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

```
# 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")
```

Boxplot Comparison of MPG by country
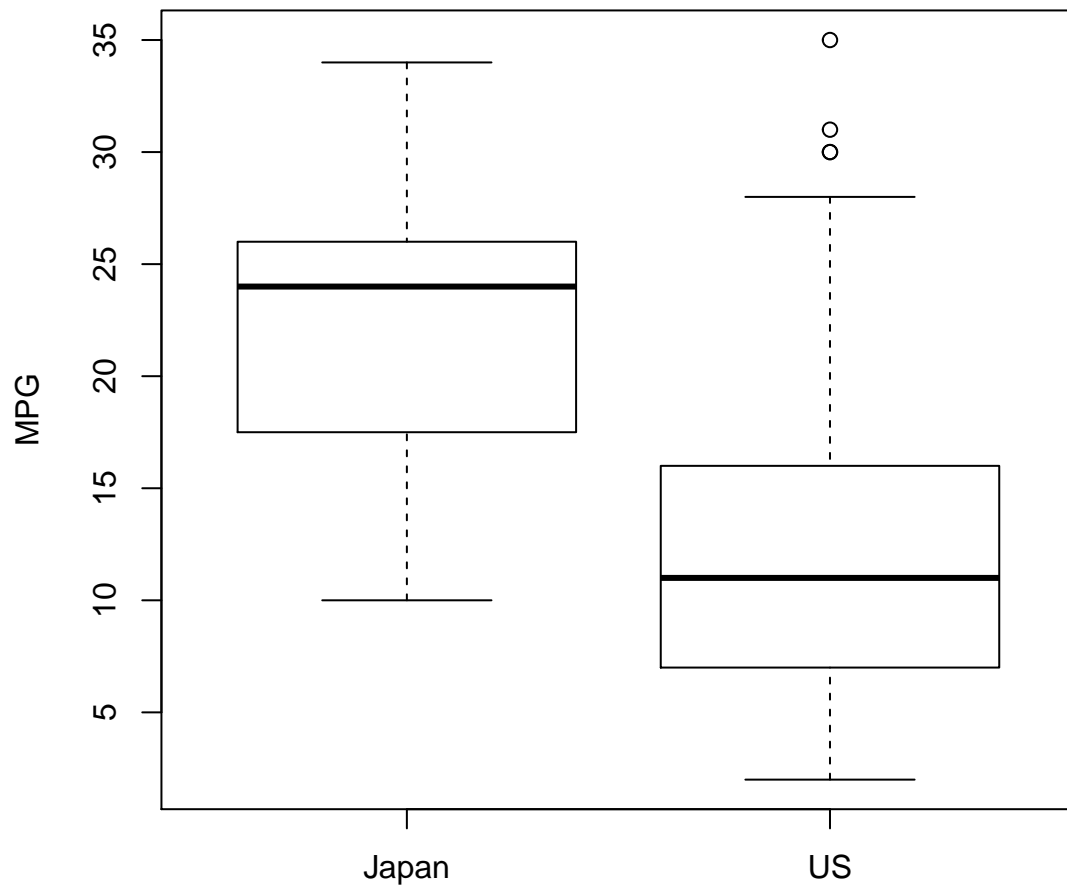
Part 1 - calculation of the confidence intervals

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 we should reject the null hypothesis (means are equal) 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

```
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 We reject the null hypothesis (means are equal) and claim that the evidence supports that the population mean mpg of Japanese and US manufactured cars are different.
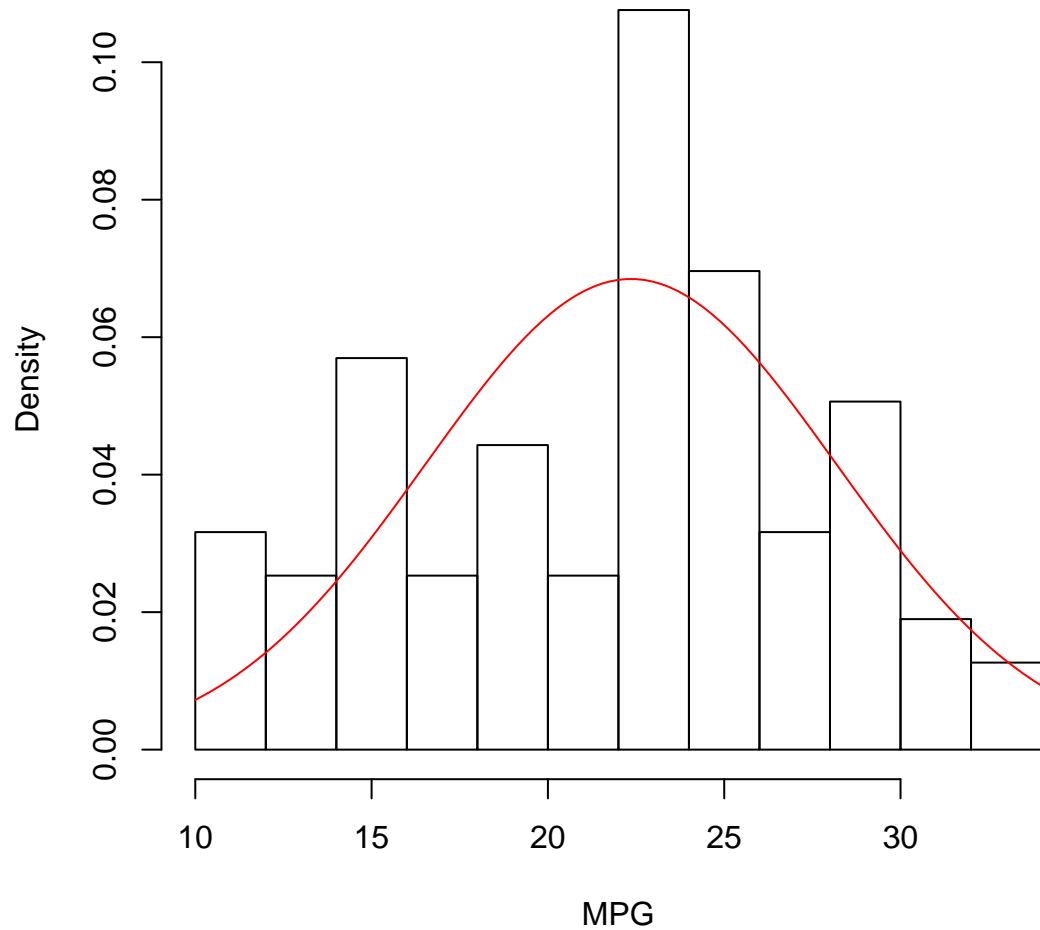

Check for normality

Here we plot the

```
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")
```
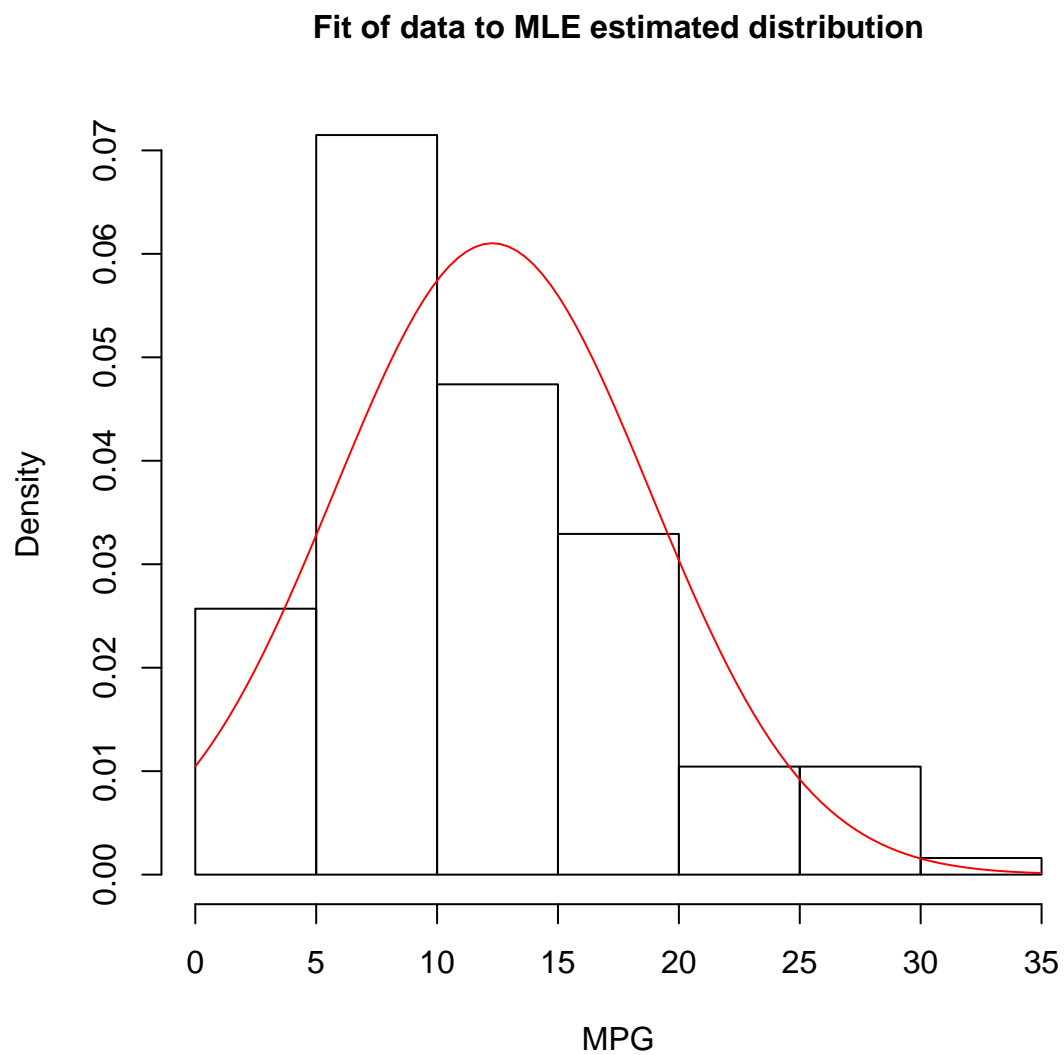
Fit of data to MLE estimated distribution


```
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")
```

Fit of data to MLE estimated distribution

We see the US MPG data is right skewed.

Part 2

Set up the parameters for the simulation study.

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

mu1 <- 0
# Delta = mu1-mu2
deltaVec <- c(-5, -3, -1, -0.5, 0, 0.5, 1, 3, 5)

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

```
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	-3	1	0.025

The first few elements of the parameter space are displayed in the table above. There are 1215 elements of the parameter space.

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

```
# The number of hypothesis test we run for each parameter configuration.
B <- 1000

results.alpha <- parameter.grid

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


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

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

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

for (b in 1:B) {
    # 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.acc

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

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


results.alpha[i, ]$pooled.proportion.rejected <- 1 - proportion.accpeted.pooled

results.alpha[i, ]$unequal.proportion.rejected <- 1 - proportion.accpeted.unequal
```

```
}

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

Function to plot the power for the t tests.

```
library(ggplot2)
library(GGally)

plotTTests <- function(alpha_val) {
    plotList <- list()
    plotColors <- rainbow(length(var1Vec))
    for (i in 1:3) {
        for (j in 1:3) {
            index = (i - 1) * 3 + j

            n1 = n1Vec[i]
            n2 = n2Vec[j]
            # Get the data for this plot
            indexVec <- results.alpha$n1 == n1 & results.alpha$n2 == n2 & results.alpha$alpha ==
                alpha_val

            plot.data <- results.alpha[indexVec, ]

            plotList[[index]] <- ggplot() + geom_line(data = plot.data[plot.data$var1 ==
                var1Vec[1], ], aes(x = delta, y = pooled.proportion.rejected,
                color = "red")) + geom_line(data = plot.data[plot.data$var1 ==
                var1Vec[2], ], aes(x = delta, y = pooled.proportion.rejected,
                color = "green")) + geom_line(data = plot.data[plot.data$var1 ==
                var1Vec[3], ], aes(x = delta, y = pooled.proportion.rejected,
                color = "blue")) + geom_line(data = plot.data[plot.data$var1 ==
                var1Vec[1], ], aes(x = delta, y = unequal.proportion.rejected,
                color = "red"), linetype = "dotted") + geom_line(data = plot.data[plot.data$var1 ==
                var1Vec[2], ], aes(x = delta, y = unequal.proportion.rejected,
                color = "green"), linetype = "dotted") + geom_line(data = plot.data[plot.data$var1 ==
                var1Vec[3], ], aes(x = delta, y = unequal.proportion.rejected,
                color = "blue"), linetype = "dotted") + geom_hline(yintercept = alpha_val,
                linetype = "dashed", color = "black", size = 0.1)

        }
    }
    alpha = as.character(alpha_val)
    title <- paste("Power Curves with alpha = ", alpha)
    # bare minimum of plotList, nrow, and ncolVec
    pm <- ggmatrix(plotList, nrow = 3, ncol = 3, xAxisLabels = n1Vec, yAxisLabels = n2Vec,
        title = title, byrow = FALSE, showStrips = TRUE, xlab = "Mu1-Mu2", ylab = "Power")
    pm <- pm + theme_bw()
    pm

}
```
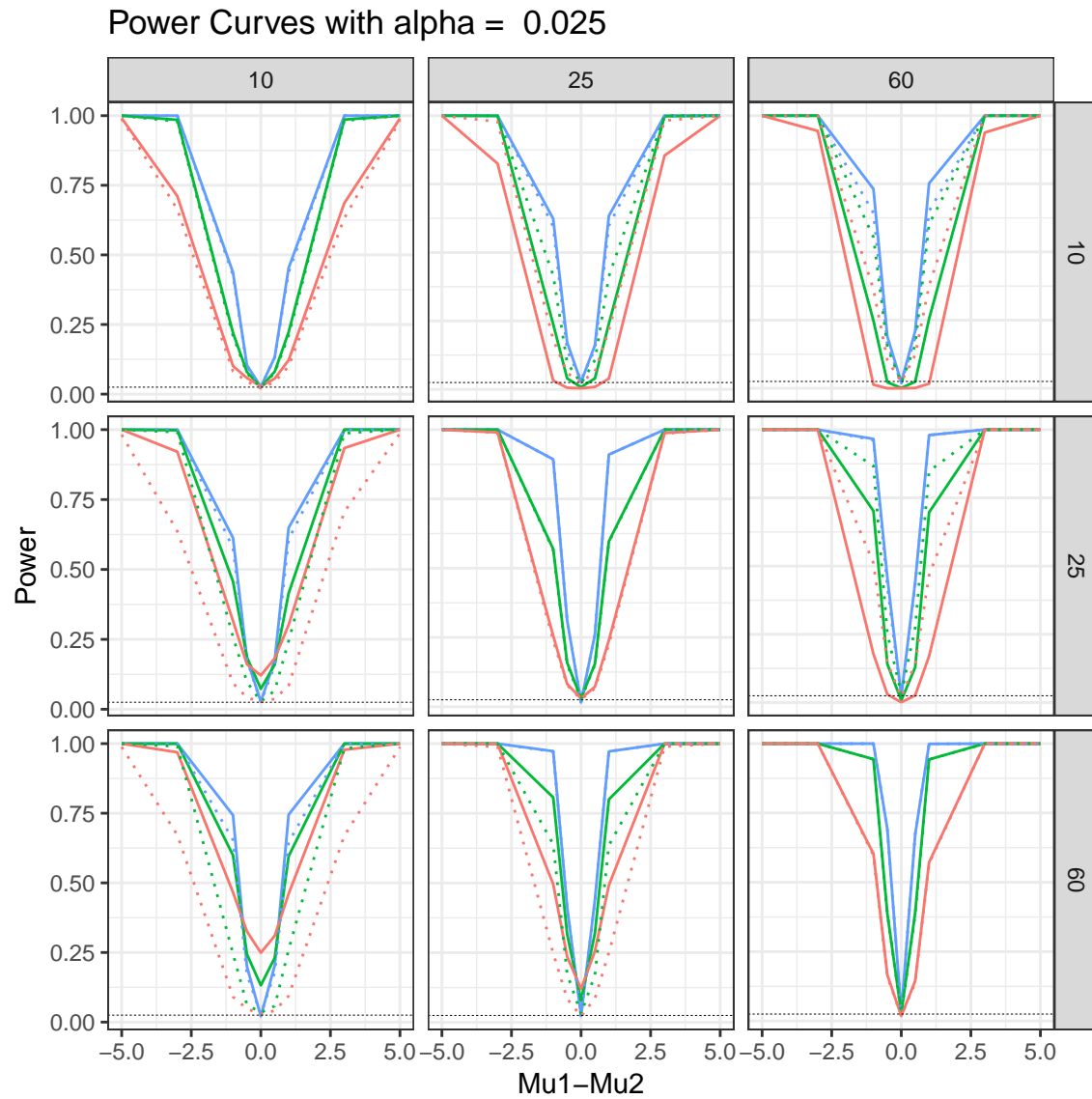
The plots below detail the results of the simulation. $\sigma_1^2 = 1$ is plotted in red, $\sigma_1^2 = 3$ is plotted in green, and

$\sigma_1^2 = 9$ is plotted in blue. The pooled t-tests are rendered in solid lines and the unpooled tests are rendered with dashed lines.

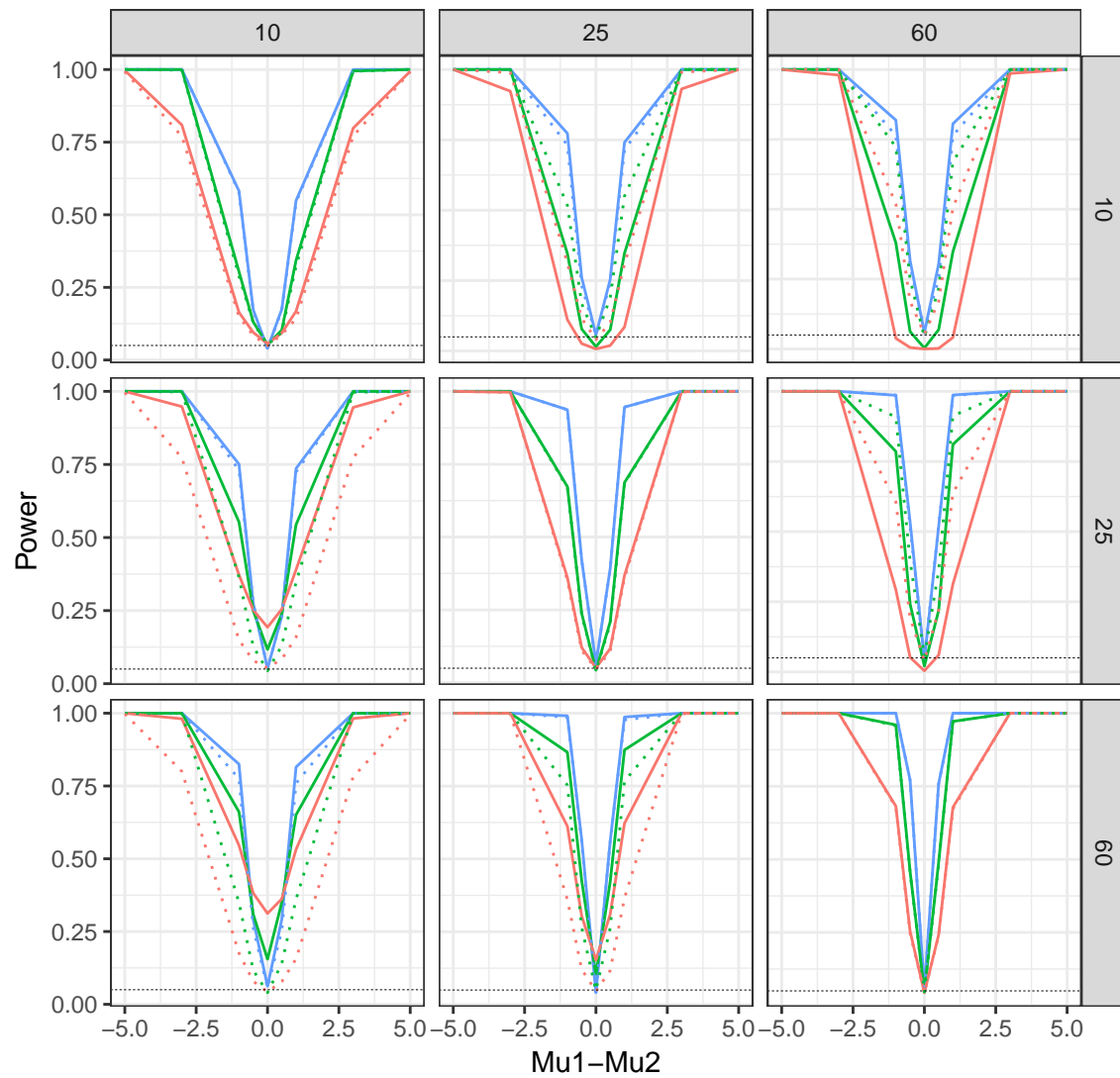## Plot the power for alpha = 0.025

```
plotTTests(0.025)
```


Power Curves with alpha = 0.025

Plot the power for alpha = 0.05
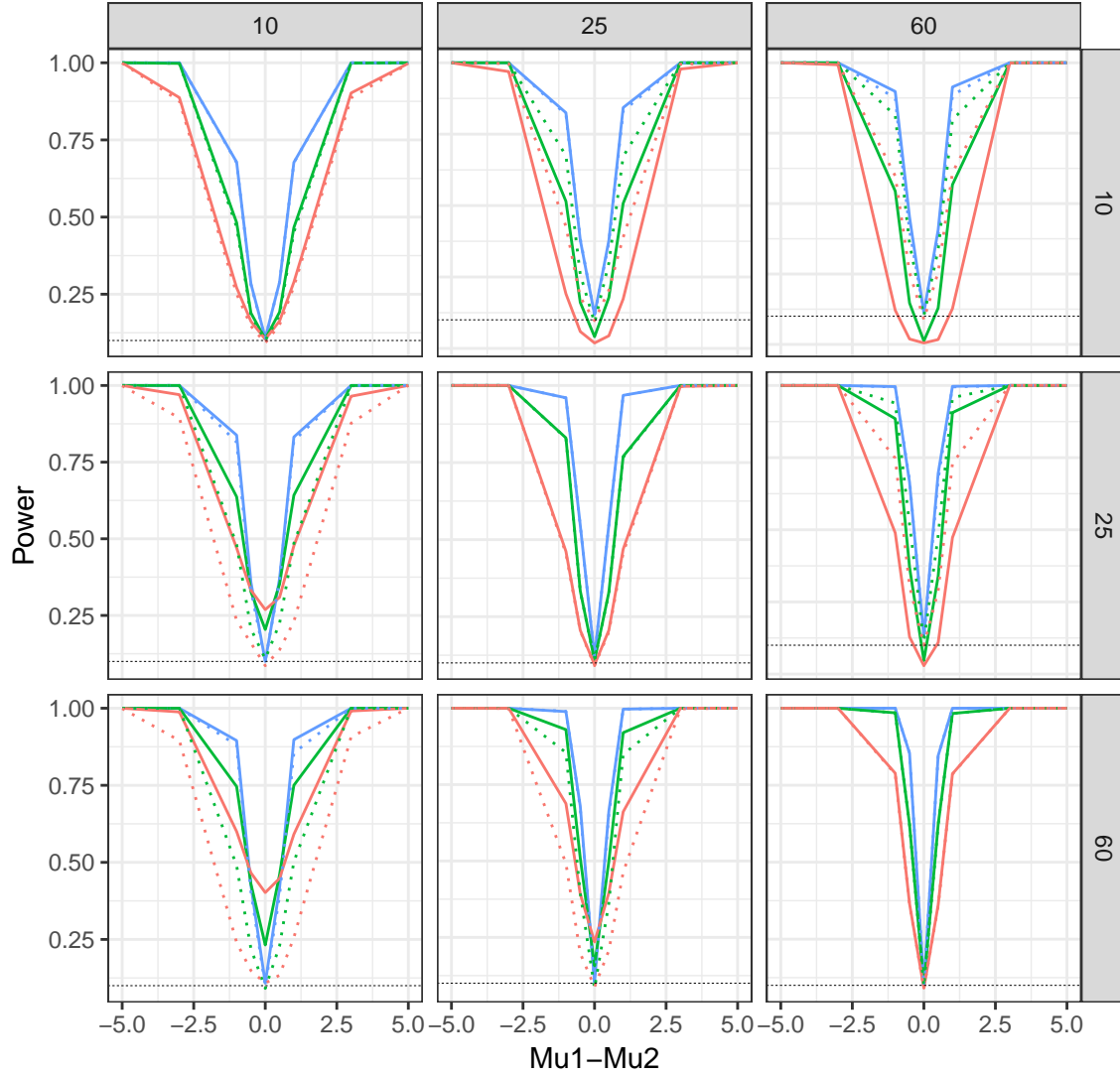
```
plotTTests(0.05)
```

Power Curves with alpha = 0.05

Plot the power for alpha = 0.1

```
plotTTests(0.1)
```

Power Curves with alpha = 0.1

Write Up

The purpose of this project was to compare the effectiveness of two types of statistical test under varying conditions. Specifically, two different type of t-tests were compared. T-tests are used to make inference on the mean between two samples under the assumptions that the data from both samples are random and that the samples follow a normal distribution. The two types of t-tests used in this comparison were the two-sample t-test assuming equal variance (pooled) and the two-sample t-test assuming unequal variances (unpooled). The formulas for these two types of t-test are as follows:

Pooled

$$\frac{\bar{Y}_1 - \bar{Y}_2 - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Where D_0 is the difference between the populations and

$$S_p{}^2 = \frac{(n_1 - 1)S_1{}^2 + (n_2 - 1)S_2{}^2}{n_1 + n_2 - 2}$$

Unpooled

$$\frac{\bar{Y}_1 - \bar{Y}_2 - D_0}{\sqrt{\frac{S_1{}^2}{n_1} + \frac{S_2{}^2}{n_2}}} \sim t_\nu$$

Where v is an estimation of the degrees of freedom given by the Satterthwaite equation.

The pooled t-test is used when the variance is similar between the two samples allowing one term to be used when referencing both variances. The unpooled t-test is used when the variance for one of your samples is much larger than the other, thus necessitating two separate terms to be used when referencing the two sample variances.

Design

To compare these two types of t-test, a simulation study was conducted using R. Two distinct data sets were created by randomly sampling from a normal population using rnorm. The t-test function was used to determine the difference in the mean between the two data sets. For each comparison, a test was performed where the equal.var parameter was set to TRUE to and a separate test was performed where equal.var was set to FALSE to get data for both the pooled and unpooled t-tests respectively. A total of 100 different pairs of randomly generated data sets were created for each simulation. The power of each test, or the probability of rejecting the null hypothesis assuming the alternative is true, was used to determine whether each test should be rejected. The proportion of times we rejected overall for each test was used as a measure for how well each test performed. The power was calculated as 1 minus proportion accepted. This entire process was repeated using varying measures for different parameters. The sample size n1 for the first sample varied between 10, 25, and 60 in different simulations. The sample size n2 for the second sample also varied between 10, 25 and 60. Simulations were conducted where the variance for the first sample ??12 was set to 1, 3 and 9. The variance of the second sample was always set to 1. Delta, the true difference in the mean between the two populations was set at -5, -1, 0, 1, and 5. All combinations of these varying parameters were used in this study. Results for all of these test comparisons were stored using the parameter.grid function. To help visualize the data, plots were generated for all test combinations using ggplot2 and GGally.

Results

The results of the simulation showed that across all cases, the power was relatively low when the true mean difference (mu1-mu2) was close to 0. Power grew towards 1 when the true mean difference deviated from 0 in either direction. Power increased as the sample size for both populations increased. Cases where the true variance for sample 1 were higher resulted in higher power. The difference in power between the two tests grew as the difference in variance between the sample populations grew. When sample 1 had a greater population size than sample 2, the unpooled t-test power was greater than the pooled t-test power. When sample 2 had a great population size than sample 1, the pooled t-test power was greater than the pooled t-test power. For the cases where the sample size for both populations were equivalent, there was almost no difference in the power between the two tests. For cases where there was a larger difference in terms of both sample size and variance, we observed a greater difference in power between the two tests.

Results also showed that the tests appeared to be roughly equivalent in cases with equal sample size. This makes sense when you consider the standard error terms for the pooled and unpooled t-tests when $n1 = n2$ we have that

$$S_p{}^2 = \frac{(n_1 - 1)S_1{}^2 + (n_2 - 1)S_2{}^2}{n_1 + n_2 - 2} = \frac{S_1{}^2 + S_2{}^2}{2}$$

So, we have $\sqrt{\frac{\frac{1}{2}S_1{}^2+S_2{}^2}{n}}$ for pooled t-test and $\sqrt{\frac{S_1{}^2+S_2{}^2}{n}}$ for unpooled t-tests.

Power being lower when the true difference in means is close to zero and growing as the true difference in the means increased makes sense considering the larger the difference in means is the greater probability of the alternative hypothesis being true. It also makes sense that the power increased as sample size increased for both populations as the relationship between greater sample size and higher power has been shown in other studies.

Conclusion

In conclusion, the results give an indication of when it's acceptable to use a pooled t-test. There is very little difference in the power of the tests when the sample sizes are equal. In fact, the results start to vary greatly when the difference in sample size increases. Overall, it can be concluded that a pooled t-test can be used when there are similar sample sizes because this leads to similar standard deviations.