Multivariate Analysis Midterm

DA 410

Marjorie Blanco

Note:

Round to the THIRD decimal place, unless otherwise noted in the instruction.

Problem 4

In the following table, we have a comparison of four reagents. The first reagent is the one presently in use and the other three are less expensive reagents that we wish to compare with the first. All four reagents are used with a blood sample from each patient.

The three variables measured for each reagent are y_1 = white blood count, y_2 = red blood count, and y_3 = hemoglobin count.

The data for twenty subject from each of four reagents:

Reagent	Subject	white.blood.count	red.blood.count	hemoglobin.count
1	1	8.0	3.96	12.5
1	2	4.0	5.37	16.9
1	3	6.3	5.47	17.1
1	4	9.4	5.16	16.2
1	5	8.2	5.16	17.0
1	6	11.0	4.67	14.3
1	7	6.8	5.20	16.2
1	8	9.0	4.65	14.7
1	9	6.1	5.22	16.3
1	10	6.4	5.13	15.9
1	11	5.6	4.47	13.3
1	12	8.2	5.22	16.0
1	13	5.7	5.10	14.9
1	14	9.8	5.25	16.1
1	15	5.9	5.28	15.8
1	16	6.6	4.65	12.8
1	17	5.7	4.42	14.5
1	18	6.7	4.38	13.1
1	19	6.8	4.67	15.6
1	20	9.6	5.64	17.0
2	1	8.0	3.93	12.7
2	2	4.2	5.35	17.2
2	3	6.3	5.39	17.5
2	4	9.4	5.16	16.7
2	5	8.0	5.13	17.5
2	6	10.7	4.60	14.7
2	7	6.8	5.16	16.7
2	8	9.0	4.57	15.0
2	9	6.0	5.16	16.9
2	10	6.4	5.11	16.4
2	11	5.5	4.45	13.6
2	12	8.2	5.14	16.5
2	13	5.6	5.05	15.3
2	14	9.8	5.15	16.6
2	15	5.8	5.25	16.4
2	16	6.4	4.59	13.2
2	17	5.5	4.31	14.9
2	18	6.5	4.32	13.4
2	19	6.6	4.57	15.8
2	20	9.5	5.58	17.5
3	1	7.9	3.86	13.0
3	2	4.1	5.39	17.2
3	3	6.0	5.39	17.2
3	4	9.4	5.17	16.7
3	5	8.1	5.10	17.4
3	6	10.6	4.52	14.6
3	7	6.9	5.13	16.8
3	8	8.9	4.58	15.0
3	9	6.1	5.14	16.9
3	10	6.4	5.11	16.4
3	11	5.3	4.46	13.6
3	12	8.0	5.14	16.5
3	13	5.5	5.02	15.4
3	14	8.1	5.10	13.8
3	15	5.7	5.26	16.4
3	16	6.3	4.58	13.1
3	17	5.5	4.30	14.9
3	18	6.5	4.32	13.6

Compare the four reagents using all four MANOVA tests. State each hypotheses clearly, and interpret the results.

```
## $`1`
##     Reagent  white.blood.count red.blood.count hemoglobin.count
##  Min.   :1   Min.   : 4.00     Min.   :3.960   Min.   :12.50   
##  1st Qu.:1   1st Qu.: 6.05     1st Qu.:4.650   1st Qu.:14.45   
##  Median :1   Median : 6.75     Median :5.145   Median :15.85   
##  Mean   :1   Mean   : 7.29     Mean   :4.954   Mean   :15.31   
##  3rd Qu.:1   3rd Qu.: 8.40     3rd Qu.:5.228   3rd Qu.:16.23   
##  Max.   :1   Max.   :11.00     Max.   :5.640   Max.   :17.10   
## 
## $`2`
##     Reagent  white.blood.count red.blood.count hemoglobin.count
##  Min.   :2   Min.   : 4.20     Min.   :3.930   Min.   :12.70   
##  1st Qu.:2   1st Qu.: 5.95     1st Qu.:4.570   1st Qu.:14.85   
##  Median :2   Median : 6.55     Median :5.120   Median :16.40   
##  Mean   :2   Mean   : 7.21     Mean   :4.899   Mean   :15.72   
##  3rd Qu.:2   3rd Qu.: 8.40     3rd Qu.:5.160   3rd Qu.:16.75   
##  Max.   :2   Max.   :10.70     Max.   :5.580   Max.   :17.50   
## 
## $`3`
##     Reagent  white.blood.count red.blood.count hemoglobin.count
##  Min.   :3   Min.   : 4.100    Min.   :3.860   Min.   :13.00   
##  1st Qu.:3   1st Qu.: 5.925    1st Qu.:4.543   1st Qu.:14.40   
##  Median :3   Median : 6.500    Median :5.100   Median :16.20   
##  Mean   :3   Mean   : 7.055    Mean   :4.881   Mean   :15.60   
##  3rd Qu.:3   3rd Qu.: 8.100    3rd Qu.:5.147   3rd Qu.:16.82   
##  Max.   :3   Max.   :10.600    Max.   :5.500   Max.   :17.40   
## 
## $`4`
##     Reagent  white.blood.count red.blood.count hemoglobin.count
##  Min.   :4   Min.   : 4.000    Min.   :3.870   Min.   :13.20   
##  1st Qu.:4   1st Qu.: 5.900    1st Qu.:4.558   1st Qu.:14.78   
##  Median :4   Median : 6.500    Median :5.095   Median :16.25   
##  Mean   :4   Mean   : 7.025    Mean   :4.891   Mean   :15.77   
##  3rd Qu.:4   3rd Qu.: 8.075    3rd Qu.:5.195   3rd Qu.:16.82   
##  Max.   :4   Max.   :10.500    Max.   :5.460   Max.   :17.50   
```

MANOVA analysis assumes both normality and homoscedasticity (equality of variance) of the experimental errors (residuals).

- Descriptive statistics by dependent variable

```
## reagents$Reagent: 1
##       median         mean      SE.mean CI.mean.0.95          var 
##    6.7500000    7.2900000    0.3976841    0.8323624    3.1630526 
##      std.dev     coef.var 
##    1.7784973    0.2439640 
## -------------------------------------------------------- 
## reagents$Reagent: 2
##       median         mean      SE.mean CI.mean.0.95          var 
##    6.5500000    7.2100000    0.3923546    0.8212075    3.0788421 
```

```
##      std.dev     coef.var
##    1.7546630    0.2433652
## --------------------------------------------------------
## reagents$Reagent: 3
##       median         mean      SE.mean CI.mean.0.95          var
##    6.5000000    7.0550000    0.3715172    0.7775944    2.7605000
##      std.dev     coef.var
##    1.6614752    0.2355032
## --------------------------------------------------------
## reagents$Reagent: 4
##       median         mean      SE.mean CI.mean.0.95          var
##    6.5000000    7.0250000    0.3773784    0.7898621    2.8482895
##      std.dev     coef.var
##    1.6876876    0.2402402

## reagents$Reagent: 1
##       median         mean      SE.mean CI.mean.0.95          var
##   5.14500000   4.95350000   0.09768444   0.20445588   0.19084500
##      std.dev     coef.var
##   0.43685810   0.08819180
## --------------------------------------------------------
## reagents$Reagent: 2
##       median         mean      SE.mean CI.mean.0.95          var
##    5.1200000    4.8985000    0.0986975    0.2065763    0.1948239
##      std.dev     coef.var
##    0.4413887    0.0901069
## --------------------------------------------------------
## reagents$Reagent: 3
##       median         mean      SE.mean CI.mean.0.95          var
##   5.10000000   4.88100000   0.09980745   0.20889939   0.19923053
##      std.dev     coef.var
##   0.44635247   0.09144693
## --------------------------------------------------------
## reagents$Reagent: 4
##       median         mean      SE.mean CI.mean.0.95          var
##   5.09500000   4.89150000   0.10021103   0.20974409   0.20084500
##      std.dev     coef.var
##   0.44815734   0.09161961

## reagents$Reagent: 1
##       median         mean      SE.mean CI.mean.0.95          var
##  15.85000000  15.31000000   0.32894568   0.68849123   2.16410526
##      std.dev     coef.var
##   1.47108982   0.09608686
## --------------------------------------------------------
## reagents$Reagent: 2
##       median         mean      SE.mean CI.mean.0.95          var
##  16.40000000  15.72500000   0.34441980   0.72087893   2.37250000
##      std.dev     coef.var
##   1.54029218   0.09795181
## --------------------------------------------------------
## reagents$Reagent: 3
##       median         mean      SE.mean CI.mean.0.95          var
##  16.20000000  15.59500000   0.34155026   0.71487291   2.33313158
```

```
##      std.dev     coef.var
##   1.52745919   0.09794544
## ------------------------------------------------------------
## reagents$Reagent: 4
##       median         mean      SE.mean CI.mean.0.95          var
##  16.25000000  15.76500000   0.32905567   0.68872142   2.16555263
##      std.dev     coef.var
##   1.47158168   0.09334486
```

- Multivariate normality

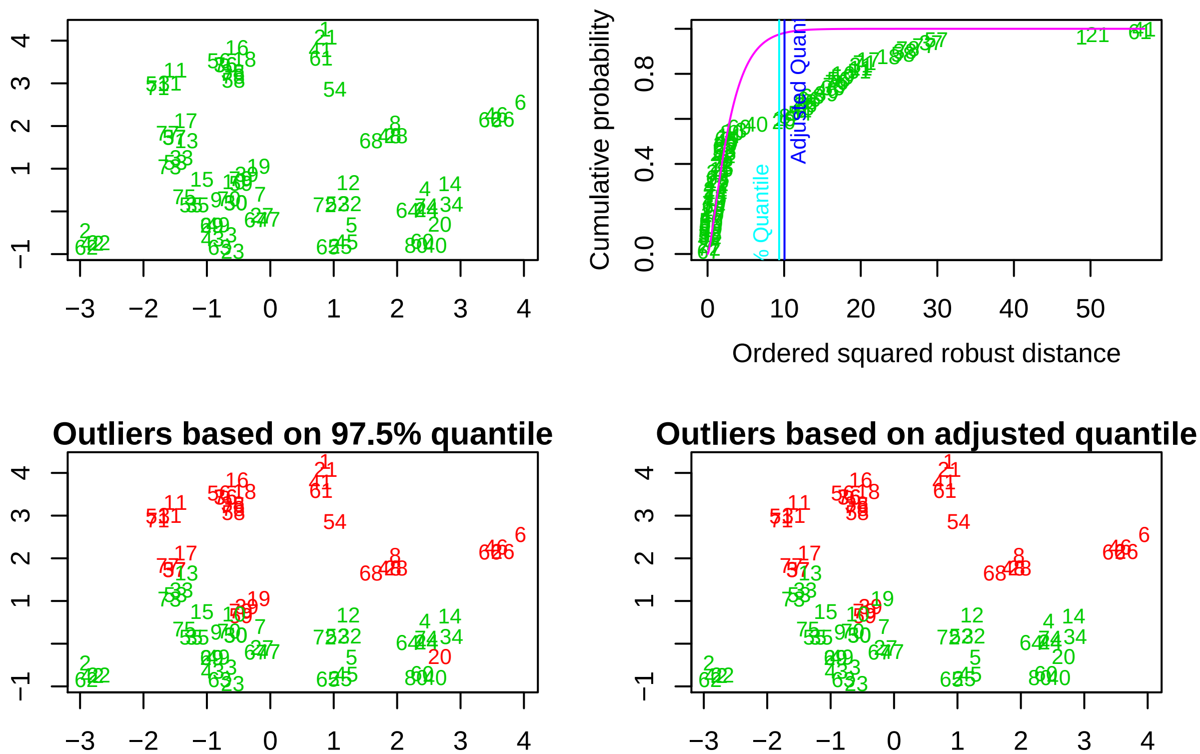
```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.62978, p-value = 0.001241
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.62978, p-value = 0.001241
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.62978, p-value = 0.001241
```

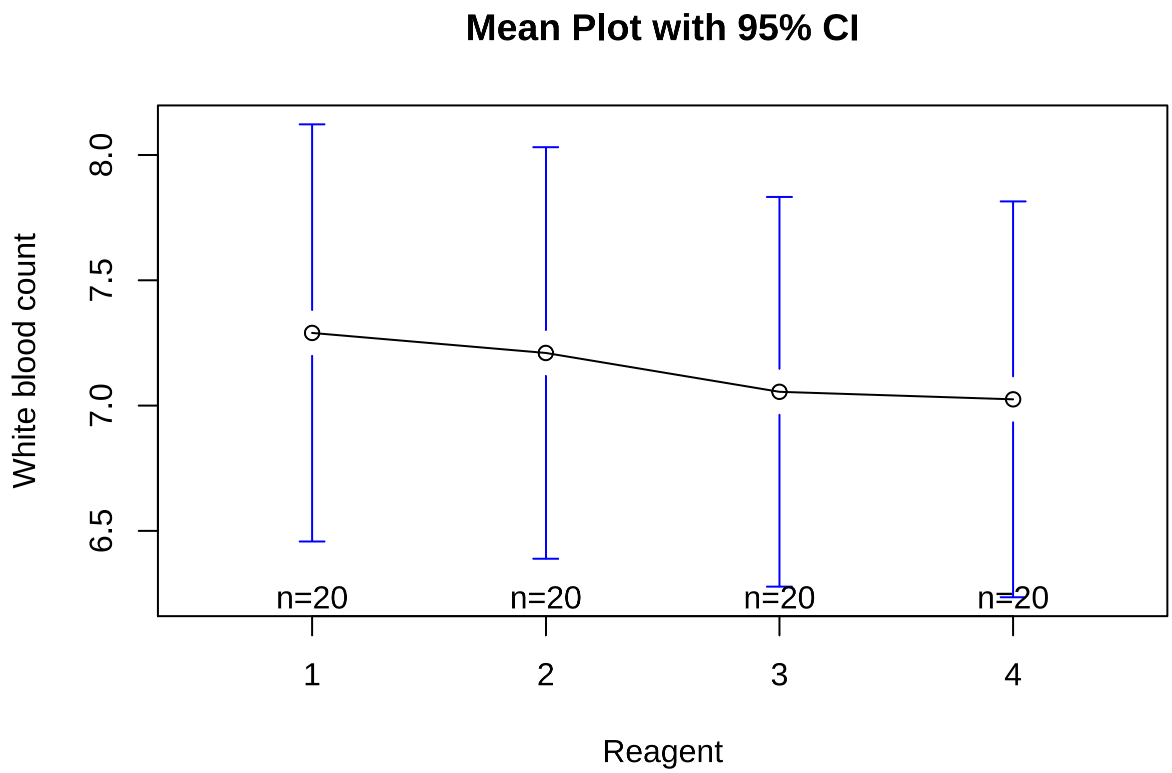
```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.62978, p-value = 0.001241
```

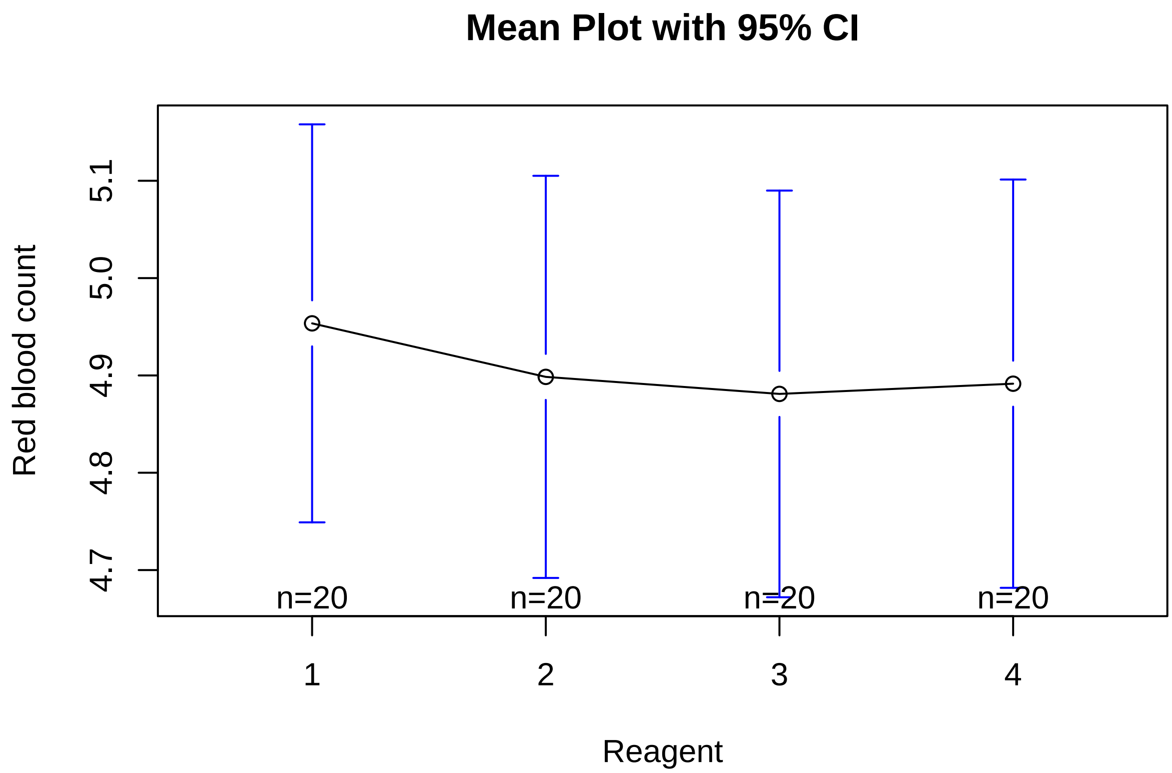
```
## Projection to the first and second robust principal components.
## Proportion of total variation (explained variance): 0.9590544
```

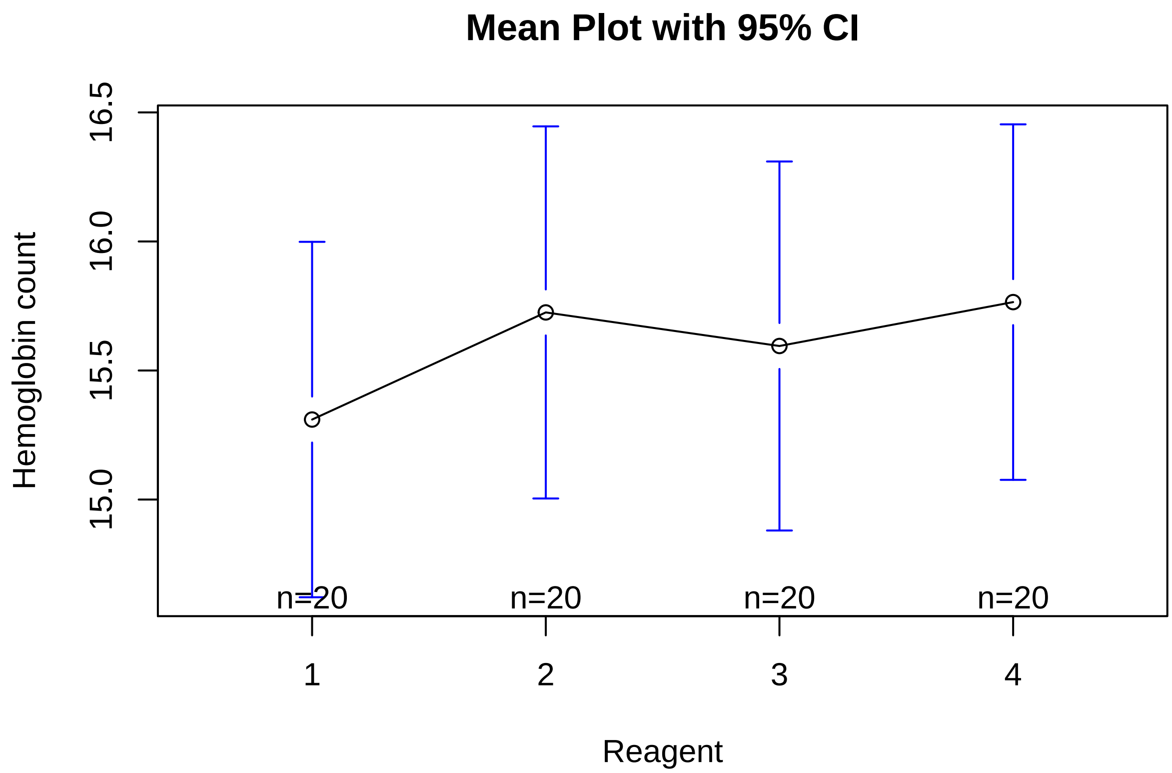



```
## $outliers
##  [1]  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE
## [12] FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE
## [23] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
## [34] FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE
## [45] FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE
## [56]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## [67] FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE
## [78]  TRUE  TRUE FALSE
```

- Mean plot by dependent variable


Mean Plot with 95% CI
White blood count
8.0
7.5
7.0
6.5
n=20
n=20
n=20
n=20
1
2
3
4
Reagent


Mean Plot with 95% CI
Red blood count
5.1
5.0
4.9
4.8
4.7
n=20
n=20
n=20
n=20
1
2
3
4
Reagent




```
n <- dim(reagents)[1] / length(unique(reagents$Reagent))
total.means <- colMeans(reagents[,3:5])
```

The overall mean vector:

white.blood.count	red.blood.count	hemoglobin.count
7.145	4.906	15.599

	1	2	3	4
white.blood.count	7.290	7.210	7.055	7.025
red.blood.count	4.954	4.899	4.881	4.891
hemoglobin.count	15.310	15.725	15.595	15.765

```
reagent1 <- reagents %>% filter(Reagent == 1)  %>% dplyr::select(-c(Reagent, Subject))
reagent2 <- reagents %>% filter(Reagent == 2)  %>% dplyr::select(-c(Reagent, Subject))
reagent3 <- reagents %>% filter(Reagent == 3)  %>% dplyr::select(-c(Reagent, Subject))
reagent4 <- reagents %>% filter(Reagent == 4)  %>% dplyr::select(-c(Reagent, Subject))

reagent1.bar <- colMeans(reagent1)
reagent2.bar <- colMeans(reagent2)
reagent3.bar <- colMeans(reagent3)
reagent4.bar <- colMeans(reagent4)
```

```r
reagent.all.bar <- (reagent1.bar+reagent2.bar+reagent3.bar+reagent4.bar)/4

reagent1.bar.diff <- reagent1.bar - reagent.all.bar
reagent2.bar.diff <- reagent2.bar - reagent.all.bar
reagent3.bar.diff <- reagent3.bar - reagent.all.bar
reagent4.bar.diff <- reagent4.bar - reagent.all.bar

H <- n * unname(reagent1.bar.diff %*% t(reagent1.bar.diff) +
                  reagent2.bar.diff %*% t(reagent2.bar.diff) +
                  reagent3.bar.diff %*% t(reagent3.bar.diff) +
                  reagent4.bar.diff %*% t(reagent4.bar.diff))
```

H =

0.955	0.208	-1.065
0.208	0.063	-0.340
-1.065	-0.340	2.539

```r
"compute.within.matrix" <-function(data, mean) {
  ret <- matrix(as.numeric(0), nrow=3, ncol=3)
  for (i in 1:20) {
    diff <- as.numeric(unname(data[i,] - mean))
    ret <- ret + diff %*% t(diff)}
  return(ret)
  }
E <- compute.within.matrix(reagent1, reagent1.bar) + compute.within.matrix(reagent2, reagent2.bar) +
  compute.within.matrix(reagent3, reagent3.bar) + compute.within.matrix(reagent4, reagent4.bar)
```

E =

225.163	-0.911	6.020
-0.911	14.929	44.549
6.020	44.549	171.671

The number of groups: k = 4

The number of variables (dimension)" p = 3

The degrees of freedom for hypothesis: $_vH = 3$

The degrees of freedom for error: $_vE = 76$

```r
# MANOVA test
reagents.manova <- manova(cbind(reagents$white.blood.count, reagents$red.blood.count,
                                reagents$hemoglobin.count) ~ Reagent, data = reagents)
reagents.summary <- summary(reagents.manova)
```

We would then like to test if the properties (white blood, red blood and hemoglobin count) are the same across the four reagents.

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

H_1 : The $\mu's$ are unequal

Wilks's test

```r
reagents.summary <- summary(manova(cbind(reagents$white.blood.count, reagents$red.blood.count,
                     reagents$hemoglobin.count) ~ reagents$Reagent), test = "Wilks")
reagents.summary
```

```
##                   Df   Wilks approx F num Df den Df  Pr(>F)
## reagents$Reagent  1 0.91344   2.4007      3     76 0.07431 .
## Residuals        78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\lambda = 0.913$

The MANOVA model reports a Wilks test statistic of 0.913 and a p-value (0.074) $>$ 0.05, thus H_0 fails to be rejected and it is concluded there are no significant differences in the means.

Roy's test

```r
reagents.summary <- summary(manova(cbind(reagents$white.blood.count, reagents$red.blood.count,
                     reagents$hemoglobin.count) ~ reagents$Reagent), test = "Roy")
reagents.summary
```

```
##                   Df      Roy approx F num Df den Df  Pr(>F)
## reagents$Reagent  1 0.094764   2.4007      3     76 0.07431 .
## Residuals        78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\Theta = 0.087$

The MANOVA model reports a Roy test statistic of 0.095 and a p-value (0.074) $>$ 0.05, thus H_0 fails to be rejected and it is concluded there are no significant differences in the means.

Hotelling-Lawley's test

```r
reagents.summary <- summary(manova(cbind(reagents$white.blood.count, reagents$red.blood.count,
                     reagents$hemoglobin.count) ~ reagents$Reagent), test = "Hotelling-Lawley")
reagents.summary
```

```
##                   Df Hotelling-Lawley approx F num Df den Df  Pr(>F)
## reagents$Reagent  1         0.094764   2.4007      3     76 0.07431 .
## Residuals        78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$U^{(s)} = 0.095$

The MANOVA model reports a Hotelling-Lawley test statistic of 0.095 and a p-value (0.074) $>$ 0.05, thus H_0 fails to be rejected and it is concluded there are no significant differences in the means.

Table 1: Math, English, and Art tests for 5 students

Math	English	Art
90	60	90
90	90	30
60	60	60
60	60	90
30	30	30

Table 2: Mean vector y

	mean
Math	66
English	60
Art	60

Pillai's test

```
reagents.summary <- summary(manova(cbind(reagents$white.blood.count, reagents$red.blood.count,
                     reagents$hemoglobin.count) ~ reagents$Reagent), test = "Pillai")
reagents.summary
```

```
##                  Df   Pillai approx F num Df den Df  Pr(>F)
## reagents$Reagent  1 0.086561   2.4007      3     76 0.07431 .
## Residuals        78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$V^{(s)} = 0.087$

The MANOVA model reports a Pillai test statistic of 0.087 and a p-value $(0.074) > 0.05$, thus H_0 fails to be rejected and it is concluded there are no significant differences in the means.

Problem 5

The table below displays scores on math, English, and art tests for 5 students. Note that data from the table is represented in matrix A, where each column in the matrix shows scores on a test and each row shows scores for a student:

```
A <- matrix(c(90,   60, 90, 90, 90, 30, 60, 60, 60, 60, 60, 90, 30, 30, 30),
            nrow = 5, ncol = 3, byrow = TRUE )
colnames(A) <- c("Math", "English", "Art")
```

A =

```
kable(A, caption = "Math, English, and Art tests for 5 students") %>%
  kable_styling(bootstrap_options = "striped")
```

$\overline{y} =$

(a) Calculate the sample covariance matrix S.

Table 3: Sample covariance matrix

	Math	English	Art
Math	630	450	225
English	450	450	0
Art	225	0	900

Table 4: Sample correlation matrix

	Math	English	Art
Math	1.0000000	0.8451543	0.2988072
English	0.8451543	1.0000000	0.0000000
Art	0.2988072	0.0000000	1.0000000

S =

```
S <- cov(A)
kable(S, caption = "Sample covariance matrix") %>%
  kable_styling(bootstrap_options = "striped")
```

Thus, 630 is the variance of the Math variable, 450 is the covariance between the Math and the English variables, 225 is the covariance between the Math and the Art variables, 450 is the variance of the English variable, 0 is the covariance between the English and Art variables and 900 is the variance of the Art variable.

(b) Calculate the sample correlation matrix R.

R =

```
R <- cor(A)
kable(R, caption = "Sample correlation matrix") %>%
  kable_styling(bootstrap_options = "striped")
```

(c) Now let's define $Z = -2y_1 + 3y_2 + y_3$, where y_1 denotes Math scores, y_2 denotes English scores, and y_3 denotes Art scores. Find the sample mean vector $\overline{z}$ and the sample variance S_z^2.

z =

```
A2 <- sweep(A, 2, c(-2, 3, 1), "*")

z <- data.frame(mean = rowSums(A2))
rownames(z) <- paste(rep(c("z"), nrow(A)), rep(1:nrow(A)), sep="")
kable(z)
```

	mean
z1	90
z2	120
z3	120
z4	150
z5	60

```r
z_bar <- sum(z) * (1/nrow(A))
```

$\overline{z} = 108$

```r
a <- c(-2, 3, 1)
s2z<- t(a) %*% as.matrix(S) %*% a
```

$s_z^2 = 1170$

Problem 6:

Use the beetle data, do the following:

(a) Find the classification function and cutoff point.

(b) Find the classification table using the nearest neighbor method by setting k = 3.

(c) Calculate misclassification rate.

Problem 7

Use the above beetle data, do the following:

(a) Use LDA by setting probability of 50% and 50% to train model.

The mean vectors:

	1	2
transverse.groove.dist	194.4737	179.55
elytra.length	267.0526	290.80
second.antennal.joint.length	137.3684	157.20
third.antennal.joint.length	185.9474	209.25

```r
beetles$Measurement.Number <- NULL
beetles.lda <- lda(Species ~ ., prior = c(0.5,0.5), data = beetles)
beetles.lda
```

```
## Call:
## lda(Species ~ ., data = beetles, prior = c(0.5, 0.5))
##
## Prior probabilities of groups:
##   1   2
## 0.5 0.5
##
## Group means:
##   transverse.groove.dist elytra.length second.antennal.joint.length
## 1               194.4737      267.0526                     137.3684
## 2               179.5500      290.8000                     157.2000
##   third.antennal.joint.length
## 1                    185.9474
```

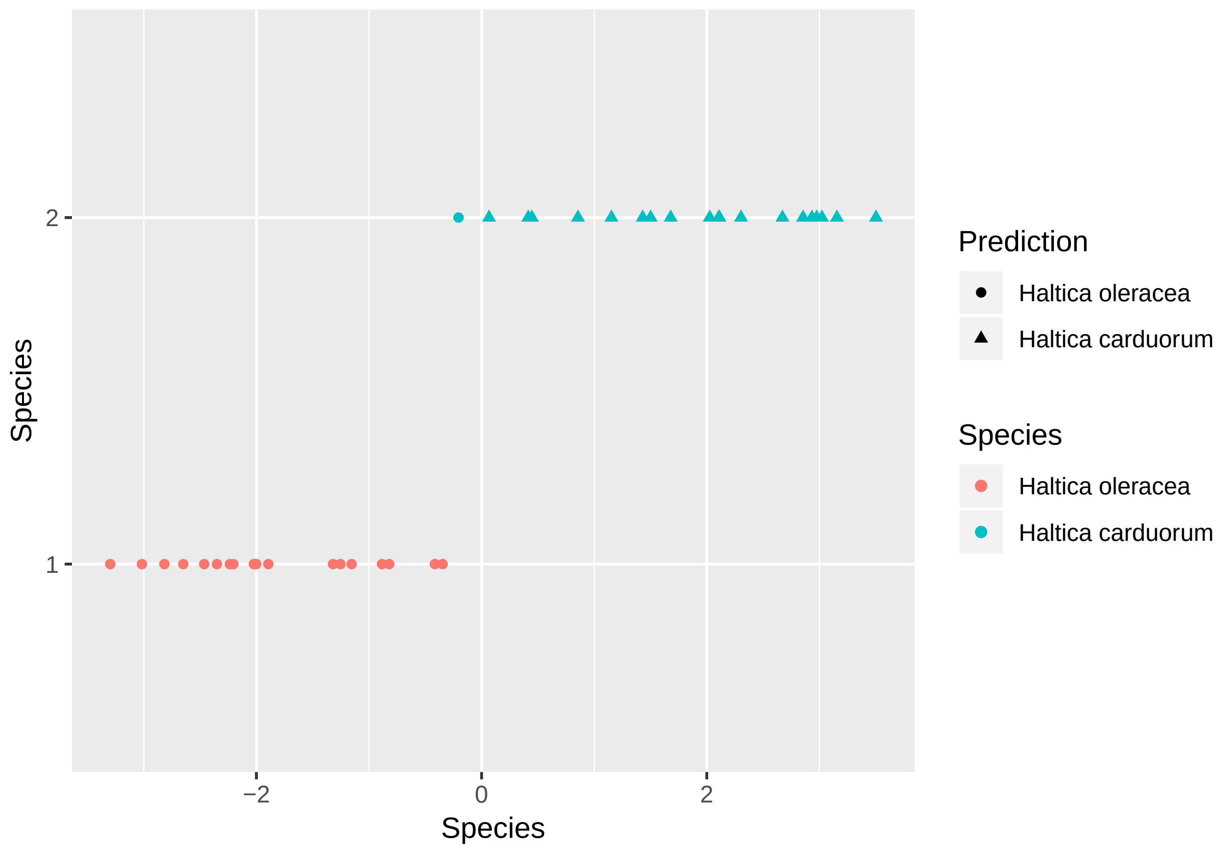
```
## 2                      209.2500
##
## Coefficients of linear discriminants:
##                                    LD1
## transverse.groove.dist       -0.09327642
## elytra.length                 0.03522706
## second.antennal.joint.length  0.02875538
## third.antennal.joint.length   0.03872998
```

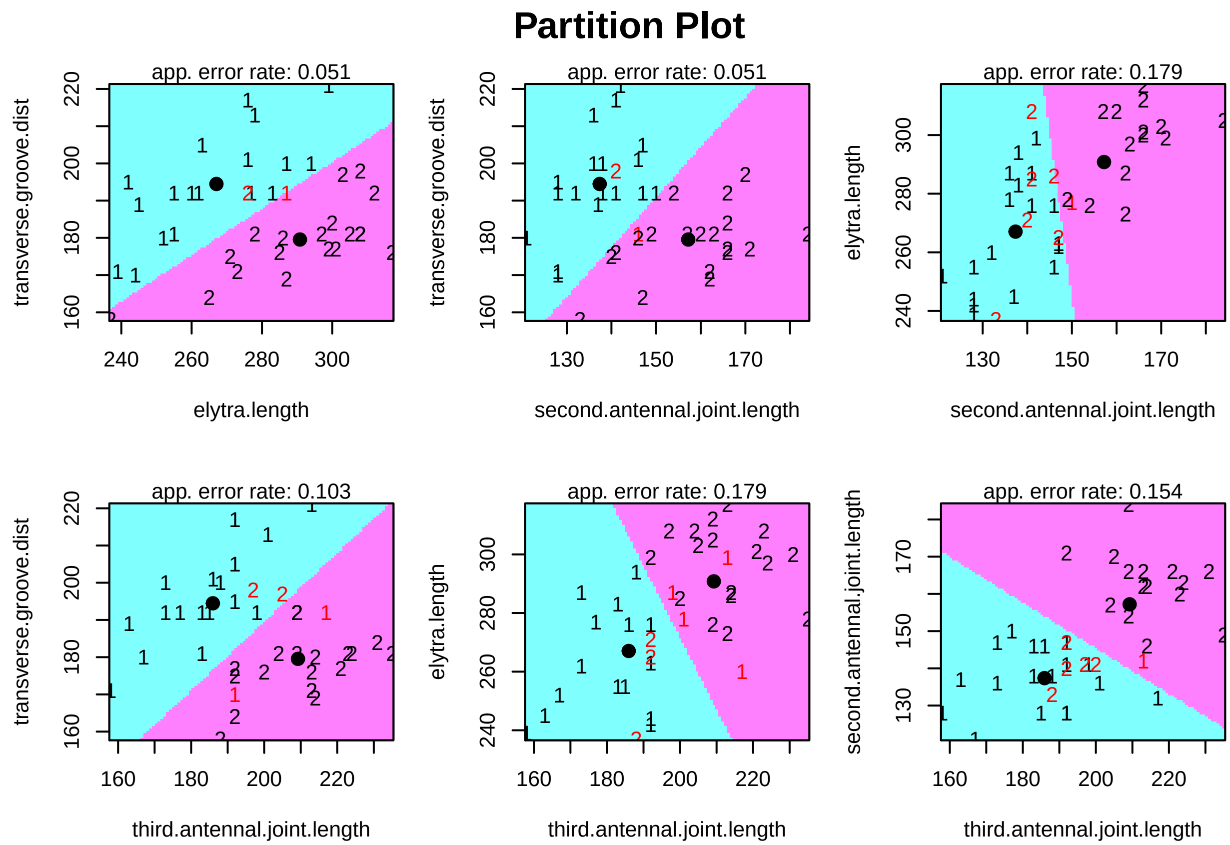
```r
lda.pred <- predict(beetles.lda)$class
```

The first discriminant function is a linear combination of the variables:

$$-0.09327642 * transverse.groove.dist + 0.03522706 * transverse.groove.dist + second.antennal.joint.length * 0.02875538 third.a$$

The LDA probability of Haltica oleracea is 50% while Haltica carduorum is 50%.





(b) Predict new observation (189,245,138,164).

```
new.data <-  data.frame(189,245,138,164)
colnames(new.data) <- c("transverse.groove.dist", "elytra.length", "second.antennal.joint.length", "thi
plda <- predict(beetles.lda, newdata = new.data)
```

The new observation LD1 is -2.9488199 and it is predicted to be assigned to Group 1 (Haltica oleracea).

(c) Calculate misclassification rate.

```
correct <- rep(0, times=nrow(beetles))
for (j in 1:nrow(beetles))
{
  mydis<- lda(grouping = beetles$Species[-j],
  x=beetles[-j, 2:5],
  prior = c(0.5, 0.5))
  mypred <- predict(mydis, newdata = beetles[j, 2:5])$class
  correct[j] <- (mypred == beetles$Species[j])
}
cv.missclass <- 1 - mean(correct)
```

The training model correctly classified 92.3% of observations.

The training model misclassification rate for LDA is 7.7%.

Problem 8

The following table contains data from O'Sullivan and Mahan with measurements of blood glucose levels on three occasions for 30 women. The y's represent fasting glucose measurements on the three occasions; the x's are glucose measurements 1 hour after sugar intake. Find the mean vector and covariance matrix for all six variables and partition them into $\binom{\bar{y}}{\bar{x}}$, and

$$\mathbf{S} = \begin{bmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{bmatrix}$$

```
blood.glucose <- read.table('data/data_problem8.txt')
blood.glucose <- cbind(blood.glucose[1:30,], blood.glucose[31:60,])
colnames(blood.glucose) <- c("y1", "y2", "y3", "x1", "x2", "x3")
```

```
y <- data.frame(mean = colMeans(blood.glucose))
```

$\binom{\bar{y}}{\bar{x}} =$

	mean
y1	72.20000
y2	72.73333
y3	73.30000
x1	108.46667
x2	102.46667
x3	108.46667

S =

	y1	y2	y3	x1	x2	x3
y1	77.614	0.986	23.731	100.076	4.869	34.317
y2	0.986	36.202	15.221	-46.457	30.370	-32.078
y3	23.731	15.221	57.459	13.407	-6.421	1.476
x1	100.076	-46.457	13.407	959.499	299.361	232.637
x2	4.869	30.370	-6.421	299.361	500.189	61.809
x3	34.317	-32.078	1.476	232.637	61.809	527.016

$S_{yy} =$

	y1	y2	y3
y1	77.614	0.986	23.731
y2	0.986	36.202	15.221
y3	23.731	15.221	57.459

$S_{yx} =$

	x1	x2	x3
y1	100.076	4.869	34.317
y2	-46.457	30.370	-32.078
y3	13.407	-6.421	1.476

$S_{xy} =$

	y1	y2	y3
x1	100.076	-46.457	13.407
x2	4.869	30.370	-6.421
x3	34.317	-32.078	1.476

$S_{xx} =$

	x1	x2	x3
x1	959.50	299.36	232.64
x2	299.36	500.19	61.81
x3	232.64	61.81	527.02

```
Syy <- matrix(c(cov(blood.glucose$y1, blood.glucose$y1),
                cov(blood.glucose$y1, blood.glucose$y2),
                cov(blood.glucose$y1, blood.glucose$y3),
                cov(blood.glucose$y2, blood.glucose$y1),
                cov(blood.glucose$y2, blood.glucose$y2),
                cov(blood.glucose$y2, blood.glucose$y3),
                cov(blood.glucose$y3, blood.glucose$y1),
                cov(blood.glucose$y3, blood.glucose$y2),
                cov(blood.glucose$y3, blood.glucose$y3)), nrow = 3, byrow = TRUE)

Sxx <- matrix(c(cov(blood.glucose$x1, blood.glucose$x1),
                cov(blood.glucose$x1, blood.glucose$x2),
                cov(blood.glucose$x1, blood.glucose$x3),
                cov(blood.glucose$x2, blood.glucose$x1),
                cov(blood.glucose$x2, blood.glucose$x2),
                cov(blood.glucose$x2, blood.glucose$x3),
                cov(blood.glucose$x3, blood.glucose$x1),
                cov(blood.glucose$x3, blood.glucose$x2),
                cov(blood.glucose$x3, blood.glucose$x3)), nrow = 3, byrow = TRUE)

Syx <- matrix(c(cov(blood.glucose$y1, blood.glucose$x1),
                cov(blood.glucose$y1, blood.glucose$x2),
                cov(blood.glucose$y1, blood.glucose$x3),
                cov(blood.glucose$y2, blood.glucose$x1),
                cov(blood.glucose$y2, blood.glucose$x2),
                cov(blood.glucose$y2, blood.glucose$x3),
                cov(blood.glucose$y3, blood.glucose$x1),
                cov(blood.glucose$y3, blood.glucose$x2),
                cov(blood.glucose$y3, blood.glucose$x3)), nrow = 3, byrow = TRUE)

Sxy <- t(Syx)

S <- cbind(rbind(Syy, Sxy), rbind(Syx, Sxx))
```

S =

```
##        [,1]   [,2]  [,3]   [,4]   [,5]   [,6]
## [1,]  77.61   0.99 23.73 100.08   4.87  34.32
## [2,]   0.99  36.20 15.22 -46.46  30.37 -32.08
## [3,]  23.73  15.22 57.46  13.41  -6.42   1.48
## [4,] 100.08 -46.46 13.41 959.50 299.36 232.64
## [5,]   4.87  30.37 -6.42 299.36 500.19  61.81
## [6,]  34.32 -32.08  1.48 232.64  61.81 527.02
```

Problem 9

Various aspects of economic cycles were measured for consumer goods and producer goods by Tintner.

The variables are:

y_1 = length of cycle

y_2 = percentage of rising prices

y_3 = cyclical amplitude

y_4 = rate of change

The data for several items are given in the following table:

```
goods <- read.table('Software-Files/T5_8_GOODS.DAT',
                    col.names = c('Item', 'Type', 'y1', 'y2', 'y3', 'y4'))
```

```
res <- t.test(y1 ~ Type, data = goods_df)
res
```

```
##
##  Welch Two Sample t-test
##
## data:  y1 by Type
## t = -3.818, df = 14.635, p-value = 0.001749
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -65.01293 -18.36485
## sample estimates:
## mean in group 1 mean in group 2
##        48.61111        90.30000
```

The p-value is 0.0017494. The consumer goods and producer goods differ in their length of cycle.

```
res <- t.test(y2 ~ Type, data = goods_df)
res
```

```
##
##  Welch Two Sample t-test
##
## data:  y2 by Type
## t = 0.57097, df = 12.427, p-value = 0.5782
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.069915 10.403248
## sample estimates:
## mean in group 1 mean in group 2
##        52.66667        50.50000
```

The p-value is 0.5782013. The consumer goods and producer goods differ in their lpercentage of rising prices.

```
res <- t.test(y3 ~ Type, data = goods_df)
res
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  y3 by Type
## t = -3.166, df = 12.152, p-value = 0.008016
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.704651  -1.984238
## sample estimates:
## mean in group 1 mean in group 2
##        11.05556        17.40000
```

The p-value is 0.0080162. The consumer goods and producer goods differ in their cyclical amplitude.

```
res <- t.test(y4 ~ Type, data = goods_df)
res
```

```
##
##  Welch Two Sample t-test
##
## data:  y4 by Type
## t = -0.71867, df = 16.083, p-value = 0.4827
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5835069  0.2879514
## sample estimates:
## mean in group 1 mean in group 2
##       0.9222222       1.0700000
```

The p-value is 0.4826578. The consumer goods and producer goods differ in their rate of change.

Use Hotelling's T^2 test to test for a difference in the mean measurements vector of the Consumers Goods and the mean vector of the Producer Goods. State each hypotheses clearly, and interpret the results.

H_0: $\mu_1 = \mu_2$

H_1: The $\mu's$ are unequal

Let G_1 = Consumer goods and G_2 = Producer goods

Mean vectors

```
y_bar1 <- colMeans(consumer)
y_bar2 <- colMeans(producer)
```

$\overline{y_1} =$

y1	y2	y3	y4
48.61111	52.66667	11.05556	0.9222222

$\overline{y_2} =$

y1	y2	y3	y4
90.3	50.5	17.4	1.07

Covariance matrices

T2 = 18.4625

```
S1 <- cov(consumer)
S2 <- cov(producer)
```

The respective sample covariances matrices for the consumer goods:

S1 =

	y1	y2	y3	y4
y1	289.673611	12.0416667	44.3680556	-1.8777778
y2	12.041667	21.7500000	8.0833333	-0.1791667
y3	44.368056	8.0833333	28.4027778	0.9048611
y4	-1.877778	-0.1791667	0.9048611	0.2244444

The respective sample covariances matrices for the producer goods:

S2 =

	y1	y2	y3	y4
y1	870.4000000	-113.277778	25.1166667	0.8044444
y2	-113.2777778	119.833333	-5.0000000	-1.7611111
y3	25.1166667	-5.000000	8.6000000	0.5188889
y4	0.8044444	-1.761111	0.5188889	0.1734444

Pooled covariance matrix

```
Spl <- (1/ (nrow(consumer) + nrow(producer) - 2)) *
  ((nrow(consumer) - 1) * S1 + (nrow(producer) - 1) * S2)
```

	y1	y2	y3	y4
y1	597.1169935	-54.303922	34.1761438	-0.4577778
y2	-54.3039216	73.676471	1.1568627	-1.0166667
y3	34.1761438	1.156863	17.9189542	0.7005229
y4	-0.4577778	-1.016667	0.7005229	0.1974444

Inverse matrix of the sample pool covariance matrix of the two samples:

S_{pl}^{-1} =

	y1	y2	y3	y4
y1	0.0022495	0.0022743	-0.0059200	0.0379301
y2	0.0022743	0.0172313	-0.0105945	0.1315883
y3	-0.0059200	-0.0105945	0.0817965	-0.3584880
y4	0.0379301	0.1315883	-0.3584880	7.1021195

```r
library(ICSNP)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: ICS
```

```r
HotellingsT2(goods_df, formula = . ~ Type)
```

```
##
##  Hotelling's one sample T2-test
##
## data:  goods_df
## T.2 = 222.77, df1 = 5, df2 = 14, p-value = 7.954e-13
## alternative hypothesis: true location is not equal to c(0,0,0,0,0)
```

```r
HotellingsT2(consumer, producer)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  consumer and producer
## T.2 = 3.8011, df1 = 4, df2 = 14, p-value = 0.02702
## alternative hypothesis: true location difference is not equal to c(0,0,0,0)
```

Table 5: Beetles

Measurement.Number	Species	transverse.groove.dist	elytra.length	second.antennal.joint.length	third.antennal.joi
1	1	189	245	137	
2	1	192	260	132	
3	1	217	276	141	
4	1	221	299	142	
5	1	171	239	128	
6	1	192	262	147	
7	1	213	278	136	
8	1	192	255	128	
9	1	170	244	128	
10	1	201	276	146	
11	1	195	242	128	
12	1	205	263	147	
13	1	180	252	121	
14	1	192	283	138	
15	1	200	294	138	
16	1	192	277	150	
17	1	200	287	136	
18	1	181	255	146	
19	1	192	287	141	
1	2	181	305	184	
2	2	158	237	133	
3	2	184	300	166	
4	2	171	273	162	
5	2	181	297	163	
6	2	181	308	160	
7	2	177	301	166	
8	2	198	308	141	
9	2	180	286	146	
10	2	177	299	171	
11	2	176	317	166	
12	2	192	312	166	
13	2	176	285	141	
14	2	169	287	162	
15	2	164	265	147	
16	2	181	308	157	
17	2	192	276	154	
18	2	181	278	149	
19	2	175	271	140	
20	2	197	303	170	

Table 6: Economic cycles measurements for consumer goods and producer goods

Item	Type	y1	y2	y3	y4
1	1	72.0	50	8.0	0.5
2	1	66.5	48	15.0	1.0
3	1	54.0	57	14.0	1.0
4	1	67.0	60	15.0	0.9
5	1	44.0	57	14.0	0.3
6	1	41.0	52	18.0	1.9
7	1	34.5	50	4.0	0.5
8	1	34.5	46	8.5	1.0
9	1	24.0	54	3.0	1.2
1	2	57.0	57	12.5	0.9
2	2	100.0	54	17.0	0.5
3	2	100.0	32	16.5	0.7
4	2	96.5	65	20.5	0.9
5	2	79.0	51	18.0	0.9
6	2	78.5	53	18.0	1.2
7	2	48.0	50	21.0	1.6
8	2	155.0	44	20.5	1.4
9	2	84.0	64	13.0	0.8
10	2	105.0	35	17.0	1.8