# Stat_641_Homework_2

Vikas_Reddy_Bodireddy

2024-02-21

**5.8: Consider a population that has a normal distribution with mean $\mu = 36$, standard deviation $\sigma = 8$.**

**a) The sampling distribution of X for samples of size 200 will have what mean, standard error, and shape?**

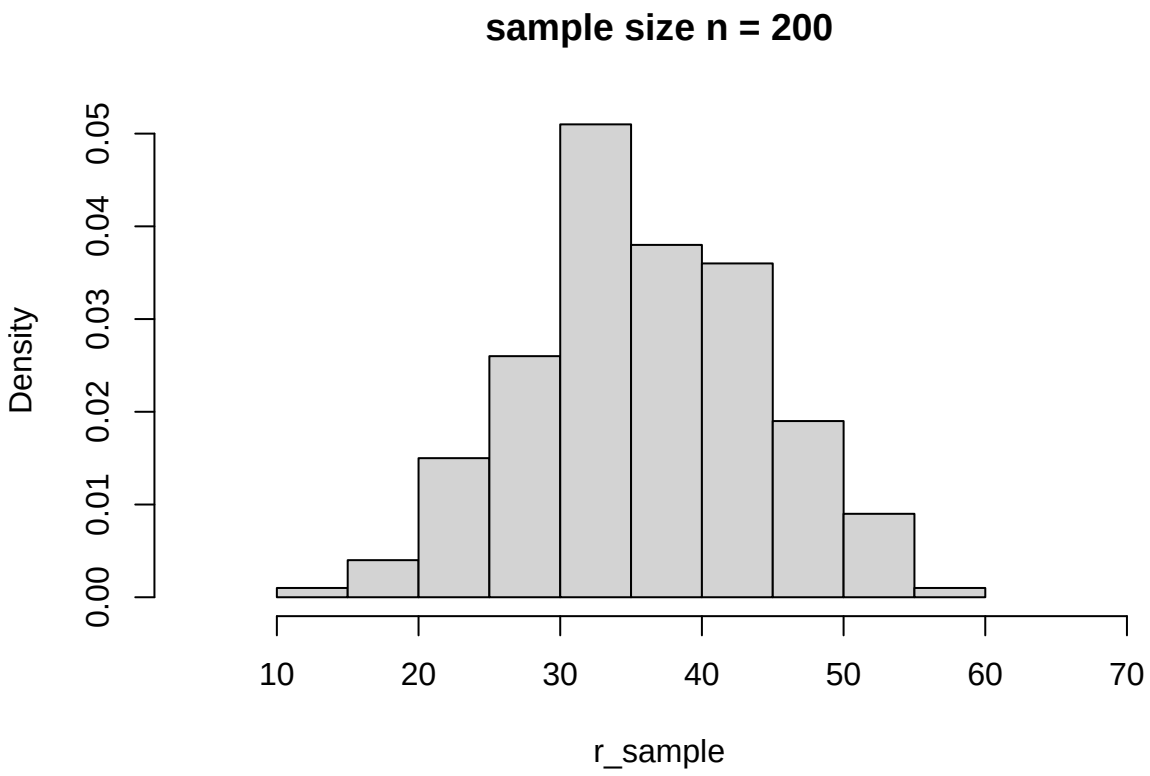
answer: The sampling distribution will be a normal distribution with the center around the mean value of 36 and the standard error of $\sigma/sqrt(n)$ i.e 8/sqrt(200) = 0.5656854 .

**b) Use R to draw a random sample of size 200 from this population. Conduct EDA on your sample.**

```
set.seed(243)
r_sample <- rnorm(200, mean = 36, sd = 8)
summary(r_sample)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.85   30.55   35.31   36.02   41.78   58.72
```

```
hist(r_sample, freq = F, main = "sample size n = 200", xlim = c(4,70))
```

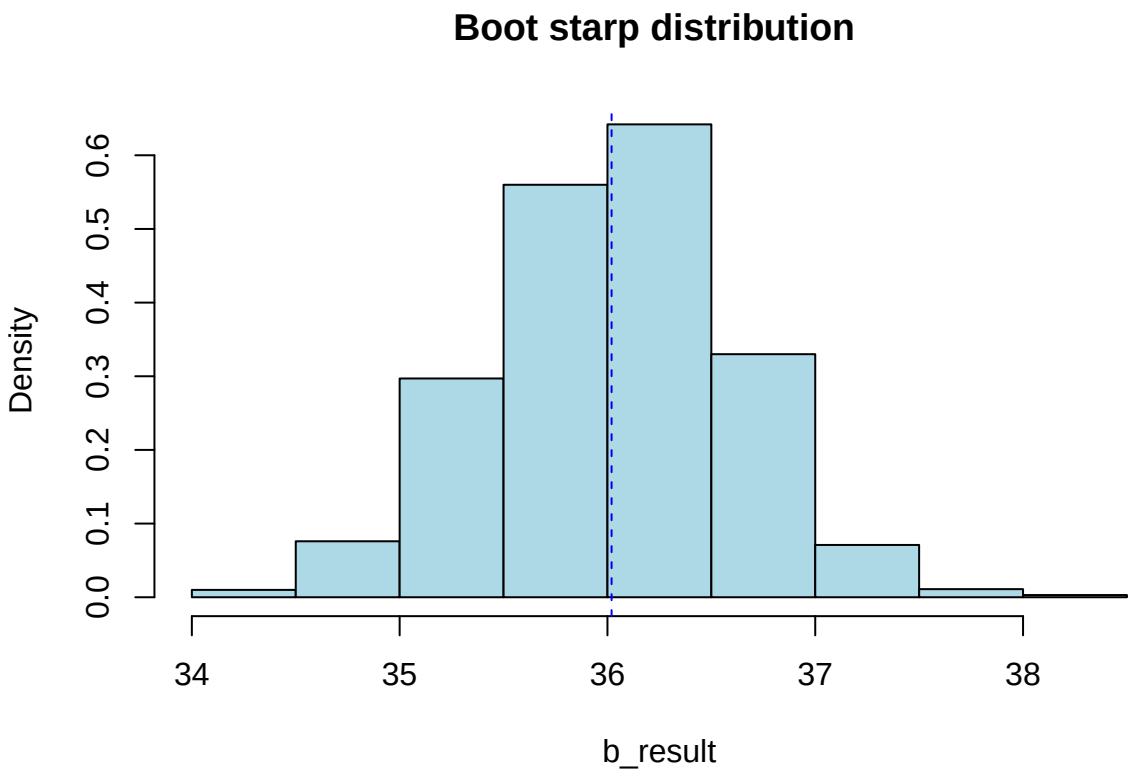



```
cat("The stadard deviation for the random sample created is ",sd(r_sample))
```

```
## The stadard deviation for the random sample created is  8.223496
```

**c) Compute the bootstrap distribution for your sample, and note the bootstrap mean and standard error.**

```
set.seed(243)
b_result <- array()
for(i in 1:2000){
  r_new <- sample(r_sample, size = 200, replace = T)
  b_result[i] <- mean(r_new)
}
hist(b_result, bins = 40,freq = F, main = "Boot starp distribution ",
     col = 'lightblue')
abline(v = mean(r_sample), lty = 2,col = "blue")
```




```
cat('The standard error for bootstrap sample is',sd(b_result))
```

```
## The standard error for bootstrap sample is 0.5843226
```

**d) Compare the bootstrap distribution to the theoretical sampling distribution by creating a table like Table 5.2.**

```
r_names <- c("Population","sample","sampling distribution of x",
             "bootstrap sample distribution")
Mean <- c(36, mean(r_sample), 36,mean(b_result))
standard_deviation <- c(8,sd(r_sample), 8/sqrt(200),sd(b_result))
df <- data.frame(r_names,Mean,standard_deviation)
kable(df, digits = 2)
```

| r_names | Mean | standard_deviation |
|---|---|---|
| Population | 36.00 | 8.00 |
| sample | 36.02 | 8.22 |
| sampling distribution of x | 36.00 | 0.57 |
| bootstrap sample distribution | 36.03 | 0.58 |

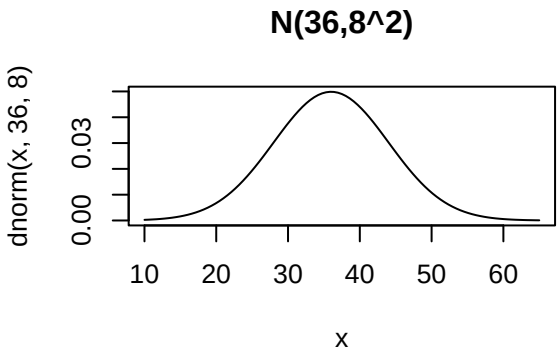
From above table we can see clearly that the sample mean and bootstrap sample mean are both nearly same and also the standard error for sampling distribution and bootstrap sampling distribution is also similar.

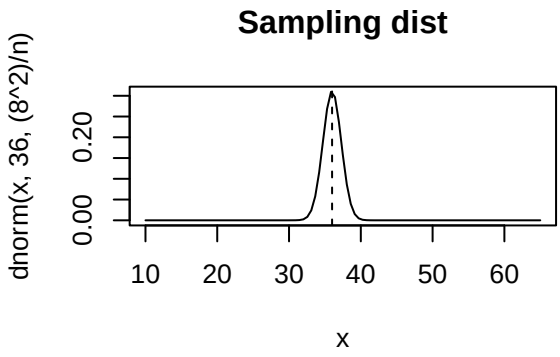
**e) Repeat for sample sizes of n=50 and n=10.Carefully describe your observations about the effects of sample size on the bootstrap distribution.**

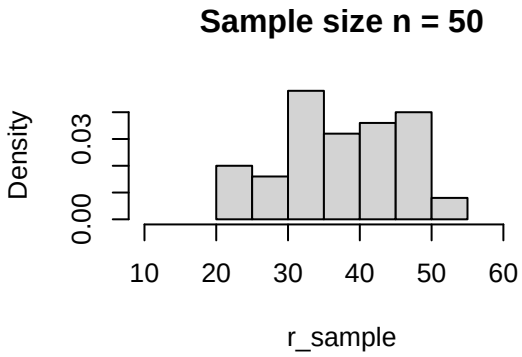
```
set.seed(143)

r_samplesizes <- function(n, mean = 36, sd = 8){
  r_sample <- rnorm(n, mean = 36, sd = 8)
  summary(r_sample)

  par(mfrow = c(2, 2))
  curve(dnorm(x,36,8),from = 10, to=65, main="N(36,8^2)")
  curve(dnorm(x,36,(8^2)/n),from = 10, to=65, main="Sampling dist")
  abline(v=36,lty=2)
  hist(r_sample, freq = F, main = paste("Sample size n =", n), xlim = c(10,65))
  cat("The standard deviation for the random sample created is ", sd(r_sample),
      "\n")

  b_result <- numeric(2000)

  for(i in 1:2000){
    r_new <- sample(r_sample, size = n, replace = TRUE)
    b_result[i] <- mean(r_new)
  }

  hist(b_result, bins = 40, freq = F,
       main = paste("Bootstrap distribution for n =", n),
       col = 'lightblue',xlim = c(10,65))
  abline(v = mean(r_sample), lty = 2, col = "blue")
  cat('The standard error for bootstrap sample is', sd(b_result), "\n")

  r_names <- c("Population", "Sample", "Sampling Distribution of x",
               "Bootstrap Sample Distribution")
  Mean <- c(36, mean(r_sample), 36, mean(b_result))
  standard_deviation <- c(8, sd(r_sample), 8 / sqrt(n), sd(b_result))

  col_names <- c("Variable", paste("Mean for n =", n), paste("SD for n =", n))

  df <- data.frame(r_names, Mean, standard_deviation)
  colnames(df) <- col_names

  kable(df, digits = 2)
}
```

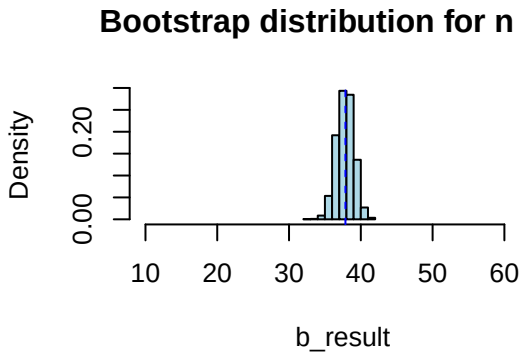
```
# For Sample size = 50
r_samplesizes(50)
```

```
## The standard deviation for the random sample created is  8.553791
```



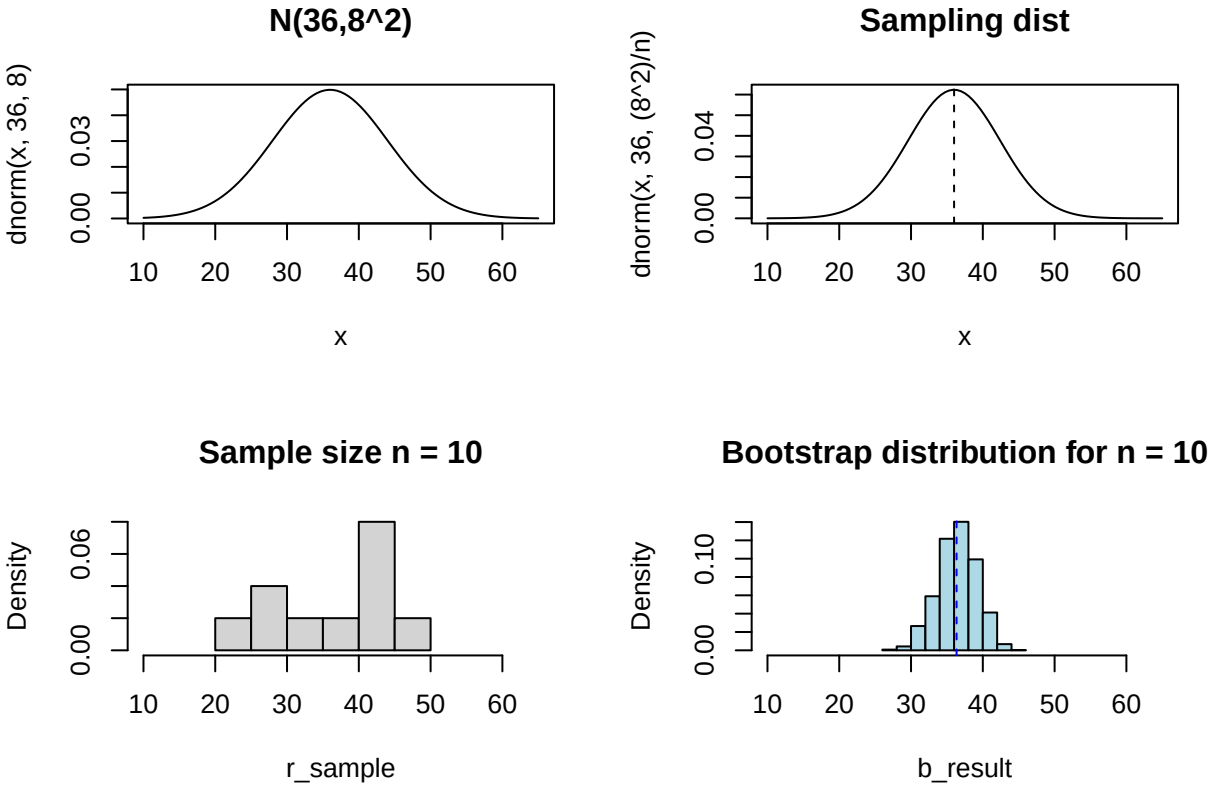




```
## The standard error for bootstrap sample is 1.205851
```

| Variable | Mean for n = 50 | SD for n = 50 |
|---|---|---|
| Population | 36.00 | 8.00 |
| Sample | 37.85 | 8.55 |
| Sampling Distribution of x | 36.00 | 1.13 |
| Bootstrap Sample Distribution | 37.82 | 1.21 |

```
# For Sample size = 10
r_samplesizes(10)
```

```
## The standard deviation for the random sample created is  9.035628
```




```
## The standard error for bootstrap sample is 2.712979
```

| Variable | Mean for n = 10 | SD for n = 10 |
|---|---|---|
| Population | 36.00 | 8.00 |
| Sample | 36.34 | 9.04 |
| Sampling Distribution of x | 36.00 | 2.53 |
| Bootstrap Sample Distribution | 36.44 | 2.71 |

As the sample size decreases from 200 to 50 and further to 10, we observe that the sampling distribution becomes progressively less normal in shape, indicating increased variability in the estimates. However, after bootstrapping the samples, we notice a remarkable shift towards a more normal distribution, suggesting that the bootstrap method effectively mitigates the effects of small sample sizes on the sampling distribution.

Examining the summary statistics, we find that for a sample size of 50, the sample mean is 37.85 and the bootstrap sample mean is 37.82, showing a close correspondence between the estimated statistics. Similarly, for a sample size of 10, the sample mean is 36.34 and the bootstrap sample mean is 36.44, indicating accurate estimation despite the small sample size.

However, it's crucial to note that as the sample size decreases, the standard error values increase substantially. For instance, for a sample size of 50, the standard error for the sampling distribution is 1.13, while for the bootstrap sampling distribution, it's 1.21. Similarly, for a sample size of 10, the standard error for the sampling distribution is 2.53, and for the bootstrap sampling distribution, it's 2.71.

This trend underscores the notion that as the sample size diminishes, the uncertainty in estimating population parameters magnifies, as evidenced by the augmented standard errors in the sampling and bootstrap distributions.

**5.18: Is there a difference in the price of groceries sold by the two retailers Target and Walmart? The data set Groceries contain, a sample of grocery items and their prices advertised on their respective websites on one specific day.**

**a) Compute summary statistics of the prices for each store.**

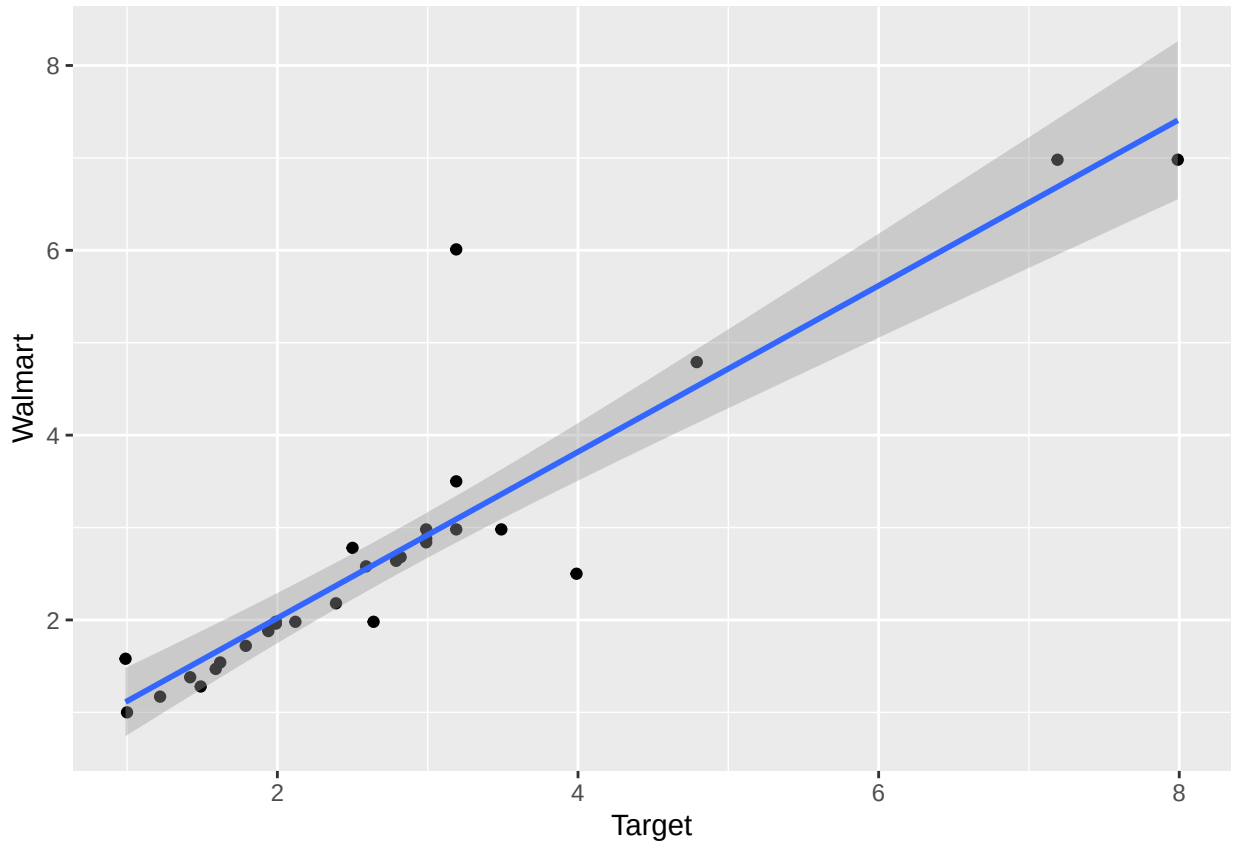
```
data("Groceries")
Groceries <- Groceries|>
  select(Product,Target,Walmart)|>
  na.omit()

summary(Groceries$Target)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.990   1.827   2.545   2.762   3.140   7.990
```

```
summary(Groceries$Walmart)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.760   2.340   2.706   2.955   6.980
```

```
ggplot(data = Groceries, aes(x = Target,y = Walmart))+
  geom_point()+
  geom_smooth(method = "lm")+
  labs(main = "Scatter plot of target vs Walmart")
```

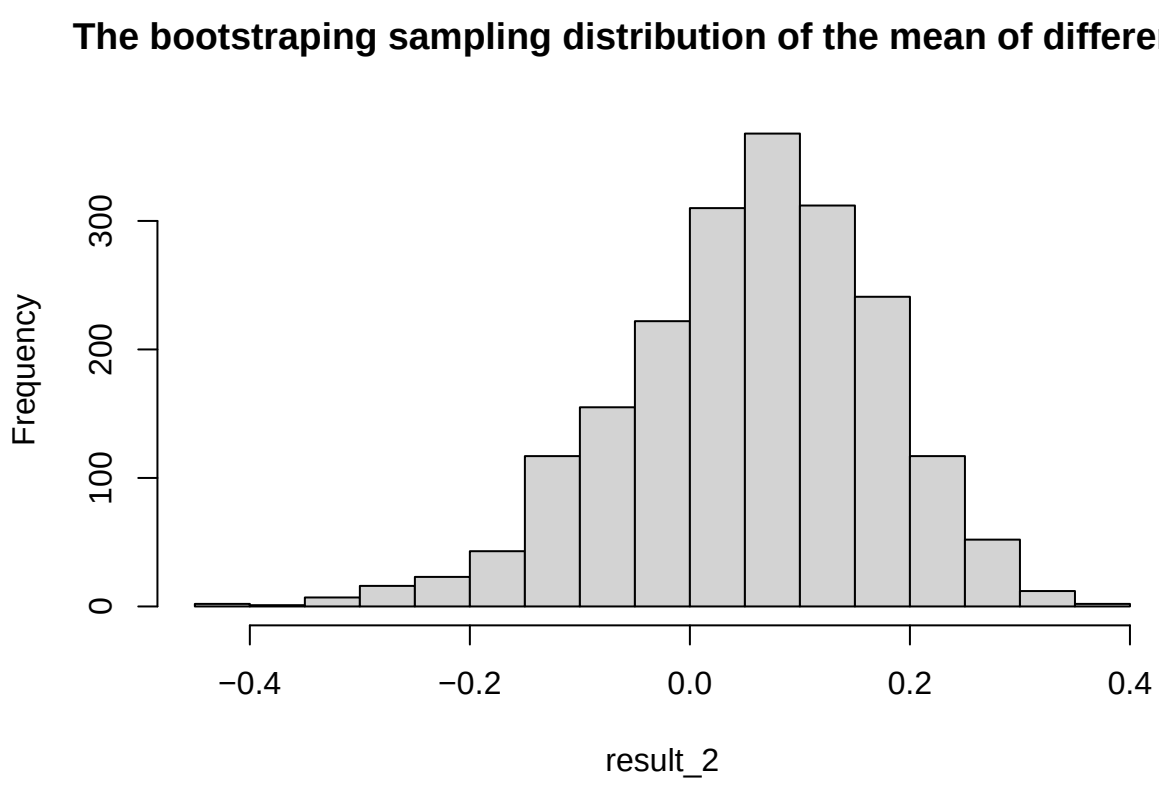


**b) Use the bootstrap to determine whether or not there is a difference in the mean prices.**

```
set.seed(123)
walmart <- Groceries$Walmart
target <- Groceries$Target

diff <- target - walmart
mean(diff)
```

```
## [1] 0.05666667
```

```
result_2 <- array()
for(i in 1:2000){
  n_diff <- sample(diff, size = length(diff), replace = T)
  result_2[i] <- mean(n_diff)
}
hist(result_2, main = "The bootstraping sampling distribution of the mean of differences ")
```




```
cat("95% Confidence interval rane is ", quantile(result_2,probs = c(0.025,0.975)),"\n")
```

```
## 95% Confidence interval rane is  -0.1996667 0.2596833
```

Answer:

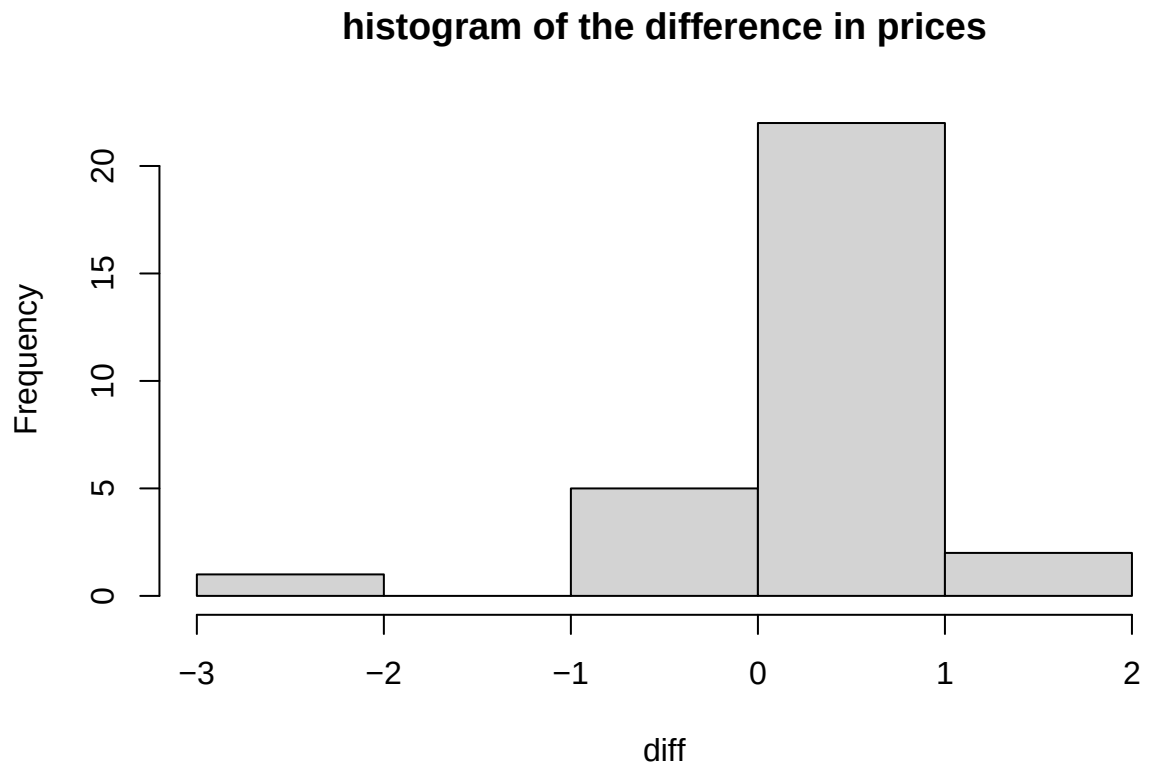
Null Hypothesis($H_0$) : $\mu_1 = \mu_2$ No mean Difference among the Groups Alt Hypothesis($H_0$) : $\mu_1 \neq \mu_2$ there is mean Difference among the Groups

Based on the provided confidence interval, which spans from -0.1996667 to 0.2596833, we fail to reject the null hypothesis. This suggests that there is no significant difference between the means of prices in the two stores. The presence of the value 0 within the confidence interval implies that the mean could potentially fall anywhere within that range, indicating a two-sided hypothesis.

Also we can see from the scatter plot that these two observations are of matched pairs, as the products the products that are being compared are same for each store.

**c) Create a histogram of the difference in prices.What is unusual about Quaker Oats Life cereal?**

```
hist(diff,main = "histogram of the difference in prices")
```




```
Quacker_cereal <- Groceries$Target[Groceries$Product ==
                                     "Quaker Oats Life Cereal  Original "] -
  Groceries$Walmart[Groceries$Product == "Quaker Oats Life Cereal  Original "]
Quacker_cereal
```

```
## [1] -2.82
```

Answer: If Quaker Oats cereal consistently maintains a price of $6.01 at Walmart, significantly higher than its $3.19 price at Target, and this trend persists across multiple iterations, it could substantially skew the mean of the bootstrap distribution. This is because the difference of -$2.82 would be repeatedly included

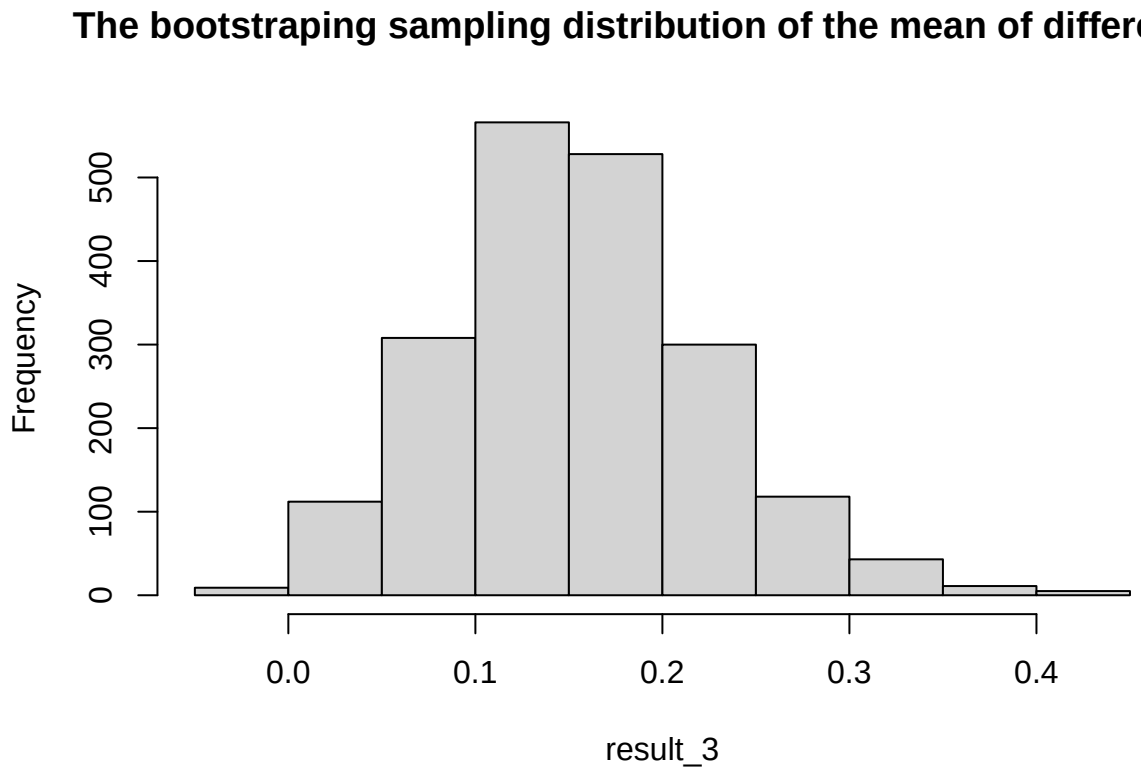
in each iteration as we allow for repetition. Consequently, the distribution would deviate significantly from accurately representing the population due to the uncertain and disproportionately large price difference observed in the dataset.

**d) Recompute the bootstrap percentile interval without this observation. What do you conclude?**

```
set.seed(123)
walmart_2 <- Groceries|>
  filter(Product != "Quaker Oats Life Cereal  Original ")|>
  pull(Walmart)
target_2 <- Groceries|>
  filter(Product != "Quaker Oats Life Cereal  Original ")|>
  pull(Target)

diff_2 <- target_2 - walmart_2
mean(diff_2)
```

```
## [1] 0.1558621
```

```
result_3 <- array()
for(i in 1:2000){
  n_diff <- sample(diff_2, size = length(diff_2), replace = T)
  result_3[i] <- mean(n_diff)
}
hist(result_3, main = "The bootstraping sampling distribution of the mean of differences ")
```



```r
cat("95% Confidence interval rane is ", quantile(result_3,probs = c(0.025,0.975)),"\n")
```

```
## 95% Confidence interval rane is  0.02689655 0.3059138
```

Answer:After excluding the row corresponding to "Quaker Oats Life Cereal Original," we observe a significant reduction in the uncertainty surrounding the mean difference. This reduction occurs because, throughout the 2000 replications, the problematic value of -2.82 is no longer present to skew the distribution. Consequently, the new bootstrap sample exhibits a near-normal distribution.

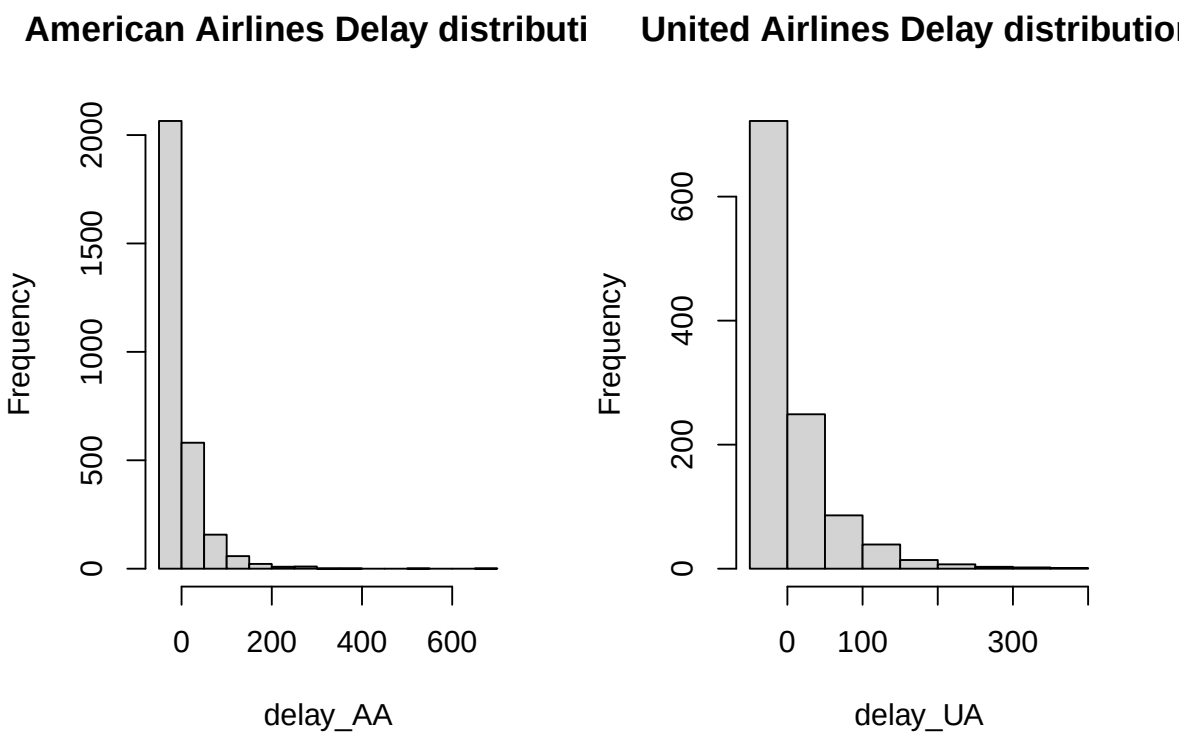
Following the removal of this specific row, the confidence interval obtained is (0.02689655, 0.3059138), which notably does not contain 0 within its range. This indicates that we can confidently assert that the mean difference lies between 0.02 and 0.30. Furthermore, as the confidence interval does not include 0, we reject the Null hypothesis , and go in favor of the Alternate hypothesis ($\mu_1 - \mu_2 > 0$), asserting that there is indeed a mean difference between the two stores, with Target charging more on average compared to Walmart.

**5.22: Import the data from flight delays case study in Section 1.1 data into R. Although the data represent all UA and AA flights in May and June of 2009, we will assume they represent a sample from a larger population of UA and AA flights flown under similar circumstances. We will consider the ratio of means of the flight delay lengths, $\mu_{UA}/\mu_{AA}$.**

a) **Perform some exploratory data analysis on flight delay lengths for each of UA and AA flights.**

```r
data(FlightDelays)
```

```r
delay_AA<- FlightDelays$Delay[FlightDelays$Carrier == "AA"]
delay_UA <- FlightDelays$Delay[FlightDelays$Carrier == "UA"]
par(mfrow = c(1,2))
hist(delay_AA,main = "American Airlines Delay distribution")
hist(delay_UA,main = "United Airlines Delay distribution")
```




```
summary(delay_AA)
```

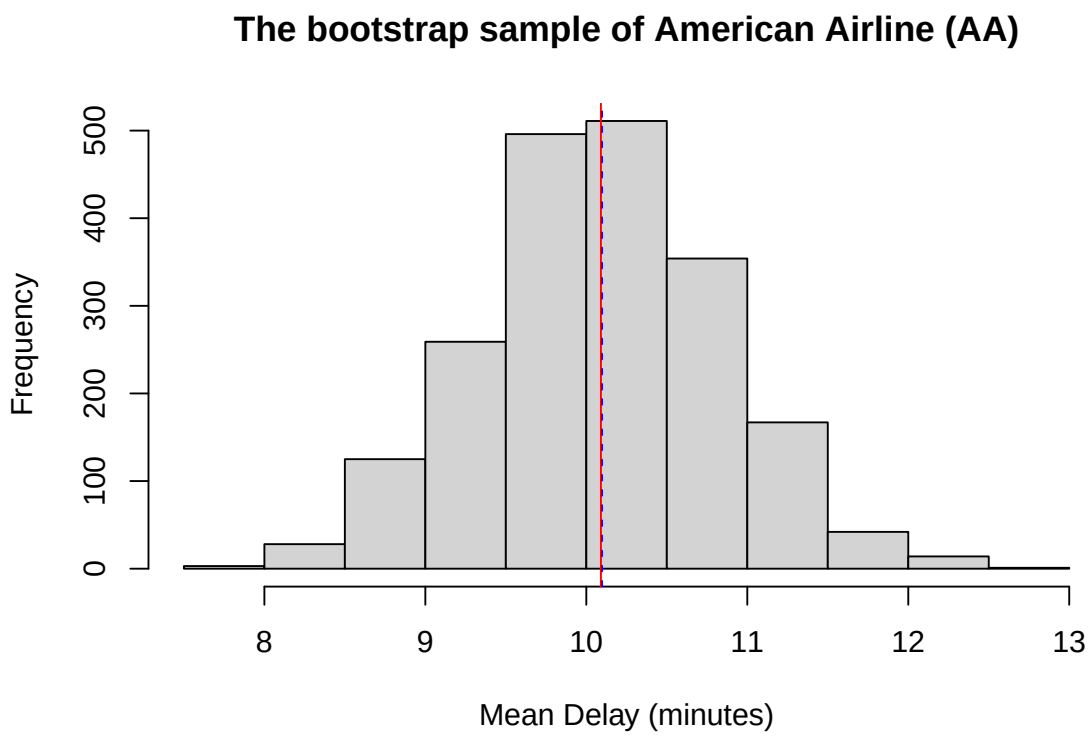
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -19.0    -6.0    -3.0    10.1     4.0   693.0
```

```
summary(delay_UA)
```

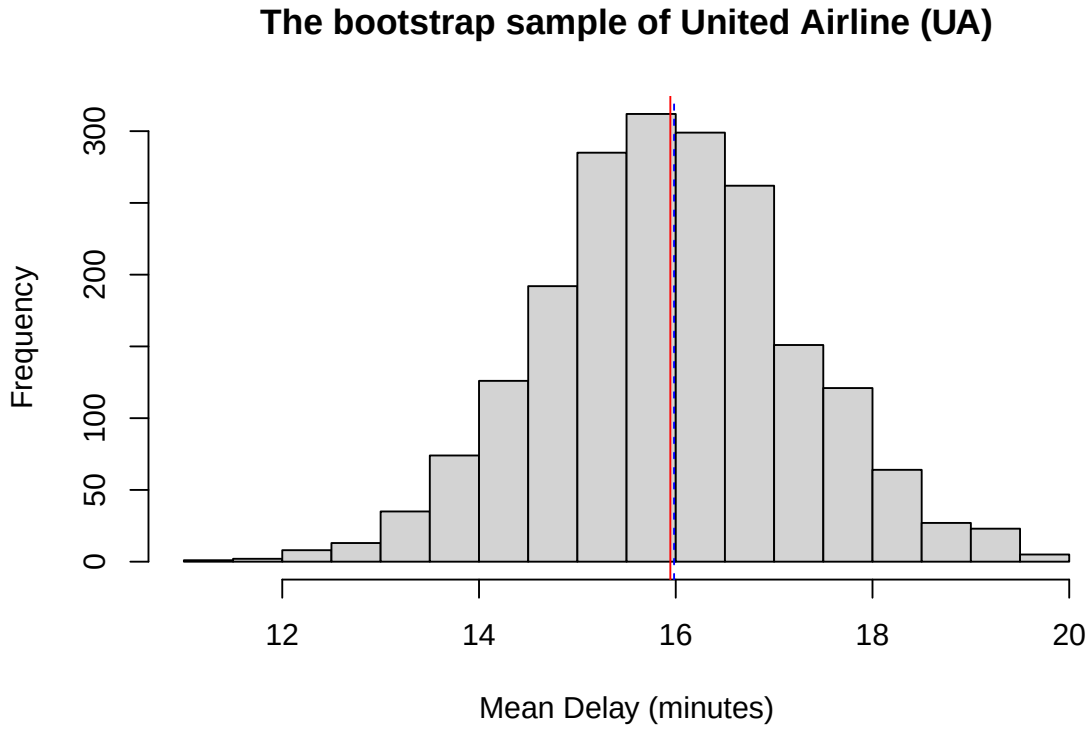
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -17.00   -5.00   -1.00   15.98   12.50  377.00
```

b) **Bootstrap the mean of flight delay lengths for each airline separately, and describe the distribution.**

```
set.seed(123)
bootstrap_function <- function(data, airline_name){
  result_boot <- numeric(2000)
  for(i in 1:2000){
    Airline_mean <- sample(data, size = length(data), replace = TRUE)
    result_boot[i] <- mean(Airline_mean)
  }
  hist(result_boot, main = paste("The bootstrap sample of", airline_name), xlab = "Mean Delay (minu
  abline(v = mean(result_boot), col = "red", lty = 1)
  abline(v = mean(data), col = "blue", lty = 2)
}

bootstrap_function(delay_AA, "American Airline (AA)")
```




```
bootstrap_function(delay_UA, "United Airline (UA)")
```



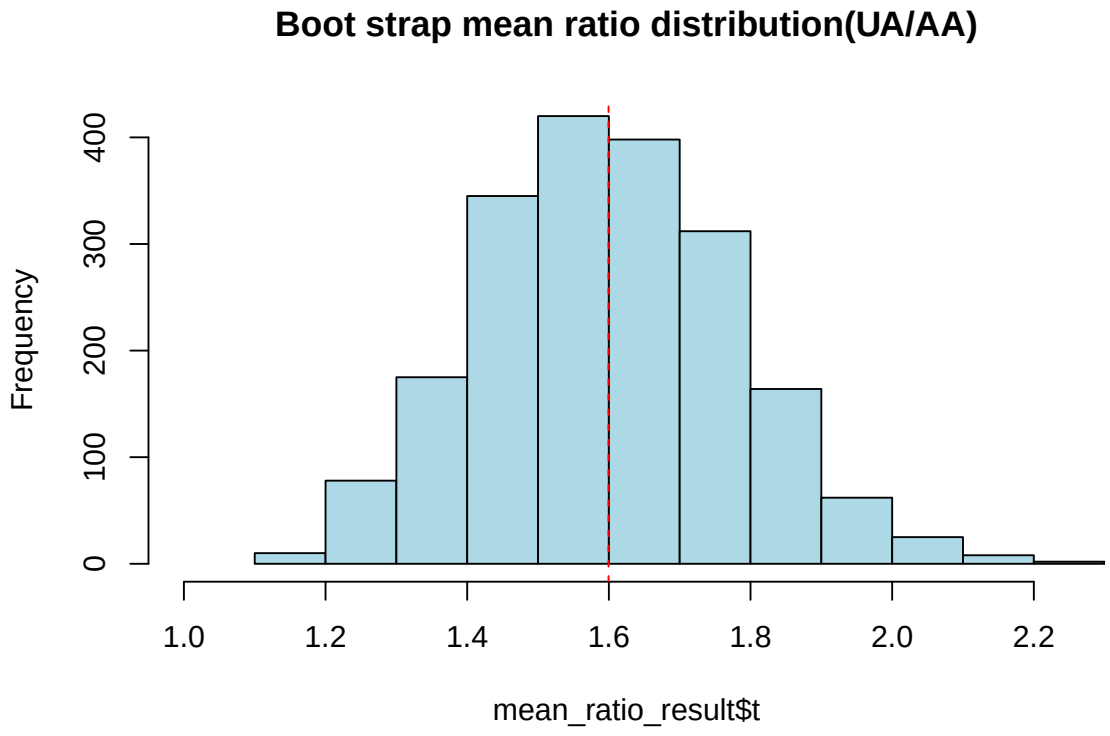
Answer: We can see that both the bootstrap samples are normally distributed. and the mean of the sapling distribution and the mean of the bootstrap sampling distribution are nearly same. but comparing the histogram and spread of the data, we can see that our bootstrap sample is more effective and more reasonable to check our assumptions. i.e If there is any difference in the mean ratio of the data.

c) **Bootstrap the ratio of means.Provide plots of the bootstrap distribution and describe the distribution.**

```
set.seed(123)
mean_ratio <- mean(delay_UA) / mean(delay_AA)
mean_ratio
```

```
## [1] 1.582893
```

```
mean_ratio_boot <- function(data, index){
  mean_UA <- data[index,]|> filter(Carrier == "UA")
  mean_AA <- data[index,]|> filter(Carrier == "AA")
  return(mean(mean_UA$Delay)/mean(mean_AA$Delay))
}

mean_ratio_result <- boot(data = FlightDelays,statistic = mean_ratio_boot, R = 2000)

hist(mean_ratio_result$t,main = "Boot strap mean ratio distribution(UA/AA)",
     col = "lightblue", xlim = c(1.0, 2.25))
abline(v = mean(mean_ratio_result$t), col = "red", lty = 2)
```



d) **Find the 95% bootstrap percentile interval for the ratio of means. Interpret this interval.**

```
set.seed(123)
boot.ci(mean_ratio_result, type = "basic")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = mean_ratio_result, type = "basic")
##
## Intervals :
## Level      Basic
## 95%   ( 1.196,  1.907 )
## Calculations and Intervals on Original Scale
```

**Answer:** The 95% bootstrap percentile interval for the ratio of means spans from 1.19 to 1.90. Remarkably, our observed true mean ratio of 1.58 falls within this interval. With such findings, we assert that in a statistical sense, 95% of the time when we repeat this analysis, the mean ratio will be encompassed within this confidence interval range.

e) **What is the bootstrap estimate of the bias? What fraction of the bootstrap standard error does it represent?**

```
set.seed(123)
boot.bias <- mean(mean_ratio_result$t) - mean_ratio_result$t0
boot.bias
```

```
## [1] 0.01666199
```

```
boot.se <- sd(mean_ratio_result$t)
boot.se
```

```
## [1] 0.1791286
```

```
cat("The Bootstrap Estimate of the bias is nearly",round(boot.bias/boot.se,3),"or",round((boot.bias
```

```
## The Bootstrap Estimate of the bias is nearly 0.093 or 9.302 % of the boot strap standard error
```

f) **For inference in this text, we assume that the observations are independent. Is that condition met here? Explain.**

Answer: We can see that there is no correlation between the points, and all the points are independent of each other. and all the observations are randomly sampled, which ensures there is no bias in selecting these observations.