Homework 2 - Multivariate

Elena Volpi

#Problem 1 a)

```
euclid_distance <- function(x,y) {
  diff <- x-y
  dist <- sqrt(t(diff)%*%diff)
  return(dist)
}

y_bar <- rbind(1,2,-2)
mu_1 <- rbind(0,0,0)
mu_2 <- rbind(3,4,-3.5)

euclid_distance(y_bar, mu_1)
```

```
##      [,1]
## [1,]    3
```

```
euclid_distance(y_bar, mu_2)
```

```
##          [,1]
## [1,] 3.201562
```

The euclidean distance for $d_E(\bar{\mathbf{y}}, \mu_{\mathbf{1}})$ is 3 and the for $d_E(\bar{\curvearrowright}, \mu_2) \approx 3.2$.

  b)

```
m_dist <- function(x,y,S) {
  diff <- x-y
  m_dist <- t(diff)%*%solve(cov_mat)%*%diff
  return(sqrt(m_dist))
}
#Aka S
cov_mat <- rbind(c(9,8.1,-3.6), c(8.1,9,-4.8), c(-3.6,-4.8,4))
m_dist(y_bar, mu_1,cov_mat)
```

```
##          [,1]
## [1,] 1.063808
```

```
m_dist(y_bar, mu_2,cov_mat)
```

```
##           [,1]
## [1,] 0.8387172
```

The Mahalanobis distance $d_M(\bar{\mathbf{y}}, \mu_{\mathbf{1}}) = 1.0638$ and $d_M(\bar{\curvearrowright}, \mu_2) = 0.8387$ for covariance matrix

$$\Sigma = \begin{bmatrix} 9 & 8.1 & -3.6 \\ 8.1 & 9 & -4.8 \\ -3.6 & -4.8 & 4 \end{bmatrix}$$

.

c) If $\bar{\mathbf{y}}$ is a sample mean from an i.i.d sample, which of $\mu_{\mathbf{1}}$ and $\mu_{\mathbf{2}}$ would you consider more plausible as the mean of the population.

Since the mahalanobis and euclidean distance did not produce similar results, we may have highly correlated variables. Therefore, we'll use the mahalanobis distance as the more accurate of the two methods. Since μ_2 has the smaller mahalanobis distance, it is the more pluasible mean of the population.

#Problem 2

a)

```r
TestScores <- read_csv("~/Desktop/Multivariate/Homework2/TestScores.csv", show_col_types = FALSE)

mu <- c(500,50,30)
T2.test(x=TestScores,mu=mu)
```

```
##
##  One-sample Hotelling test
##
## data:  TestScores
## T2 = 223.310, F = 72.706, df1 = 3, df2 = 84, p-value < 2.2e-16
## alternative hypothesis: true mean vector is not equal to (500, 50, 30)'
##
## sample estimates:
##                  SocSciHist   Verbal  Science
## mean x-vector   526.5862 54.68966 25.12644
```

```r
n <- nrow(TestScores)
p <- ncol(TestScores)

F_crit_val <- qf(0.95, df1 =p,df2 = n-p)
```

We see that our p-value is much less than our $\alpha = 0.05$ and the null hypothesis is rejected so we have reason to believe that for 2011 the scoring is different.

b) Determine the lengths and directions for the axes of the 95% confidence ellipsoid for μ.

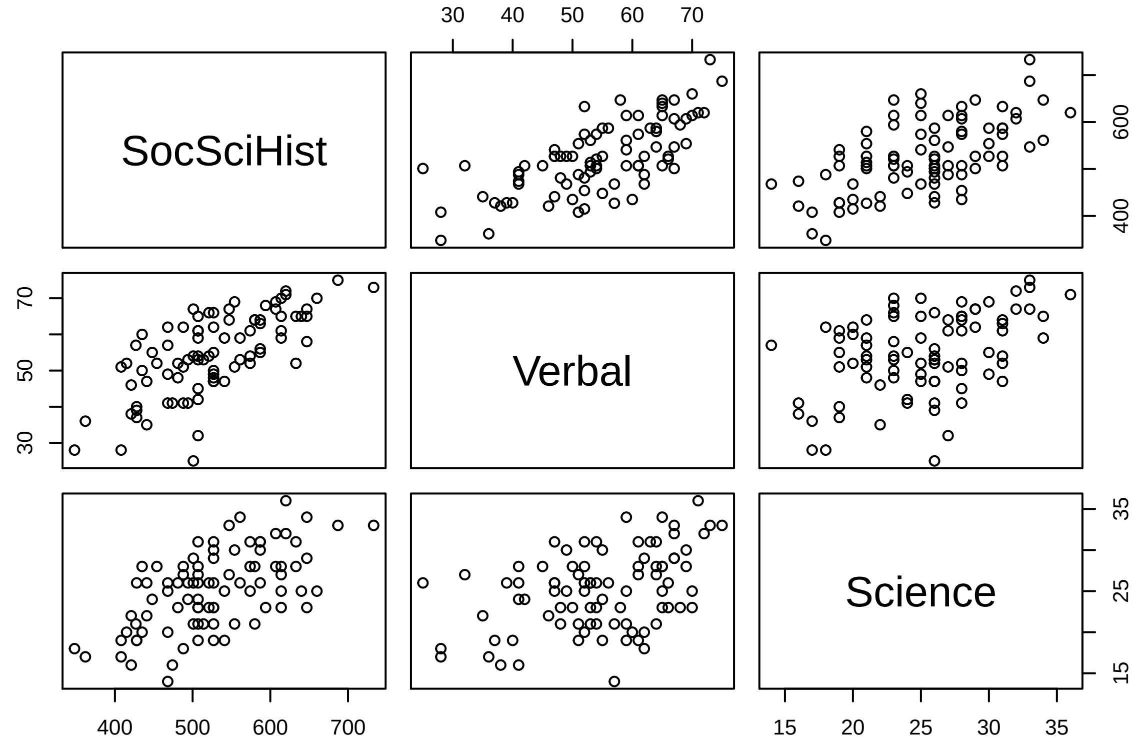
```r
eigen_test_scores <- cov(TestScores) %>% eigen()
n <- nrow(TestScores)
p <- ncol(TestScores)

F_crit_val <- qf(0.95, df1 =p,df2 = n-p)
region <- sqrt(eigen_test_scores$values*(p*(n-1)/(n*(n-p)))*F_crit_val)  #formula lecture8,slide 33
(directions <- t(eigen_test_scores$vectors))
```

```
##             [,1]         [,2]         [,3]
## [1,]  0.9939054  0.103443390  0.038099056
## [2,]  0.1037315 -0.994589227 -0.005660238
## [3,] -0.0373074 -0.009577815  0.999257936
```

Our 95% confidence ellipsoid has lengths 23.73, 2.47, 1.18 with directions with directions $[0.994, 0.104, -0.037]^T$, $[0.103, -0.995, -0.010]^T$, and $[0.038, -0.006, 0.999]^T$ respectively. c)

```
pairs(TestScores)
```



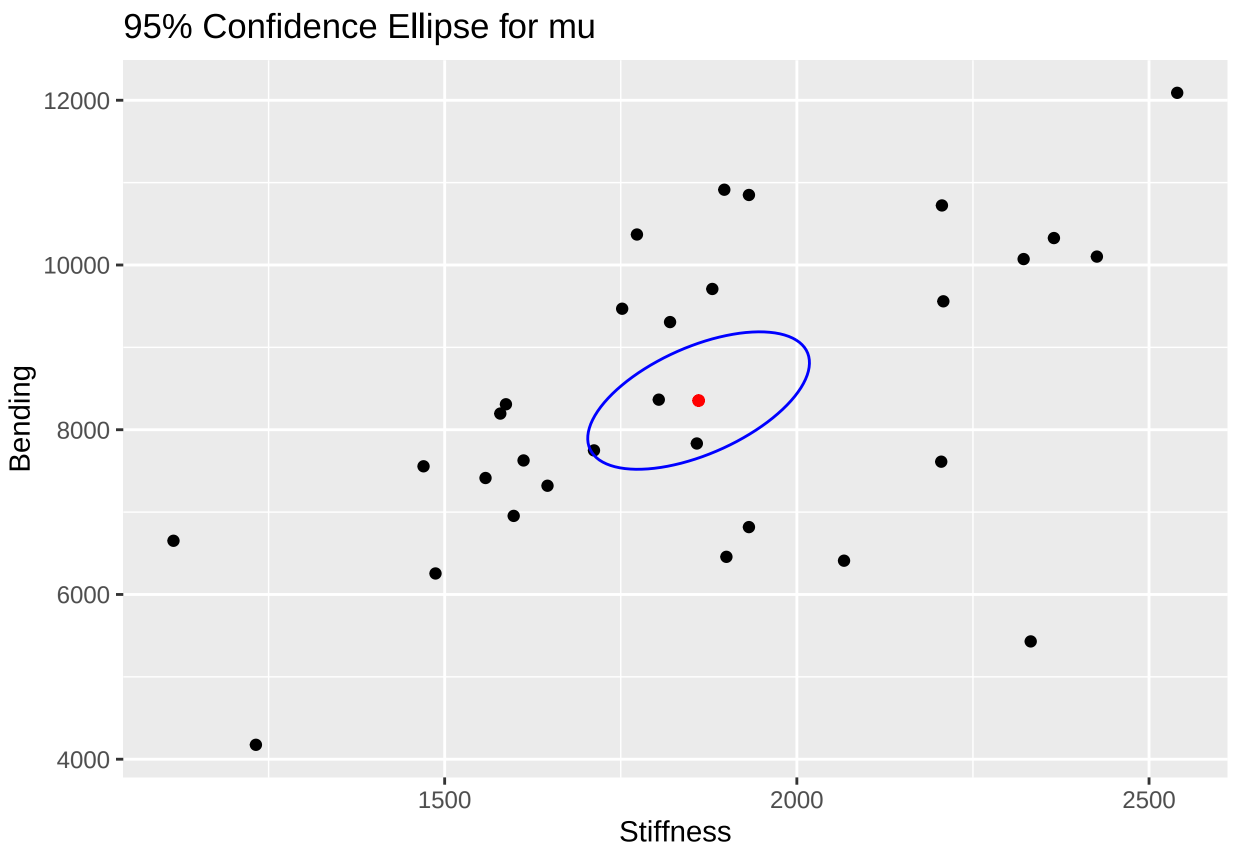
The results are consistent with approximate normality. Our sample size is large enough that the CLT applies.

#Problem 3

```
LumberData <- read_csv("~/Desktop/Multivariate/Homework2/LumberData.csv",show_col_types = FALSE)
```

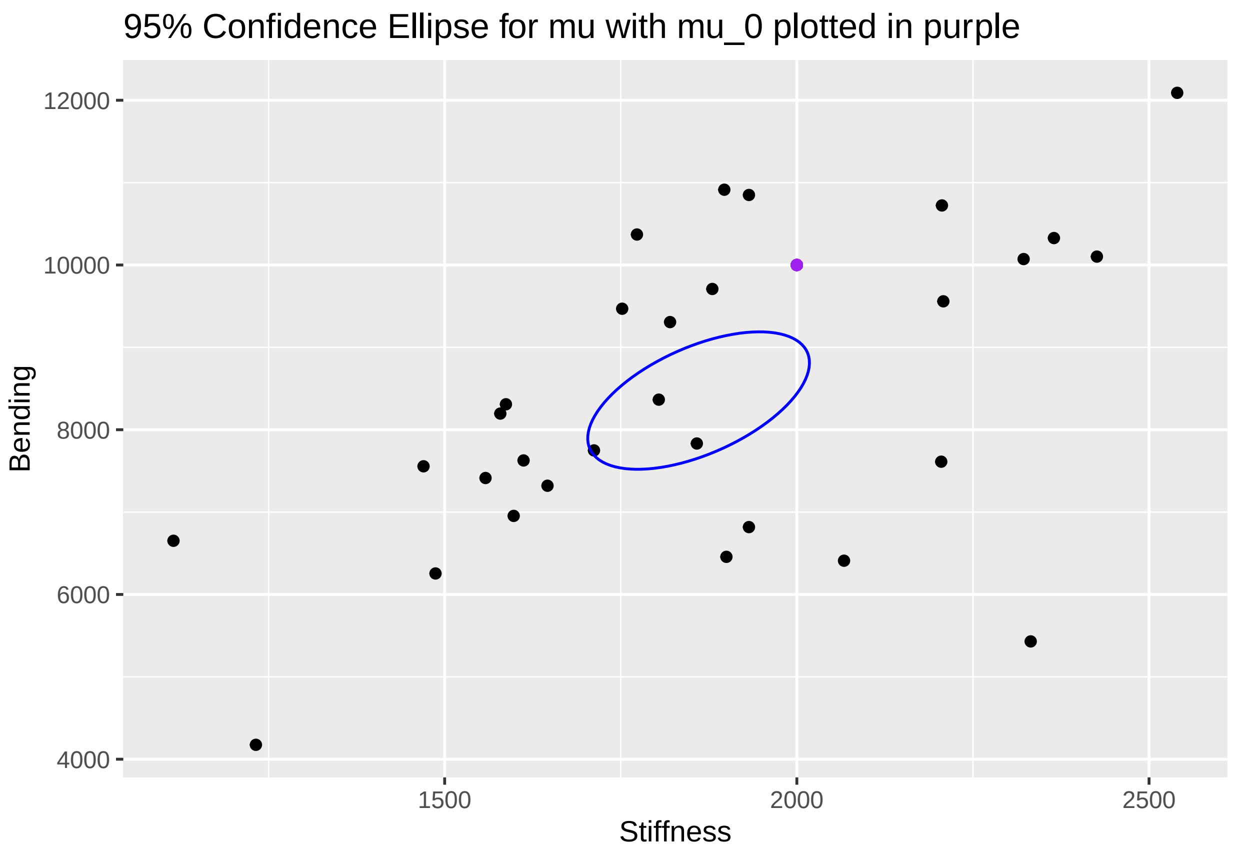
a)

```
ellipse1 <- ggplot(aes(x= Stiffness, y=Bending), data= LumberData)+
  geom_point() +
  geom_point(aes(x=mean(Stiffness),y=mean(Bending)), color = "red")+
  stat_conf_ellipse(color="blue")
ellipse1 + labs(title = "95% Confidence Ellipse for mu")
```



b) If we plot the suggested vector, $\mu_0 = [2000, 10000]^T$, in purple :

```
ellipse2 <- ggplot(aes(x= Stiffness, y=Bending), data= LumberData)+
  geom_point() +
  geom_point(aes(x=2000,y=10000), color = "purple")+
  stat_conf_ellipse(color="blue")
ellipse2 + labs(title = "95% Confidence Ellipse for mu with mu_0 plotted in purple")
```



We see that $\mu_0 = [2000, 10000]^T$ is outside the 95% confidence ellipse for μ which suggests that μ_0 is outside the plausible range of values for the lumber data set.

#Problem 4

```
BoneMineral <- read_csv("~/Desktop/Multivariate/Homework2/BoneMineral.csv", show_col_types = FALSE)
```

a)

```
alpha <- 0.05
n <- nrow(BoneMineral)
p <- ncol(BoneMineral)
alpha_star <- alpha/p

bone_xbar <- apply(BoneMineral,2,mean)
bone_var <- apply(BoneMineral,2,var)
bone_cov <- cov(BoneMineral)

c1 <- qt(1-alpha_star,n-1)*sqrt(bone_var/n)
bon_ub <- bone_xbar+c1
bon_lb<- bone_xbar -c1

bon_CI <- cbind(bon_lb,bon_ub)
bon_CI
```

```
##               bon_lb    bon_ub
## dRadius  0.7851084 0.9024916
## nRadius  0.7633190 0.8733210
## dHumerus 1.6467682 1.9385918
## nHumerus 1.5991581 1.8705219
## dUlna    0.6490376 0.7597624
## nUlna    0.6408476 0.7468324
```

The Bonferroni Confidence Intervals for $\alpha = 0.05$ is shown above.

b)

```
c2 <- sqrt((p*(n-1)/(n*(n-p)))*qf(1-alpha, p, n-p)*bone_var)
hot_ub <- bone_xbar+c2
hot_lb <- bone_xbar-c2
(hot_CI <- cbind(hot_lb,hot_ub))
```

```
##             hot_lb    hot_ub
## dRadius  0.7420179 0.9455821
## nRadius  0.7229380 0.9137020
## dHumerus 1.5396419 2.0457181
## nHumerus 1.4995425 1.9701375
## dUlna    0.6083914 0.8004086
## nUlna    0.6019414 0.7857386
```

The Hotelling's T^2 confidence interval is wider than the Bonferroni, that is that the Bonferroni confidence interval falls within the Hotelling's T^2 confidence interval. This makes sense since the Bonferroni has an adjusted α of $\alpha/2$.

#Problem 5

```
FlourBags <- read_csv("~/Desktop/Multivariate/Homework2/FlourBags.csv",show_col_types = FALSE)
```

a) Hotelling's T^2 method

```
mu_1 <- c(10,10,10)
T2.test(x=FlourBags,mu=mu_1)
```

```
##
##  One-sample Hotelling test
##
## data:  FlourBags
## T2 = 15.4265, F = 4.6009, df1 = 3, df2 = 17, p-value = 0.0156
## alternative hypothesis: true mean vector is not equal to (10, 10, 10)'
##
## sample estimates:
##                 Scale1 Scale2 Scale3
## mean x-vector   9.73  10.02  9.863
```

With a p-value of 0.0156, at level $\alpha = 0.05$ we have fairly strong evidence to reject the null hypothesis in favor of the alternative, suggesting that the average weight for the three scales is not equal.

b) Bonferroni Method

```
alpha <- 0.05
alpha_star <- alpha/3

p_values <- rep(0,3)
for(i in 1:3){
  p_values[i] <- t.test(FlourBags[,i], mu = mu_1[i], conf.level = alpha_star)$p.value
}
p_values
```

```
## [1] 0.01701377 0.76802686 0.15068959
```

Our p-values are 0.017, 0.77, and 0.15 and thus we have insufficient evidence to reject the null hypothesis that all three scales are the same average weight at an α =0.05 level.

c) Since our conclusion from the Bonferonni method was that we would fail to reject the null hypothesis, the T^2 simultaneous intervals would contain u_0.

d)

```
flour_new <- apply(FlourBags,1,mean)
t.test(flour_new,mu=10)
```

```
##
##  One Sample t-test
##
## data:  flour_new
## t = -1.8113, df = 19, p-value = 0.08594
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.721932 10.020068
## sample estimates:
## mean of x
##     9.871
```

We fail to reject the null hypothesis,$H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 = 10$, at an α =0.05 level with a p-value of 0.086.

#Problem 6

```
ReadingTest <- read_csv("~/Desktop/Multivariate/Homework2/ReadingTest.csv",show_col_types = FALSE)
```

a) A paired test would be appropriate. Each row corresponds to one subject. Each subject took the same two tests, and we have data for before and after treatment.

b)

```
diff_u1u2 <- cbind(ReadingTest$PRE1-ReadingTest$PRE2, ReadingTest$POST1-ReadingTest$POST2)
alpha <- 0.05
n <- nrow(diff_u1u2)
p <- ncol(diff_u1u2)
alpha_star <- 0.05/p
t.test(diff_u1u2[,1])
```

```
##
##  One Sample t-test
##
## data:  diff_u1u2[, 1]
## t = 12.312, df = 65, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  3.922390 5.441246
## sample estimates:
## mean of x
##  4.681818
```

```
t.test(diff_u1u2[,2])
```

```
##
##  One Sample t-test
##
## data:  diff_u1u2[, 2]
## t = 2.6626, df = 65, p-value = 0.009764
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.3407949 2.3864778
## sample estimates:
## mean of x
##  1.363636
```

At α =0.05, with p-values 0f 2.2 x 10^{-}16 and 0.009764, we have strong evidence to reject the null hypothesis in favor of the alternative that instruction impacted average population test scores.

c)

```
T2.test(x = ReadingTest[,2:3], y= ReadingTest[,4:5])
```

```
##
##  Two-sample Hotelling test
##
## data:  ReadingTest[, 2:3] and ReadingTest[, 4:5]
## T2 = 28.862, F = 14.320, df1 = 2, df2 = 129, p-value = 2.418e-06
## alternative hypothesis: true difference in mean vectors is not equal to (0,0)
## sample estimates:
##                     PRE1     PRE2
## mean x-vector 9.787879 5.106061
## mean y-vector 8.075758 6.712121
```

We have strong evidence at a 95% confidence level to suggest reading instruction produces a difference in performance on the two tests.

d)

```
mean_difference <- apply(diff_u1u2,2,mean)
delta_vec <- c(1,1)
reading_cov <- cov(diff_u1u2)
scale <- (n-p)/((n-1)*p)

delta <- scale*n*t(mean_difference - delta_vec)%*%
  solve(reading_cov)%*%(mean_difference - delta_vec)

F_crit <- qt(0.95,p,n-1)

delta < F_crit
```

```
##      [,1]
## [1,] TRUE
```

Δ_0 will be in our 95% confidence interval.