Homework 6

```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))
```

Point Distribution:

50 points for total homework assignment

Question 1: 5 points total (be generous)

Question 2: 10 points total

- 3 points if they correctly compute p
- 2 points if correctly compute standard error
- 4 points for confidence intervals
- 1 pt for part B

Question 3: 15 points total

- 6 points for part A (3 for each odds)
- 4 points for part B (2 if they did ratio with wrong numbers). They should say it does look like there is association, but no points off if they don't
- 5 points. No evidence for association since 1 IS in the interval

Question 4: 20 points total

- 5 points for histograms in Part A
- 3 points for correct CR ratio. Maybe skewed right, don't remove points for their comment
- 7 points for part C, they need to run bootstrap function and find the quantiles (numbers will not be exact)

- 3 points for part D (they should recognize its standard deviation of bootstrap)
- Part E, 3 points
- Part F, 4 points, minus 1 if they say its normal

Question 1

Below is the equation used to construct confidence intervals using critical values:

$$\overline{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}}$$

Explain what impact each term has on the size and location of confidence intervals and explain how the sample size impacts the critical values for a given confidence interval.

$\overline{x}$ is sample mean and tells me where CI will be centered. $\hat{\sigma}/\sqrt{n}$ is the standard error. As $\hat{\sigma}$ increases, so does the the width of the interval. As n increases, the size of the interval will decrease. The critical value, C, is based off the t distribution. It will be larger for a higher percentage interval (OK if this fact omitted), and it will also be larger when n is smaller

Question 2

As we saw in class, the Central Limit Theorem tells us that, for a population with mean μ and standard deviation σ, the sample mean will follow an approximately normal distribution with

$$\overline{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

In fact, the CLT also applies to our estimates of *proportions* which, as can be seen, are a type of mean. For example, consider ten coin flips in which four of them result in heads. We understand this proportion to be

$$\hat{p} = \frac{4}{10} = 0.4$$

If we were to record these values as 1s and 0s instead and take the mean, we would find the same

```r
## Ten flips, four heads
flips <- c(0, 1, 1, 0, 0, 0, 1, 0, 0, 1)
mean(flips)
```

```
## [1] 0.4
```

In fact, there is a special formula for the distribution of a proportion, p, based on the calculation for its variance. It can be shown that, for "large enough" n, the CLT tells us that the distribution of a proportion is given as

$$\hat{p} \sim N\left(p, \quad \sqrt{\frac{p(1-p)}{n}}\right)$$

Given a sample proportion, we can use the CLT to construct a confidence interval for the proportion, just as we did for the sample mean. Use this information to answer the following problem:

In a study conducted by Johns Hopkins University researchers investigated the survival of babies born prematurely. They searched their hospital's medical records and found that of 39 babies born at 25 weeks gestation (15 weeks early), 31 of these babies went on to survive at least 6 months.

Part A Using a normal approximation, construct a 95% confidence interval to estimate the true proportion of babies born at 25 weeks that are expected to survive at least 6 months.

Part B An article on Wikipedia suggests that of babies born at 25 weeks, 72% are expected to survive. Is this estimate consistent with what we found in Part A?

```r
## Part A
p <- c(rep(0, 8), rep(1, 31))
(mm <- mean(p)) # mean
```

```
## [1] 0.79487
```

```r
(ss <- se(p)) # se
```

```
## [1] 0.065504
```

```r
mm + c(-1, 1)*1.96*ss #CI
```

```
## [1] 0.66648 0.92326
```

```r
## Part B
# yes this is consistent
```

Question 3

The Center for Disease Control conducted a study on 6,168 women in hopes of finding factors related to breast cancer. In particular, they considered at which age a woman gave birth to her first child and computed the following contingency table:

```
##      Cancer
## Age     No  Yes
##    <25 4475   65
##    25+ 1597   31
```

- **Part A:** Compute the odds for both groups of women (those <25 years old at the delivery of their first child and those 25+ years old) for developing breast cancer.
- **Part B:** Compute the *odds ratio* of getting breast cancer for those younger than 25 compared to those older than 25. Based on this *point estimate*, would you say that women <25 have increased odds of getting breast cancer?
- **Part C:** Unlike the mean, a statistic which has a measure of departure such as the standard deviation, this is no immediate deviation metric available for the odds ratio. Use the function defined below, `bootstrap_or()` to compute a 95% confidence interval for the odds ratio. Based on this, would you say that there is evidence to suggest there is increased risk associated with early child birth?

```r
################################
### COPY AND RUN FOR PART C ###
################################

## Here is data.frame containing data,
df <- expand.grid(Cancer = c("Yes", "No"),
                  Age = c("<25", "25+"))
df <- df[rep(1:4, times = c(4475, 65, 1597, 31)), ]
```

```
bootstrap_or <- function(x, n = 1000L) {
  bs <- replicate(n, {
    # Randomly select rows of data frame
    idx <- sample(nrow(x), replace = TRUE)
    samp_df <- x[idx, ]
    tt <- with(samp_df, table(Age, Cancer))
    (tt[1] * tt[4]) / (tt[2] * tt[3])
  })
  data.frame(Sample = seq_len(n),
             Statistic = bs)
}


bs <- bootstrap_or(df)
```

```
## Odds (under 25)
(u25 <- 65/4475)
```

```
## [1] 0.014525
```

```
## Odds (over 25)
(o25 <- 31/1597)
```

```
## [1] 0.019411
```

```
## Odds ratio
## Odds of cancer 1.33 times higher in over 25 group
# based on this, it seems as if there is association between age of first birth
# and breast cancer
o25/u25
```

```
## [1] 1.3364
```

```
## Part C
## Because 1 is within plausible values, we do not have enough
# evidence to suggest that there is definitely relationship
bs <- bootstrap_or(df)
quantile(bs$Statistic, probs = c(0.025, 0.975))
```

```
##    2.5%   97.5%
## 0.82036 2.01564
```

Question 4

This question will involve the Grinnell rainfall data from 2014-2024, with a sample size collected from this data of size $n = 25$

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

# replace = FALSE because this is our sample, not a bootstrap estimate
set.seed(1)
idx <- sample(nrow(rain), size = 40, replace = FALSE)

# rs = rain sample
rs <- rain[idx, ]
```

- **Part A:** Create a histogram showing the distribution of precipitation in the population dataset `rain` and compare it with that drawn from our random sample in `rs`. Do they appear to be reasonably close? If we were not able to verify the population distribution, what can we do to ensure the quality of our sample?
- **Part B:** When distributions of data are skewed, such as they are in the Grinnell rain data, we know that the mean and the median will not be equal to one another. One statistic that can capture this information is the *ratio* between the mean and the median. We will call this the *centrality ratio*, denoted CR, where

$$CR = \frac{\text{mean}}{\text{median}}$$

If the mean value is *greater* than the median, we expect to find $CR > 1$, and if the median is greater than the mean, we should expect $CR < 1$; $CR = 1$ when the mean and median are equal. Find the centrality ratio of our sample data. Does there appear to be compelling evidence that our data are skewed? If so, in which direction?

- **Part C:** Instead of relying on a point estimate, we might instead consider creating a 95% confidence interval of this skew. Run the code below to create bootstrapped samples and describe the 95% confidence interval. Does it appear symmetric about our point estimate? In other words, is the upper bound of our confidence interval the same distance away from our point estimate as the lower bound of the confidence interval?

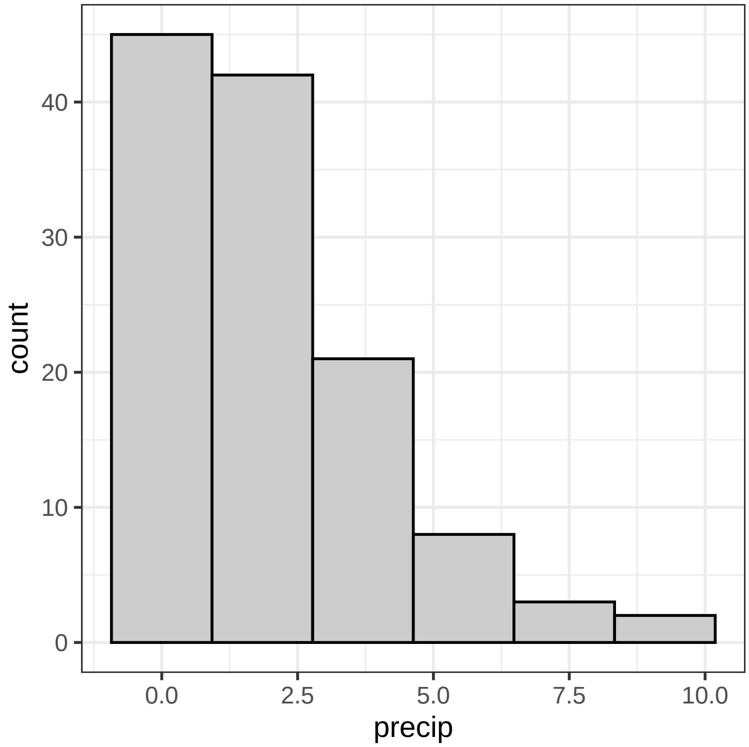
```r
mm_ratio <- function(x) mean(x) / median(x)

mm_boot <- bootstrap(rs$precip, mm_ratio)
```

- **Part D:** How would you find the *standard error* for this distribution, given what you have in Part C? What is it?

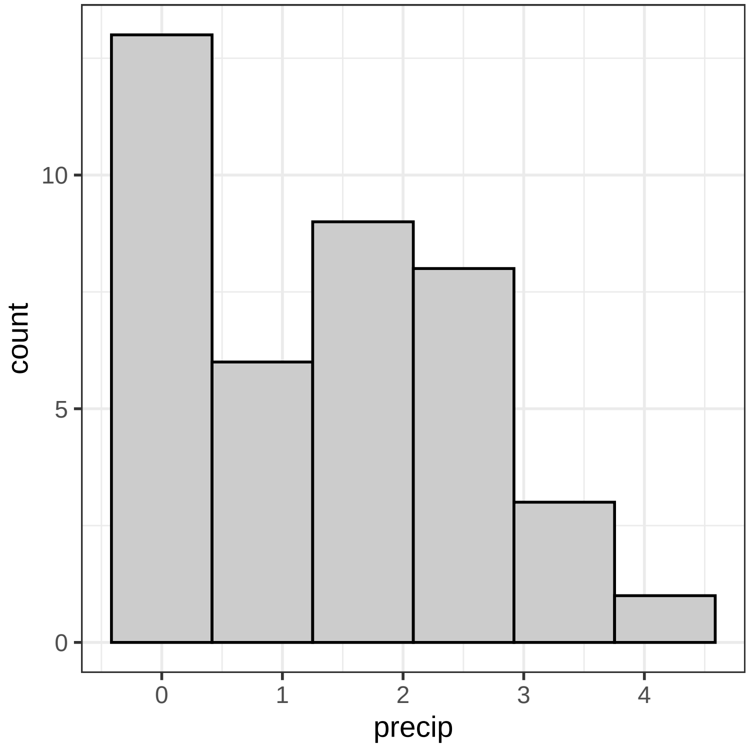
- **Part E:** Based on your bootstrapped interval, would you say there is compelling evidence from our sample that our data is skewed? What do you think in light of this having seen the distribution of the entire population?

- **Part F:** Using `mm_boot` that you computed in the previous part, create a histogram showing the *sampling distribution* of the centrality ratio statistic. What do you notice about this distribution? Would the normal approximation be appropriate here? Why or why not?

```r
# Part A
## They appear to be reasonably close
## There is nothing we can do once a sample is collected.
ggplot(rain, aes(precip)) + gh(6)
```



```
ggplot(rs, aes(precip)) + gh(6)
```



```
## Part B
## Over 1 so evidence of skew (skewed right)
mean(rs$precip) / median(rs$precip)
```

```
## [1] 1.0874
```

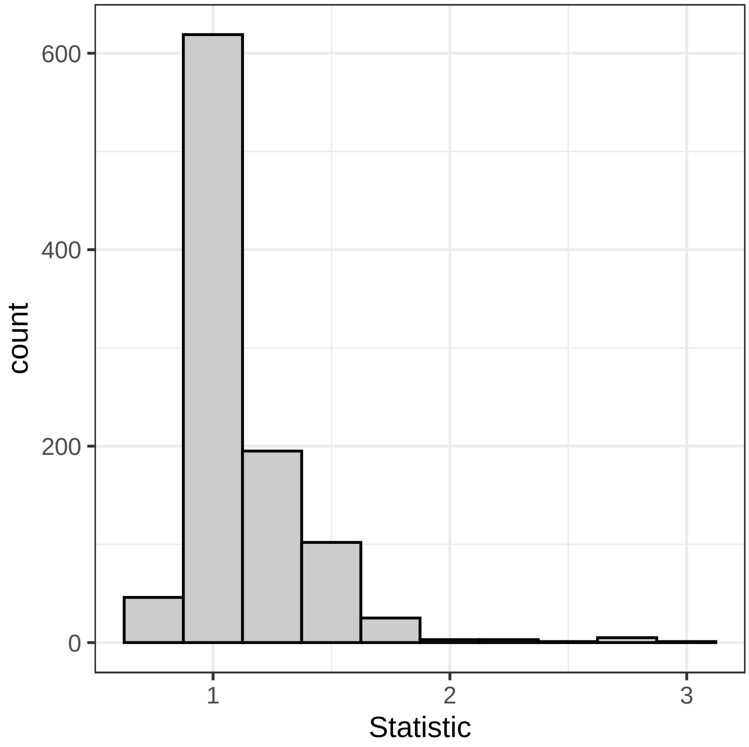
```r
## Part C
mm_ratio <- function(x) mean(x) / median(x)

## Not symmetric around point estimate
mm_boot <- bootstrap(rs$precip, mm_ratio)
quantile(mm_boot$Statistic, probs = c(0.025, 0.975))
```

```
##    2.5%   97.5%
## 0.84818 1.67353
```

```r
## Part D
# We can find std. error by taking standard deviation of bootstrap
sd(mm_boot$Statistic)
```

```
## [1] 0.24936
```

```r
## Part E
# There is not compelling evidence since 1 is a plausible value

## Part F
## Normal dist probably not appropriate here
ggplot(mm_boot, aes(Statistic)) + gh(bins= 10)
```


count
Statistic