Homework 1

View it online: http://acsweb.ucsd.edu/~djc035/Assignment1.html (http://acsweb.ucsd.edu/~djc035/Assignment1.html)

Investigation

What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?

Analysis 1

Summarize numerically the two distributions of birth weight for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy.

```
#read in "babies.txt" into data
data <- read.table("babies.txt", header=TRUE)
head(data)
```

```
##   bwt gestation parity age height weight smoke
## 1 120       284      0  27     62    100     0
## 2 113       282      0  33     64    135     0
## 3 128       279      0  28     64    115     1
## 4 123       999      0  36     69    190     0
## 5 108       282      0  23     67    125     1
## 6 136       286      0  25     62     93     0
```

Check if unknown value 999 is in bwt column

```
999 %in% data$bwt
```

```
## [1] FALSE
```

Check the unique values in the smoke column

```
unique(data$smoke)
```

```
## [1] 0 1 9
```

Show the number of rows in data

```
nrow(data)
```

```
## [1] 1236
```

Cleaning the data by removing rows with 999 in bwt and 9 in smoke

```
#cleaning unknown data rows
data <- data[!(data$bwt==999 | data$smoke==9),]
```

Again check if unknown value 999 is in bwt column

```
999 %in% data$bwt
```

```
## [1] FALSE
```

Again check the unique values in the smoke column

```
unique(data$smoke)
```

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

Show the number of rows in clean data

```
nrow(data)
```

```
## [1] 1226
```

Make two seperate data version one with rows that are only smokers (smoke==1) and non-smokers (smoke==0)

```
#use which() function to select rows which contain observation of smokers
smoker.ind <- which(data['smoke'] == 1)
#we pass in the vector of indices and use setdiff() function to get the non-smokers
data.smoker <- data[smoker.ind,]
nonsmoker.ind <- which(data['smoke'] == 0)
data.nonsmoker <- data[nonsmoker.ind,]
```

Summaries

Numerical summary of the distribution of birth weight for babies born to women who **smoked** during their pregnancy.

```
summary(data.smoker$bwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    58.0   102.0   115.0   114.1   126.0   163.0
```

```
var(data.smoker$bwt)
```

```
## [1] 327.5718
```

Numerical summary of the distribution of birth weight for babies born to women who **did not smoke** during their pregnancy.

```
summary(data.nonsmoker$bwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      55     113     123     123     134     176
```

```
var(data.nonsmoker$bwt)
```

```
## [1] 302.7144
```

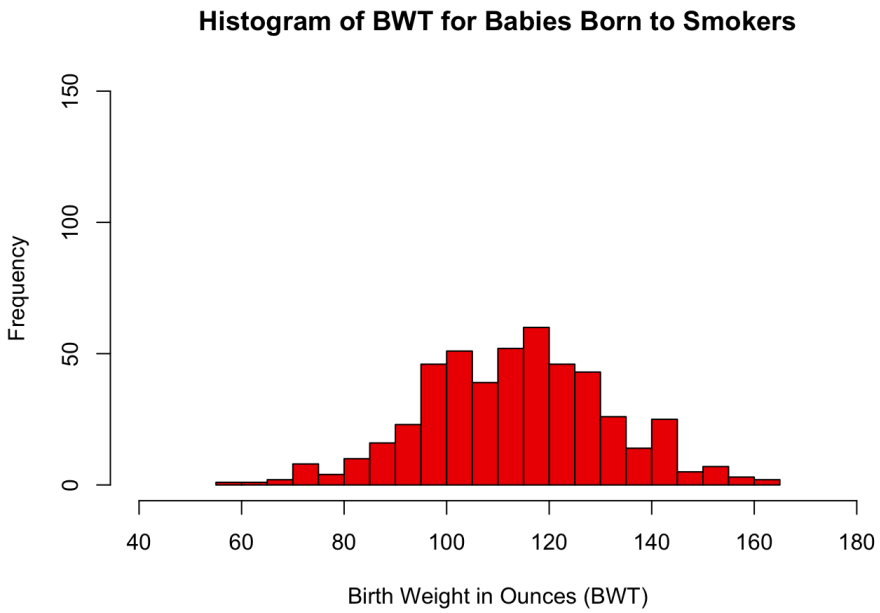
Analysis 2

Use graphical methods to compare the two distributions of birth weight.

Histograms

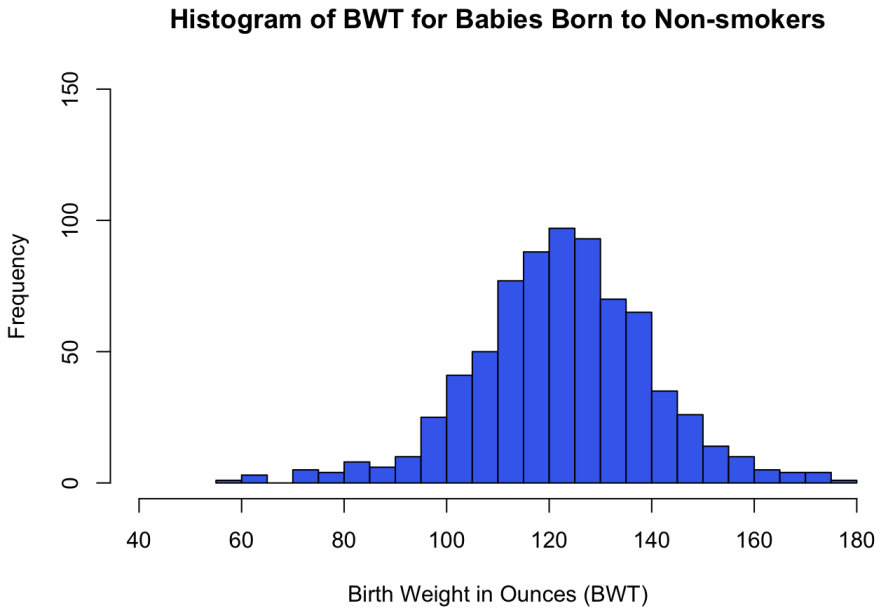
Histogram of distribution of birth weight for babies born to women who **smoked** during their pregnancy.

```
hist(data.smoker$bwt, xlim=c(40,180), ylim=c(0,150), breaks=30, col='red2', main="Histogram of BWT for Babies Bor
n to Smokers", xlab="Birth Weight in Ounces (BWT)")
```



Histogram of distribution of birth weight for babies born to women who **did not smoke** during their pregnancy.

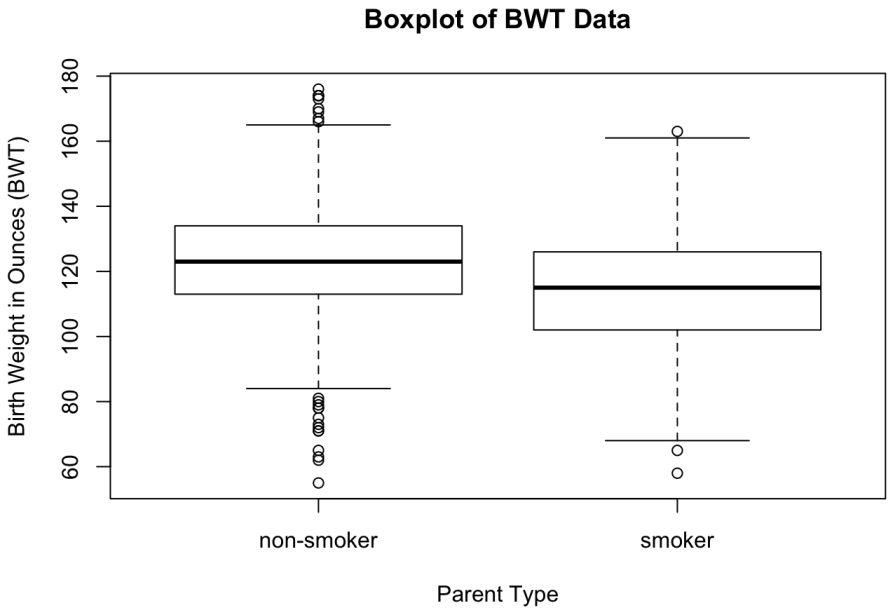
```
hist(data.nonsmoker$bwt, xlim=c(40,180), ylim=c(0,150), breaks=30, col='royalblue2', main="Histogram of BWT for B
abies Born to Non-smokers", xlab="Birth Weight in Ounces (BWT)")
```



Box Plots

Box plots for **both** smoker and non-smoker mothers

```
#change data$smoke labels for the graph
data$smoke[data$smoke == 0] <- "non-smoker"
data$smoke[data$smoke == 1] <- "smoker"
boxplot(bwt~smoke, data, main="Boxplot of BWT Data", xlab="Parent Type", ylab="Birth Weight in Ounces (BWT)", par
s = list(boxwex = 0.8, staplewex = 0.5, outwex = 0.5))
```



Rereading and cleaning the data since the labels were changed to make the box plots.

```
#reread and clean the data
data <- read.table("babies.txt", header=TRUE)
data <- data[!(data$bwt==999 | data$smoke==9),]

#use which() function to select rows which contain observation of smokers
smoker.ind <- which(data['smoke'] == 1)
#we pass in the vector of indices and use setdiff() function to get the non-smokers
data.smoker <- data[smoker.ind,]
nonsmoker.ind <- which(data['smoke'] == 0)
data.nonsmoker <- data[nonsmoker.ind,]
```

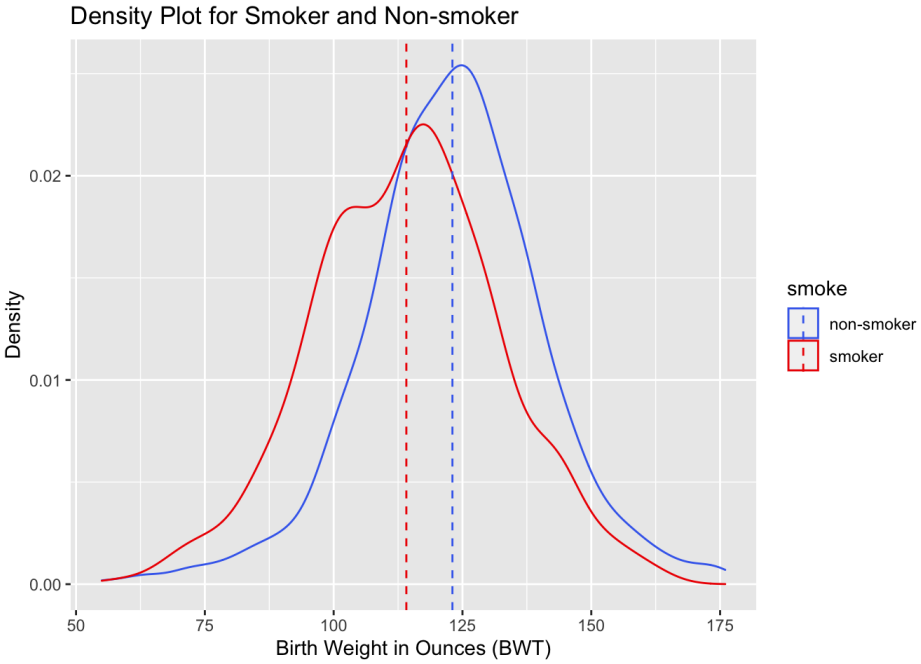
Density Plots

Density plot which includes the distribution for **both** smoker and non-smoker mothers. The dashed line are the means of each distribution.

```
#loading packages
library(ggplot2)
library(plyr)

#TEST code
#p <- ggplot(data, aes(x=bwt)) + geom_density()
#p + geom_vline(aes(xintercept=mean(bwt)), color="blue", linetype="dashed", size=1)
#change line color and fill color
#ggplot(data, aes(x=bwt)) + geom_density(color="darkblue", fill="lightblue")
#change line type
#ggplot(data, aes(x=bwt)) + geom_density(linetype="dashed")

#change data$smoke labels for the graph
data$smoke[data$smoke == 0] <- "non-smoker"
data$smoke[data$smoke == 1] <- "smoker"
mu <- ddply(data, "smoke", summarise, grp.mean=mean(bwt))
head(mu)
```

```
##        smoke grp.mean
## 1 non-smoker 123.0472
## 2     smoker 114.1095
```

```
#change density plot line colors by groups
#ggplot(data, aes(x=bwt, color=smoke)) + geom_density()
#add mean lines
p <- ggplot(data, aes(x=bwt, color=smoke)) + geom_density() + geom_vline(data=mu, aes(xintercept=grp.mean, color=
smoke), linetype="dashed")
p + scale_color_manual(values=c("royalblue2", "red2")) + labs(y="Density", x="Birth Weight in Ounces (BWT)") + gg
title("Density Plot for Smoker and Non-smoker")
```



Rereading and cleaning the data since the labels were changed to make the density plots.

```
#reread and clean the data
data <- read.table("babies.txt", header=TRUE)
data <- data[!(data$bwt==999 | data$smoke==9),]
head(data)
```

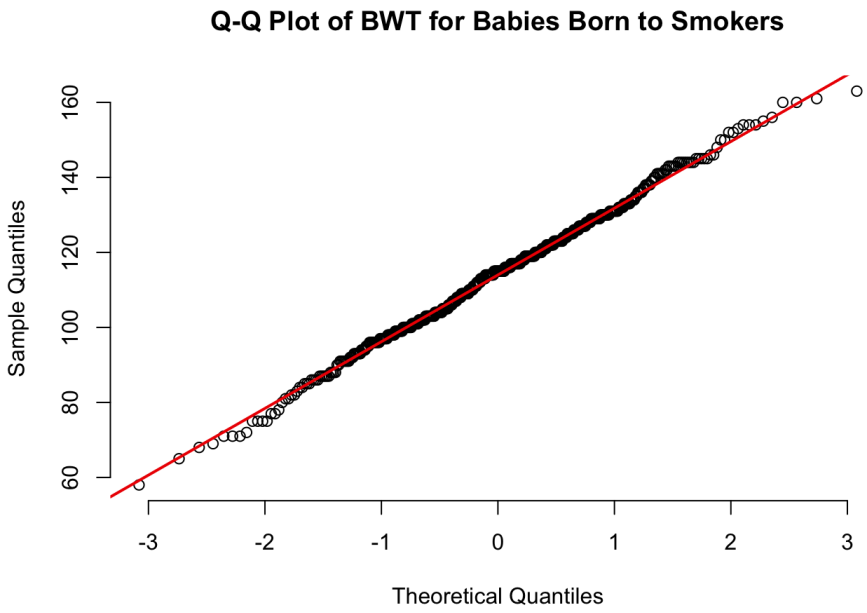
```
##   bwt gestation parity age height weight smoke
## 1 120       284      0  27     62    100     0
## 2 113       282      0  33     64    135     0
## 3 128       279      0  28     64    115     1
## 4 123       999      0  36     69    190     0
## 5 108       282      0  23     67    125     1
## 6 136       286      0  25     62     93     0
```

```
#use which() function to select rows which contain observation of smokers
smoker.ind <- which(data['smoke'] == 1)
#we pass in the vector of indices and use setdiff() function to get the non-smokers
data.smoker <- data[smoker.ind,]
nonsmoker.ind <- which(data['smoke'] == 0)
data.nonsmoker <- data[nonsmoker.ind,]
```

Q-Q Plots

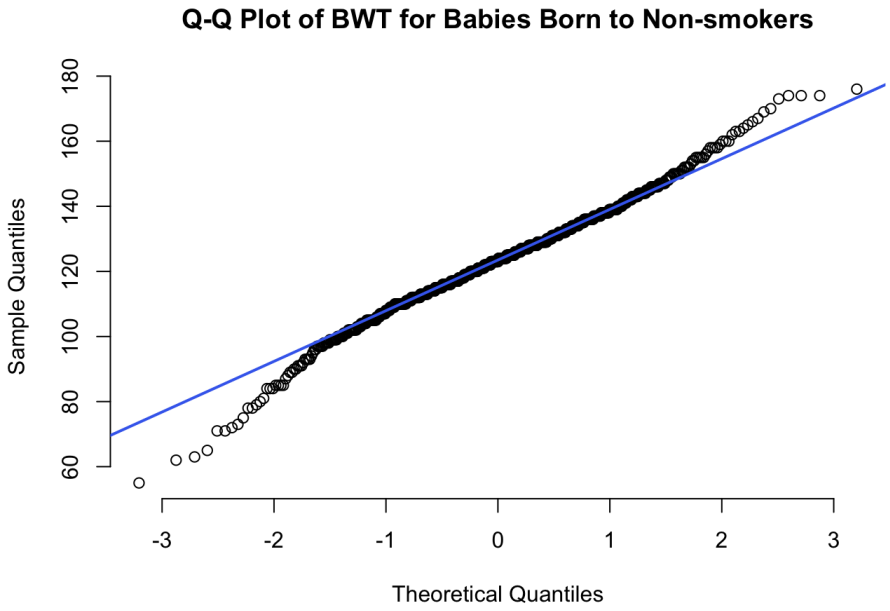
Q-Q Plot of birth weight for babies born to women who **smoked** during their pregnancy.

```
qqnorm(data.smoker$bwt, pch=1, frame=FALSE, main="Q-Q Plot of BWT for Babies Born to Smokers")
qqline(data.smoker$bwt, col="red2", lwd=2)
```



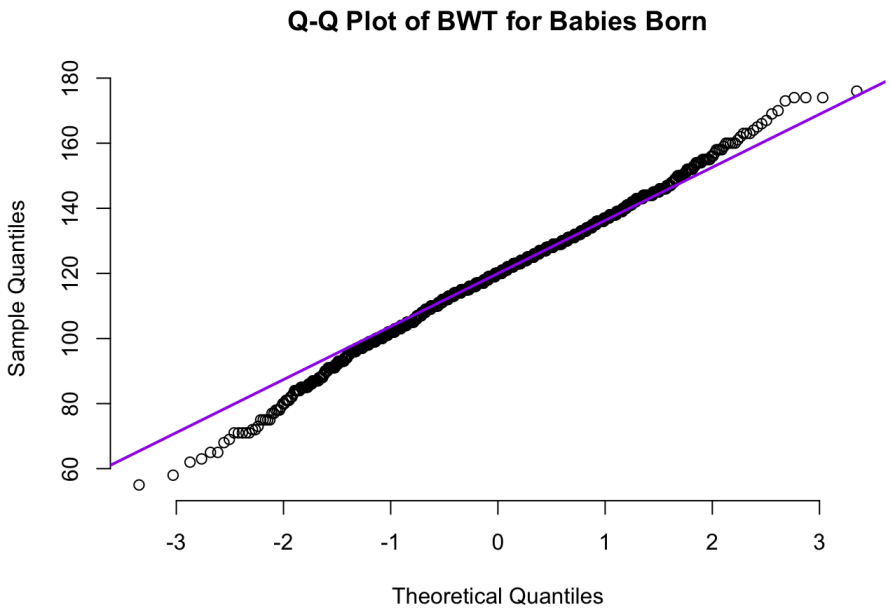
Q-Q Plot of birth weight for babies born to women who **did not smoke** during their pregnancy.

```
qqnorm(data.nonsmoker$bwt, pch=1, frame=FALSE, main="Q-Q Plot of BWT for Babies Born to Non-smokers")
qqline(data.nonsmoker$bwt, col="royalblue2", lwd=2)
```



Q-Q Plot of birth weight for babies born to **all** women.

```
qqnorm(data$bwt, pch=1, frame=FALSE, main="Q-Q Plot of BWT for Babies Born")
qqline(data$bwt, col="purple", lwd=2)
```



Analysis 3

Compare the frequency, or incidence, of low-birth-weight babies for the two groups. How reliable do you think your estimates are? That is, how would the incidence of low birth weight change if a few more or fewer babies were classified as low birth weight?

Simulation Test

Below is our code for the simulation test. We first calculate the observation difference.

```
non_smoker_under <- data[(data$smoke==0 & data$bwt<=88),]
smoker_under <- data[(data$smoke==1 & data$bwt<=88),]
nrow(non_smoker_under)
```

```
## [1] 23
```

```
nrow(smoker_under)
```

```
## [1] 40
```

```
non_smoker<- data[(data$smoke==0),]
smoker <- data[(data$smoke==1),]
nrow(non_smoker)
```

```
## [1] 742
```

```
nrow(smoker)
```

```
## [1] 484
```

```
nrow(non_smoker_under)/nrow(non_smoker)
```

```
## [1] 0.0309973
```

```
nrow(smoker_under)/nrow(smoker)
```

```
## [1] 0.08264463
```

```
smoker.ind <- which(data['smoke'] == 1)
#we pass in the vector of indices and use setdiff() function to get the non-smokers
data.smoker <- data[smoker.ind,]
nonsmoker.ind <- which(data['smoke'] == 0)
data.nonsmoker <- data[nonsmoker.ind,]
nrow(data.smoker)
```

```
## [1] 484
```

Observation Difference

```
obs_diff = (nrow(smoker_under)/nrow(smoker)) - (nrow(non_smoker_under)/nrow(non_smoker))
obs_diff
```

```
## [1] 0.05164732
```

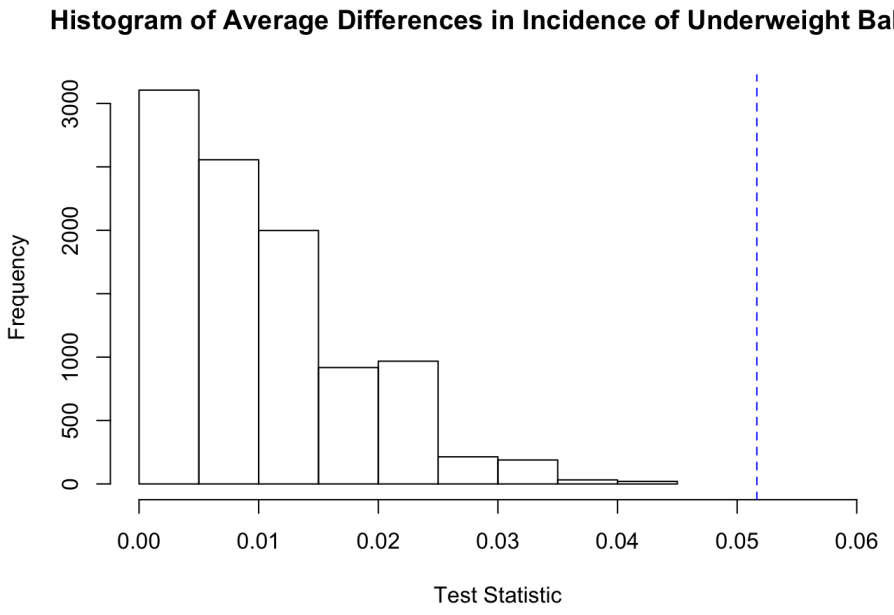
Simulation

Number of time simulated is 10,000.

```
numcases <- 10000

res <- list()

for (i in 1:numcases) {
  test_data <- data.frame(bwt=sample(data$bwt), gestation=data$gestation,  parity=data$parity,  age=data$age,  he
ight=data$height,  weight=data$weight,  smoke=data$smoke)

  non_smoker_under <- test_data[(test_data$smoke==0 & test_data$bwt<=88),]
  smoker_under <- test_data[(test_data$smoke==1 & test_data$bwt<=88),]

  non_smoker<- test_data[(test_data$smoke==0),]
  smoker <- test_data[(test_data$smoke==1),]

  test_stat = (nrow(smoker_under)/nrow(smoker)) - (nrow(non_smoker_under)/nrow(non_smoker))

  res <- c(res, test_stat)
}
df <- do.call(rbind.data.frame, res)
```

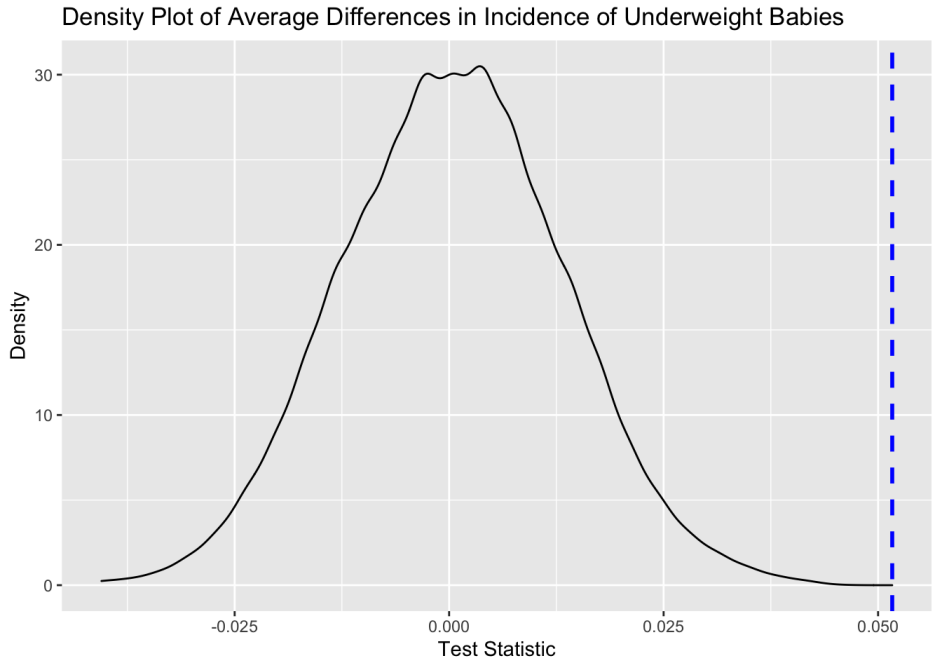
Histogram which shows the distribution of test statistics from the simualtion. Dotted blue line is the observed difference.

```
names(df)[names(df) == colnames(df)[1]] <- "test_stats"
hist(abs(df)$test_stats, xlim=c(0,0.06), main="Histogram of Average Differences in Incidence of Underweight Babie
s", xlab="Test Statistic")
abline(v=obs_diff, col="blue", lty=2)
```



Density plot which shows the distribution of test statistics from the simualtion. Dotted blue line is the observed difference.

```
p <- ggplot(df, aes(x=test_stats)) + geom_density()
p + geom_vline(aes(xintercept=obs_diff), color="blue", linetype="dashed", size=1) + labs(y="Density", x="Test Sta
tistic") + ggtitle("Density Plot of Average Differences in Incidence of Underweight Babies")
```



P-value Test

```
p_value <- 0.01
df <- df[(df$test_stats >= obs_diff),]
df
```

```
## numeric(0)
```

```
nrow(df)
```

```
## NULL
```

```
p_test <- nrow(df)/10000
p_test
```

```
## numeric(0)
```

We see that from our simulation test there is never a test difference as extreme as our observed difference. Which means our observed difference is not due to random chance.