Lab 8 – t-Distribution and Bootstrapping

2024-11-04

Contents

Introduction

```r
library(ggplot2)
library(dplyr)

theme_set(theme_bw())

## Better histograms
gh <- function(bins = 8) {
  geom_histogram(color = 'black', fill = 'gray80', bins = bins)
}

## Bootstrap Function
bootstrap <- function(x, statistic, n = 1000L) {
  bs <- replicate(n, {
    sb <- sample(x, replace = TRUE)
    statistic(sb)
  })
  data.frame(Sample = seq_len(n),
             Statistic = bs)
}

## Standard Error
se <- function(x) sd(x) / sqrt(length(x))

## College dataset
college <- read.csv("https://collinn.github.io/data/college2019.csv")
```

Question 1 Using the `qnorm()` function, find the critical values associated with an 80% confidence interval. How do these compare to the critical values of a 95% confidence interval? Explain.

```r
## 95 has larger CV
qnorm(c(0.1, 0.9))
```

```
## [1] -1.2816  1.2816
```

```r
qnorm(c(0.025, 0.975))
```

```
## [1] -1.96  1.96
```

Question 2 This question will use the `hawks` data. The code below will load the `hawks` dataset which we will use for the first parts of this problem. It will also create `hawks2`, a randomly sampled subset of the data.

```
hawks <- read.csv("https://collinn.github.io/data/hawks.csv")
```

(a) Subset the `hawks` data using dplyr to include only Red-tailed hawks (`Species == "RT"`). How many observations are in this dataset?

(b) Create a histogram of the variable `Weight`. What does it look like? Based on your experience with the CLT lab, how well do you think a normal approximation will fit the sampling distribution?

(c) Find the sample mean and standard error for the `Weight` variable. Use the `qt()` function to find the appropriate quantiles to create a 95% confidence interval for the sample mean. Use these to construct a 95% confidence interval

(d) Repeat the previous step, this time using the `hawks2` subset created in the code chunk below. How does the sample mean in `hawks2` compare with the sample mean in `hawks`? How does the size of the 95% confidence interval compare? Does changing the size of the sample appear to have more of an impact on the sample mean or on the confidence interval?

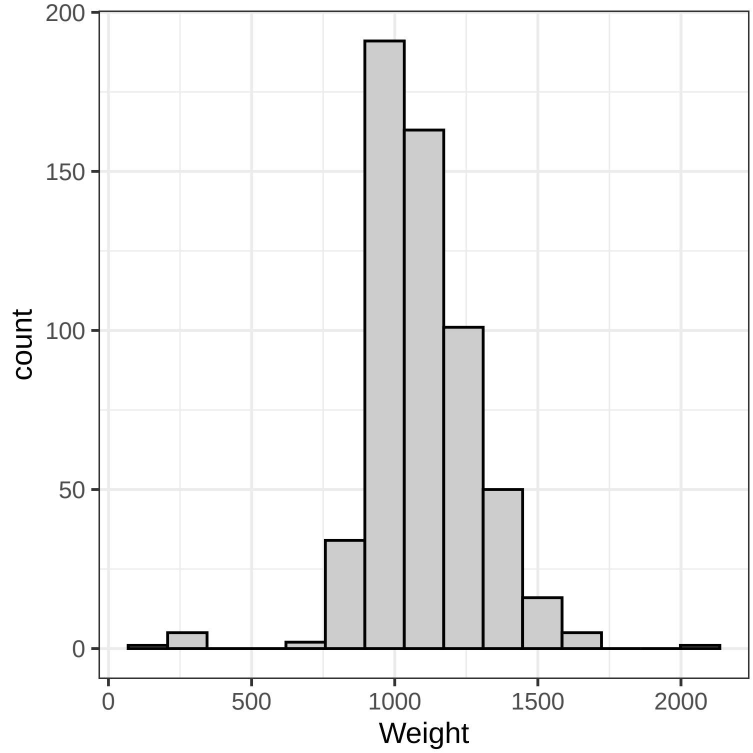
```
## First subset Hawks data with RT
# hwk <- filter(hawks, ...)

## Create subset of size n = 20
set.seed(89)
idx <- sample(seq_len(nrow(hwk)), size = 20)
hawks2 <- hwk[idx, ]
```

```
## Part A
hawks <- read.csv("https://collinn.github.io/data/hawks.csv")
hwk <- subset(hawks, Species == "RT")
dim(hwk)
```

```
## [1] 569  19
```

```
## Part B  (looks normal, probably fit alright)
ggplot(hwk, aes(Weight)) + gh(bins = 15)
```




```r
## Part C
mm <- mean(hwk$Weight)
ss <- se(hwk$Weight)
qt(c(0.025, 0.975), df = nrow(hwk)-1)
```

```
## [1] -1.9641  1.9641
```

```r
mm + c(-1.9641, 1.9641) * ss
```

```
## [1] 1079.1 1110.3
```

```r
## Part D
set.seed(89)
idx <- sample(seq_len(nrow(hwk)), size = 20)
hwk2 <- hwk[idx, ]

## Mean same, CI larger because n smaller
mm2 <- mean(hwk2$Weight)
ss2 <- se(hwk2$Weight)
qt(c(0.025, 0.975), df = nrow(hwk2)-1)
```

```
## [1] -2.093  2.093
```

```r
mm2 + c(-2.093, 2.093) * ss2
```

```
## [1] 1011.7 1177.7
```

Question 3 For this question, we are going to again use the `college` dataset

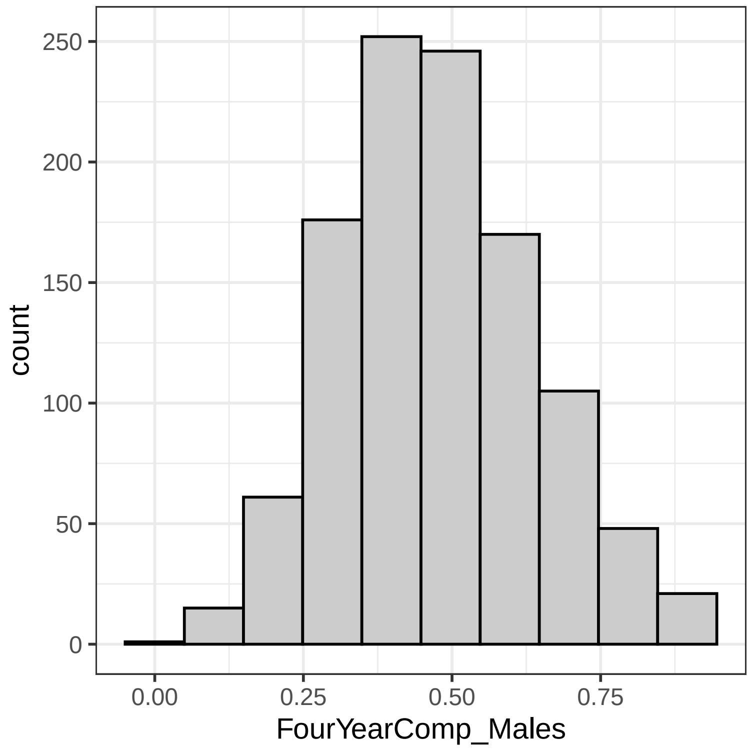
(a) Create a histogram of the variable `FourYearComp_Males`, the four year graduation rate for males

(b) Find the mean and *standard deviation* of the variable `FourYearComp_Males`. Use the `qnorm()` function to find the the 0.025 and 0.975 quantiles.

(c) Using the `quantile()` function the 0.025 and 0.975 quantiles of the variable `college$FourYearComp_Males`. Would you expect these to be similar to what you found in part (b)? Why or why not?

(d) Why are the quantiles you found in this problem so different than the ones we found in the example above using the same data? Be as specific as you can be.

```
## Part A
ggplot(college, aes(FourYearComp_Males)) + gh(bins = 10)
```




```
## These should be similar since the samp dist is probably normalish
pp <- c(0.025, 0.975)
qnorm(pp, mean(college$FourYearComp_Males), sd(college$FourYearComp_Males))
```

```
## [1] 0.15300 0.79944
```

```
quantile(college$FourYearComp_Males, pp)
```

```
##    2.5%   97.5%
## 0.18722 0.82688
```

```
## Will be different from above as these are quantiles for the actual variable
## rather than the sampling distribution (point is to differentiate between distributions)
```

Question 4 Based on the histogram of the bootstrapped sampling distribution above, do you think that the 95% confidence interval constructed with bootstrapping should match what we would find using the point estimate ± margin of error method? Explain your answer.

Yes because it looks approximately normal

Question 5 Verify using the `qt()` function what you answered in Question 4.

```
mm <- mean(USArrests$Murder)
ss <- se(USArrests$Murder)

mm + qt(c(0.025, 0.975), df = nrow(USArrests)-1)*ss
```

```
## [1] 6.5502 9.0258
```

Question 6 This question uses the Grinnell Rain dataset. Typically, this dataset includes precipitation data on 121 months; here, we will collect a sample of size $n = 20$ instead

```
## Load data
rain <- read.csv("https://collinn.github.io/data/grinnell_rain.csv")

## Subset
set.seed(10)
idx <- sample(1:nrow(rain), size = 20)
rainsub <- rain[idx, ]
```

Part A Using your sample `rainsub` and the `qt()`, attempt to create an 80% confidence interval using the point estimate $\pm$ method (i.e., median $\pm C \times \hat{\sigma}/\sqrt{n}$)

Part B Use the `bootstrap()` function to bootstrap 1,000 samples of the `median` statistic. With your resulting data frame, create a histogram of the sampling distribution. Based on this, does it seem like the confidence interval you found in Part A is appropriate? Why or why not?

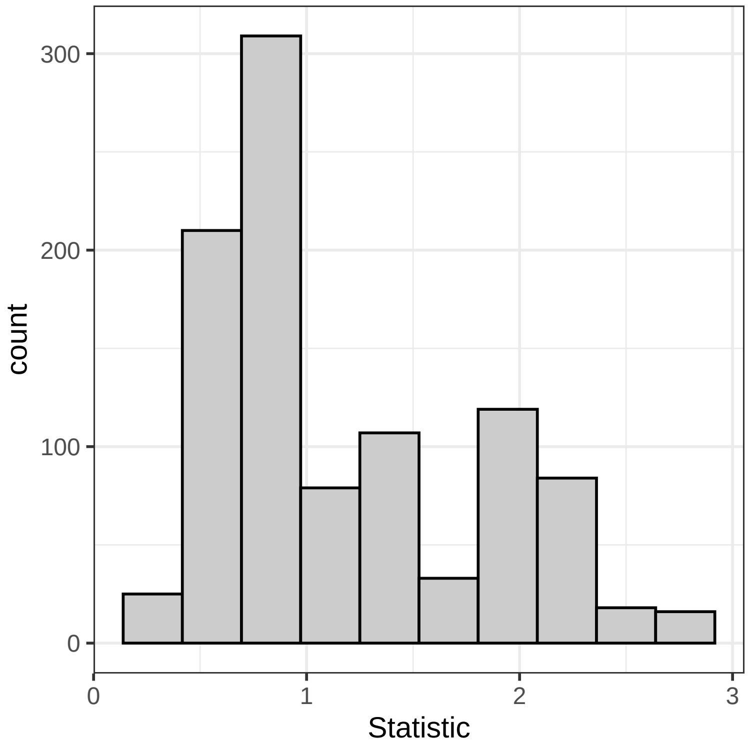
Part C Use the `quantile()` function to create an 80% confidence interval for the median. How does this compare with what you found in Part A?

Part D Now using the full rain dataset, find the true median value of the population. Does it fall within the intervals you constructed in Part A? How about Part B? Why did it work for one and not the other?

```
## Part A
mm <- median(rainsub$precip)
ss <- se(rainsub$precip)
mm + qt(c(0.1, 0.9), df = nrow(rainsub) - 1)*ss
```

```
## [1] 0.51033 1.25967
```

```
## Pat B (not valid because not normal)
bs <- bootstrap(rainsub$precip, median)
ggplot(bs, aes(Statistic)) + gh(10)
```




```
## Part C (much larger than A)
quantile(bs$Statistic, probs = c(0.1, .9))
```

```
##   10%   90%
## 0.515 2.090
```

```
## Part D -- falls in C, not A because CLT not appropriate
median(rain$precip)
```

```
## [1] 1.68
```