Homework 4

Anishka Chauhan

2022-07-25

1a)

```
library(pacman)
p_load(nycflights13, dplyr)
flights <- inner_join(flights, airlines, by="carrier")
flights <- flights[,-20] # drop extra copy of 'name'
colnames(flights)[19] <- "name" # simplify variable name
```

1b)

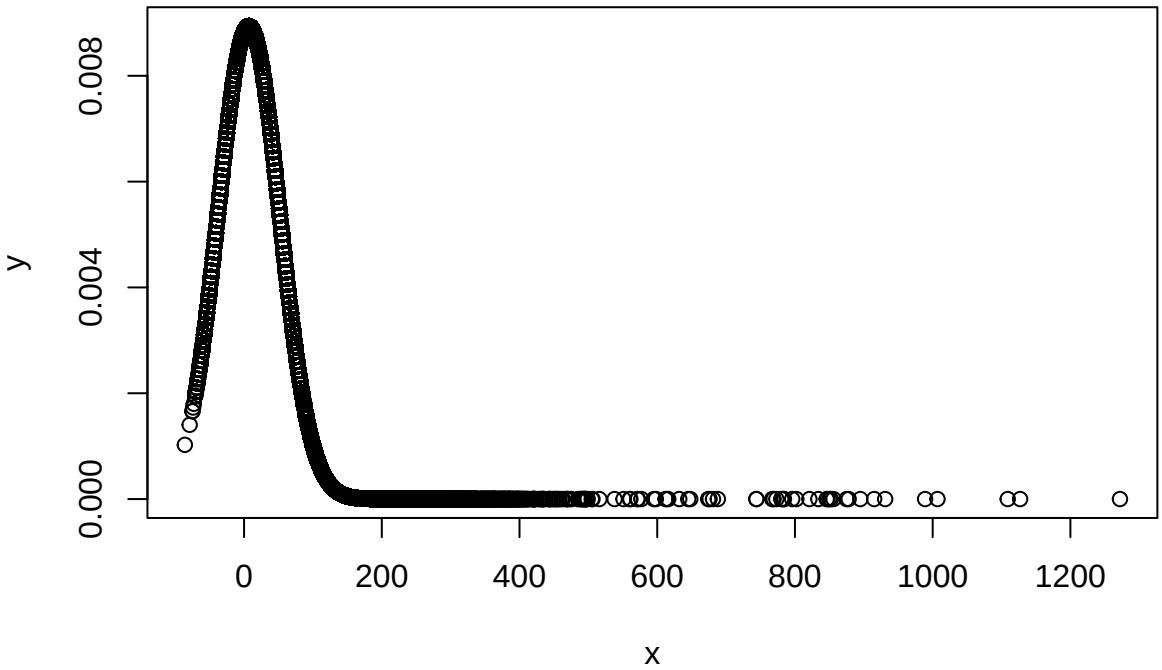
```
x = flights$arr_delay
paste0("mean: ", mean(x, na.rm = TRUE))
```

```
## [1] "mean: 6.89537675731489"
```

```
paste0("standard deviation: ", sd(x, na.rm = TRUE))
```

```
## [1] "standard deviation: 44.633291690194"
```

```
y = dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE))
plot(x, y)
```



CLT states that the distribution of x with a sample size n = 100 will be nearly normally distributed.

1c)

```r
B = 100

s = sample(x, 100, replace = TRUE)

paste0("mean: ", mean(s, na.rm = TRUE))
```

```
## [1] "mean: 5.02127659574468"
```

```r
paste0("standard deviation: ", sd(s, na.rm = TRUE))
```

```
## [1] "standard deviation: 46.6134824566871"
```

1d)

```r
SE = sd(x, na.rm = TRUE)/sqrt(length(s))
paste0("Standard Error: ", SE)
```

```
## [1] "Standard Error: 4.4633291690194"
```

1e)

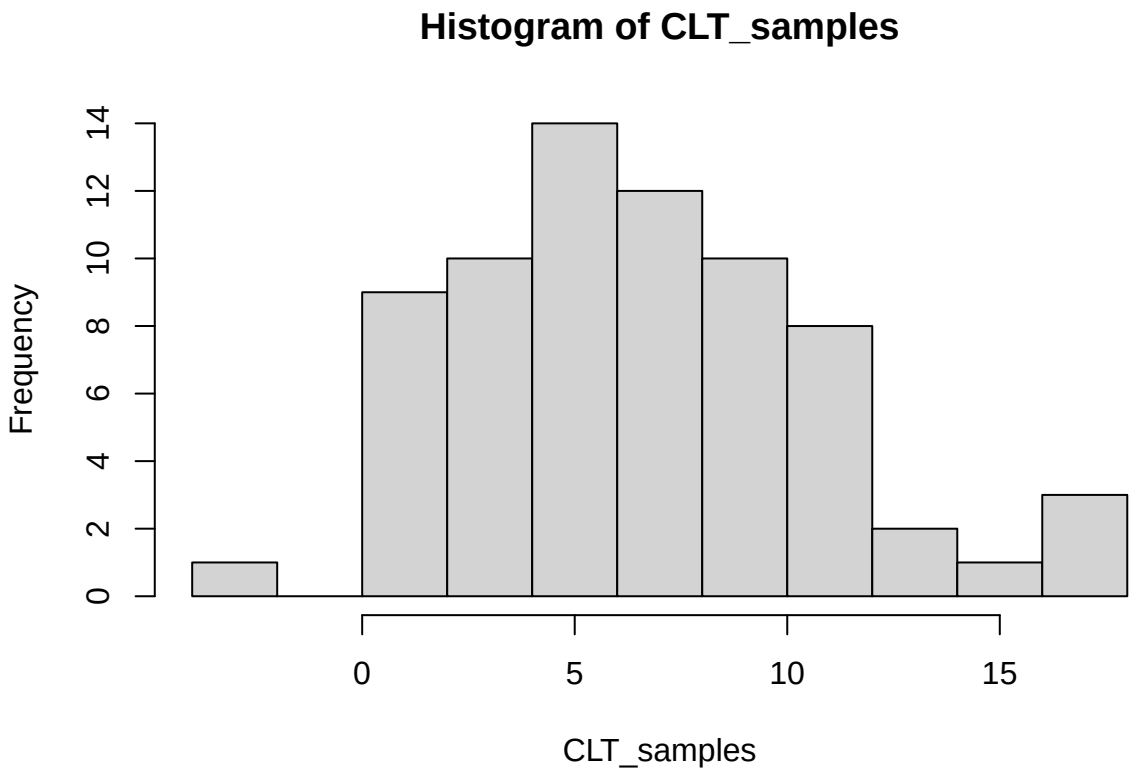
```r
boot_mean <- function(x, n, B){

  means = c()

  for(i in 1:B){
    boot_sample <- sample(x, n, replace=TRUE)
    s <- mean(boot_sample, na.rm = TRUE)
    means = c(means, s)
   # print(i)
  }

  return(means)
}

mean.sample.delays = boot_mean(s, length(s), 10000)
SE.sample.delays = sd(mean.sample.delays)/sqrt(length(s))

paste0("Standard Error: ", SE.sample.delays)
```

```
## [1] "Standard Error: 0.471365434501443"
```

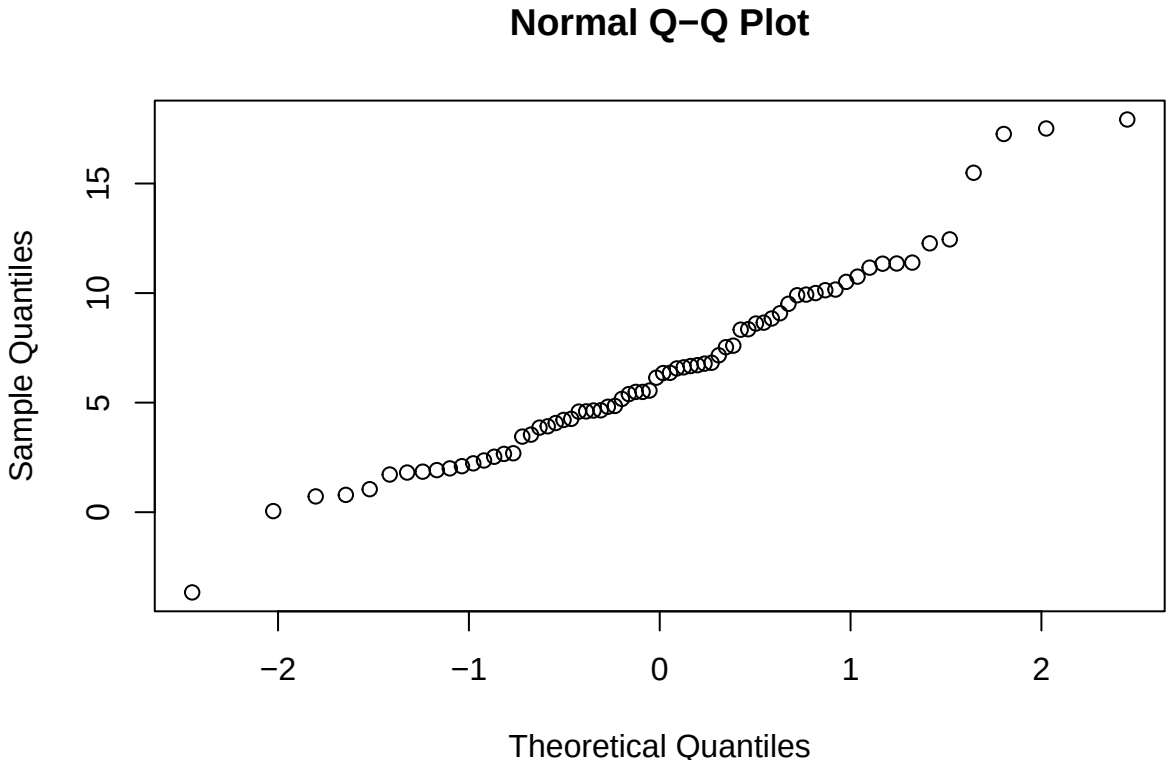
The standard error is significantly lower than the one obtained by the CLT

1f)

```r
CLT_samples = c()
for (i in 1:1000) {
  y = sample(x, 100, replace = TRUE)
  CLT_samples = c(CLT_samples, mean(y))
}
hist(CLT_samples)
```



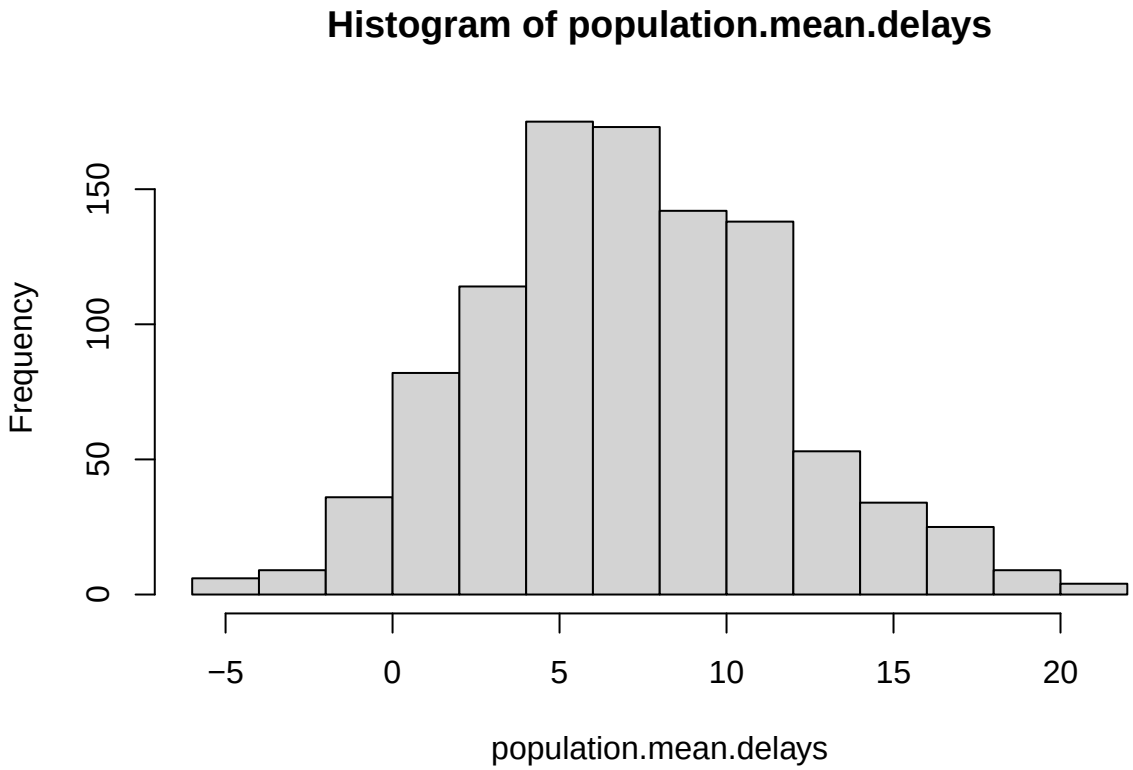

```
qqnorm(CLT_samples)
```



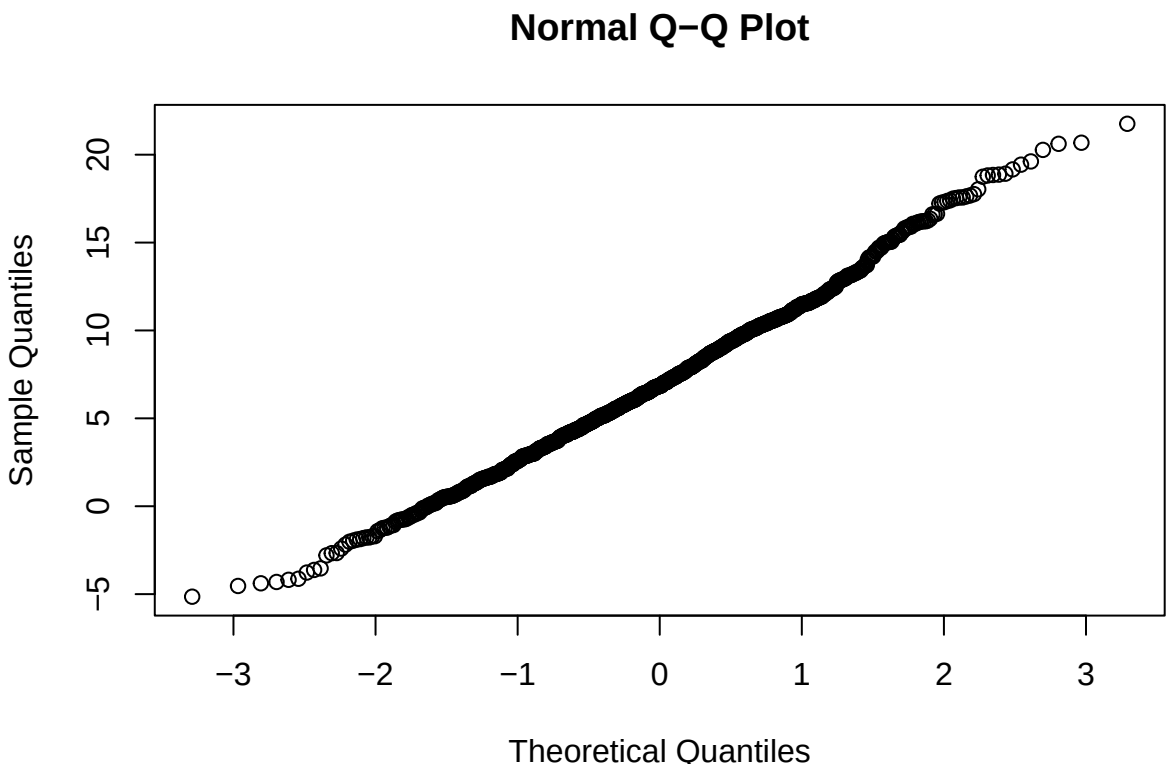
It is rather inconclusive as to whether or not the means are normally distributed, as each collection of means results in a different distribution shown in the histogram. This is also seen with the qq plot, which displays a weak linear correllation between the theoretical and sample quantiles.

1g)

```
population.mean.delays = boot_mean(x, 100, 1000)
hist(population.mean.delays)
```




```
qqnorm(population.mean.delays)
```



Compared to the CLT method, the bootstrap histogram shows that the sampling distribution is approxiamtely normal and it remains consistently normal each time the samples are collected. The bootstrap QQ plot also shows a stronger linear correlation between the theoretical and sample quantiles compared to f, indicating that it is very likely the data is normally distributed. OVerall, the bootstrap does provide a better approximation to the shape of the data.

2)

```
library(boot)
data("claridge")
```

2a)

```
x = claridge$dnan
y = claridge$hand
paste0("Sample correlation of x and y: ", cor(x, y))
```

```
## [1] "Sample correlation of x and y: 0.508775820060637"
```

2b)

```
boot_cor <- function(w, h, B=1000){

  n <- length(w)
  boot_stats <- matrix(nrow=B)

  for(i in 1:B){
    indices <- sample(n, replace=TRUE)
    boot_stats[i] <- cor(w[indices], h[indices])
  }

  return(boot_stats)
}
bootstrap_correllation = boot_cor(x, y, 10000)
```

Since x and y are dependent, they must be sampled using the same indicies in each variable vector in order for the same x and y pairs to be selected to calculate the correlation.

2c)

```
SE_cor = sd(bootstrap_correllation)
c(cor(x,y) - 1.96*SE_cor, cor(x,y) + 1.96*SE_cor)
```

```
## [1] 0.1059115 0.9116401
```

2d)

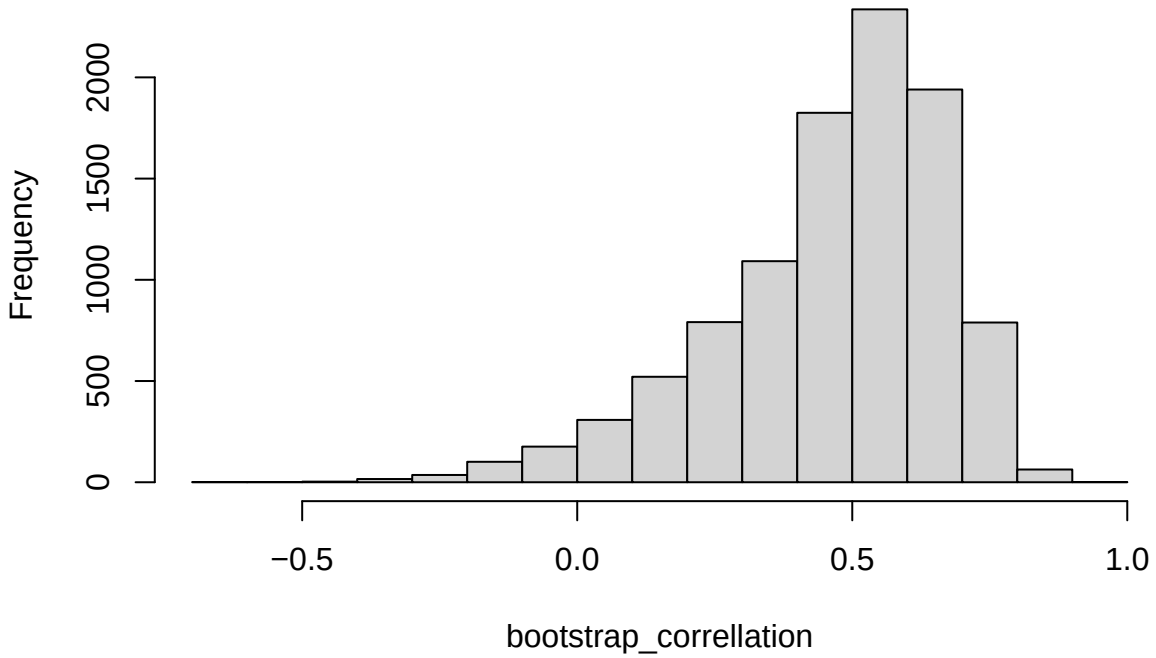
```
quantile(bootstrap_correllation, c(0.025, 0.975))
```

```
##        2.5%       97.5%
## -0.03956109  0.75864124
```

The confidence intervals are lower in the bootstrap method than the CLT method

2e)

```
hist(bootstrap_correllation)
```



The samples don't look normal, meaning CLT does not apply. Therefore I'd trust the bootstrap confidence interval more than the CLT confidence interval because the CLT interval assumes the data is normally distributed, which in this case is false.

Bootstrap True or False

a) TRUE
b) FALSE
c) TRUE
d) FALSE