QTM 220 HW #2

Exercise #2

Estimated Sampling Distributions

```
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
```

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

```
head(nba.sample.data)
```

```
    X POS Team Age GP  W  L    Min  PTS FGM
1 354  PF  WAS  30 59 21 38  624.6  195  71
2  19  SG  GSW  23 82 44 38 2458.1 1675 550
3 427   F  MIA  19 15  7  8  204.5   82  28
4  55  SF  DET  34 59 15 44 1892.9 1273 430
5 368   F  NYK  25 42  9 33  473.6  172  67
6 132  PG  ORL  22 60 28 32 1551.8  781 277
```

```
summary(nba.sample.data)
```

```
       X             POS                Team                Age       
 Min.   :  5.0   Length:100         Length:100         Min.   :19.00  
 1st Qu.:101.8   Class :character   Class :character   1st Qu.:22.00  
 Median :274.0   Mode  :character   Mode  :character   Median :25.00  
 Mean   :261.5                                         Mean   :25.24  
 3rd Qu.:425.5                                         3rd Qu.:27.25  
 Max.   :537.0                                         Max.   :35.00  
       GP              W               L              Min        
 Min.   : 1.00   Min.   : 0.00   Min.   : 1.00   Min.   :   5.0  
 1st Qu.:24.50   1st Qu.:10.00   1st Qu.:12.75   1st Qu.: 213.6  
 Median :53.00   Median :24.00   Median :24.00   Median :1179.6  
 Mean   :47.25   Mean   :24.48   Mean   :22.77   Mean   :1100.9  
 3rd Qu.:68.00   3rd Qu.:38.00   3rd Qu.:32.00   3rd Qu.:1913.1  
 Max.   :82.00   Max.   :57.00   Max.   :51.00   Max.   :2746.4  
      PTS              FGM        
 Min.   :   0.00   Min.   :  0.00  
 1st Qu.:  84.25   1st Qu.: 28.75  
 Median : 371.50   Median :143.00  
 Mean   : 579.83   Mean   :210.22  
 3rd Qu.: 973.00   3rd Qu.:358.00  
 Max.   :1959.00   Max.   :707.00  
```

(a) Sampling Summary Statistics

```
mean(nba.sample.data$PTS)
```

```
[1] 579.83
```

```
sd(nba.sample.data$PTS)
```

```
[1] 561.8942
```

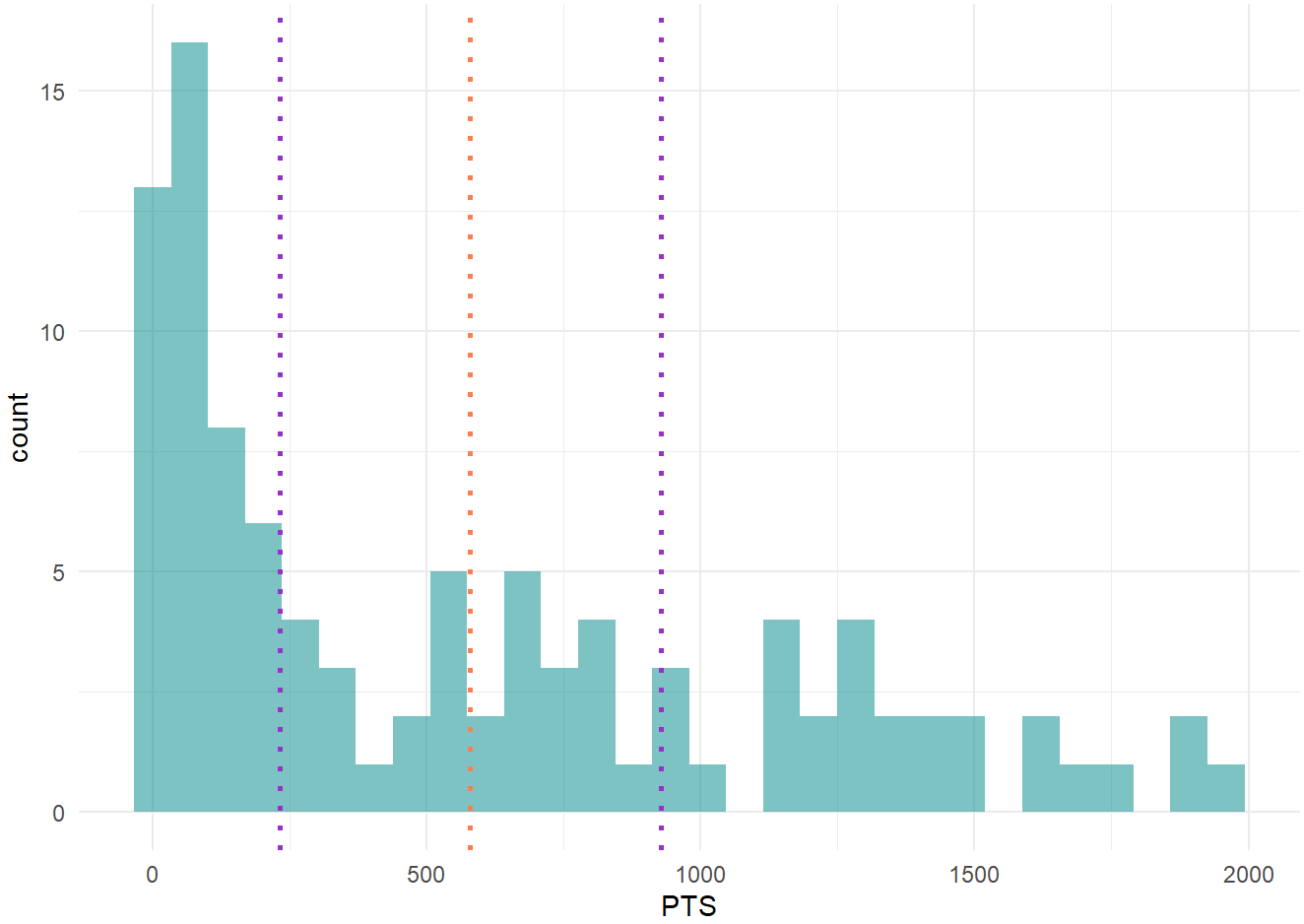
```
## plug-in method
sample.mean <- mean(nba.sample.data$PTS)
sample.sd <- sd(nba.sample.data$PTS)

est.mean.se <- sample.sd/sqrt(length(nba.sample.data))
q <- qnorm(1 - 0.05/2)
n <- length(nba.sample.data)

lower.bound <- sample.mean - q*est.mean.se
upper.bound <- sample.mean + q*est.mean.se
```

```
ggplot(data = nba.sample.data, aes(x = PTS)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = 'identity') +
   geom_vline(xintercept = sample.mean, linetype="dotted",
                color = "coral", linewidth=1) +
  geom_vline(xintercept = lower.bound, linetype = 'dotted',
                color = "darkorchid", linewidth = 1) +
    geom_vline(xintercept = upper.bound, linetype = "dotted",
                color = "darkorchid", linewidth=1) +
  theme_minimal()
```

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



(b) Calibrating 95% CI w/ Bootstrapped Sample

```
set.seed(42)

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

for(i in 1:n){
  sampled.PTS <- sample(nba.sample.data$PTS, 1000, replace = T)
  x_bar[i] <- mean(sampled.PTS)
}
```

```
mean(x_bar)
```

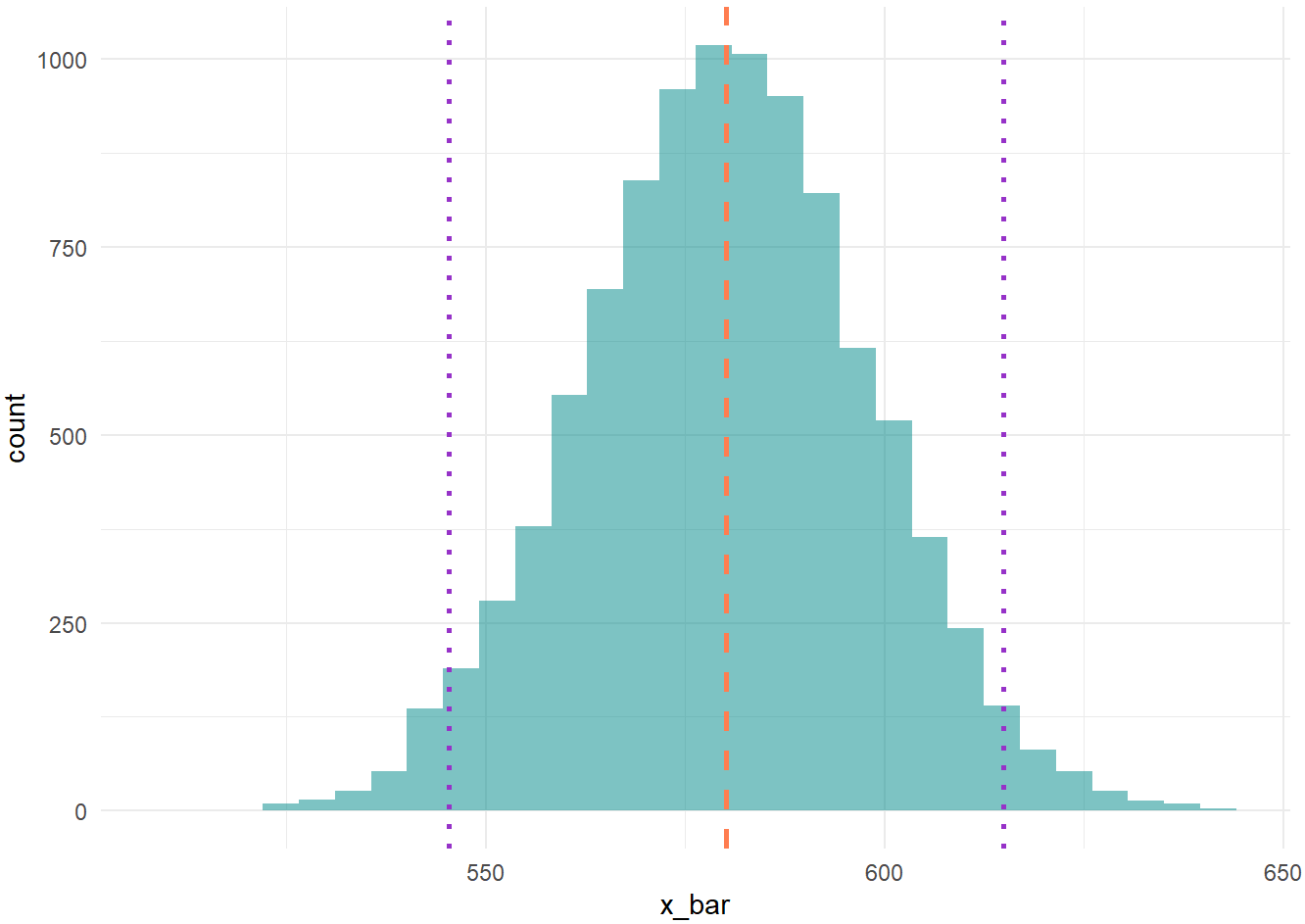
```
[1] 580.0755
```

```
sd(x_bar)
```

```
[1] 17.73286
```

```
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 {545.319095634198, 614.831902365802}"
```

(c) Calculating 90% CI, 95% CI, and 99% CI

```
## 90% CI
quantile(x_bar, c(0.05,0.95))
```

```
      5%      95% 
550.3322 608.9783 
```

```
## 95% CI
quantile(x_bar, c(0.05, 0.975))
```

```
      5%    97.5% 
550.3322 614.6186 
```

```
## 99% CI
quantile(x_bar, c(0.005, 0.995))
```

```
    0.5%    99.5% 
535.2928 626.4292 
```

As the interval increases as the CI gets larger. For instance, the interval for the 90% CI is smaller than the interval for the 95% CI and the interval for the 99% CI is larger than the interval or the 95% CI.

(d) Plug-In Method

```
sample.mean <- mean(nba.sample.data$PTS)
sample.sd <- sd(nba.sample.data$PTS)

est.mean.se <- sample.sd/sqrt(length(nba.sample.data))
q <- qnorm(1 - 0.05/2)
n <- length(nba.sample.data)

lower.bound <- sample.mean - q*est.mean.se
upper.bound <- sample.mean + q*est.mean.se

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

```
[1] "The Plug-in 95% CI is {231.570755002728, 928.089244997272}"
```

Here, the interval for the plug-in method is much wider than the bootstrap estimated version.

(e) Repeating w/ Alternate Sample

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

head(nba.sample.data.alt)
```

```
    X POS Team Age GP  W  L    Min  PTS FGM
1 491  SF  PHI  24  1  1  0   28.8   20   8
2  70  SG  WAS  29 50 24 26 1672.9 1160 444
3 453   G  BKN  25 15  7  8  158.4   44  18
4 318   C  PHX  30 61 33 28  874.1  263 119
5 435   F  LAC  21 22 10 12  195.5   59  24
6 167   G  SAS  19 66 15 51 1549.6  673 269
```

```
summary(nba.sample.data.alt)
```

```
       X             POS                Team                Age
 Min.   :  6.0   Length:100         Length:100         Min.   :19.00
 1st Qu.:117.8   Class :character   Class :character   1st Qu.:23.00
 Median :298.0   Mode  :character   Mode  :character   Median :25.00
 Mean   :278.2                                         Mean   :25.78
 3rd Qu.:444.0                                         3rd Qu.:28.00
 Max.   :539.1                                         Max.   :42.00
       GP              W               L              Min
 Min.   : 1.00   Min.   : 0.00   Min.   : 0.00   Min.   :   5.0
 1st Qu.:22.75   1st Qu.: 9.00   1st Qu.:11.50   1st Qu.: 203.5
 Median :49.50   Median :23.50   Median :23.50   Median : 894.1
 Mean   :44.22   Mean   :22.38   Mean   :21.84   Mean   :1064.0
 3rd Qu.:65.00   3rd Qu.:36.00   3rd Qu.:31.00   3rd Qu.:1915.5
 Max.   :82.00   Max.   :53.00   Max.   :58.00   Max.   :2841.5
      PTS              FGM
 Min.   :   0.0   Min.   :  0.0
 1st Qu.:  51.0   1st Qu.: 21.0
 Median : 297.5   Median :113.0
 Mean   : 541.0   Mean   :197.2
 3rd Qu.: 853.2   3rd Qu.:301.2
 Max.   :1946.0   Max.   :707.0
```

```
mean(nba.sample.data.alt$PTS)
```

```
[1] 540.98
```

```
sd(nba.sample.data.alt$PTS)
```

```
[1] 560.7308
```

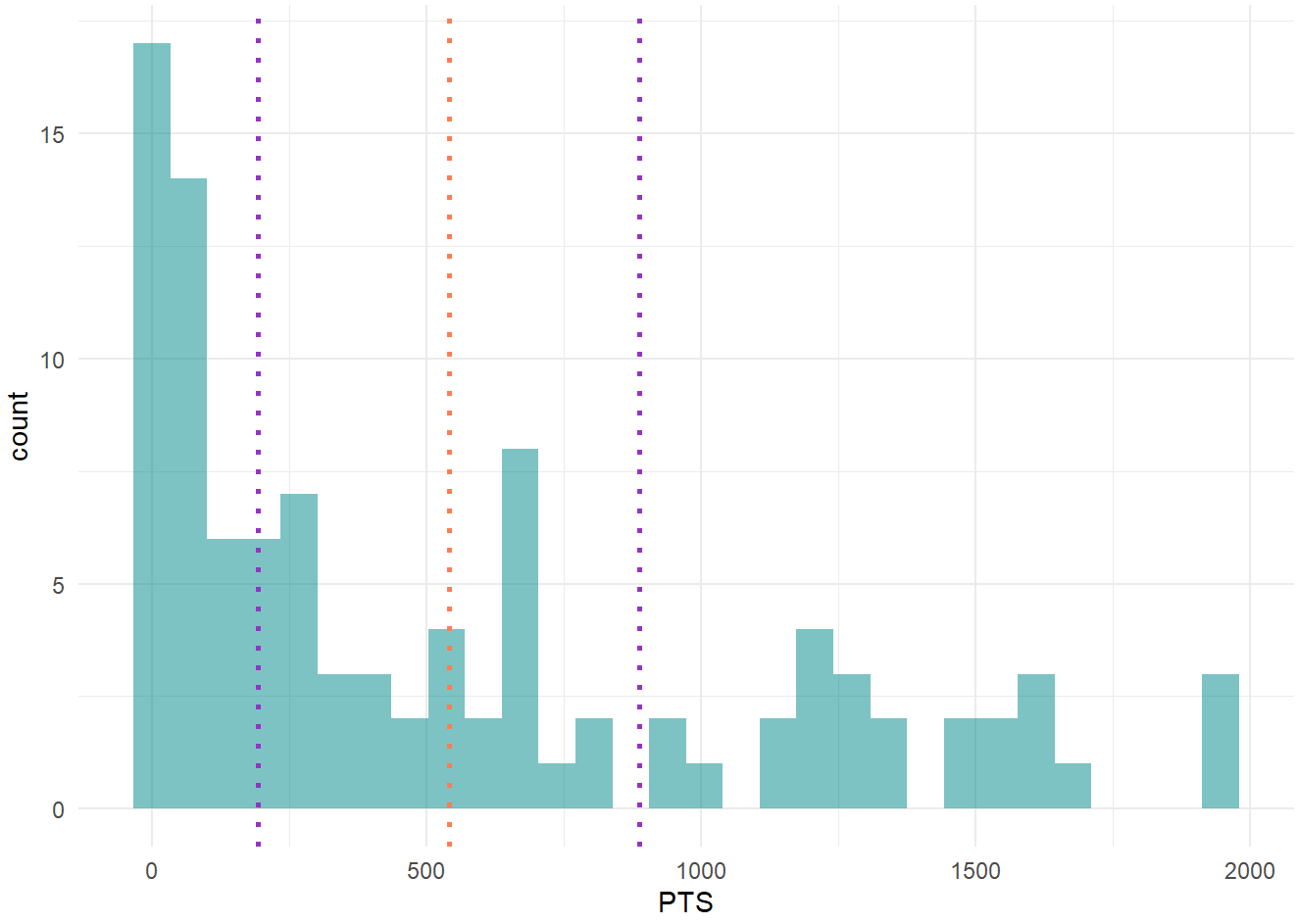
```
## plug-in method
sample.mean <- mean(nba.sample.data.alt$PTS)
sample.sd <- sd(nba.sample.data.alt$PTS)

est.mean.se <- sample.sd/sqrt(length(nba.sample.data.alt))
q <- qnorm(1 - 0.05/2)
n <- length(nba.sample.data.alt)

lower.bound <- sample.mean - q*est.mean.se
upper.bound <- sample.mean + q*est.mean.se
```

```
ggplot(data = nba.sample.data.alt, aes(x = PTS)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = 'identity') +
   geom_vline(xintercept = sample.mean, linetype="dotted",
              color = "coral", linewidth=1) +
  geom_vline(xintercept = lower.bound, linetype = 'dotted',
              color = "darkorchid", linewidth = 1) +
    geom_vline(xintercept = upper.bound, linetype = "dotted",
              color = "darkorchid", linewidth=1) +
  theme_minimal()
```

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



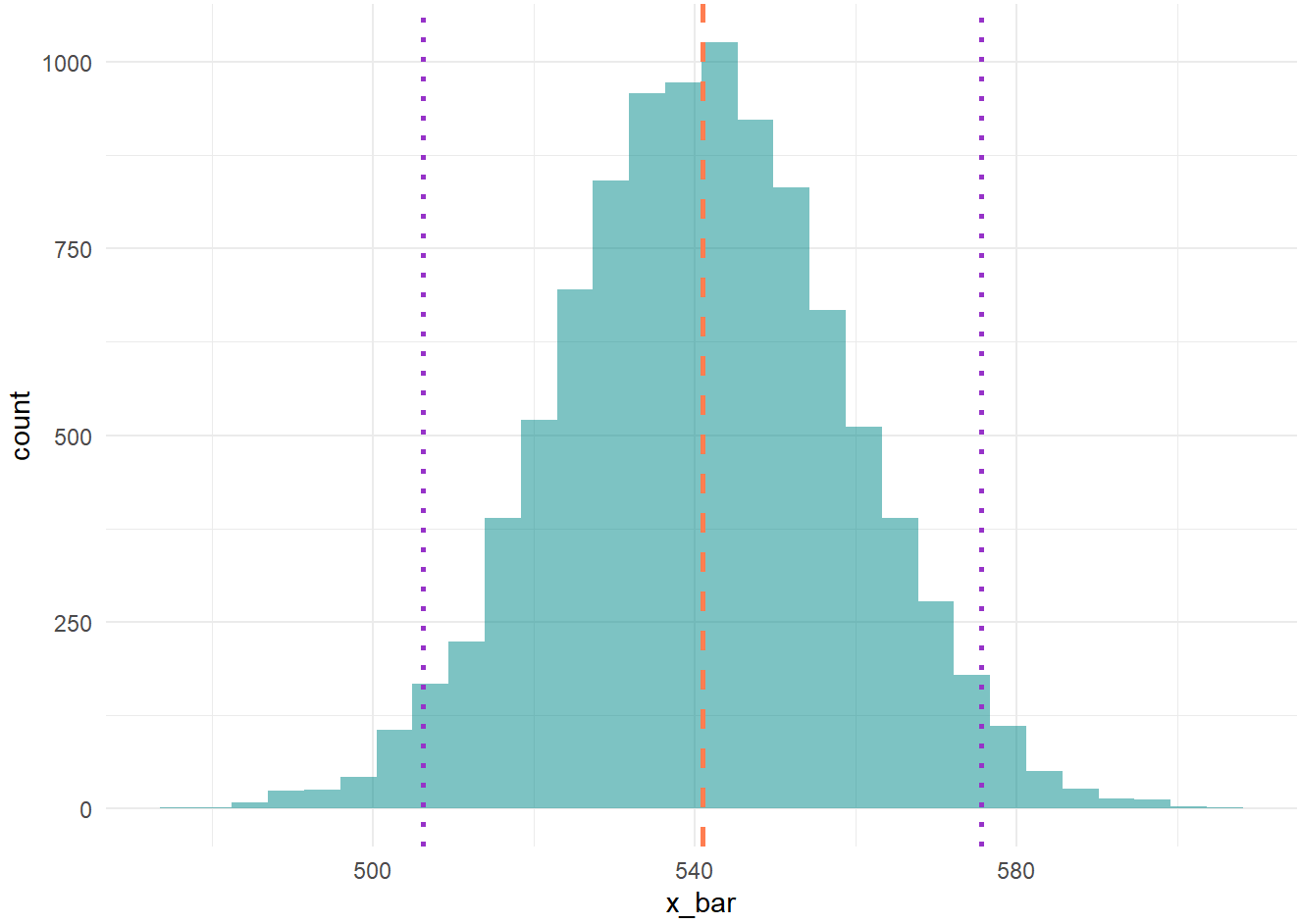
```
set.seed(42)

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

for(i in 1:n){
  sampled.PTS <- sample(nba.sample.data.alt$PTS, 1000, replace = T)
  x_bar[i] <- mean(sampled.PTS)
}

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 {506.270665662571, 575.693480737429}"
```

```
sample.mean <- mean(nba.sample.data.alt$PTS)
sample.sd <- sd(nba.sample.data.alt$PTS)

est.mean.se <- sample.sd/sqrt(length(nba.sample.data.alt))
q <- qnorm(1 - 0.05/2)
n <- length(nba.sample.data.alt)

lower.bound <- sample.mean - q*est.mean.se
upper.bound <- sample.mean + q*est.mean.se

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

```
[1] "The Plug-in 95% CI is {193.441846958591, 888.518153041409}"
```

For the mean in the original sample is larger than the mean in the alternative sample. Conversely, the standard deviation of the original sample is very similar that of the alternative sample. Finally, the 95% CI for both samples is roughly the same, but the internal is slightly skewed to the left for the alternative sample. This is likely because the mean of the alternative sample is smaller than the mean of the original sample.

Exercise #3

Comparing Sample and Population

```
nba.data <- read.csv("C:/Users/13015/OneDrive - Emory University/Documents/Fall 2024/QTM
         220/nba.data.csv")

head(nba.data)
```

```
  X                  Player POS Team Age GP  W  L    Min  PTS
1 1            Jayson Tatum  SF  BOS  25 74 52 22 2732.2 2225
2 2             Joel Embiid   C  PHI  29 66 43 23 2284.1 2183
3 3             Luka Doncic  PG  DAL  24 66 33 33 2390.5 2138
4 4 Shai Gilgeous-Alexander  PG  OKC  24 68 33 35 2416.0 2135
5 5   Giannis Antetokounmpo  PF  MIL  28 63 47 16 2023.6 1959
6 6         Anthony Edwards  SG  MIN  21 79 40 39 2841.5 1946
```

```
summary(nba.data)
```

```
       X           Player              POS                Team
 Min.   :  1.0   Length:539         Length:539         Length:539
 1st Qu.:135.5   Class :character   Class :character   Class :character
 Median :270.0   Mode  :character   Mode  :character   Mode  :character
 Mean   :270.0
 3rd Qu.:404.5
 Max.   :539.0
      Age              GP              W               L
 Min.   :19.00   Min.   : 1.00   Min.   : 0.00   Min.   : 0.00
 1st Qu.:23.00   1st Qu.:30.50   1st Qu.:12.00   1st Qu.:14.00
 Median :25.00   Median :54.00   Median :25.00   Median :25.00
 Mean   :25.97   Mean   :48.04   Mean   :24.02   Mean   :24.02
 3rd Qu.:29.00   3rd Qu.:68.00   3rd Qu.:36.00   3rd Qu.:34.00
 Max.   :42.00   Max.   :83.00   Max.   :57.00   Max.   :60.00
      Min              PTS
 Min.   :   1.0   Min.   :   0.0
 1st Qu.: 329.0   1st Qu.: 120.5
 Median : 970.2   Median : 374.0
 Mean   :1103.6   Mean   : 523.4
 3rd Qu.:1845.9   3rd Qu.: 769.5
 Max.   :2963.2   Max.   :2225.0
```

(a) Calculating Population Mean

```
mean(nba.data$PTS)
```

```
[1] 523.4267
```

```
sd(nba.data$PTS)
```

```
[1] 498.0844
```

(b) Population Sampling Distribution

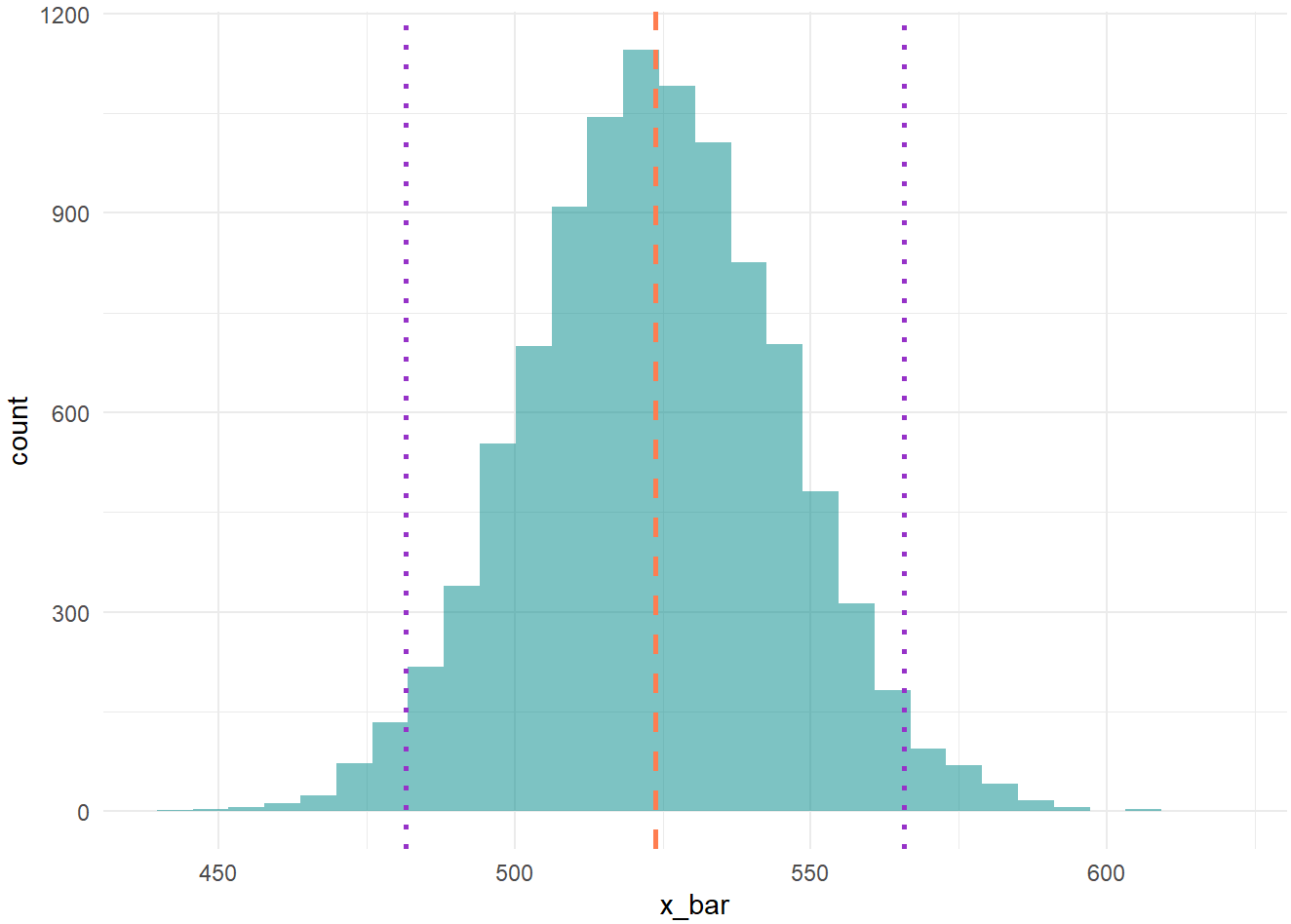
```
set.seed(42)

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

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

```
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 {481.634403647392, 565.822677985262}"
```

The width of the bootstrapped 95% CI from the population is much wider than the intervals created by either of the previous samples.