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?

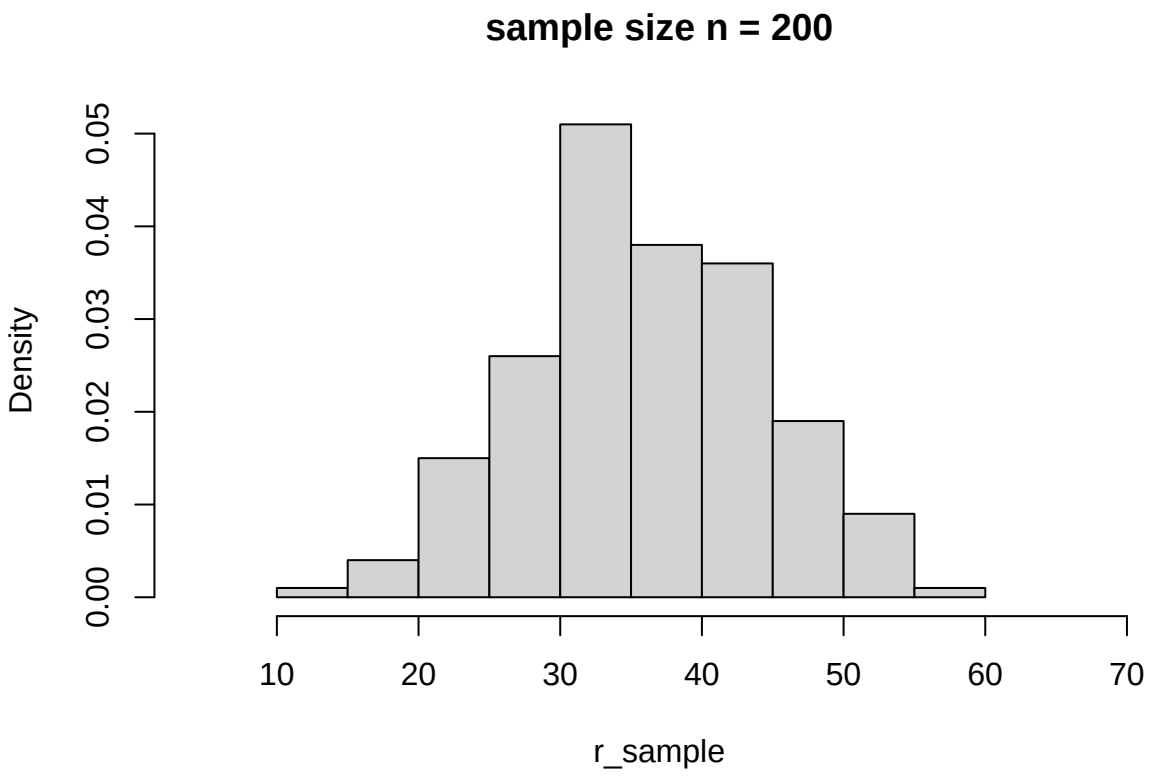
answer: The sampling distribution will be a normal distribution with the center around the mean value of 36 and the standard error of $\sigma/sqrt(n)$ i.e 8/sqrt(200) = 0.5656854 .

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))
```

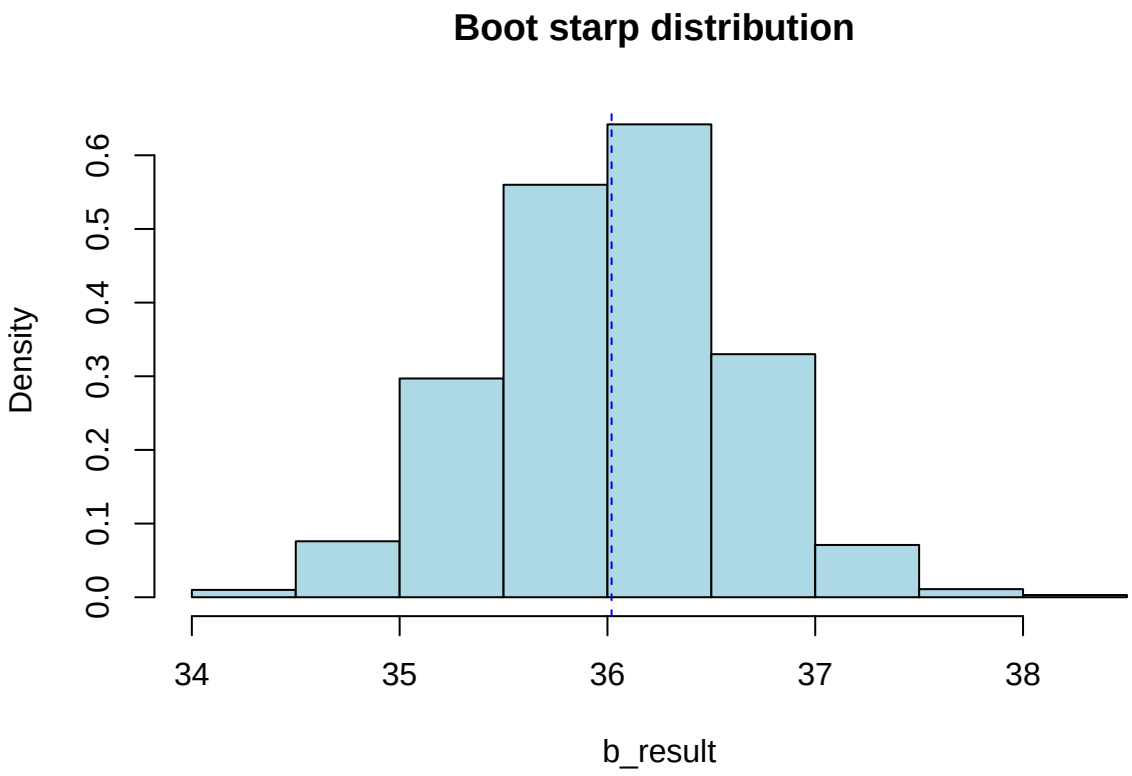



```
cat("The stadard deviation for the random sample created is ",sd(r_sample))
```

```
## 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

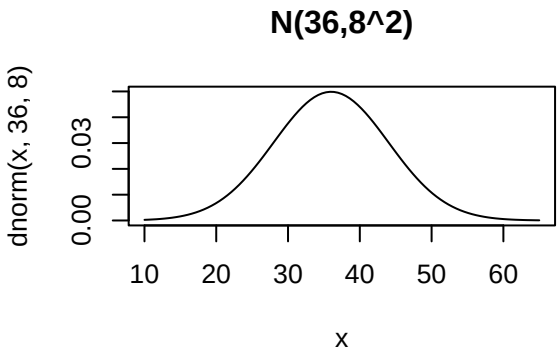
From above table we can see clearly that the sample mean and bootstrap sample mean are both nearly same and also the standard error for sampling distribution and bootstrap sampling distribution is also similar.

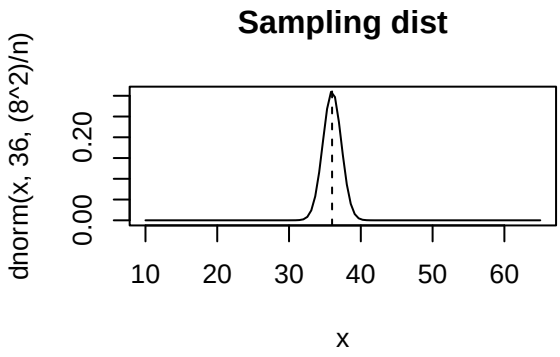
e) Repeat for sample sizes of n=50 and n=10.Carefully describe your observations about the effects of sample size on the bootstrap distribution.

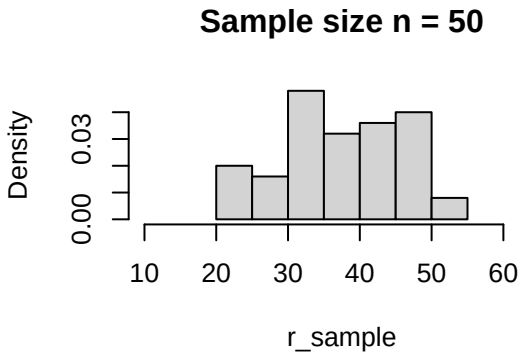
```
set.seed(143)

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)
}
```

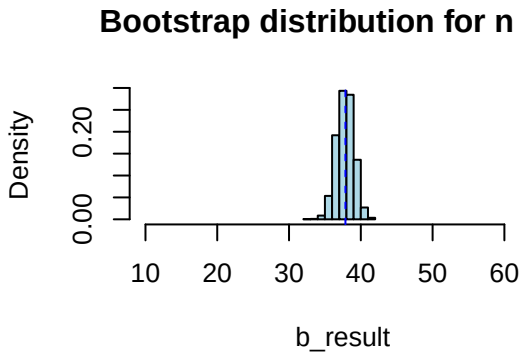
```
# For Sample size = 50
r_samplesizes(50)
```

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







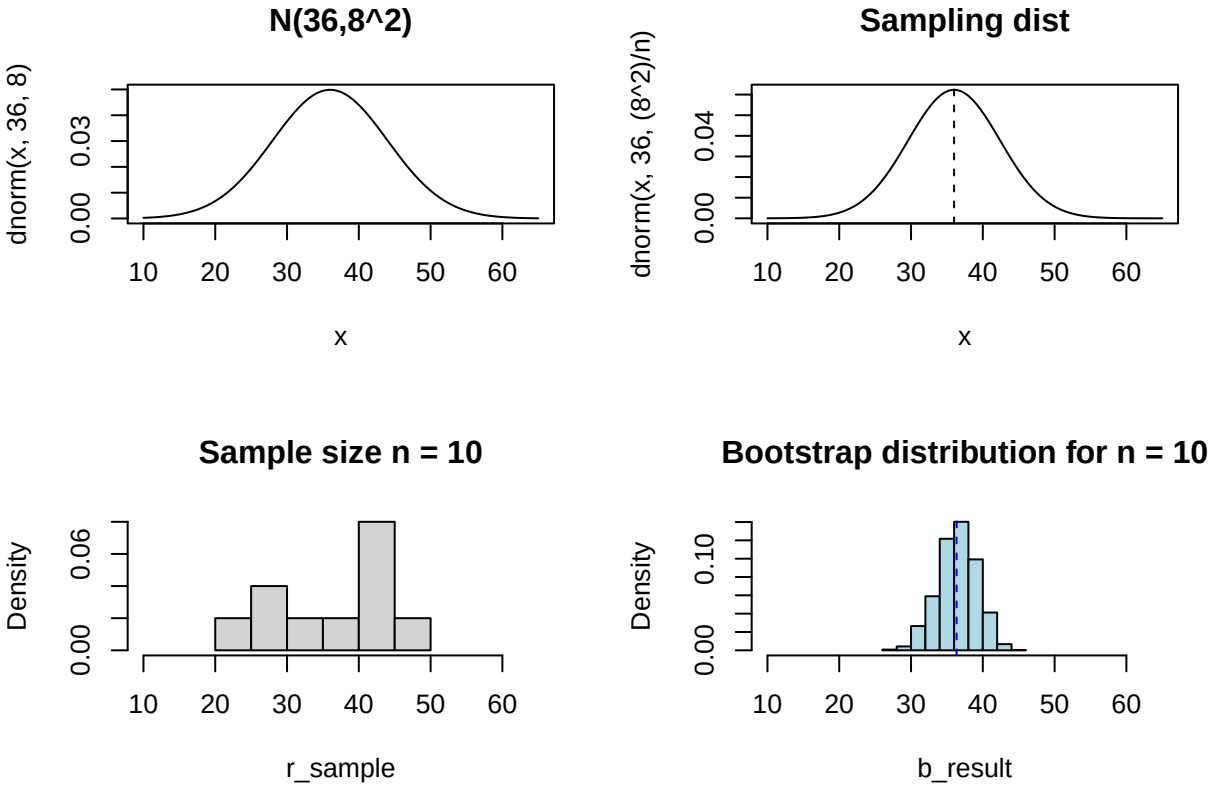



```
## The standard error for bootstrap sample is 1.205851
```

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()

cat("The summary statistics for Target store are",summary(Groceries$Target))
```

```
## The summary statistics for Target store are 0.99 1.8275 2.545 2.762333 3.14 7.99
```

```
cat("The summary statistics for Walmart store are",summary(Groceries$Walmart))
```

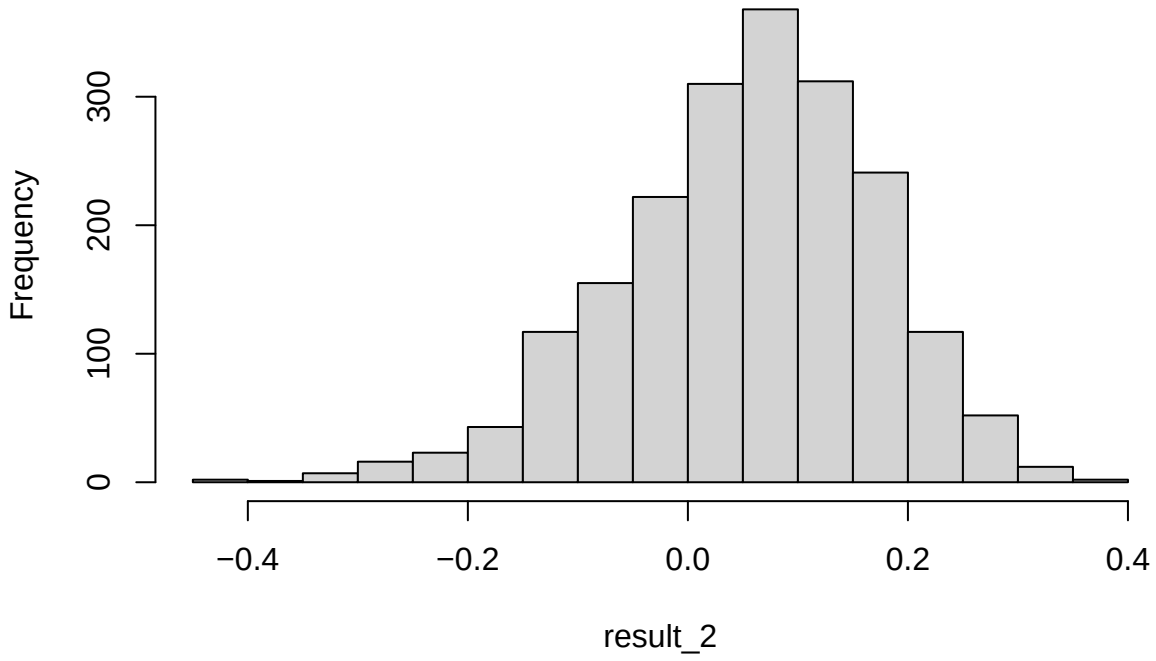
```
## The summary statistics for Walmart store are 1 1.76 2.34 2.705667 2.955 6.98
```

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
```

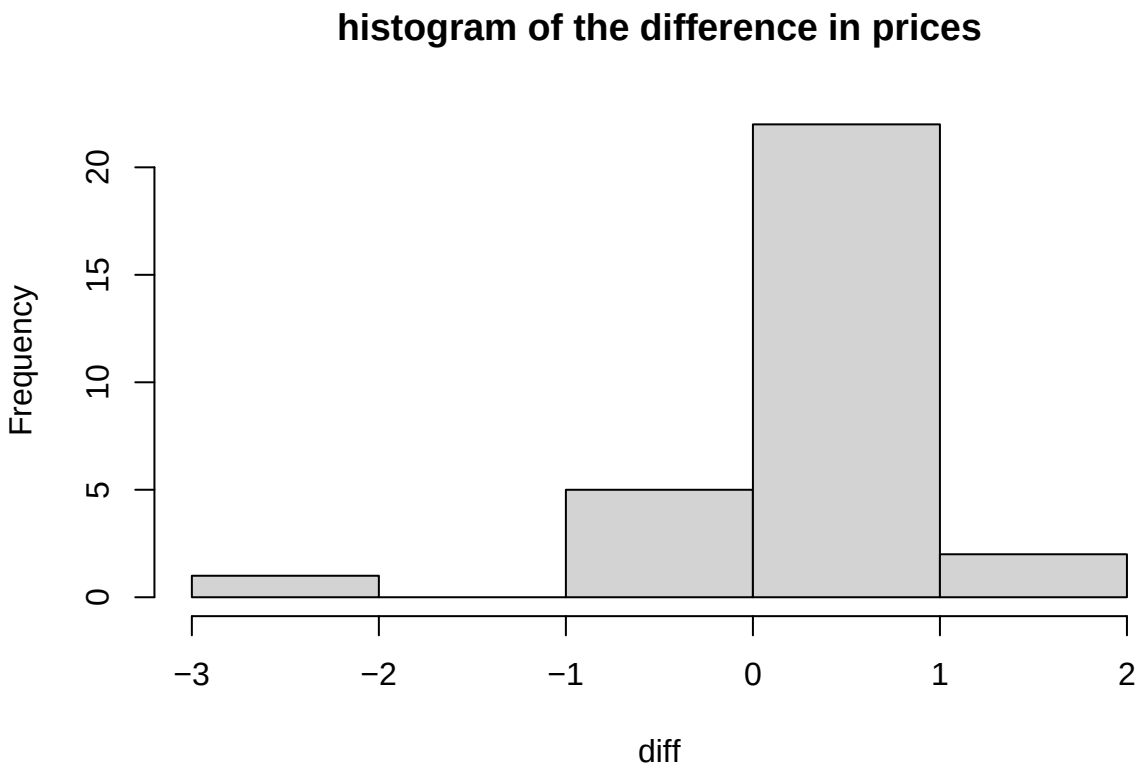
Answer:

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

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.

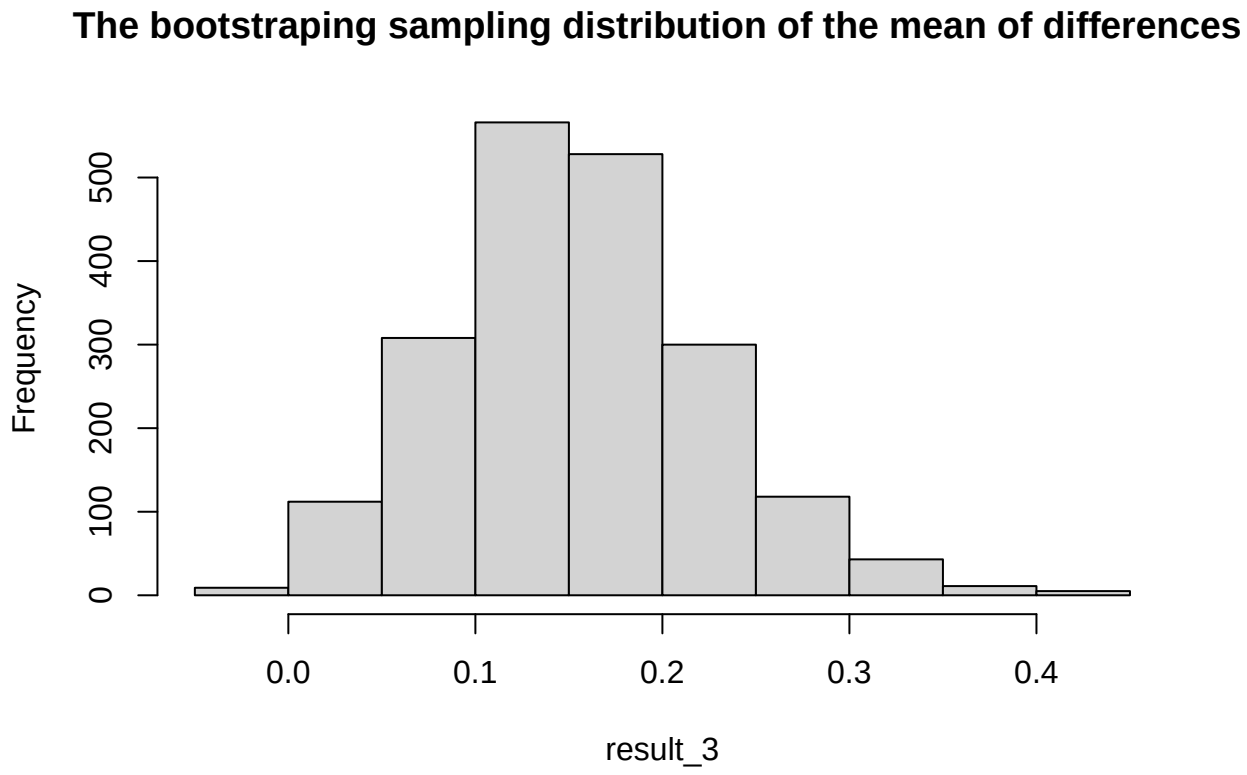
d) Recompute the bootstrap percentile interval without this observation. What do you conclude?

```
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)
```

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

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

```r
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.