ST 502 R Project 2

Elizabeth Burke Bruce Campbell, John Davidson

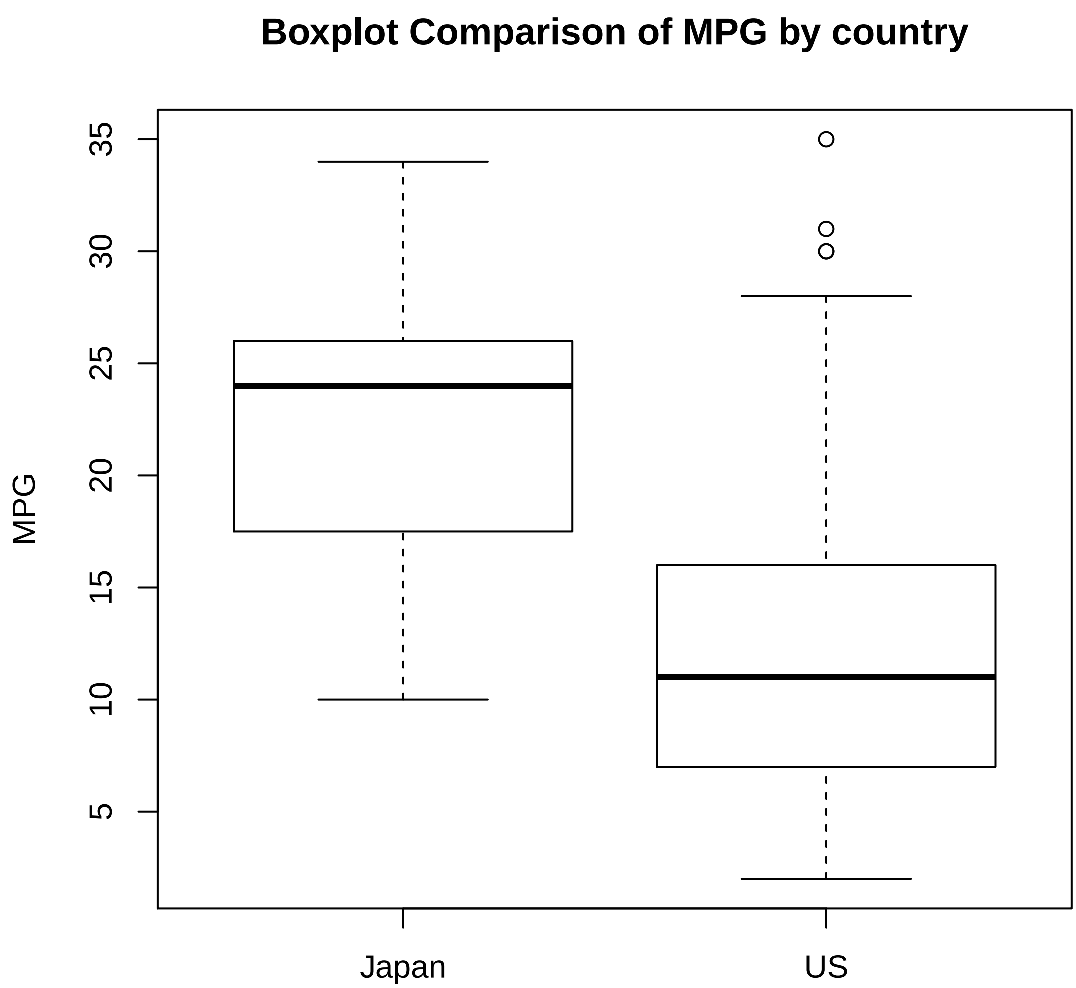
April 14, 2017

This report considers independent samples of miles per gallon measurements of US and Japanese manufactured cars. For this analysis we assume the measurements come from normally distributed populations. To test the hypothesis that means of the two populations are different we will perform two-sample t-tests. We will be performing a test where equal variance is assumed (pooled) and one where unequal variances are assumed.

Plot the data

```r
# install.packages('pander') install.packages('plyr')
# install.packages('dplyr') install.packages('readr')
# install.packages('ggplot2')
library(pander)
library(plyr)
library(dplyr)
library(readr)
library(ggplot2)

mpg <- read.table("mpg.txt")
# This makes the variables available in the name space
attach(mpg)
boxplot(MPG ~ Country, main = "Boxplot Comparison of MPG by country", ylab = "MPG")
```



Part 1 - calculation of the confidence intervals

a) Conduct both two-sample t-tests

In this section we calculate the 2 sample t-test on the data at the $\alpha = 0.05$ signicance level.

2-Sample t-test equal variance

```
alpha <- 0.05

pooled.var <- t.test(x = mpg[Country == "Japan", ]$MPG, y = mpg[Country == "US",
    ]$MPG, alternative = "two.sided", var.equal = TRUE, conf.level = alpha)
pooled.var
```

```
##
##  Two Sample t-test
##
## data:  mpg[Country == "Japan", ]$MPG and mpg[Country == "US", ]$MPG
## t = 12.238, df = 326, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 5 percent confidence interval:
##  10.02165 10.12496
## sample estimates:
## mean of x mean of y
##  22.35443  12.28112
```

We see that for the 2 sample t-test with pooled variance the $p - value < 2.2e - 16$ this means there is not enough evidence to support the null hypothesis that the means are the same. We reject the null hypothesis and claim that the evidence supports that the population mean mpg of Japanese and US manufactured cars are different.

2-Sample t-test unequal variances

```r
unequal.var <- t.test(x = mpg[Country == "Japan", ]$MPG, y = mpg[Country ==
    "US", ]$MPG, alternative = "two.sided", var.equal = FALSE, conf.level = alpha)
unequal.var
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg[Country == "Japan", ]$MPG and mpg[Country == "US", ]$MPG
## t = 12.99, df = 145.45, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 5 percent confidence interval:
##  10.02459 10.12202
## sample estimates:
## mean of x mean of y
##  22.35443  12.28112
```

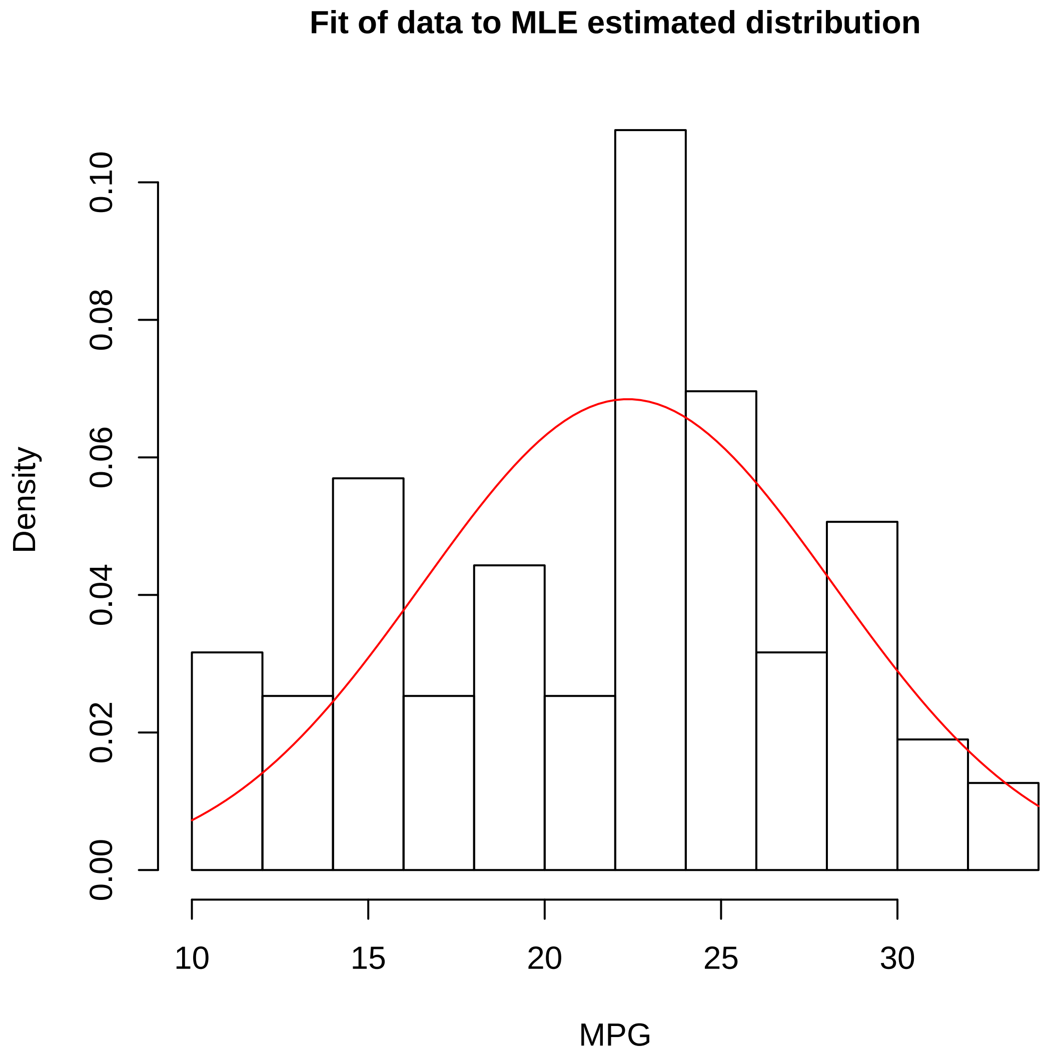
We see that for the 2 sample t-test with unequal variances assumet that the $p - value < 2.2e - 16$ this means there is not enough evidence to support the null hypothesis that the means are the same. We reject the null hypothesis and claim that the evidence supports that the population mean mpg of Japanese and US manufactured cars are different.

Check for normality

Here we plot the

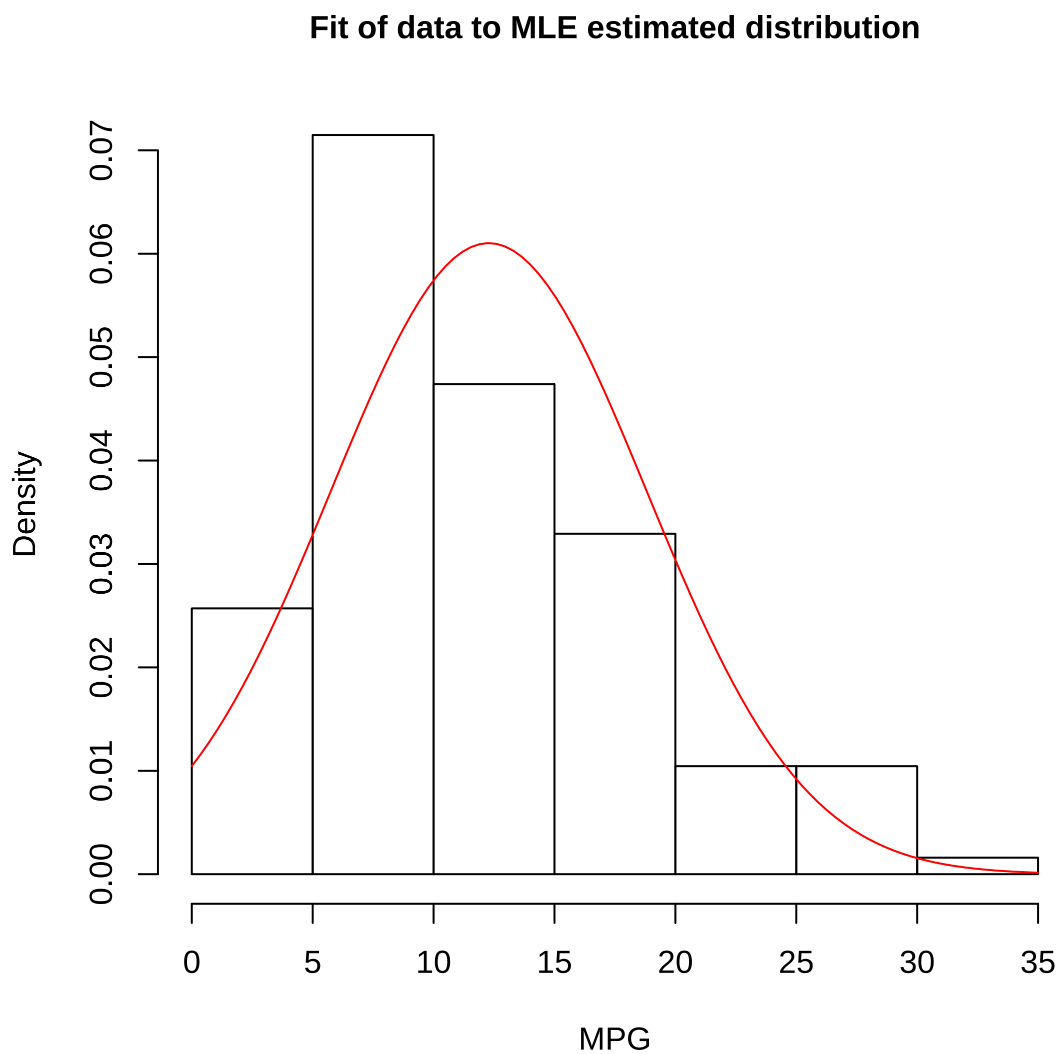
```r
Japan <- mpg[Country == "Japan", ]$MPG
Japan.MLE.mean <- mean(Japan)
Japan.MLE.SD <- sd(Japan)

hist(Japan, 10, freq = FALSE, cex.main = 1, main = "Fit of data to MLE estimated distribution",
    xlab = "MPG")
curve(dnorm(x, Japan.MLE.mean, Japan.MLE.SD), add = TRUE, col = "red")
```




```r
US <- mpg[Country == "US", ]$MPG
US.MLE.mean <- mean(US)
US.MLE.SD <- sd(US)

hist(US, 10, freq = FALSE, cex.main = 1, main = "Fit of data to MLE estimated distribution",
    xlab = "MPG")
curve(dnorm(x, US.MLE.mean, US.MLE.SD), add = TRUE, col = "red")
```



We see the US MPG data is right skewed.

Part 2

Set up the parameters for the simulation study.

```
n1Vec <- c(10, 25, 60)
n2Vec <- c(10, 25, 60)

mu1 <- 0
# Delta = mu1-mu2
deltaVec <- c(-5, -1, 0, 1, 5)

var1Vec <- c(1, 3, 9)
var2 <- 1
```

```r
alphaVec <- c(0.025, 0.05, 0.1, 0.15, 0.2)

parameter.grid <- expand.grid(n1Vec, n2Vec, deltaVec, var1Vec, alphaVec)

colnames(parameter.grid) <- c("n1", "n2", "delta", "var1", "alpha")

pander(parameter.grid[1:10, ], caption = "First Few elements of parmaters space")
```

Table 1: First Few elements of parmaters space

n1	n2	delta	var1	alpha
10	10	-5	1	0.025
25	10	-5	1	0.025
60	10	-5	1	0.025
10	25	-5	1	0.025
25	25	-5	1	0.025
60	25	-5	1	0.025
10	60	-5	1	0.025
25	60	-5	1	0.025
60	60	-5	1	0.025
10	10	-1	1	0.025

The first few elements of the parameter space are displayed in the table above. There are 675 elements of the parameter space.

Simulate from the null distribution and check the empirical acceptance probability.

```r
# The number of hypothesis test we run for each parameter configuration.
B <- 30

results.alpha <- parameter.grid

results.alpha$pooled.proportion.accepted <- numeric(nrow(results.alpha))
results.alpha$unequal.proportion.accepted <- numeric(nrow(results.alpha))

for (i in 1:nrow(parameter.grid)) {
    parameters <- parameter.grid[i, ]

    n1 <- parameters$n1
    n2 <- parameters$n2
    delta <- parameters$delta
    var1 <- parameters$var1
    alpha <- parameters$alpha

    pooled.var.simulated.accepted <- rep(0, B)

    unequal.var.simulated.accepted <- rep(0, B)

    for (b in 1:B) {
```

```r
        # configuration.key <-
        # paste('n1:',n1,'n2:',n2,'delta:',delta,'var1:',var1,sep = '')

        sample1 <- data.frame(MPG = rnorm(n1, mean = mu1, sd = sqrt(var1)),
            Country = as.numeric(rep(0, times = n1)))

        mu2 <- mu1 + delta

        sample2 <- data.frame(MPG = rnorm(n2, mean = mu2, sd = sqrt(var2)),
            Country = as.numeric(rep(1, times = n2)))

        # Form a data frame - for consistency we
        DF = rbind(sample1, sample2)

        pooled.var <- t.test(x = DF[DF$Country == 0, ]$MPG, y = DF[DF$Country ==
            1, ]$MPG, alternative = "two.sided", var.equal = TRUE, conf.level = alpha)

        if (pooled.var$p.value < alpha)
            pooled.var.simulated.accepted[b] <- FALSE

        if (pooled.var$p.value >= alpha)
            pooled.var.simulated.accepted[b] <- TRUE


        unequal.var <- t.test(x = DF[DF$Country == 0, ]$MPG, y = DF[DF$Country ==
            1, ]$MPG, alternative = "two.sided", var.equal = FALSE, conf.level = alpha)


        if (unequal.var$p.value < alpha)
            unequal.var.simulated.accepted[b] <- FALSE

        if (unequal.var$p.value >= alpha)
            unequal.var.simulated.accepted[b] <- TRUE

    }

    proportion.accpeted.pooled <- sum(pooled.var.simulated.accepted)/length(pooled.var.simulated.accepte

    proportion.accpeted.unequal <- sum(unequal.var.simulated.accepted)/length(unequal.var.simulated.acce

    results.alpha[i, ]$pooled.proportion.accepted <- proportion.accpeted.pooled

    results.alpha[i, ]$unequal.proportion.accepted <- proportion.accpeted.unequal
}

save(results.alpha, file = "results.alpha.Rdata")
```