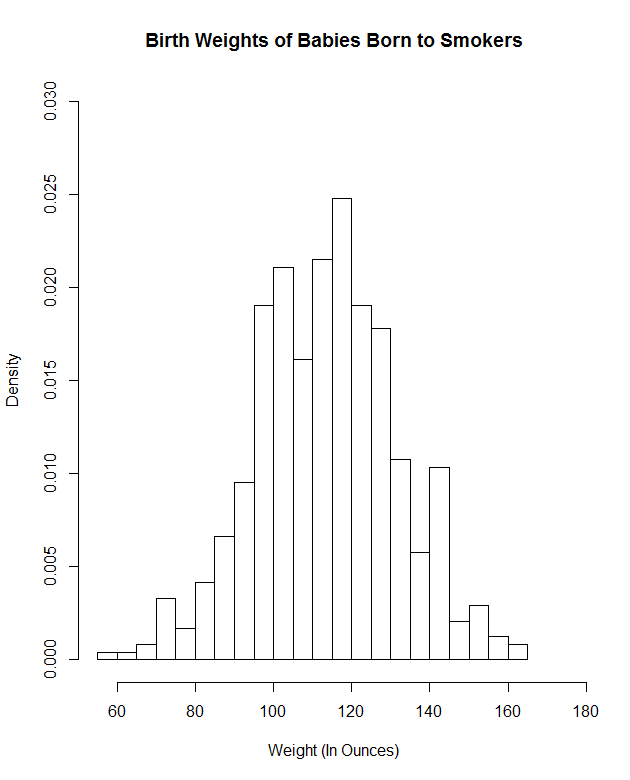
Stat 135 Homework 1

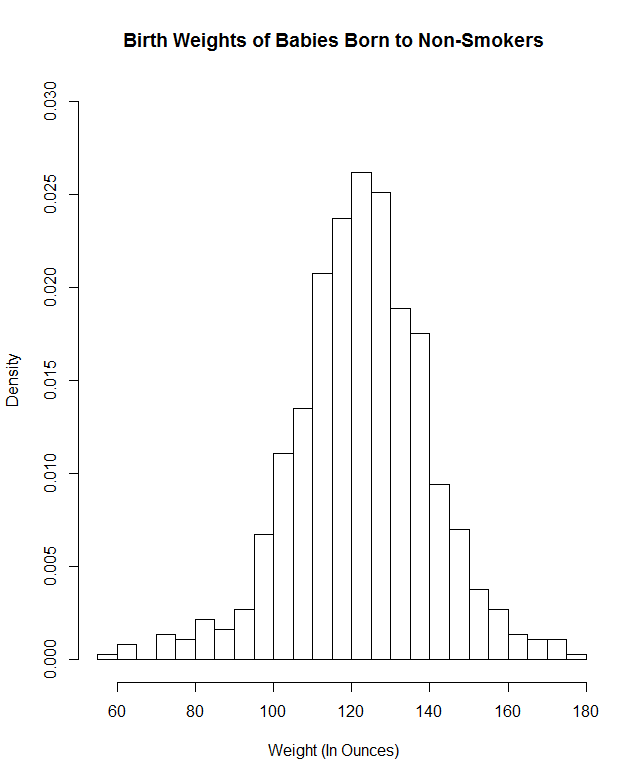
1. The output of the R code is as follows:
   1. Summary of Babies born to Smokers:

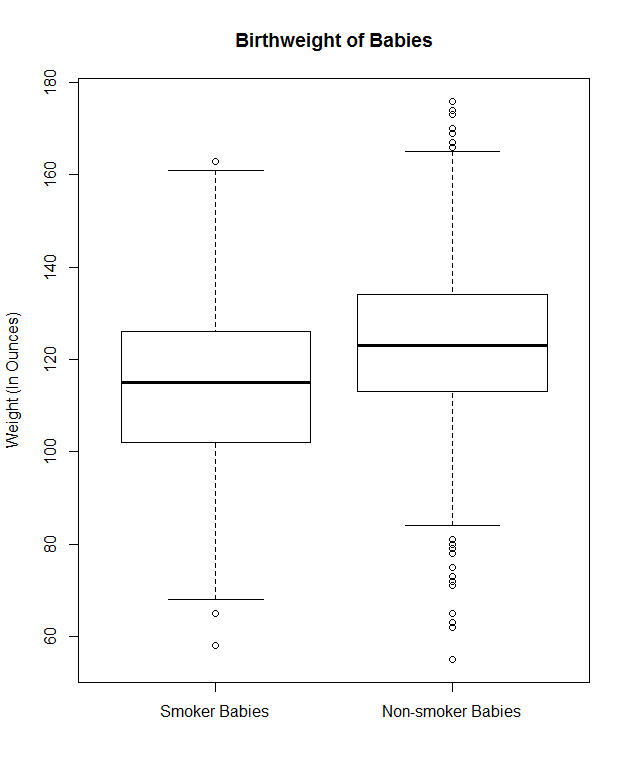
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max | Standard Dev. | IQR |
| 58.0 | 102.0 | 115.0 | 114.1 | 126.0 | 163.0 | 18.09895 | 24 |

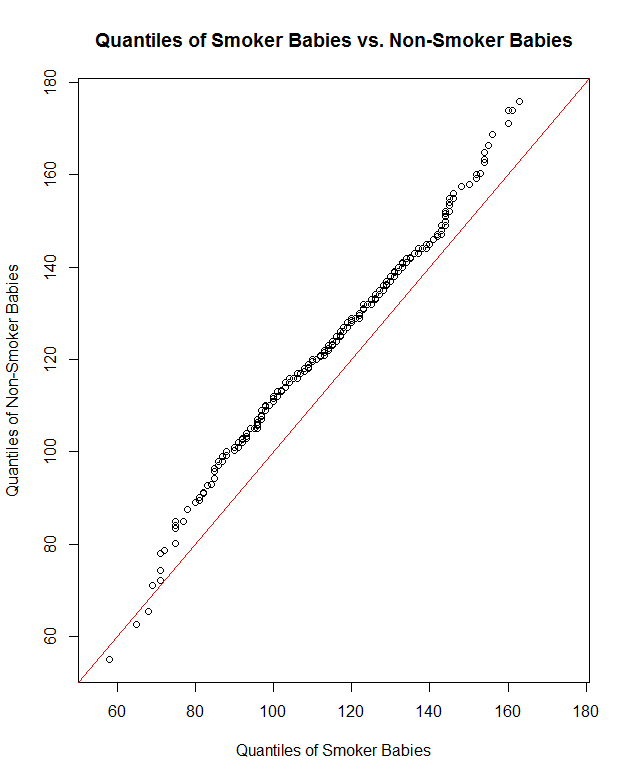
Summary of Babies born to Non-Smokers:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max | Standard Dev. | IQR |
| 55 | 113 | 123 | 123 | 134 | 176 | 17.39869 | 21 |

* 1. Graphs to compare both distributions:







* 1. One important thing to note is that there are 742 non-smoker data points versus 484 smoker points, which makes the smoker data relatively unreliable. This is reflected in the standard deviation, which is higher for the smoker data. Therefore, slight caution should be used as these two datasets are compared.

The summaries of both distributions give us the first clue that the Surgeon General’s warning is correct – smoking by pregnant women may result in low birthweight. In particular, we see that the minimum, maximum, quartiles and mean weights of the non-smokers group is higher than the smokers group, possibly indicating that smoking may reduce birthweight.

The box plots of both distributions cement this sentiment, as we can clearly see that the non-smoker box plot is shifted higher than the smoker group’s plot. It is interesting to note that the smoker group has fewer ‘outliers’ than the non-smoker group, which is likely due to fewer data points and a looser IQR.

Lastly, the Q-Q Plot of both groups’ quantiles furthers the idea that the non-smoker quantiles are always higher than the smoker quantiles, except for the first few quantiles. This implies that smoking during pregnancy shifts the birthweight down consistently, however this cannot be said for certain due to the different sample sizes.

Overall, the differences in the distributions indicate a strong correlation between smoking and lower birthweights, especially the Q-Q plot which emphasizes how the non-smoker quantiles are almost always higher. Despite the fact that both datasets have different sizes, there is clearly some kind of underlying effect caused by smoking, which could be confirmed by another analysis with a greater subset of data.

Appendix – R Code:

setwd("C:/Users/Naval/Dropbox/Berkeley Documents/Fall 2015/Stat 135")

# Imports data from Source Table into the 'babies' variable

babies <- read.table("babiesI.data", header=TRUE)

# creates a vector with birthweights of babies born to mothers who smoke

bwt.smokers <- babies$bwt[babies$smoke == 1]

# creates a vector with birthweights of babies born to mothers who don't smoke

bwt.nonsmokers <- babies$bwt[babies$smoke == 0]

#Part A: This following block of code summarises the two distributions.

summary(bwt.smokers)

sd(bwt.smokers)

IQR(bwt.smokers)

summary(bwt.nonsmokers)

sd(bwt.nonsmokers)

IQR(bwt.nonsmokers)

#Part B: This following block of code creates graphs to compare both distributions.

hist(bwt.smokers, freq = FALSE, breaks = 20, xlim = range (bwt.smokers, bwt.nonsmokers), ylim = c(0,0.03), xlab = "Weight (In Ounces)", main = "Birth Weights of Babies Born to Smokers")

hist(bwt.nonsmokers, freq = FALSE, breaks = 20, xlim = range (bwt.smokers, bwt.nonsmokers), ylim = c(0,0.03), xlab = "Weight (In Ounces)", main = "Birth Weights of Babies Born to Non-Smokers")

boxplot(bwt.smokers, bwt.nonsmokers, names = c("Smoker Babies", "Non-smoker Babies"), ylab = "Weight (In Ounces)", main = "Birthweight of Babies")

qqplot(bwt.smokers, bwt.nonsmokers, xlim = range (bwt.smokers, bwt.nonsmokers), ylim = range (bwt.smokers, bwt.nonsmokers), xlab = "Quantiles of Smoker Babies", ylab = "Quantiles of Non-Smoker Babies", main = "Quantiles of Smoker Babies vs. Non-Smoker Babies")

abline(a=0, b=1, col = "red")