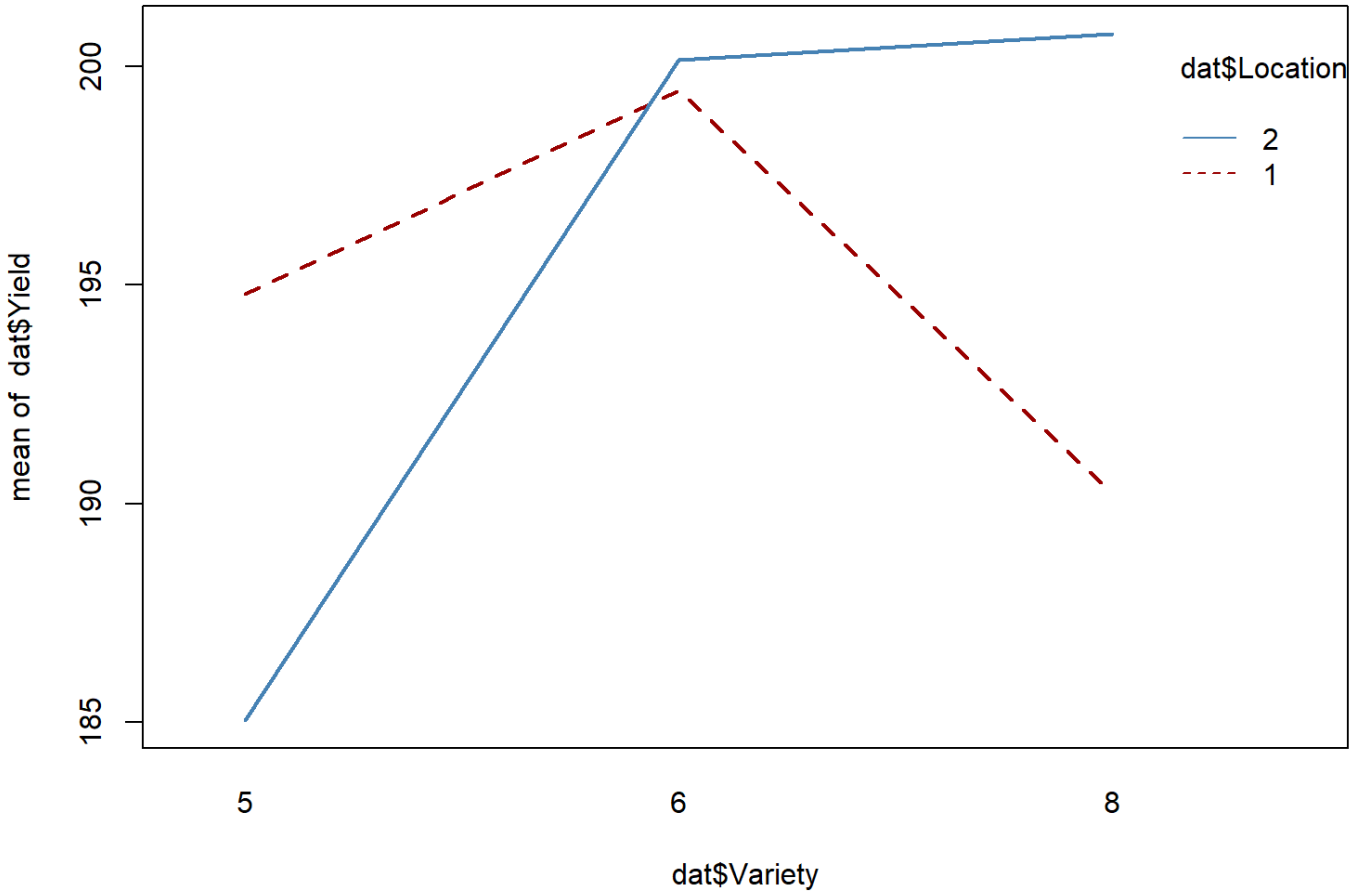
```
summary.aov( manova(lmres) )
```

```
##  Response Yield :
##                  Df  Sum Sq Mean Sq F value  Pr(>F)
## Location          1   0.701   0.701  0.0201 0.89063
## Variety           1  30.857  30.857  0.8871 0.37382
## Location:Variety  1 196.301 196.301  5.6436 0.04485 *
## Residuals         8 278.264  34.783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response SdMatKer :
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## Location          1 162.07  162.07  1.4682 0.26020
## Variety           1 835.71  835.71  7.5709 0.02500 *
## Location:Variety  1 503.01  503.01  4.5569 0.06531 .
## Residuals         8 883.09  110.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Size :
##                  Df  Sum Sq Mean Sq F value   Pr(>F)
## Location          1  72.521  72.521  3.7159 0.090044 .
## Variety           1 269.800 269.800 13.8243 0.005887 **
## Location:Variety  1  38.957  38.957  1.9961 0.195404
## Residuals         8 156.131  19.516
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems the interaction effect shows up for the `Yeild` variable, and gives a borderline p-value for `SdMatKer` variable.

d. Explain in words (in the context of the problem) what the location-variety interaction effect means.

It means that the difference between the average yield at the two locations will change depending on the variety of peanuts.



We can see that for peanut variety `5`, location 1 gives more yield than location 2; however for peanut variety `8`, location 2 gives more yield than location 1.