SDA Group Submission Assignment Assign1

Group Gr18

MengliFeng (2720589) and PepijnVanOostveen (2801582)

Exercise 1

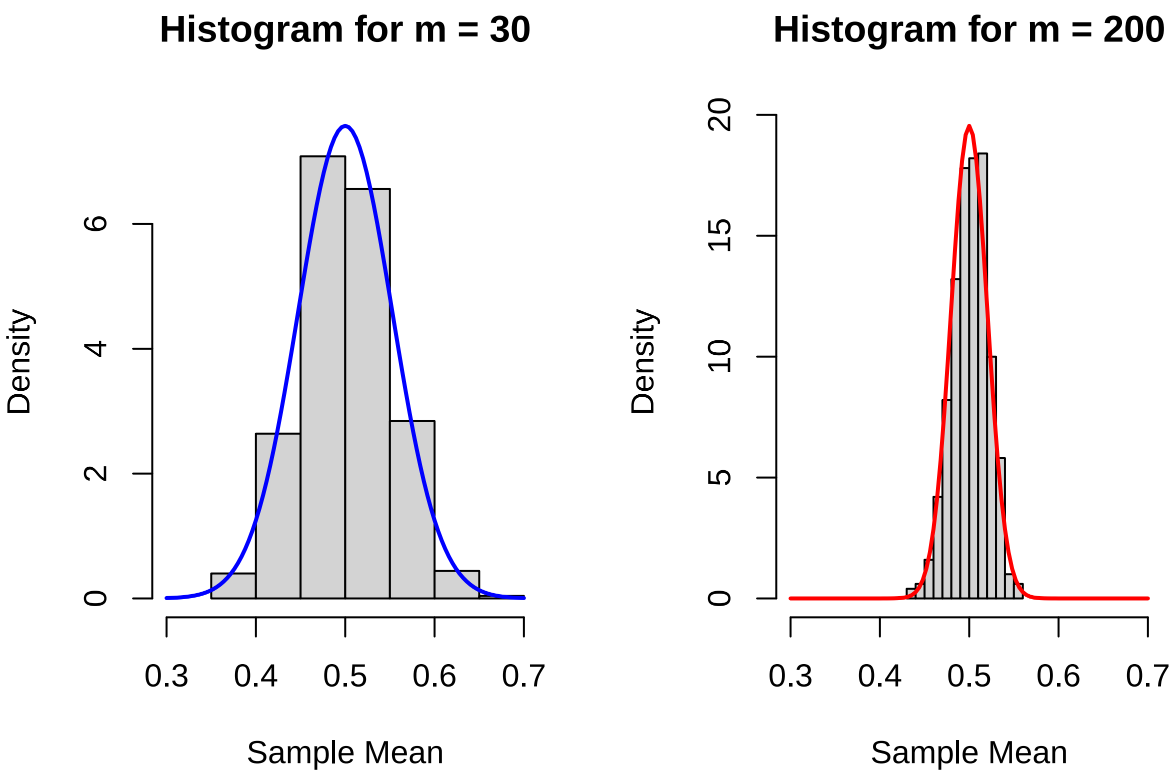
a.

```
CTL_unif <- function(n,m){
  # Use replicate to generate n samples, each consisting of m draws from U(0, 1)
  sample_means <- replicate(n, mean(runif(m)))

  # Return the vector of sample means
  return(sample_means)
}
```

b.

```
# Set seed for reproducibility
set.seed(42)

# Generate sample means for n = 500, m = 30
means_30 <- CTL_unif(n = 500, m = 30)

# Generate sample means for n = 500, m = 200
means_200 <- CTL_unif(n = 500, m = 200)

# Define the parameters for the normal distribution
mean_theoretical <- 0.5
sd_30 <- sqrt(1 / (12 * 30))  # Standard deviation for m = 30
sd_200 <- sqrt(1 / (12 * 200))  # Standard deviation for m = 200

# Get dynamic ylim values based on the density range
ylim_30 <- c(0, max(density(means_30)$y, dnorm(mean_theoretical, mean = mean_theoretical,
↪  sd = sd_30)))
ylim_200 <- c(0, max(density(means_200)$y, dnorm(mean_theoretical, mean =
↪  mean_theoretical, sd = sd_200)))

# Plot the histograms side-by-side
par(mfrow = c(1, 2))  # Set up a 1x2 plotting layout

# Plot for m = 30
hist(means_30, prob = TRUE, xlim = c(0.3, 0.7), ylim = ylim_30,
     main = "Histogram for m = 30", xlab = "Sample Mean")
curve(dnorm(x, mean = mean_theoretical, sd = sd_30), col = "blue", lwd = 2, add = TRUE)
```

```r
# Plot for m = 200
hist(means_200, prob = TRUE, xlim = c(0.3, 0.7), ylim = ylim_200,
     main = "Histogram for m = 200", xlab = "Sample Mean")
curve(dnorm(x, mean = mean_theoretical, sd = sd_200), col = "red", lwd = 2, add = TRUE)
```




```r
# Reset the plotting layout
par(mfrow = c(1, 1))
```

c.

we can observe that when we take more samples in each draw, the variance of the distribution of the sample means is smaller. And overall, we can see the distribution of the sample means converges to the theoretical distribution as the sample size in each draw increases. This can be explained by central limit theorem.

Exercise 2

```r
data("airquality")
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

a.

```r
# It isn't a good idea to use 'airquality$Ozone == NA' since NA represents a missing
↪  value and comparing with NA always gives NA. Thus we use .is.na()
NA_count <- sum(is.na(airquality$Ozone))

# calculate the proportion of missing values
NA_proportion <- NA_count / nrow(airquality)
NA_proportion
```

```
## [1] 0.2418301
```

```r
# na.rm = TRUE ensures a numerical output since it removes the missing values as it
↪  stands for na.remove
mean_ozone <- mean(airquality$Ozone, na.rm = TRUE)
mean_ozone
```

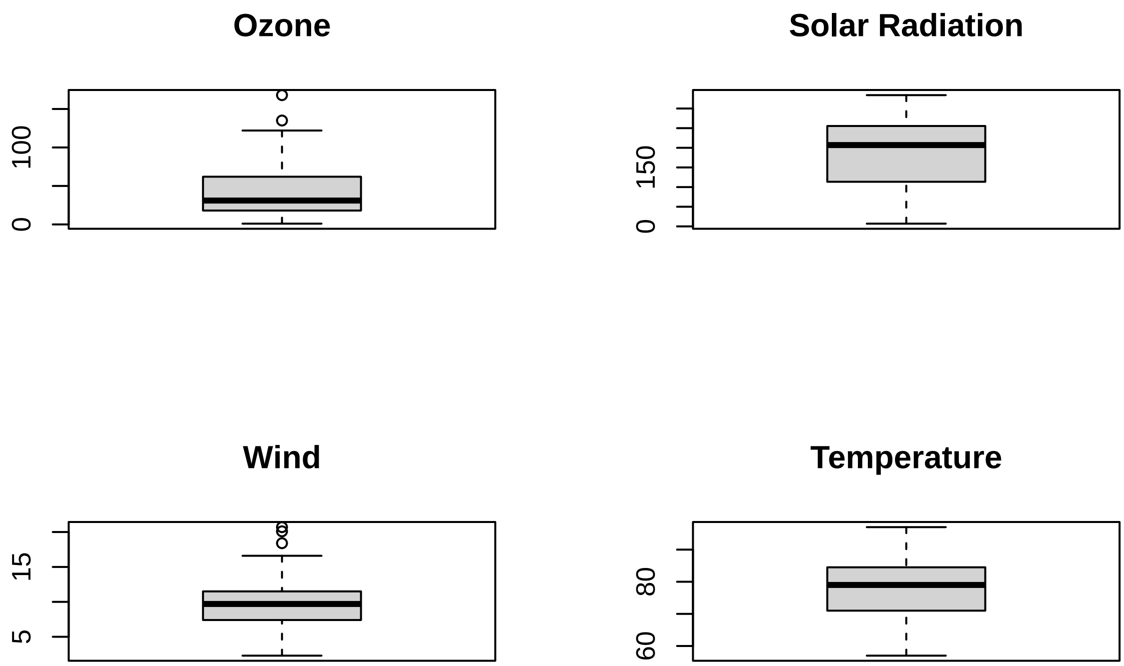
```
## [1] 42.12931
```

b.

```r
# remove all rows with a missing value
airquality_clean <- na.omit(airquality)

# remove the Day and Month columns
airquality_clean <- subset(airquality_clean, select = -c(Month, Day))

# make a summarry for every column
summary(airquality_clean)
```

```
##      Ozone          Solar.R           Wind            Temp
##  Min.   :  1.0   Min.   :  7.0   Min.   : 2.30   Min.   :57.00
##  1st Qu.: 18.0   1st Qu.:113.5   1st Qu.: 7.40   1st Qu.:71.00
##  Median : 31.0   Median :207.0   Median : 9.70   Median :79.00
##  Mean   : 42.1   Mean   :184.8   Mean   : 9.94   Mean   :77.79
##  3rd Qu.: 62.0   3rd Qu.:255.5   3rd Qu.:11.50   3rd Qu.:84.50
##  Max.   :168.0   Max.   :334.0   Max.   :20.70   Max.   :97.00
```

```r
# make the layout of boxplots a 2x2 grid
par(mfrow = c(2, 2))

# one by one plot all 4 of the boxplots
boxplot(airquality_clean$Ozone, main = "Ozone")
boxplot(airquality_clean$Solar.R, main = "Solar Radiation")
boxplot(airquality_clean$Wind, main = "Wind")
boxplot(airquality_clean$Temp, main = "Temperature")
```




```
# reset layout
par(mfrow = c(1, 1))
```

c.

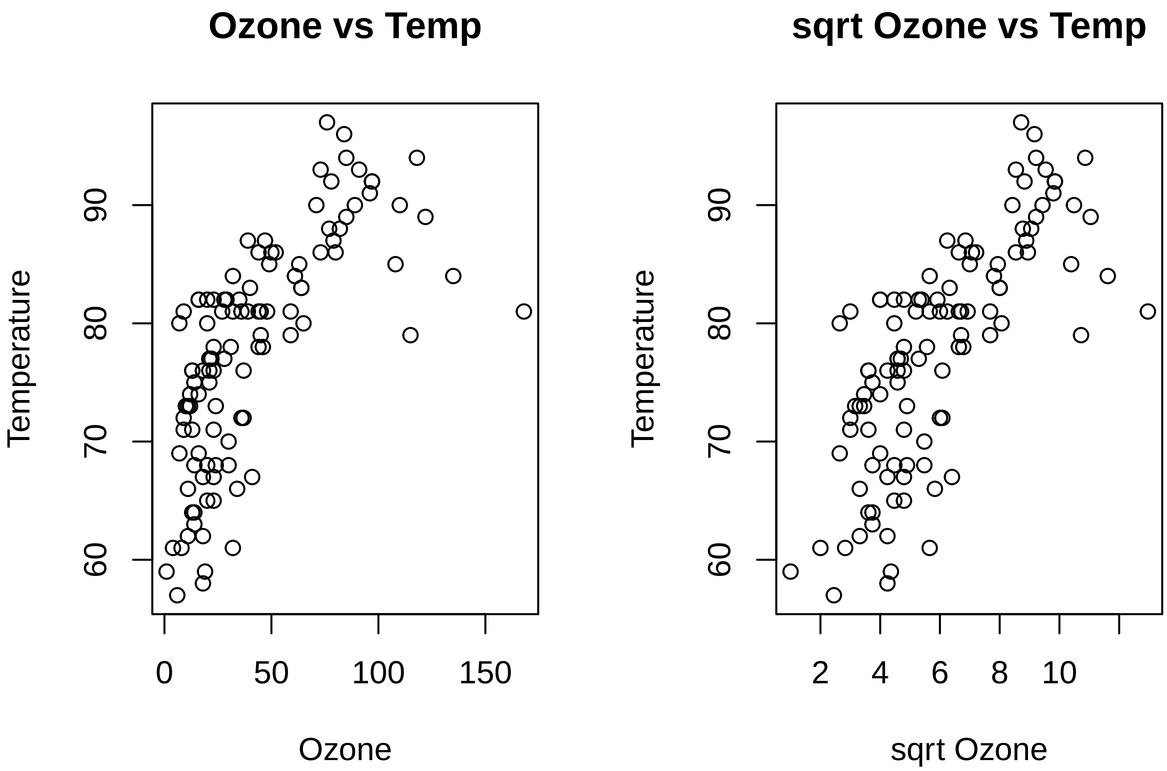
```
# correlation between Ozone and Temp
cor(airquality_clean$Ozone, airquality_clean$Temp)
```

```
## [1] 0.6985414
```

```
# correlation between square root of Ozone and Temp
cor(sqrt(airquality_clean$Ozone), airquality_clean$Temp)
```

```
## [1] 0.7458552
```

```
# The correlation changes between these two and it is higher for sqrt Ozone vs Temp, thus
↪  that has a better correlation

# use a 1x2 grid so the plots are beside each other
par(mfrow = c(1, 2))

# plot Ozone vs Temp
plot(airquality_clean$Ozone, airquality_clean$Temp,
     main = "Ozone vs Temp", xlab = "Ozone", ylab = "Temperature")

# plot sqrt Ozone vs Temp
plot(sqrt(airquality_clean$Ozone), airquality_clean$Temp,
     main = "sqrt Ozone vs Temp", xlab = "sqrt Ozone", ylab = "Temperature")
```




```
# Both have a lot of variation but the second plot follows a line better so that one more
↪  closely resembles a linear relationship
```