QTM 220 HW #3

Author

Veronica Vargas

Exercise #1

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.3
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.3     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.4.3     ✓ tibble    3.2.1
✓ lubridate 1.9.2     ✓ tidyr     1.3.0
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
pokemon.sample <- read.csv("C:/Users/13015/OneDrive - Emory University/Documents/Fall 2024/QTM
         220/Pokemon.Sample.csv")
```

```
head(pokemon.sample)
```

```
  attack base_egg_steps base_happiness base_total capture_rate
1     50           5120              70        320          190
2    120           6400              70        455           45
3     53          30720               0        570           45
4     60           5120              70        305          255
5    103          30720             100        600           45
6     80           3840              70        455           90
      classfication defense height_m  hp      name percentage_male
1    Lonely Pokémon      95      0.4  50    Cubone              50
2   Kicking Pokémon      53      1.5  50 Hitmonlee             100
3  Parasite Pokémon      47      1.2 109  Nihilego              NA
4      Numb Pokémon      40      0.7  60     Numel              50
5 Gratitude Pokémon      75      0.2 100   Shaymin              NA
6       Bat Pokémon      70      1.6  75    Golbat              50
  pokedex_number sp_attack sp_defense speed    type1  type2 weight_kg
1            104        40         50    35   ground             6.5
2            106        35        110    87 fighting            49.8
3            793       127        131   103     rock poison      55.5
4            322        65         45    35     fire ground      24.0
5            492       120         75   127    grass  grass       2.1
6             42        65         75    90   poison flying      55.0
  generation is_legendary
1          1            0
2          1            0
3          7            1
4          3            0
5          4            1
6          1            0
```

```
summary(pokemon.sample)
```

```
     attack       base_egg_steps  base_happiness    base_total
 Min.   : 30.00   Min.   : 2560   Min.   :  0.0   Min.   :195.0
 1st Qu.: 60.00   1st Qu.: 5120   1st Qu.: 70.0   1st Qu.:338.8
 Median : 80.00   Median : 5120   Median : 70.0   Median :474.0
 Mean   : 81.08   Mean   : 7091   Mean   : 63.3   Mean   :442.4
 3rd Qu.:100.00   3rd Qu.: 5440   3rd Qu.: 70.0   3rd Qu.:507.8
 Max.   :160.00   Max.   :30720   Max.   :100.0   Max.   :680.0

 capture_rate       classfication         defense          height_m
 Length:100         Length:100         Min.   : 15.00   Min.   :0.10
 Class :character   Class :character   1st Qu.: 51.50   1st Qu.:0.50
 Mode  :character   Mode  :character   Median : 70.00   Median :0.90
                                       Mean   : 70.93   Mean   :1.11
                                       3rd Qu.: 81.25   3rd Qu.:1.50
                                       Max.   :180.00   Max.   :5.80
                                                        NA's   :7
      hp             name           percentage_male  pokedex_number
 Min.   :  1.00   Length:100         Min.   :  0.00   Min.   :  8.0
 1st Qu.: 55.00   Class :character   1st Qu.: 50.00   1st Qu.:185.5
 Median : 68.00   Mode  :character   Median : 50.00   Median :392.0
 Mean   : 71.39                      Mean   : 54.04   Mean   :401.2
 3rd Qu.: 86.00                      3rd Qu.: 50.00   3rd Qu.:626.5
 Max.   :150.00                      Max.   :100.00   Max.   :800.0
                                     NA's   :18
   sp_attack        sp_defense         speed           type1
 Min.   : 10.00   Min.   : 25.00   Min.   : 10.00   Length:100
 1st Qu.: 55.00   1st Qu.: 55.00   1st Qu.: 50.00   Class :character
 Median : 70.00   Median : 65.50   Median : 78.00   Mode  :character
 Mean   : 74.51   Mean   : 69.67   Mean   : 74.85
 3rd Qu.: 95.00   3rd Qu.: 80.00   3rd Qu.: 97.25
 Max.   :137.00   Max.   :131.00   Max.   :150.00

    type2            weight_kg        generation    is_legendary
 Length:100         Min.   :  0.20   Min.   :1.00   Min.   :0.00
 Class :character   1st Qu.: 10.90   1st Qu.:2.00   1st Qu.:0.00
 Mode  :character   Median : 28.50   Median :4.00   Median :0.00
                    Mean   : 48.96   Mean   :3.71   Mean   :0.09
                    3rd Qu.: 51.50   3rd Qu.:5.00   3rd Qu.:0.00
                    Max.   :291.00   Max.   :7.00   Max.   :1.00
                    NA's   :7
```

New Estimator

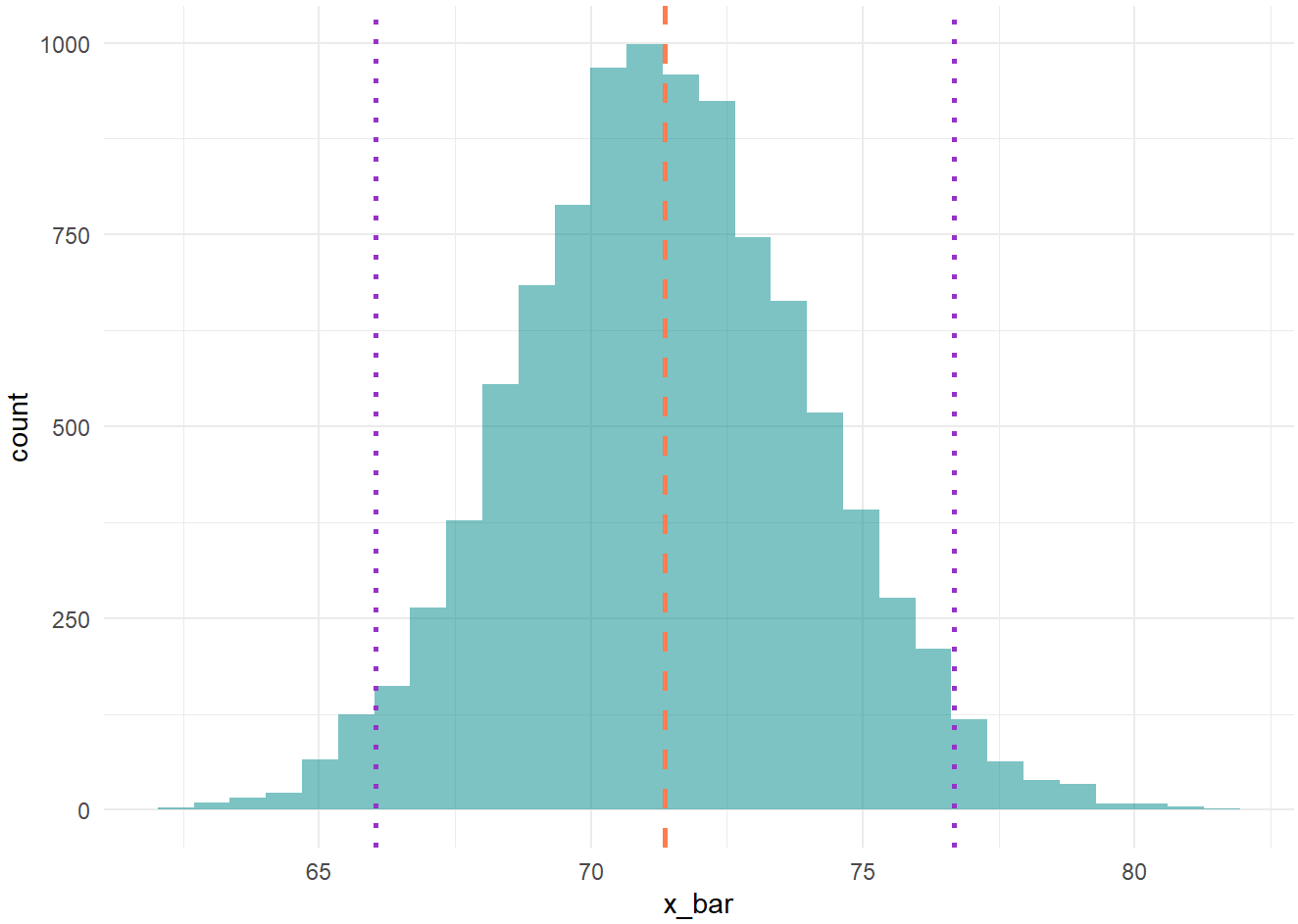
(c) Bootstrap Estimated Sampling Distribution

```
set.seed(42)

n <- 10000
x_bar <-  rep(NA, n)

for(i in 1:n){
  sampled.hp <- sample(pokemon.sample$hp, length(pokemon.sample$hp), replace = T)
  x_bar[i] <- mean(sampled.hp)
}
```

```
ggplot(data = data.frame(x_bar = x_bar), aes(x = x_bar)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = "identity") +
  geom_vline(xintercept = mean(x_bar), linetype="dashed",  #x_bar mean
                color = "coral", linewidth=1) +
  geom_vline(xintercept = mean(x_bar) + (1.96 * sd(x_bar)), linetype = 'dotted',
                color = "darkorchid", linewidth = 1) + # plus 1.96 stdev
```

```
    geom_vline(xintercept = mean(x_bar) - (1.96 * sd(x_bar)), linetype = "dotted",
               color = "darkorchid", linewidth=1) + # minus 1.96 stdev
  theme_minimal()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
lower.bound <- mean(x_bar) - (1.96 * sd(x_bar))
upper.bound <- mean(x_bar) + (1.96 * sd(x_bar))

print(paste0("The Bootstrapped 95% CI is {", lower.bound,", ",upper.bound,"}"))
```

```
[1] "The Bootstrapped 95% CI is {66.0389322577152, 76.6707877422848}"
```

(d) Population Estimated Sampling Distribution

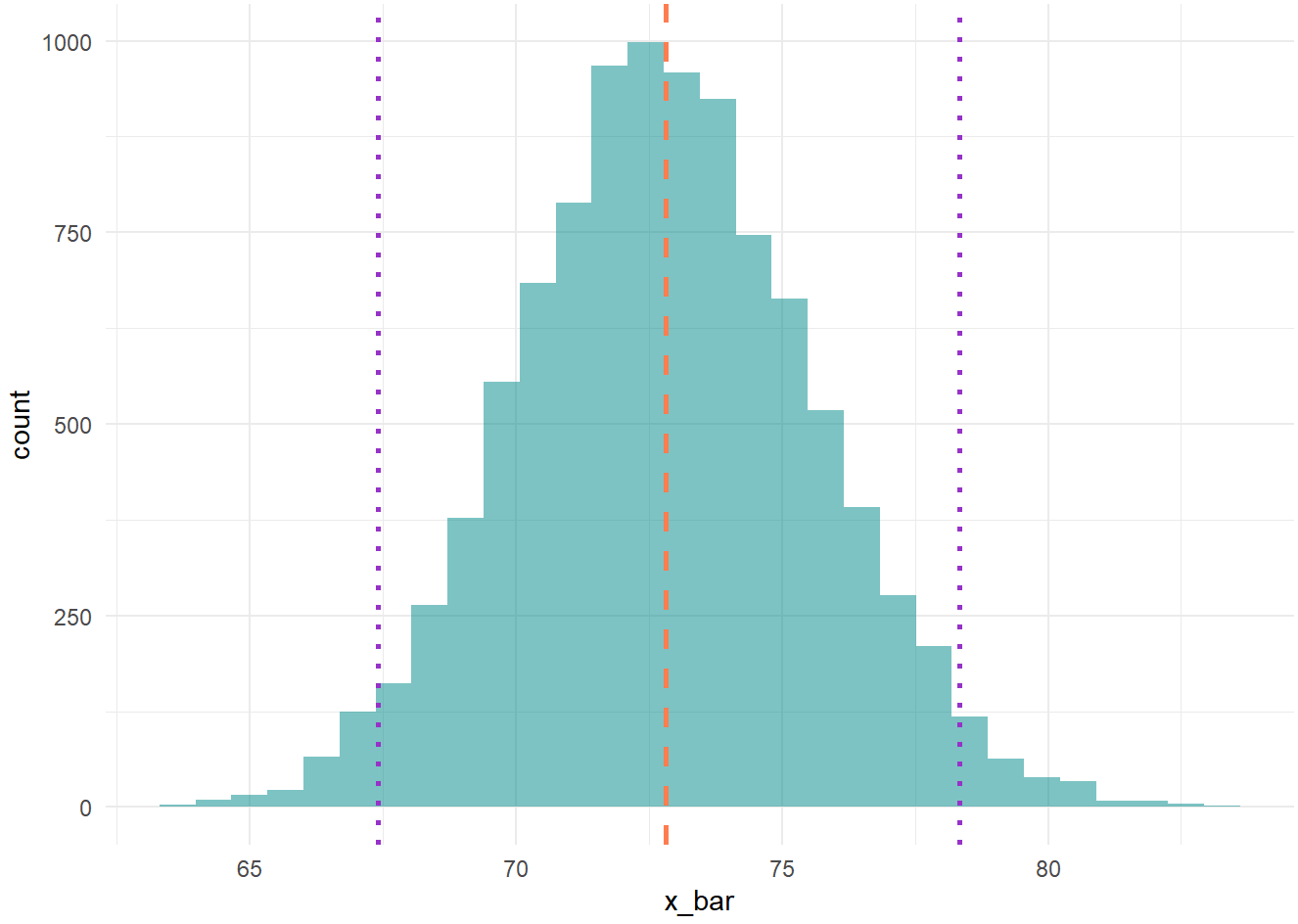
If we had the population, we could calculate the true sampling distribution of the estimator on the population. We would not know whether the expected value of the estimator would be equal to the estimand in the population.

Exercise #2

New Estimator

(c) Bootstrap Estimated Sampling Distribution

```
set.seed(42)

n <- 10000
x_bar <-  rep(NA, n)

for(i in 1:n){
  sampled.hp <- sample(pokemon.sample$hp, length(pokemon.sample$hp), replace = T)
  x_bar[i] <- (1/(length(sampled.hp) - 2)) * sum(sampled.hp)
}

ggplot(data = data.frame(x_bar = x_bar), aes(x = x_bar)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = "identity") +
  geom_vline(xintercept = mean(x_bar), linetype="dashed",  #x_bar mean
             color = "coral", linewidth=1) +
  geom_vline(xintercept = quantile(x_bar, 0.025), linetype = 'dotted',
             color = "darkorchid", linewidth = 1) + # plus 1.96 stdev
    geom_vline(xintercept = quantile(x_bar, 0.975), linetype = "dotted",
             color = "darkorchid", linewidth=1) + # minus 1.96 stdev
  theme_minimal()

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
lower.bound <- quantile(x_bar, 0.025)
upper.bound <- quantile(x_bar, 0.975)

print(paste0("The Bootstrapped 95% CI is {", lower.bound,", ",upper.bound,"}"))

[1] "The Bootstrapped 95% CI is {67.4081632653061, 78.3267857142857}"
```

(d) Comparing Estimator w/ Population

If I had the population and could plot the sampling distribution of my estimator it would still likely contain the estimand if I were to repeat the procedure. That said, since the interval is skewed, the bias from the estimator would reduce the likelihood in which the estimand would be found using our estimator.

Exercise #3

A Bimodal Situation

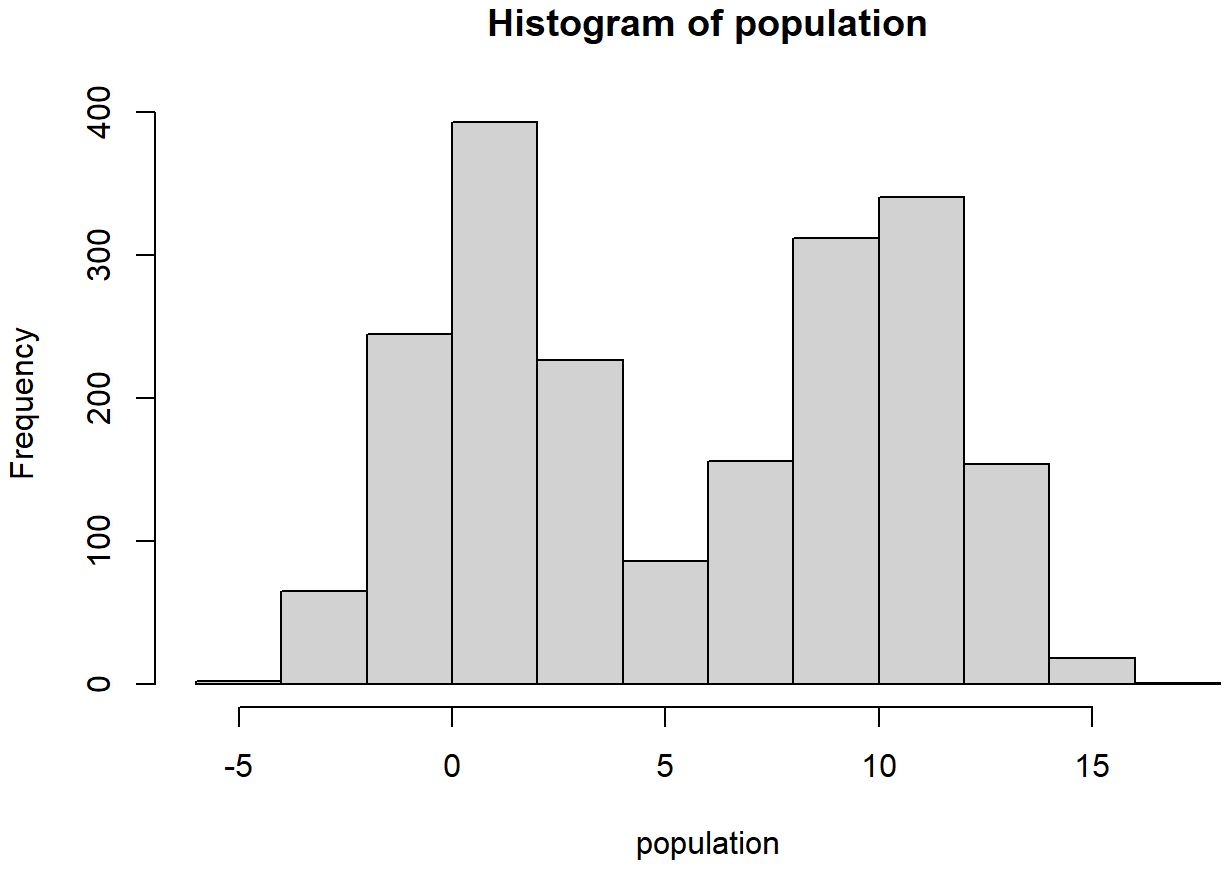
```
population <- c(rnorm(n = 1000, mean = 1, sd = 2), rnorm(n = 1000, mean = 10, sd = 2))
```

(a) Population Mean & Histogram

```
mean(population)
```

```
[1] 5.491986
```

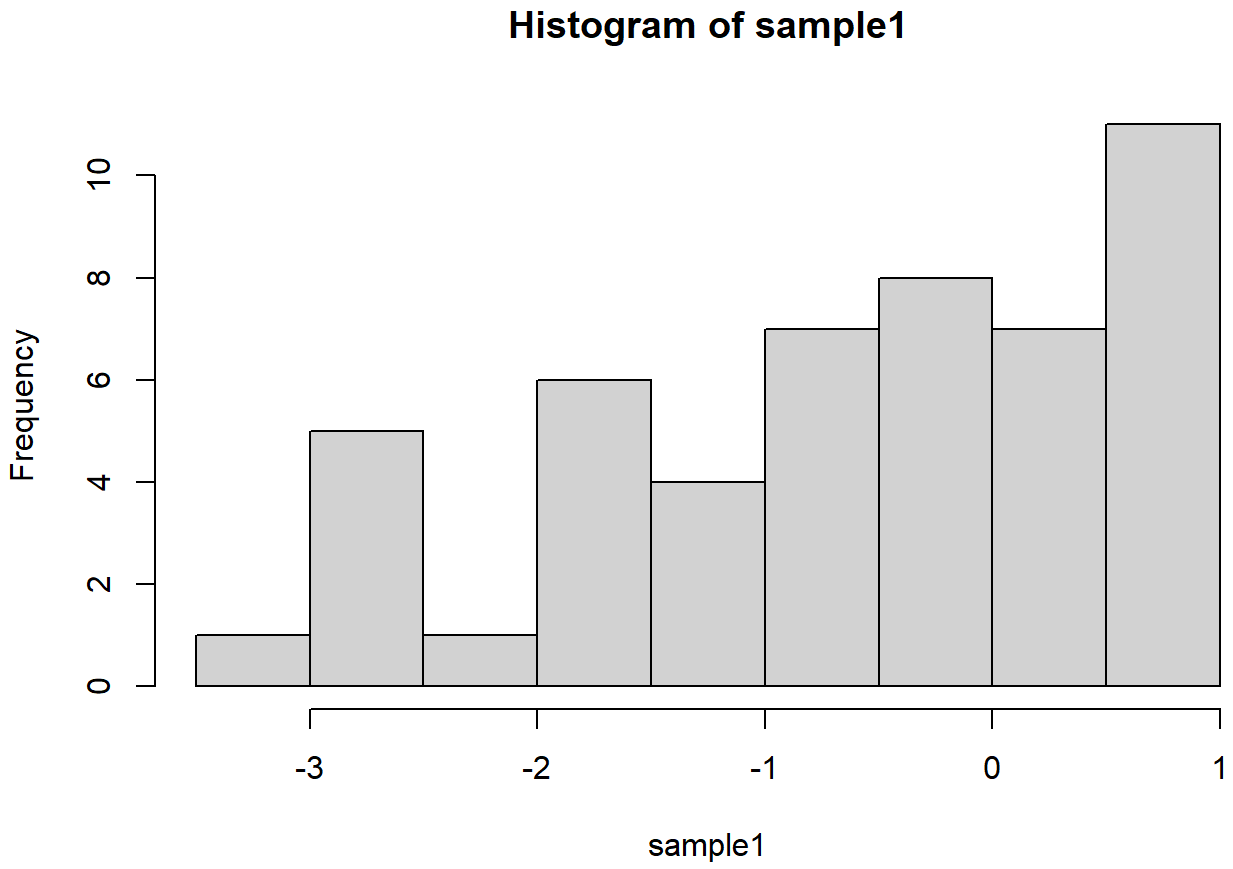
```
hist(population)
```



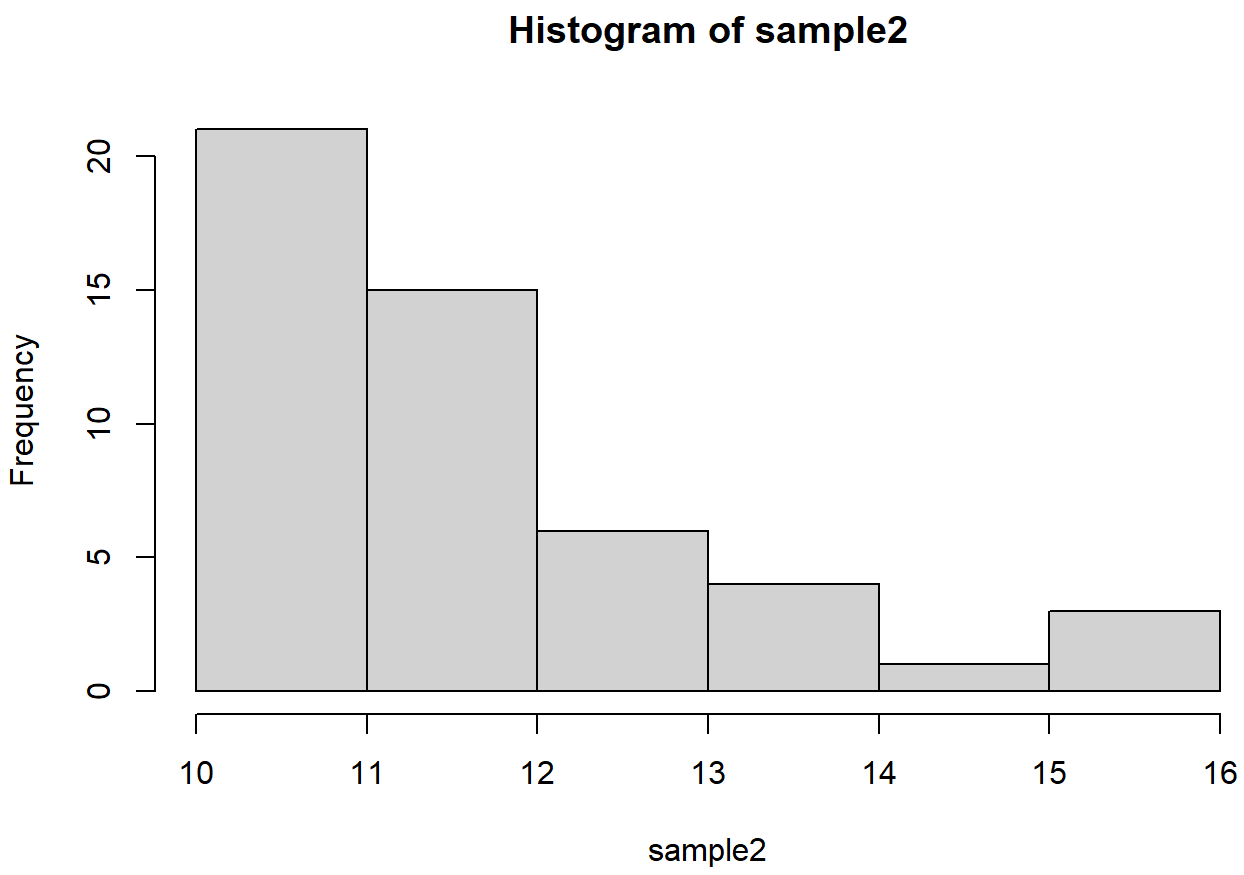
(b) Generating Samples w/ Histogram

```
sample1 <- sample(population[population <1], 50, replace = T )
sample2 <- sample(population[population > 10], 50, replace = T)
```

```
hist(sample1)
```




```
hist(sample2)
```



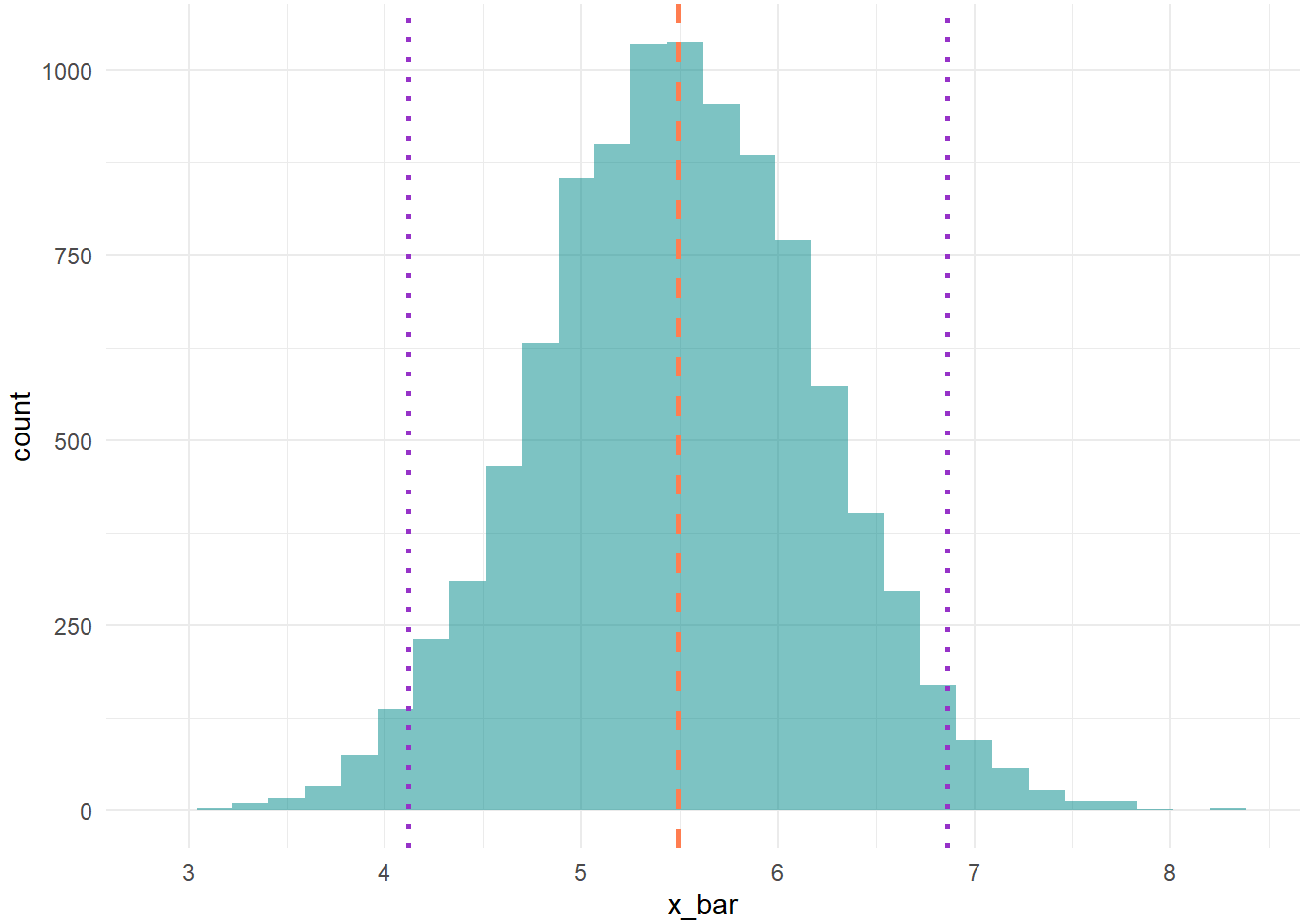
Both of these histograms are unimodal, unlike the bimodal histogram in part (a). This is not the usual way we sample.

(c) Population Sampling Distribution

```
set.seed(42)

n <- 10000
x_bar <-  rep(NA, n)

for(i in 1:n){
  sample <- sample(population, 50, replace = T)
  x_bar[i] <- mean(sample)
}
```

```
ggplot(data = data.frame(x_bar = x_bar), aes(x = x_bar)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = "identity") +
  geom_vline(xintercept = mean(x_bar), linetype="dashed",
                color = "coral", linewidth=1) +
  geom_vline(xintercept = quantile(x_bar, 0.025), linetype = 'dotted',
                color = "darkorchid", linewidth = 1) +
    geom_vline(xintercept = quantile(x_bar, 0.975), linetype = "dotted",
                color = "darkorchid", linewidth=1) +
  theme_minimal()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
lower.bound <- quantile(x_bar, 0.025)
upper.bound <- quantile(x_bar, 0.975)

print(paste0("The Bootstrapped 95% CI is {", lower.bound,", ",upper.bound,"}"))
```

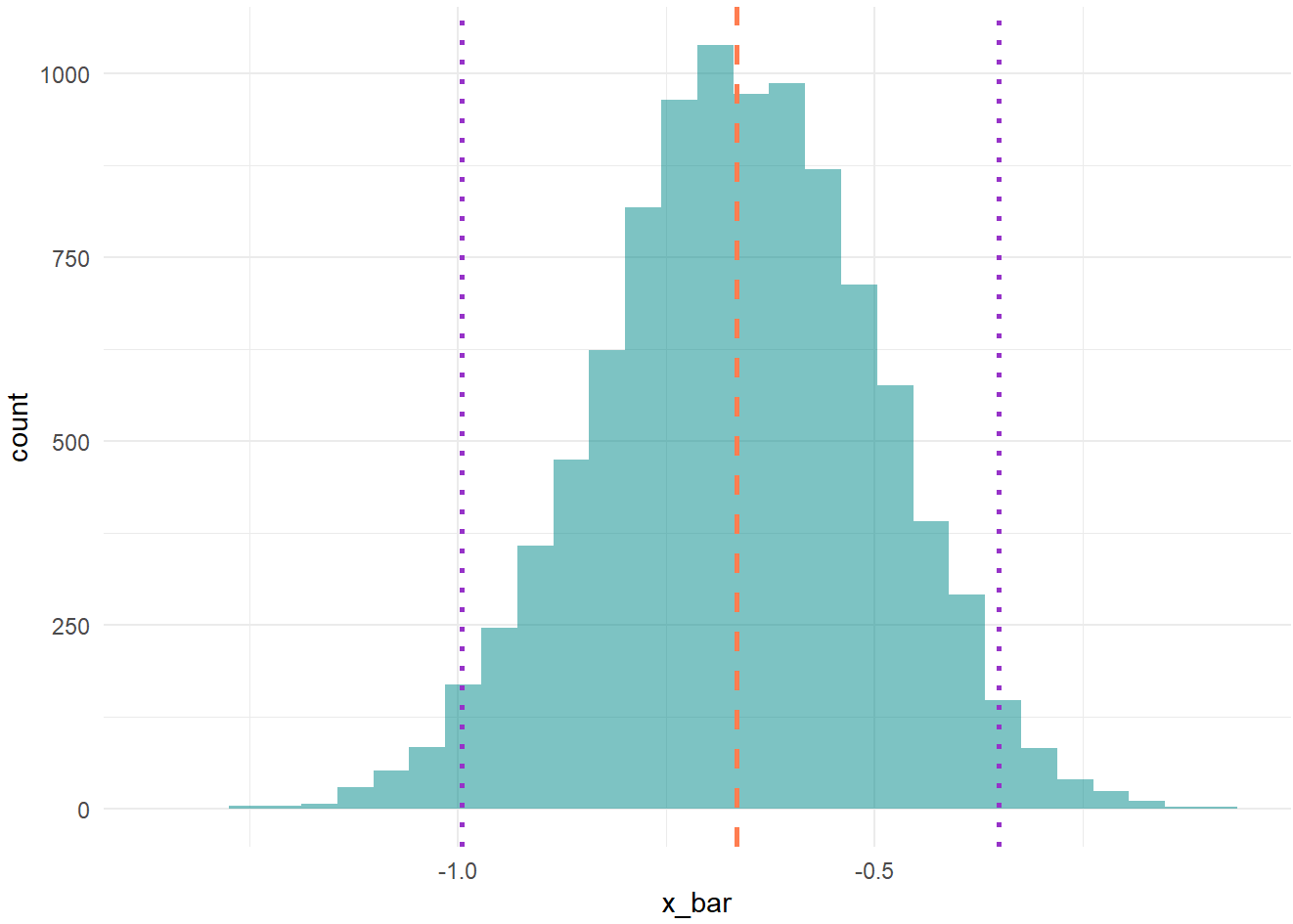
```
[1] "The Bootstrapped 95% CI is {4.11985850663474, 6.86410160372402}"
```

(d) Sample #1 Sampling Distribution

```
set.seed(42)

n <- 10000
x_bar <-  rep(NA, n)

for(i in 1:n){
  sample <- sample(sample1, 50, replace = T)
  x_bar[i] <- mean(sample)
}
```

```
ggplot(data = data.frame(x_bar = x_bar), aes(x = x_bar)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = "identity") +
  geom_vline(xintercept = mean(x_bar), linetype="dashed",
                color = "coral", linewidth=1) +
  geom_vline(xintercept = quantile(x_bar, 0.025), linetype = 'dotted',
                color = "darkorchid", linewidth = 1) +
    geom_vline(xintercept = quantile(x_bar, 0.975), linetype = "dotted",
                color = "darkorchid", linewidth=1) +
  theme_minimal()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
lower.bound <- quantile(x_bar, 0.025)
upper.bound <- quantile(x_bar, 0.975)

print(paste0("The Bootstrapped 95% CI is {", lower.bound,", ",upper.bound,"}"))
```

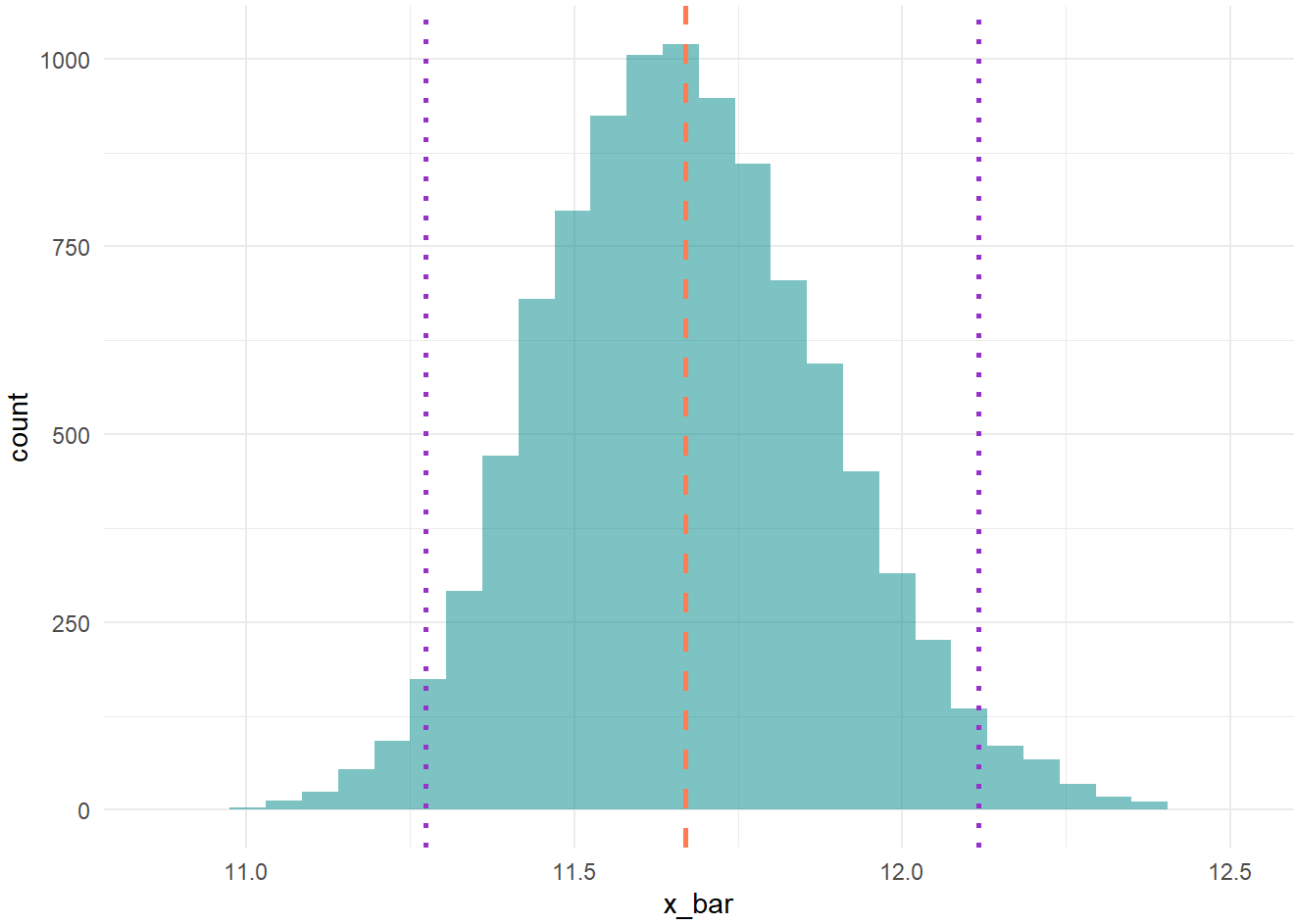
```
[1] "The Bootstrapped 95% CI is {-0.995374406226446, -0.350846463575224}"
```

(e) Sample #2 Sampling Distribution

```
set.seed(42)

n <- 10000
x_bar <-  rep(NA, n)

for(i in 1:n){
  sample <- sample(sample2, 50, replace = T)
  x_bar[i] <- mean(sample)
}

ggplot(data = data.frame(x_bar = x_bar), aes(x = x_bar)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = "identity") +
  geom_vline(xintercept = mean(x_bar), linetype="dashed",
                color = "coral", linewidth=1) +
  geom_vline(xintercept = quantile(x_bar, 0.025), linetype = 'dotted',
                color = "darkorchid", linewidth = 1) +
    geom_vline(xintercept = quantile(x_bar, 0.975), linetype = "dotted",
```

```
                color = "darkorchid", linewidth=1) +
  theme_minimal()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
lower.bound <- quantile(x_bar, 0.025)
upper.bound <- quantile(x_bar, 0.975)

print(paste0("The Bootstrapped 95% CI is {", lower.bound,", ",upper.bound,"}"))
```

```
[1] "The Bootstrapped 95% CI is {11.2735157574396, 12.1160365678011}"
```

(f) Comparing Histograms

Only the histogram in part (c) is centered with respect to the population mean. This is because it is taking samples from the population, whereas the other histograms are taking from the biased samples.

Exercise #4

Difference in Means

```
chihuahua.sample <- read.csv("C:/Users/13015/OneDrive - Emory University/Documents/Fall 2024/QTM
        220/Chihuahua.Sample.csv")

head(chihuahua.sample)
```

```
     sex   weight
1 female 4.425425
2 female 5.830472
3   male 7.681156
4 female 6.020102
5 female 5.526363
6 female 6.510369
```

```
summary(chihuahua.sample)
```

```
     sex                 weight      
 Length:150         Min.   :4.425  
 Class :character   1st Qu.:6.075  
 Mode  :character   Median :6.873  
                    Mean   :6.797  
                    3rd Qu.:7.507  
                    Max.   :8.831  
```

(a) Calculating Difference in Mean Weight

```
mean_female <- mean(chihuahua.sample$weight[chihuahua.sample$sex == "female"])
mean_male <- mean(chihuahua.sample$weight[chihuahua.sample$sex == "male"])

mean_diff <- mean_male - mean_female
mean_diff
```

```
[1] 1.50516
```

```
table(chihuahua.sample$sex)
```

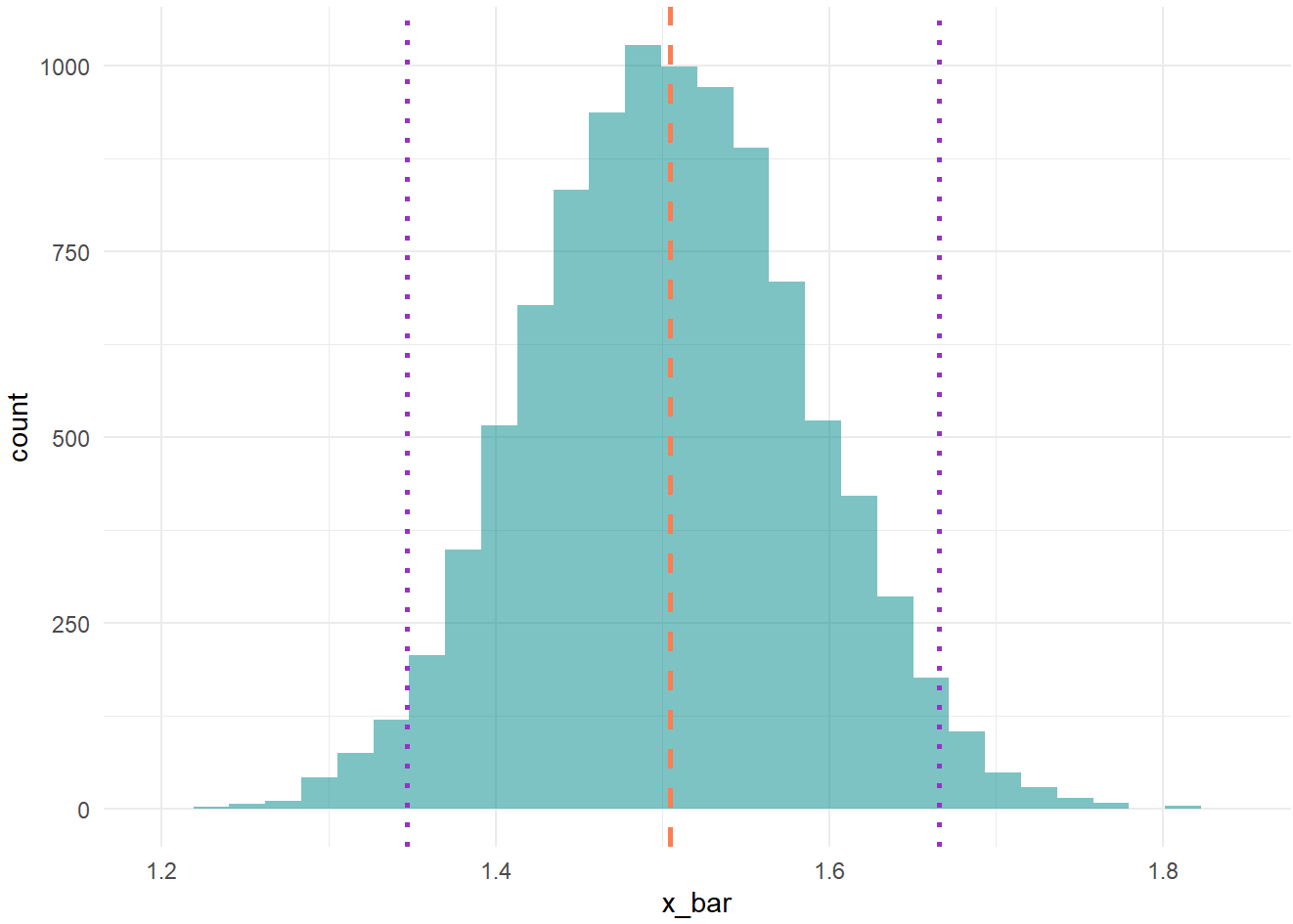
```
female   male 
    72     78 
```

(b) Creating 95% CI

```
set.seed(42)

n <- 10000
x_bar <-  rep(NA, n)

for(i in 1:n) {
  sample <- chihuahua.sample[sample(nrow(chihuahua.sample), nrow(chihuahua.sample), replace = TRUE),
         ]

  mean_female <- mean(sample$weight[sample$sex == "female"])
  mean_male <- mean(sample$weight[sample$sex == "male"])

  x_bar[i] <- mean_male - mean_female
}
```

```
ggplot(data = data.frame(x_bar = x_bar), aes(x = x_bar)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = "identity") +
  geom_vline(xintercept = mean(x_bar), linetype="dashed",
                color = "coral", linewidth=1) +
  geom_vline(xintercept = quantile(x_bar, 0.025), linetype = 'dotted',
                color = "darkorchid", linewidth = 1) +
    geom_vline(xintercept = quantile(x_bar, 0.975), linetype = "dotted",
                color = "darkorchid", linewidth=1) +
  theme_minimal()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
lower.bound <- quantile(x_bar, 0.025)
upper.bound <- quantile(x_bar, 0.975)

print(paste0("The Bootstrapped 95% CI is {", lower.bound,", ",upper.bound,"}"))

[1] "The Bootstrapped 95% CI is {1.34671924713184, 1.66557198219821}"
```

(c) Repeating w/ Alternative Sample

```
chihuahua.sample.alt <- read.csv("C:/Users/13015/OneDrive - Emory University/Documents/Fall 2024/QTM
        220/Alternate.Chihuahua.sample.csv")

head(chihuahua.sample.alt)

   sex   weight
1 male 6.860251
2 male 7.591237
3 male 7.139886
4 male 7.095968
5 male 7.591438
6 male 6.981313

summary(chihuahua.sample.alt)

     sex                weight
 Length:150         Min.   :4.680
 Class :character   1st Qu.:5.649
 Mode  :character   Median :6.097
```

```
                  Mean   :6.259
                  3rd Qu.:6.671
                  Max.   :8.621
```

```
mean_female <- mean(chihuahua.sample.alt$weight[chihuahua.sample.alt$sex == "female"])
mean_male <- mean(chihuahua.sample.alt$weight[chihuahua.sample.alt$sex == "male"])

mean_diff <- mean_male - mean_female
mean_diff
```

```
[1] 1.726183
```

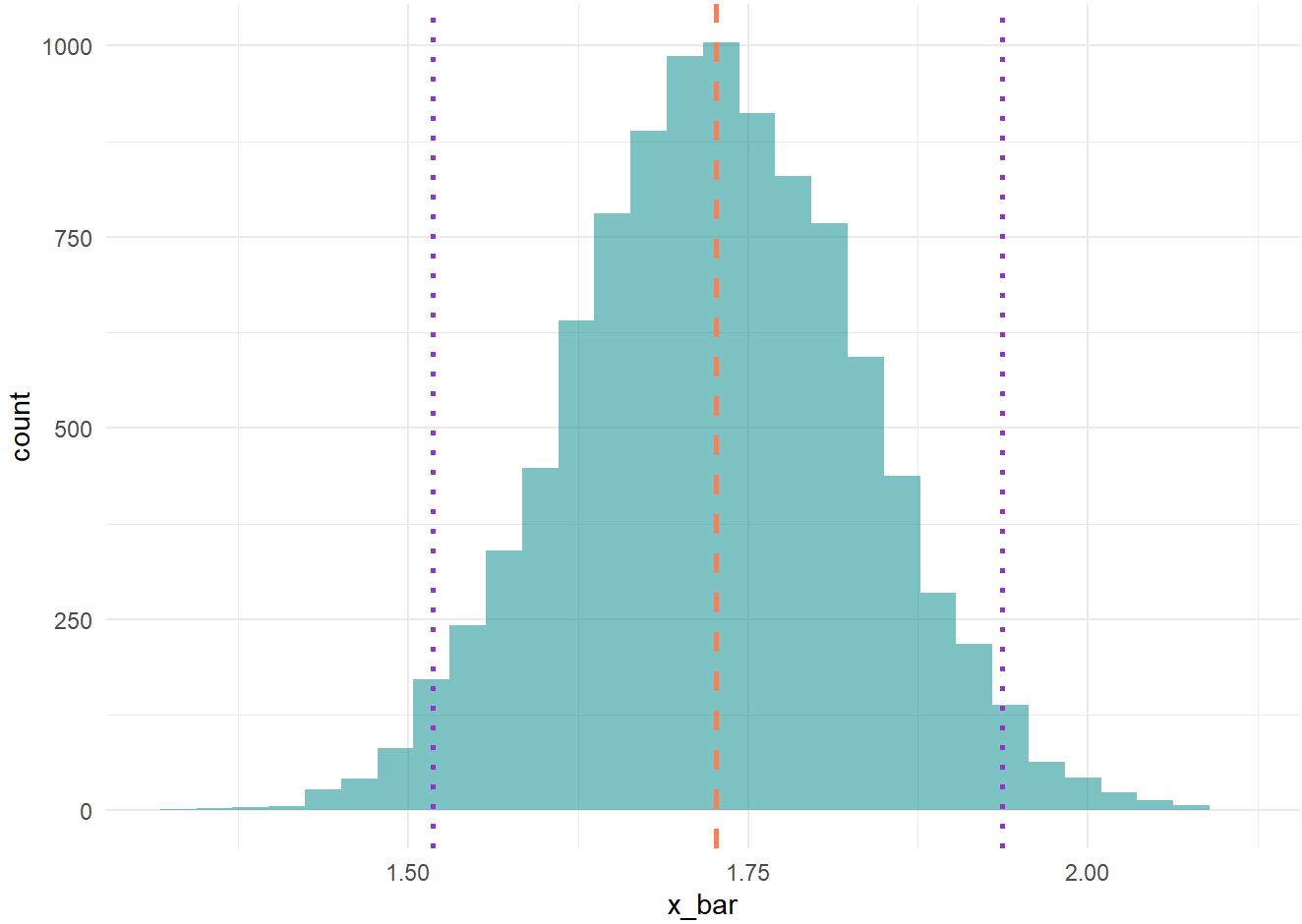
```
table(chihuahua.sample.alt$sex)
```

```
female   male
   120     30
```

```
set.seed(42)

n <- 10000
x_bar <-  rep(NA, n)

for(i in 1:n) {
  sample <- chihuahua.sample.alt[sample(nrow(chihuahua.sample.alt), nrow(chihuahua.sample.alt),
         replace = TRUE), ]

  mean_female <- mean(sample$weight[sample$sex == "female"])
  mean_male <- mean(sample$weight[sample$sex == "male"])

  x_bar[i] <- mean_male - mean_female
}
```

```
ggplot(data = data.frame(x_bar = x_bar), aes(x = x_bar)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = "identity") +
  geom_vline(xintercept = mean(x_bar), linetype="dashed",
                color = "coral", linewidth=1) +
  geom_vline(xintercept = quantile(x_bar, 0.025), linetype = 'dotted',
                color = "darkorchid", linewidth = 1) +
    geom_vline(xintercept = quantile(x_bar, 0.975), linetype = "dotted",
                color = "darkorchid", linewidth=1) +
  theme_minimal()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
lower.bound <- quantile(x_bar, 0.025)
upper.bound <- quantile(x_bar, 0.975)

print(paste0("The Bootstrapped 95% CI is {", lower.bound,", ",upper.bound,"}"))
```

```
[1] "The Bootstrapped 95% CI is {1.51790712855283, 1.93714170810816}"
```

Comparing the original sample to the alternative sample, the mean of the original sample is larger than that of the alternative sample. Furthermore, the original sample features an almost balanced set between males and females whereas the alternative sample is majority female. Finally, the 95% CI is slightly smaller in the original sample and the interval is skewed to the right in the alternative sample. This is likely because there is more weight being placed on females, resulting in higher values for the difference in means estimator.

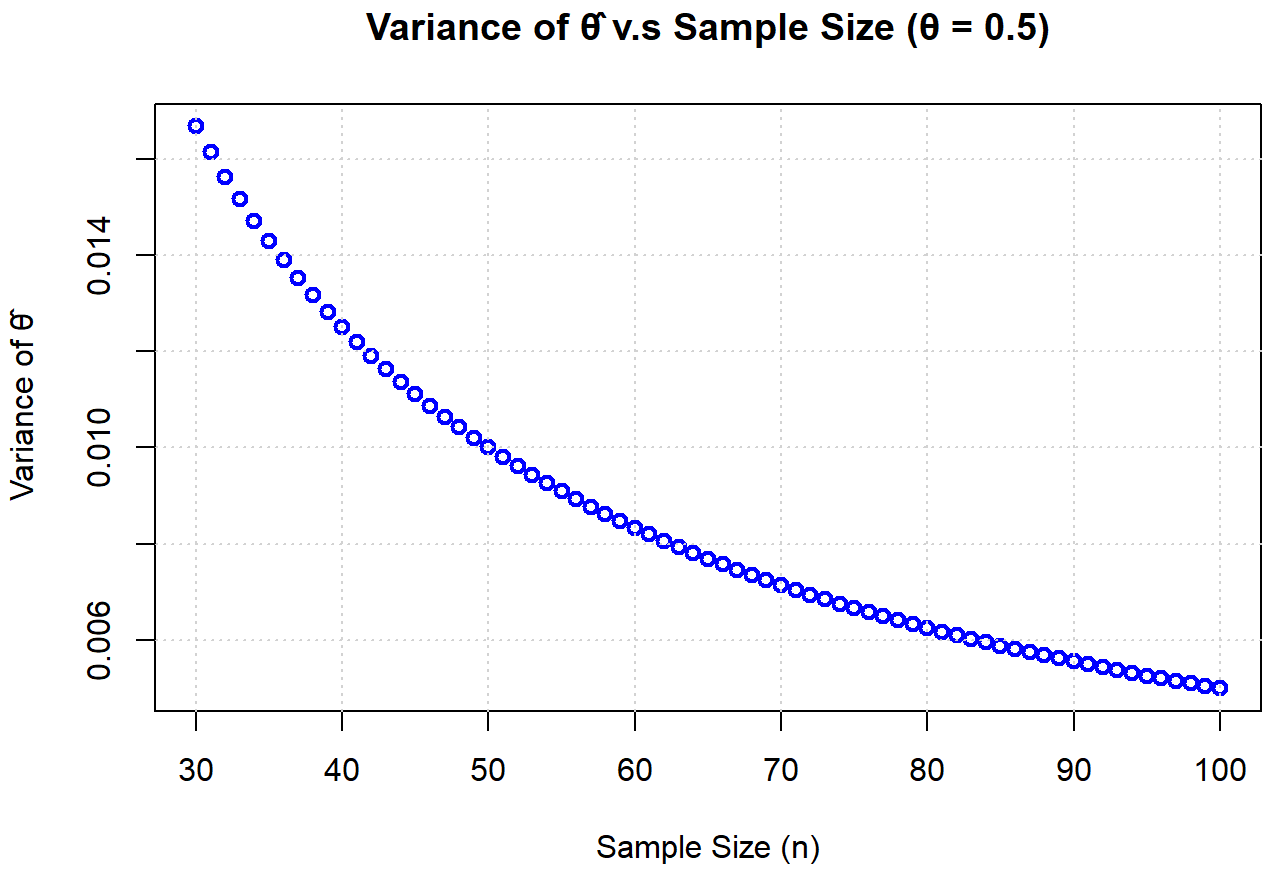
Exercise #5

Two Estimators

(d) Plotting Variance #1 vs. Sample Size

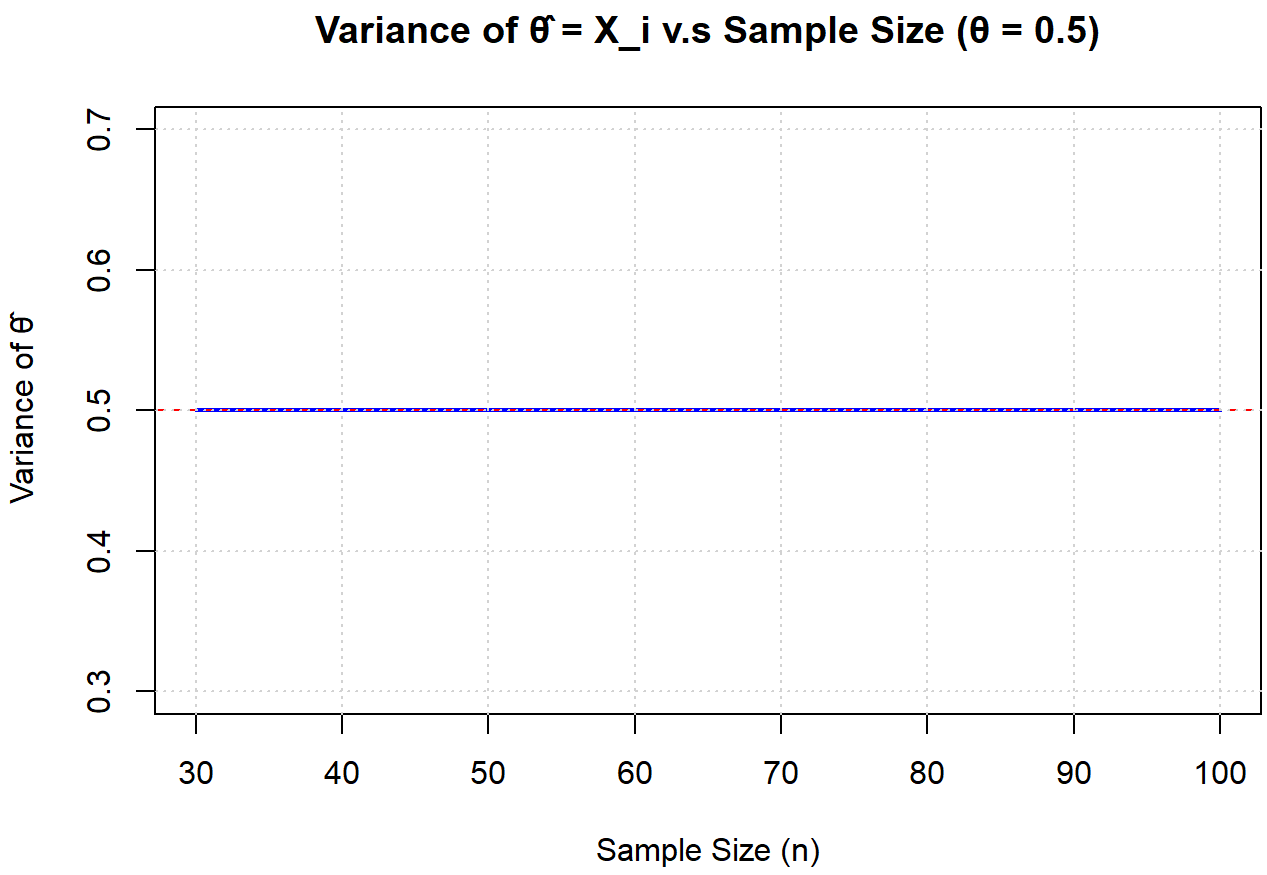
```
theta <- 0.5
n_values <- 30:100

var_theta_hat <- theta/n_values

plot(n_values, var_theta_hat, type = "b", col = "blue",
     xlab = "Sample Size (n)", ylab = "Variance of θ̂",
```

```
     main = "Variance of θˆv.s Sample Size (θ = 0.5)",
     lwd = 2)
grid()
```



(g) Plotting Variance #2 vs. Sample Size

```
theta <- 0.5
n_values <- 30:100

var_theta_hat <- rep(0.5, length(n_values))

plot(n_values, var_theta_hat, type = "l", col = "blue",
     xlab = "Sample Size (n)", ylab = "Variance of θˆ",
     main = "Variance of θˆ= X_i v.s Sample Size (θ = 0.5)",
     lwd = 2)
abline(h = 0.5, col = "red", lty = 2)
grid()
```


Variance of θ̂ = X_i v.s Sample Size (θ = 0.5)
Variance of θ̂
0.3
0.4
0.5
0.6
0.7
30
40
50
60
70
80
90
100
Sample Size (n)