GR5223 - HW3

MATHIEU SAUTEREY – UNI: MJS2364

26 February 2018

Problem 1

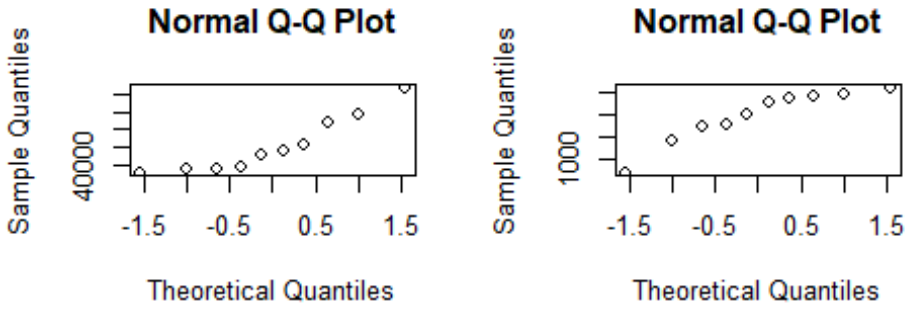
Consider again the corporations data from the last exercise on the previous homework assignment. Let x1 = Sales, x2 = Profits and x3 = Assets.

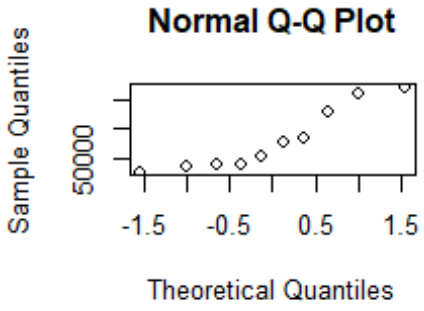
```
# Loads dataset from the data CSV file
data <- read.csv("Companies.csv", as.is = TRUE)

head(data)
```

```
##    Sales Profits Assets
## 1 126974    4224 173297
## 2  96933    3835 160893
## 3  86656    3510  83219
## 4  63438    3758  77734
## 5  55265    3939 128344
## 6  50976    1809  39080
```

(a) Construct normal probability plots (Q-Q plots) for each variable individually. Does treating these data as a random sample from a multivariate normal distribution seem reasonable? Explain.

```
# Plots the QQ-plot for each univariate variable
par(mfrow=c(2,2))
qqnorm(data$Sales)
qqnorm(data$Profits)
qqnorm(data$Assets)
```

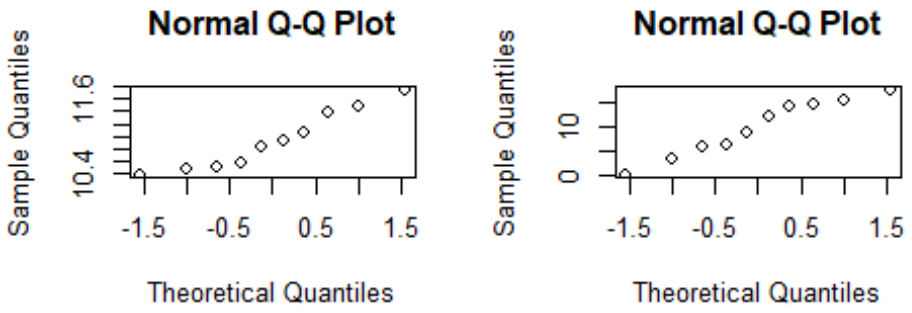


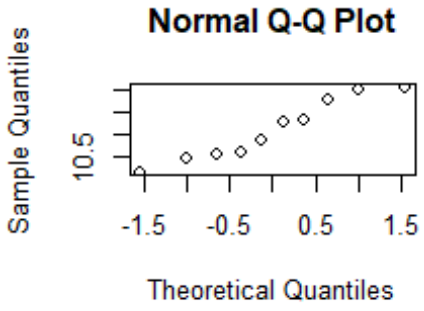


Looking at the QQ-plots we observe that the univariate distributions deviate significantly from a straight line representing a normal distribution. Therefore it is not reasonbale to consider that these variables are samples from a multivariate normal distribution.

(b) Consider the transformations y1 = ln(x1), y2 = (x2/1000)^2, and y3 = ln(x3). Do the normal probability plots look better?

```
# Creates a new dataframe with transformed data
data_tr <- data.frame("Y1"=log(data$Sales), "Y2"=(data$Profits/1000)^2,
"Y3"=log(data$Assets))

# Plots the QQ-plot for each transformed univariate variable
par(mfrow=c(2,2))
qqnorm(data_tr$Y1)
qqnorm(data_tr$Y2)
qqnorm(data_tr$Y3)
```



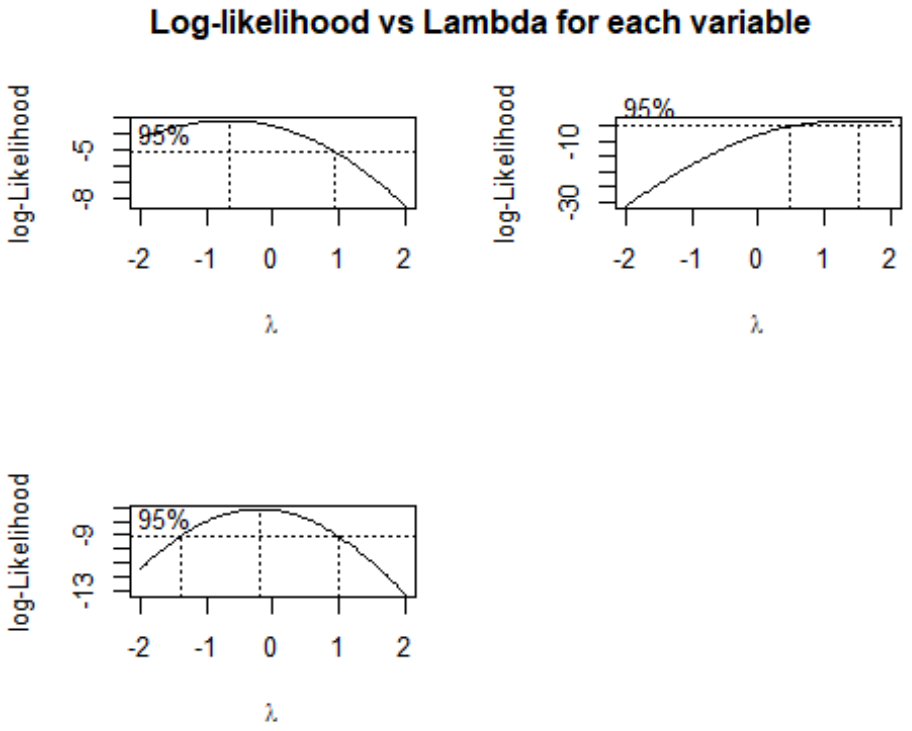


We observe that the transformed univariate data have a QQ-plot closer to a straight line than before, and so they are closer to a normal distribution. Therefore the transformation was helpful but the QQ-plots are still somewhat far from displaying a straight line. We conclude that we need to search for an optimal transformations.

(c) Use the Box-Cox method to find optimal power transformations of x1, x2 and x3 individually. Are the power transformations suggested in part (b) within 95% confidence limits for the optimal transformation?

```
library(MASS)

# Uses Box-Cox method to calculate and plot log-likelihood vs lambda
par(mfrow=c(2,2))
bc.1 <- boxcox(lm(data$Sales~1))
bc.2 <- boxcox(lm(data$Profits~1))
bc.3 <- boxcox(lm(data$Assets~1))
title("Log-likelihood vs Lambda for each variable", line = -1.5, outer=TRUE,
font=2)

# Finds the optimal lambda for "Sales"
max_loc.1 <- which.max(bc.1$y)
lambda.1 <-bc.1$x[max_loc.1]
print(lambda.1)

## [1] -0.6666667
```

```r
# Finds the optimal lambda for "Profits"
max_loc.2 <- which.max(bc.2$y)
lambda.2 <-bc.2$x[max_loc.2]
print(lambda.2)

## [1] 1.515152

# Finds the optimal lambda for "Assets"
max_loc.3 <- which.max(bc.3$y)
lambda.3 <-bc.3$x[max_loc.3]
print(lambda.3)

## [1] -0.1818182
```



The power transformations in (b) are the following for each univariate variable:

- "Sales" undergoes a log transformation equivalent to $\lambda = 0$, which is within the 95% confidence interval as seen on the log-Likelihood vs λ
- "Profits" undergoes a power transformation equivalent to $\lambda = 2$, which is within the 95% confidence interval as seen on the log-Likelihood vs λ
- "Assets" undergoes a log transformation equivalent to $\lambda = 0$, which is within the 95% confidence interval as seen on the log-Likelihood vs λ

(d) Use the Box-Cox method to find the optimal power transformation of x = (x1, x2, x3) simultaneously. Obtain the p-value for a test of the null hypothesis that the optimal vector of powers is lambda = (0, 2, 0), as suggested in part (b).

```
library(car)

# Uses Box-Cox method to find the optimal lambda vector
bc.multi <- powerTransform(data)
print(powerTransform(data))

## Estimated transformation parameters
##       Sales     Profits      Assets
## -0.61580687  2.59756500  0.02403237

# Likelihood Ratio Test for optimal lambda vector = (0, 2, 0)
testTransform(bc.multi, c(0, 2, 0))

##                                LRT df      pval
## LR test, lambda = (0 2 0) 1.705446  3 0.6357238
```

According to the Box-Cox method, the optimal vector of lambda powers for the multivariate transformation is: λ = (-0.6158, 2.5976, 0.0240).

We then test the null hypothesis H_0: $\lambda = (0,2,0)$ using the likelihood ratio test and we find a p-value of 0.6357 which supports the conlusion that $\lambda = (0,2,0)$

Problem 4

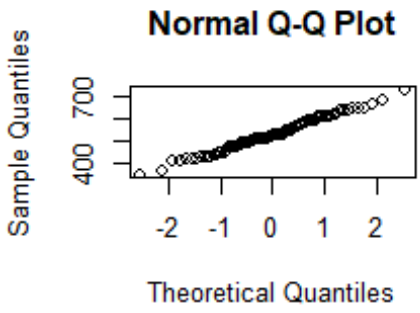
The scores obtained by n = 87 college students on the College Level Examination Program social science & history subtest (x1), and the College Qualification Test verbal (x2) and science (x3) subtests are given in the data file College.csv, in the Data folder on Courseworks.

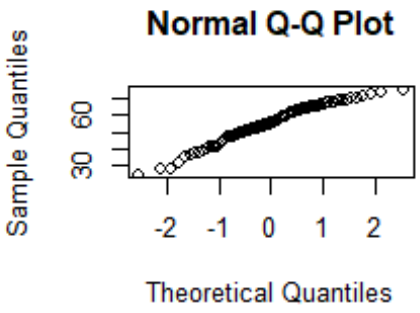
```
# Loads dataset from the data CSV file and renames columns
grades <- read.csv("College.csv", as.is = TRUE)
colnames(grades) <- c("SS_History","Verbal","Science")
head(grades)
```

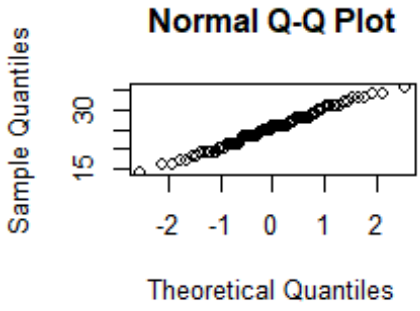
```
##   SS_History Verbal Science
## 1        468     41      26
## 2        428     39      26
## 3        514     53      21
## 4        547     67      33
## 5        614     61      27
## 6        501     67      29
```

(a) Construct Q-Q plots from the marginal distributions of social science & history, verbal, and science scores. Also, construct the three possible scatter diagrams for the pairs of observations of different variables. Do these data appear to be normally distributed? Discuss.

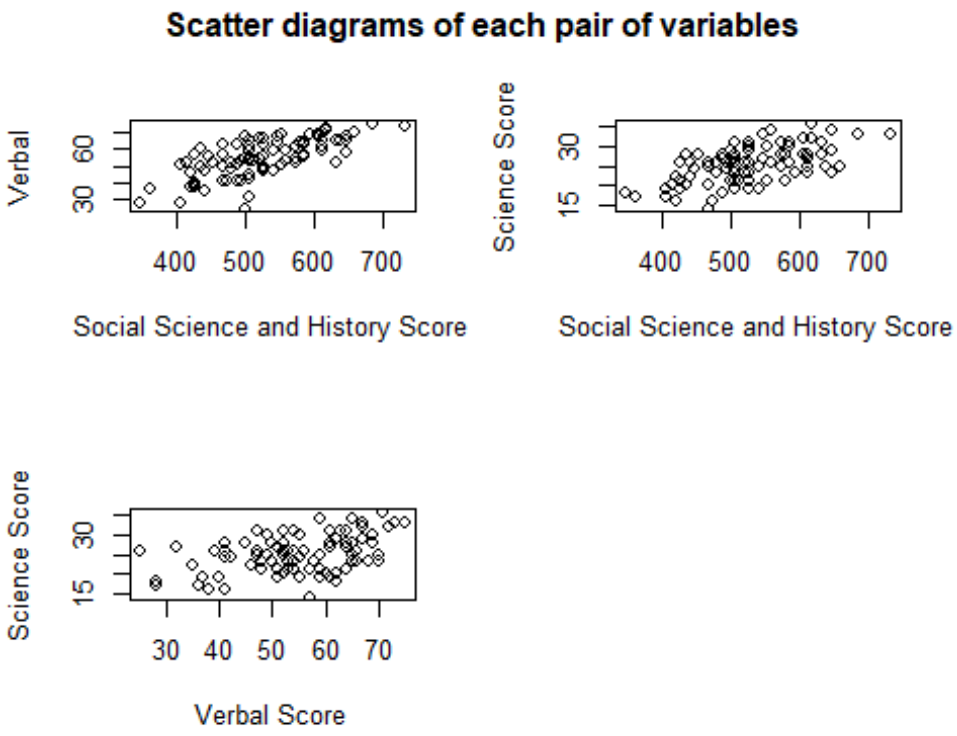
```
# Plots the QQ-plot for each univariate variable
par(mfrow=c(2,2))
qqnorm(grades$SS_History)
qqnorm(grades$Verbal)
qqnorm(grades$Science)

# Plots the scatter diagrams of each two pair of variables
par(mfrow=c(2,2))
```








```
plot(grades$SS_History, grades$Verbal, xlab="Social Science and History
Score",
     ylab="Verbal")
plot(grades$SS_History, grades$Science, xlab="Social Science and History
Score",
     ylab="Science Score")
plot(grades$Verbal, grades$Science, xlab="Verbal Score",
     ylab="Science Score")
title("Scatter diagrams of each pair of variables", line = -1.5, outer=TRUE,
font=2)
```



Looking at the QQ-plots we observe that the marginal distributions are close to a straight line representing the normal distribution. Thus the data appear to be normally distrbuted. Additionally, when we plot the scatter diagrams for each pair of variables we observe that the data clusters look like ellipses, which further supports normality.

(b) Suppose the average scores for thousands of college students over the last 10 years are 500 for social science & history, 50 for verbal, and 30 for science. Is there reason to believe that the group of students represented by the scores in College.csv is scoring differently? Support your answer with the p-value for a test of H0 : (mu1, mu2, mu3) = (500, 50, 30).

```
library(ellipse)

# Loads and calculates basic parameters
n_g     <- nrow(grades)
p_g     <- 3
xbar_g  <- colMeans(grades)
alpha_g <- 0.05
mu0_g   <- c(500, 50, 30)
S_g     <- cov(grades)
```

```r
# Calculates Hotelling's T2-statistic
T2_g <- n_g * sum( (xbar_g - mu0_g) * solve(S_g, xbar_g - mu0_g) )

# Finds the p-value by comparing T2 statistic to the F-distribution(p;n-p)
p.value_g <- 1 - pf((n_g-p_g)/((n_g-1)*p_g) * T2_g, df1=p_g, df2=n_g-p_g)
p.value_g
```

```
## [1] 0
```

The p-value is exactly zero so with absolute certainty we reject the null hypothesis $H0$: $(\mu 1, \mu 2, \mu 3) = (500,50,30)$. Therefore the data in Lumber.csv are different from typical scores, which means that the student represented in the dataset are scoring differently than average students in the last 10 years.

Problem 5

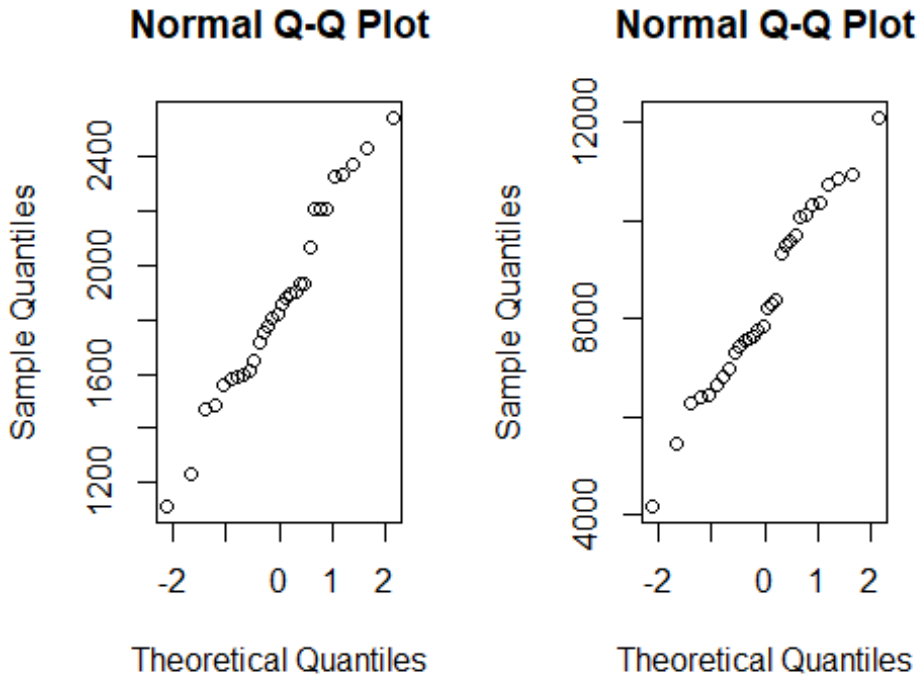
Measurements of x1 = stiffness and x2 = bending strength for a sample of n = 30 pieces of a particular grade of lumber are given in the data file Lumber.csv, in the Data folder on Courseworks; the units are pounds per square inch.

```
# Loads dataset from the data CSV file and renames columns
lumber <- read.csv("Lumber.csv", as.is = TRUE)
colnames(lumber) <- c("Stiffness","Strength")
head(lumber)
```

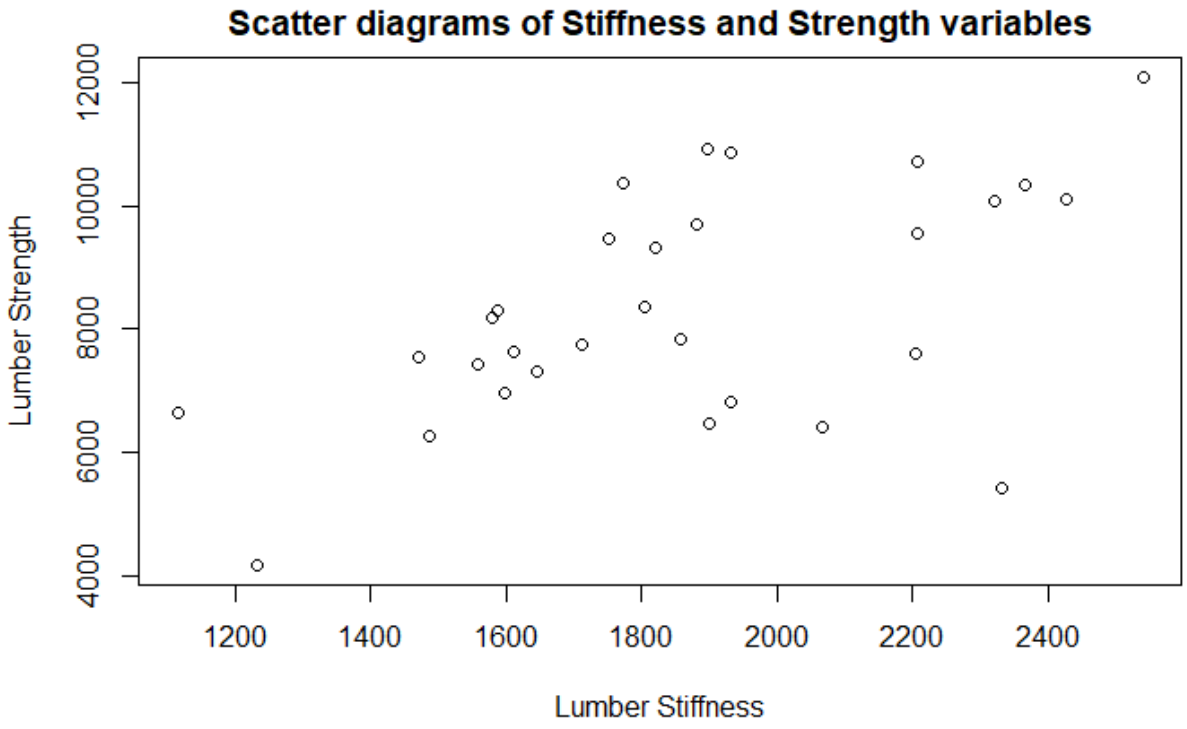
```
##   Stiffness Strength
## 1      1232     4175
## 2      1115     6652
## 3      2205     7612
## 4      1897    10914
## 5      1932    10850
## 6      1612     7627
```

(a) Is the bivariate normal distribution a viable population model? Explain with reference to Q-Q plots and a scatter diagram.

```
# Plots the QQ-plot for each univariate variable
par(mfrow=c(1,2))
qqnorm(lumber$Stiffness)
qqnorm(lumber$Strength)
```



```r
# Plots the scatter diagram of the two variables
plot(lumber$Stiffness, lumber$Strength, xlab="Lumber Stiffness", ylab="Lumber
Strength",
     main="Scatter diagrams of Stiffness and Strength variables")
```

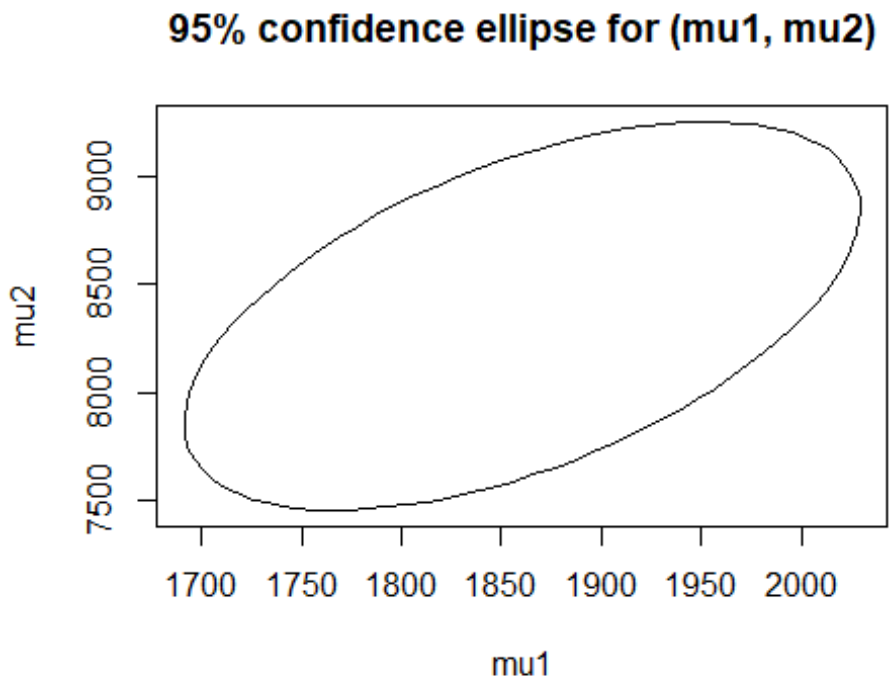


Looking at the QQ-plots we observe that the univariate distributions do not deviate deviate significantly from a straight line representing the normal distribution. If the joint distribution of the data is bivariate normal, then we expect each marginal univariate to be normally distributed. This is exactly what the QQ-plots shows -each univariate variable approximately follows a normal distribution-. Thus our results indicate that the bivariate normal distribution a viable population model. Additionally, when we plot the scatter diagrams for the pair of variables we observe that the data cluster look like an ellipse, which further supports normality.

(b) Find the equation of a 95% confidence ellipse for the pair (mu1, mu2), the true mean stiffness and bending strength for this grade of lumber. Sketch this confidence region.

```r
# Loads and calculates basic parameters
n <- nrow(lumber)
p <- 2
xbar <- colMeans(lumber)
alpha <- 0.05
```

```r
# Confidence ellipse for means (mu1, mu2)
size <- sqrt((n-1)*p/(n-p) * qf(.95, p, n-p)) # p=2 for two variables
plot(ellipse(x=(1/n)*cov(lumber[,1:2]), centre=xbar[1:2], t=size), type="l",
     xlab="mu1", ylab="mu2", main="95% confidence ellipse for (mu1, mu2)")
```




```r
# Bonferroni method to calculate a 95% simultaneous confidence interval for
(mu1, mu2)
q <- 2
mean(lumber$Stiffness) + c(-1,1) * qt(1-alpha/(2*q), df=n-1) *
sd(lumber$Stiffness)/sqrt(n)

## [1] 1708.492 2012.508

mean(lumber$Strength) + c(-1,1) * qt(1-alpha/(2*q), df=n-1) *
sd(lumber$Strength)/sqrt(n)

## [1] 7548.304 9159.963
```

Therefore, we are 95% confident that $1708.492 \leq \text{mu1} \leq 2012.508$ and $7548.304 \leq \text{mu2} \leq 9159.963$.

(c) Suppose mu1 = 2000 and mu2 = 10,000 represent "typical" values for stiffness and bending strength, respectively. Are the data in Lumber.csv consistent with these values? Explain.

First we notice that the point mu(2000, 10000) is outside of the 95% ellipse of mean values so the data are not consistent with these values. We are now going to test the null hypothesis $H_0: \mu = (2000,10000)$:

```
# Loads and calculates basic parameters needed to perform a test
mu0 <- c(2000,10000)
S <-cov(lumber)

# Calculates Hotelling's T2-statistic
T2 <- n * sum( (xbar - mu0) * solve(S, xbar - mu0) )

# Finds the p-value by comparing statistic to the F-distribution(p;n-p)
p.value <- 1 - pf((n-p)/((n-1)*p) * T2, df1=p, df2=n-p)
p.value

## [1] 0.0002367286
```

The p-value is approximately zero so with almost absolute certainty we reject the null hypothesis $H_0: \mu = (2000,10000)$. Therefore the data in Lumber.csv are not consistent with the statement that mu1 = 2000 and mu2 = 10,000 represent "typical" values for stiffness and bending strength, respectively.