MSDS 5043 - Assignment 2

Olivia Samples, Lipscomb University

10/2/2019

The LBWunicef Data for Questions 1-5

The data at a UNICEF data site, that described the percentage of low birth weight (less than 2,500 grams) infants for a number of nations (actually, counties and territories, but we'll refer to them as nations here) around the world. LBWunicef.csv data set is built based on this which includes the following elements.

```
-nation = the nation's name

-lbw.pct = the nation's low birth weight percentage

-least.dev = whether or not the nation is regarded by the United Nations Population Division as one
```

The LBWunicef.csv file is part of the general Data and Code materials. Import the LBWunicef.csv file into R Studio, answer questions 1-5.

```
lbw = read.csv("/Users/osamples/Desktop/MSDS5043/LBWunicef.csv", head = TRUE, sep = ",")
head(lbw)
```

```
##                nation lbw.pct least.dev
## 1             Albania       4         0
## 2             Algeria       6         0
## 3              Angola      12         1
## 4 Antigua and Barbuda       5         0
## 5           Argentina       7         0
## 6             Armenia       8         0
```

```
summary(lbw)
```

```
##                  nation       lbw.pct        least.dev
##  Albania            :  1   Min.   : 0.00   Min.   :0.00
##  Algeria            :  1   1st Qu.: 6.00   1st Qu.:0.00
##  Angola             :  1   Median : 9.00   Median :0.00
##  Antigua and Barbuda:  1   Mean   :10.08   Mean   :0.25
##  Argentina          :  1   3rd Qu.:12.00   3rd Qu.:0.25
##  Armenia            :  1   Max.   :35.00   Max.   :1.00
##  (Other)            :174
```

Question 1

How many nations have non-missing low birth weight percentage estimates?

```
#Number of missing values for Low Birth Weight Percentage Estimates
sum(is.na(lbw$lbw.pct))
```

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

```
#Number of missing values for all columns
sapply(lbw, function(x) sum(is.na(lbw$lbw.pct)))
```

```
##    nation   lbw.pct least.dev
##         0         0         0
```

```
#Number of total entries
nrow(lbw)
```

```
## [1] 180
```

```
#Check and make sure there are no duplicates
length(unique(lbw$nation)) == nrow(lbw)
```

```
## [1] TRUE
```

We have that there are no missing values for any column, which implies the total number of nations have non-missing low birth weight percentage estimates.

Question 2

Which nations have the three largest low birth weight percentages? Are each of these considered by the UN to be "least developed" nations or not?

```
head(lbw[rev(order(lbw$lbw.pct)),], 3)
```

```
##         nation lbw.pct least.dev
## 103 Mauritania      35         1
## 122   Pakistan      32         0
## 73       India      28         0
```

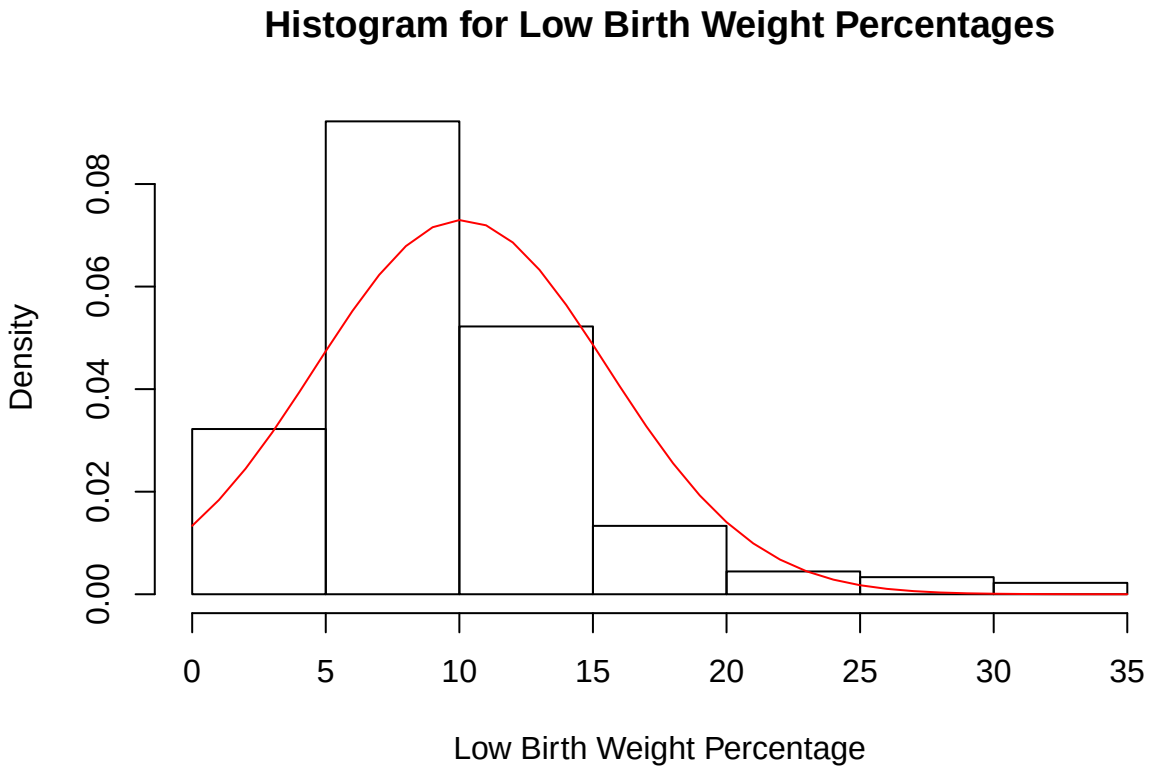
We can see that the nations with the three largest low birth weight percentages are Mauritania, Pakistan, and India. Of these, Mauritania is the only nation considered by the UN to classify as a "least developed" nation.

Question 3

Create a histogram of the low birth weight percentages, then superimpose a normal density function with the same mean and standard deviation in red. Based on your plot, is the standard deviation or the inter-quartile range a more appropriate measure of variation in the low birth weight rates? Why?

```
pctmean <- mean(lbw$lbw.pct)
pctsd   <- sd(lbw$lbw.pct)

hist(lbw$lbw.pct, probability = TRUE, main="Histogram for Low Birth Weight Percentages", xlab="Low Birt
x <- 0:35
y <- dnorm(x = x, mean = pctmean, sd = pctsd)
lines(x = x, y = y, col = "red")
```



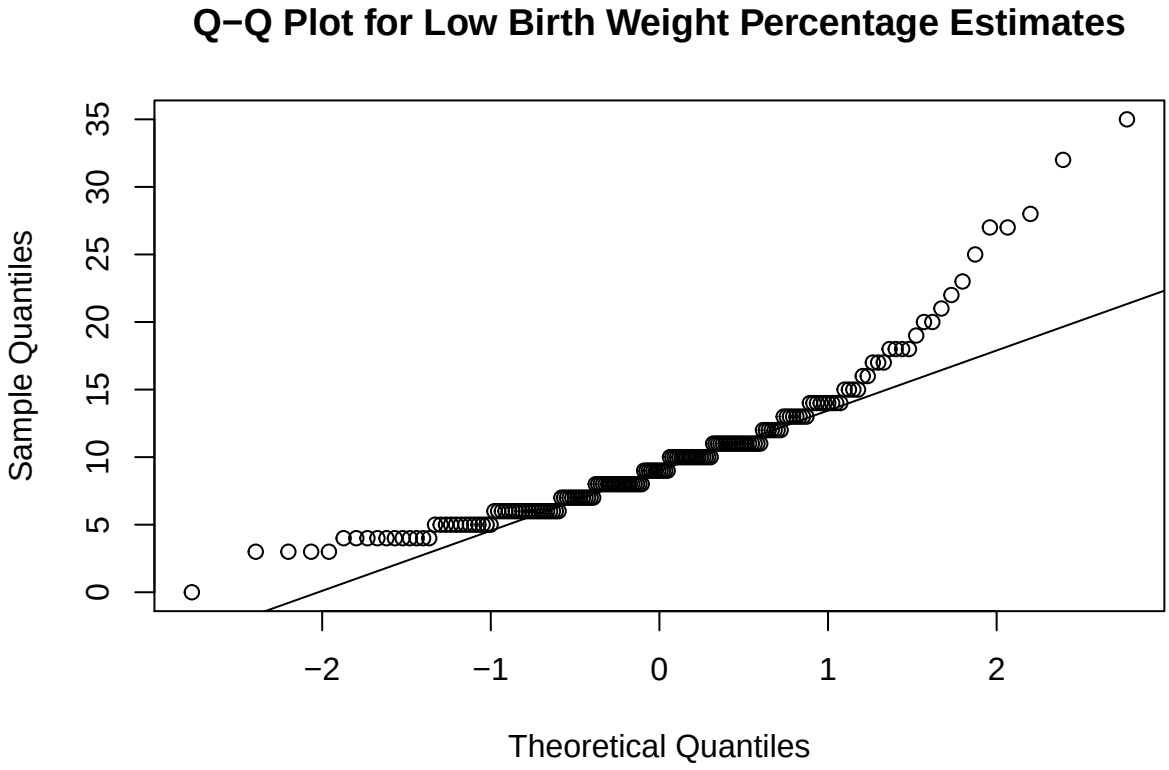
We have that the standard deviation represents a better measure of variation within this plot because it provides a more thorough look into the data. While the inter-quartile range only looks at the Q3-Q1 difference, the standard deviation can describe every point on the historgram in relation to the mean.

Question 4

Create a normal Q-Q plot for the low birth weight percentage estimates. Would you say that the data are approximately Normally distributed, or not appproximately Normally distributed. Justify your answer by interpreting what you see in your plot, and whatever summary statistics you deem to be useful in making your decision.

```
library("ggplot2")
```

```
qqnorm(lbw$lbw.pct, main='Q-Q Plot for Low Birth Weight Percentage Estimates')
qqline(lbw$lbw.pct)
```



Based on the Q-Q Plot for Low Birth Weight Percentage Estimates, it seems that the data are not approximately normally distributed. When data are normally distributed, they lie on the straight line, but we have an obvious curve which suggests otherwise.

```
mean(lbw$lbw.pct)
```

```
## [1] 10.07778
```

```
median(lbw$lbw.pct)
```

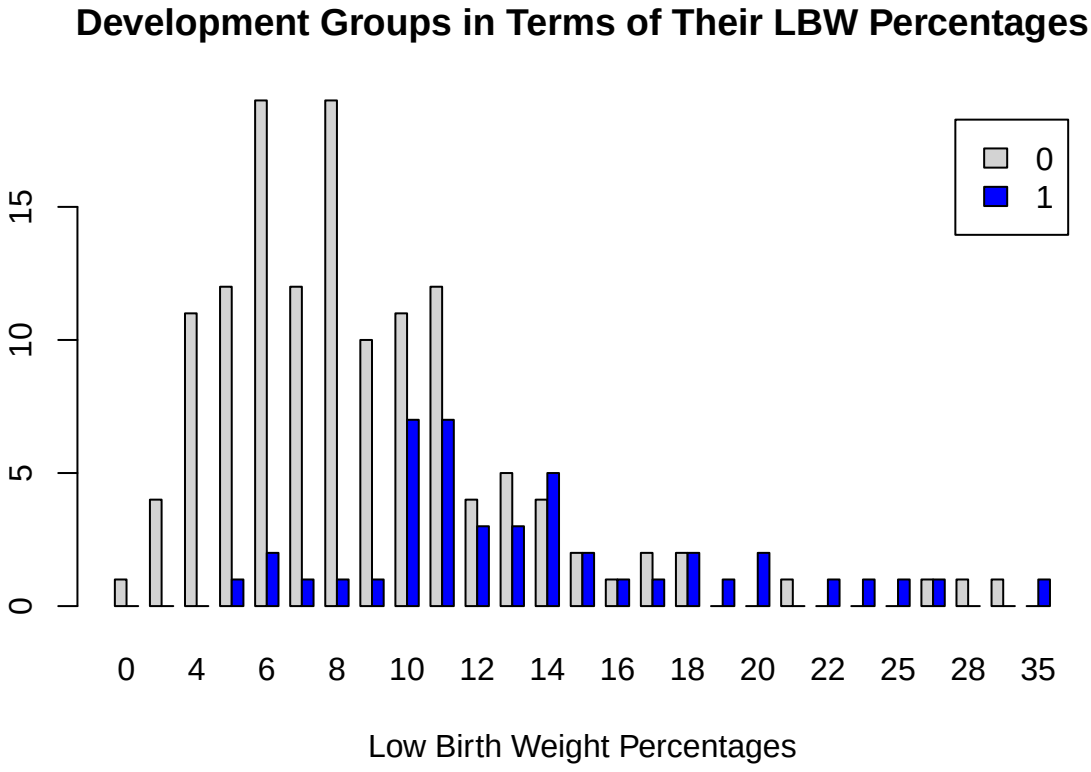
```
## [1] 9
```

We also have that the mean is greater than the median, which generally implies that the data is skewed right. This coincides with the histogram in Question 3.

Question 5

Display an effective graph comparing the two development groups (least developed nations vs. all other nations) in terms of their percentages of low birth weight births. What conclusions can you draw about the distribution of low birth weight rates across the two development groups? Be sure to label your graph so it stands alone, and also supplement your graph with separate text discussing your conclusions.

```
co.table <- table(lbw$least.dev, lbw$lbw.pct)
barplot(co.table, main="Development Groups in Terms of Their LBW Percentages",
  xlab = "Low Birth Weight Percentages", col=c("light grey","blue"),
  legend =  rownames(co.table), beside = TRUE)
```

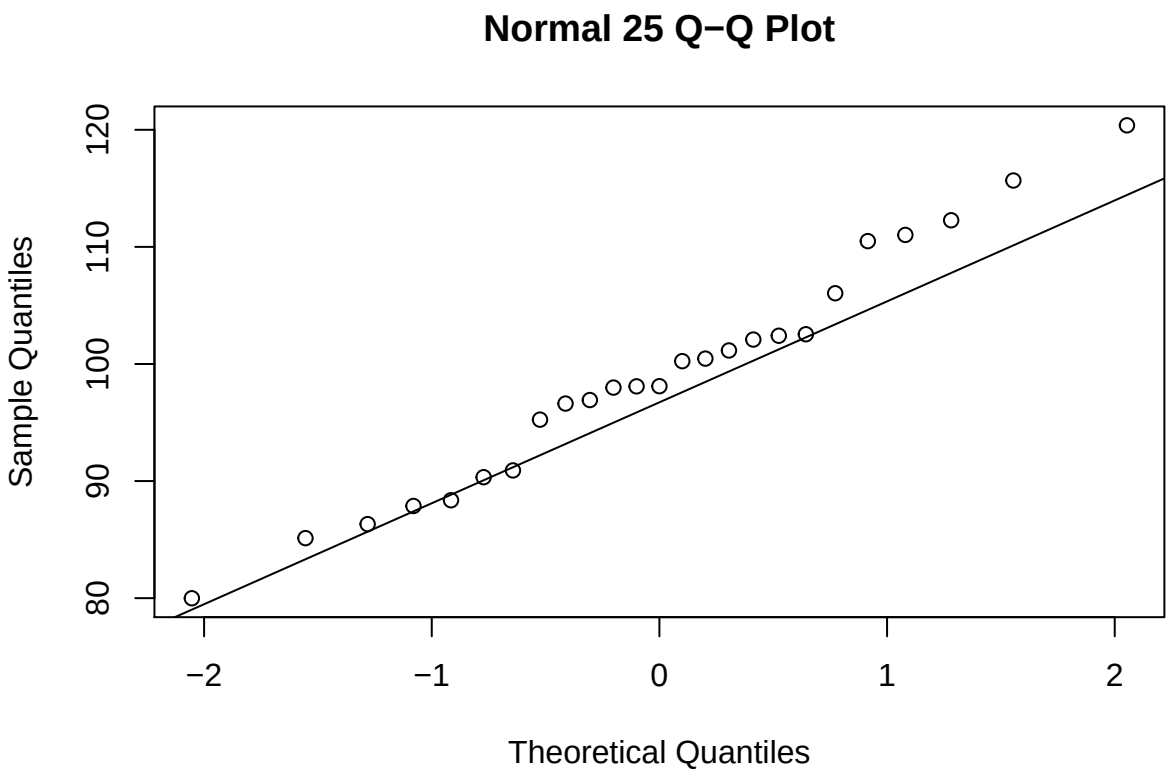


In this graph, we can see that the least developed countries (in blue) are largely represented by the higher percentage of babies born (10-35%) with a low birth weight on the right tail of the histogram. Comparitively, most of the developed countries have lower percentages, within the 4-11% range. And so the data coincides with the idea that babies born in least developed countries have a larger chance of having a low birth weight.

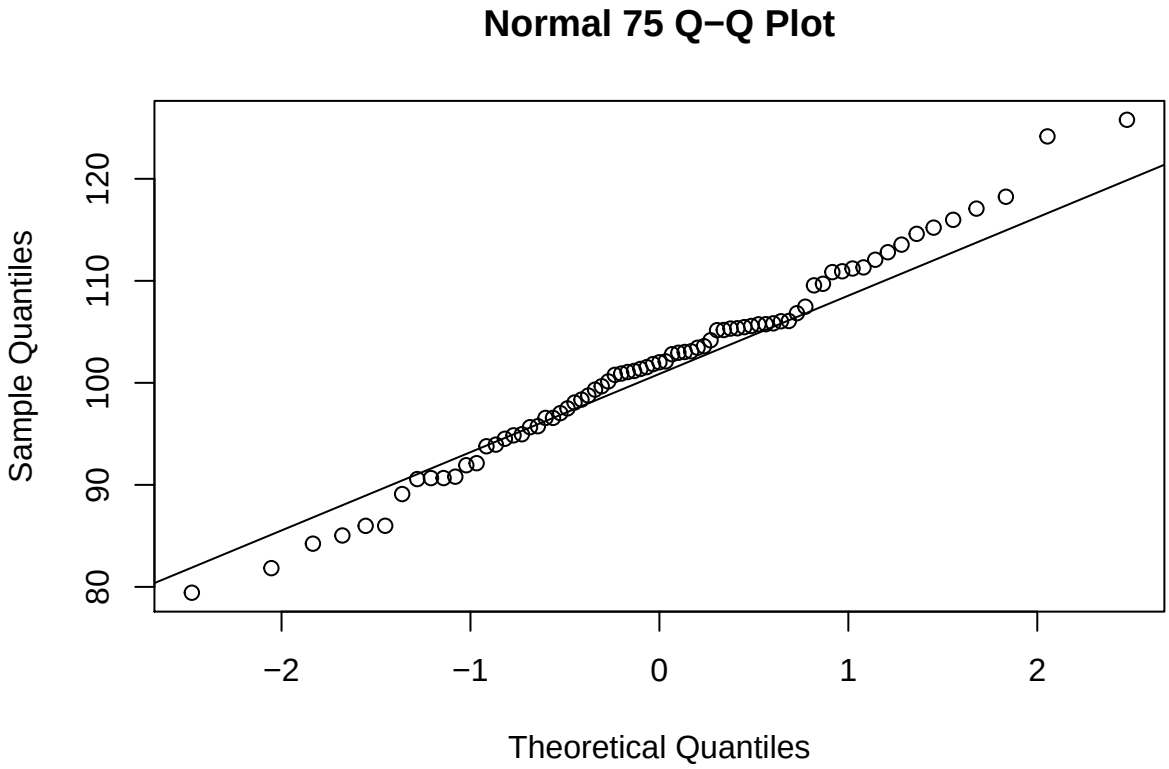
Question 6 (Optional)

Generate a "random" samples of 75 observations from a Normal distrubtion with mean 100 and standard deviation 10 using R. The rnorm function is likely to be helpful. Now, display a normal Q-Q plot of these data, using ggplot2 package from the tidyverse. How well does the Q-Q plot approximate a straight line? Repeat this task for a second sample of 150 Normally distributed observations, again with a mean of 100 and a standard deviation of 10. Then repeat again for samples of 25 to 225 Normally distributed observations with a different mean and variance. Which of the four Q-Q plots you have developed better approximates a straight line and what should we expect the relationship of sample size with this phenomenon to be?

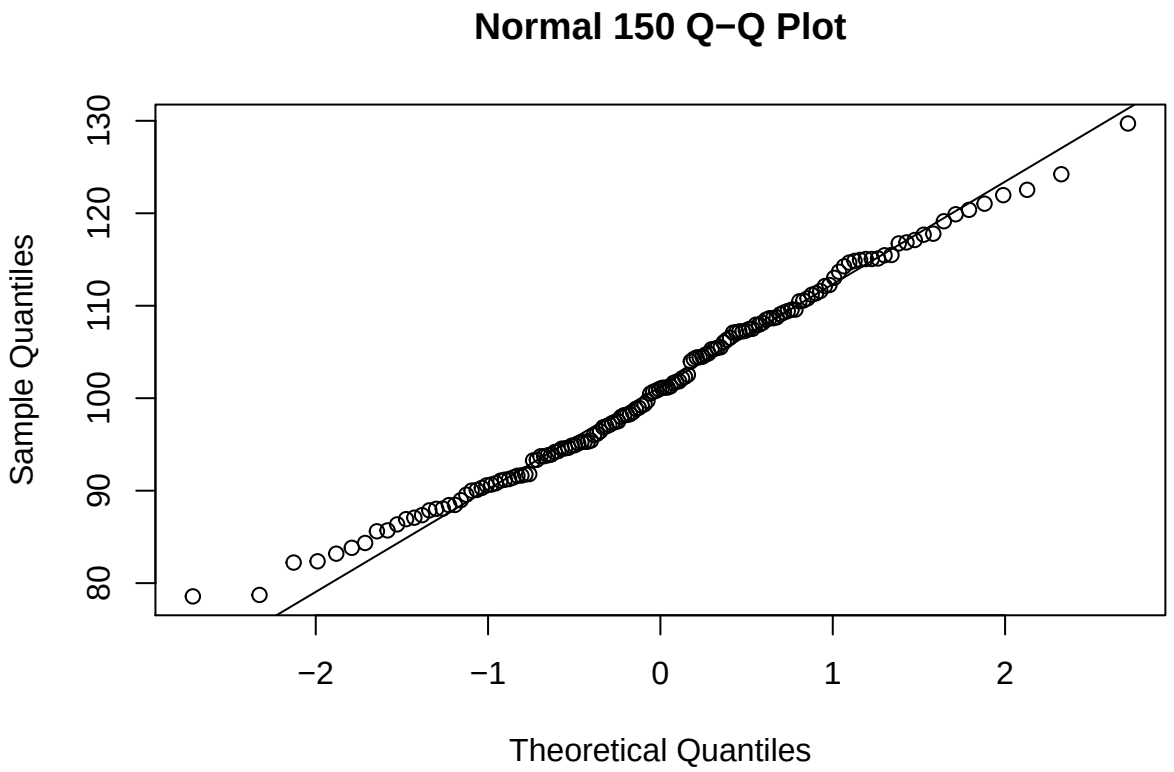
```
norm25 <- rnorm(25, 100, 10)
qqnorm(norm25, main = "Normal 25 Q-Q Plot")
qqline(norm25)
```



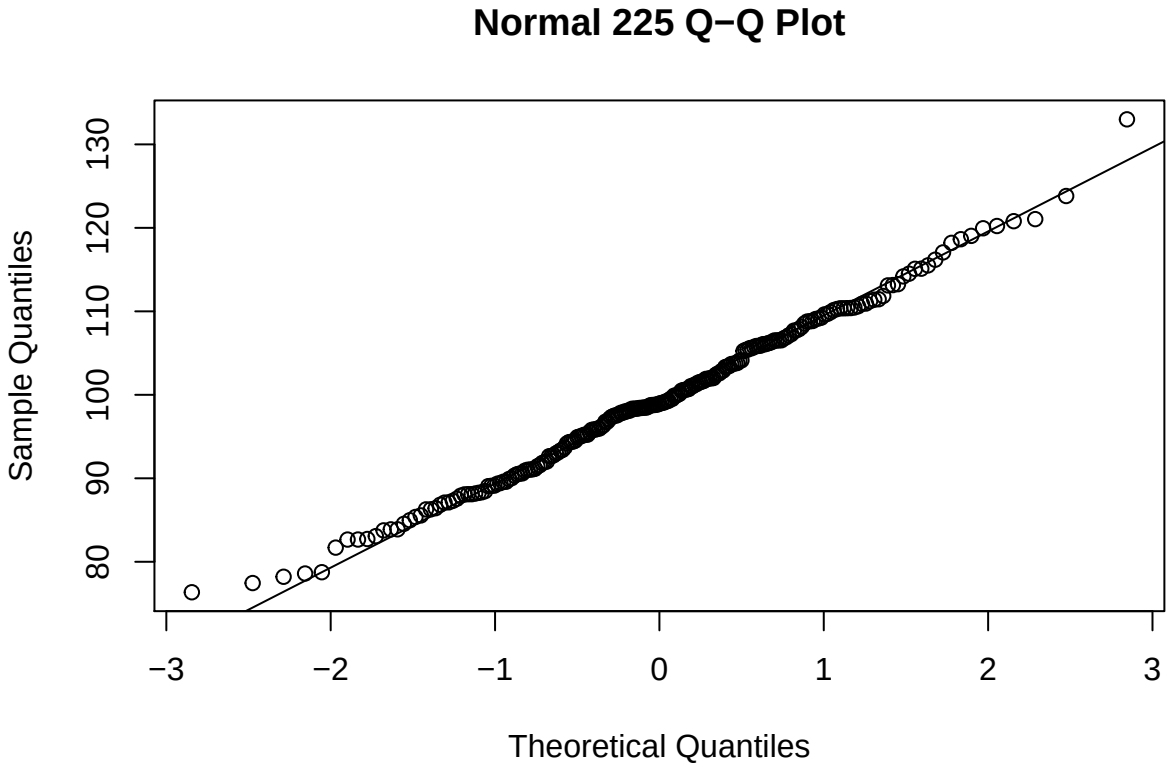

```r
norm75 <- rnorm(75, 100, 10)
qqnorm(norm75, main = "Normal 75 Q-Q Plot")
qqline(norm75)
```



```r
norm150 <- rnorm(150, 100, 10)
qqnorm(norm150, main = "Normal 150 Q-Q Plot")
qqline(norm150)
```




```r
norm225 <- rnorm(225, 100, 10)
qqnorm(norm225, main = "Normal 225 Q-Q Plot")
qqline(norm225)
```



It seems that the approximation improves as the sample size increases. For instance, the Normal 25 Q-Q Plot has more outliers and is two tailed, while the Normal 225 Q-Q plot is concentrated on the normal line with few outliers. As the sample size increases, the Q-Q plot gets closer to a normal distribution.