Lab 2: Key

Group Member Names - here

Lab Overview

All students attending class in the group can turn in a single document with each participants name. Students not attending class will need to complete their own lab.

This lab is focused on understanding the uncertainty in a SRS from a hypothetical population. We will use *fake* data so that we can compare point estimates with the true population parameter. This also allows us to take multiple samples to visualize and understand the sampling distribution. However, in most scenarios we will only have the ability to take a single sample for making statistical inferences.

1. (4 points)

The code below creates a *fake* data set. Briefly describe what this code is doing.

```
set.seed(09052019)
sampling_frame <- tibble(student_id = 1:20,
                         snow_days = base::sample(x = 0:50, size = 20,
                                                  prob = c(.20,rep(.80/50, 50)), replace = T))
```

This creates a data set with 20 students. To create the truth*, each student is randomly assigned a value between 0 and 50. No snow_days (days spent skiing) is selected with probability .2 and the other values, 1 to 50 are selected with probability 0.016.*

2. (4 points)

Suppose you take a random sample of 2 students, what is the probability that a student is selected in the sample?

Analytically the value is 0.1. However, we can take repeated samples to determine how frequently this event occurs. The following code selects one million samples of size 2 and computes the proportion of times id # 1 is selected.

```
find_val <- function(samples, val){
  return(val %in% samples)
}
samples_large <- replicate(1000000, sample(20, 2, replace = F))
apply(samples_large, 2, find_val, val = 1) %>% mean()
```

```
## [1] 0.100302
```

3. (4 points)

Suppose you take a random sample of 12 students, what is the probability that a student is selected in the sample?

The same procedure is conducted with a large number of samples, where .6 contain id #1.

```
samples_large <- replicate(1000000, sample(20, 12, replace = F))
apply(samples_large, 2, find_val, val = 1) %>% mean()
```

```
## [1] 0.600077
```

4. (4 points)

Take a single sample of 12 students and construct a *point estimate* for the mean number of days skiied by students in the class. Describe your results, how well do they match the truth?

```
point_estimate <- mean(sample(sampling_frame$snow_days, 12, replace = F))
truth <- sampling_frame %>% summarize(mean(snow_days)) %>% pull()
```

Based on this single sample, our point estimate is 13, while the truth is 12.5.

5.

Take several more (~1000) samples of 12 students and construct *point estimates* for each sample.

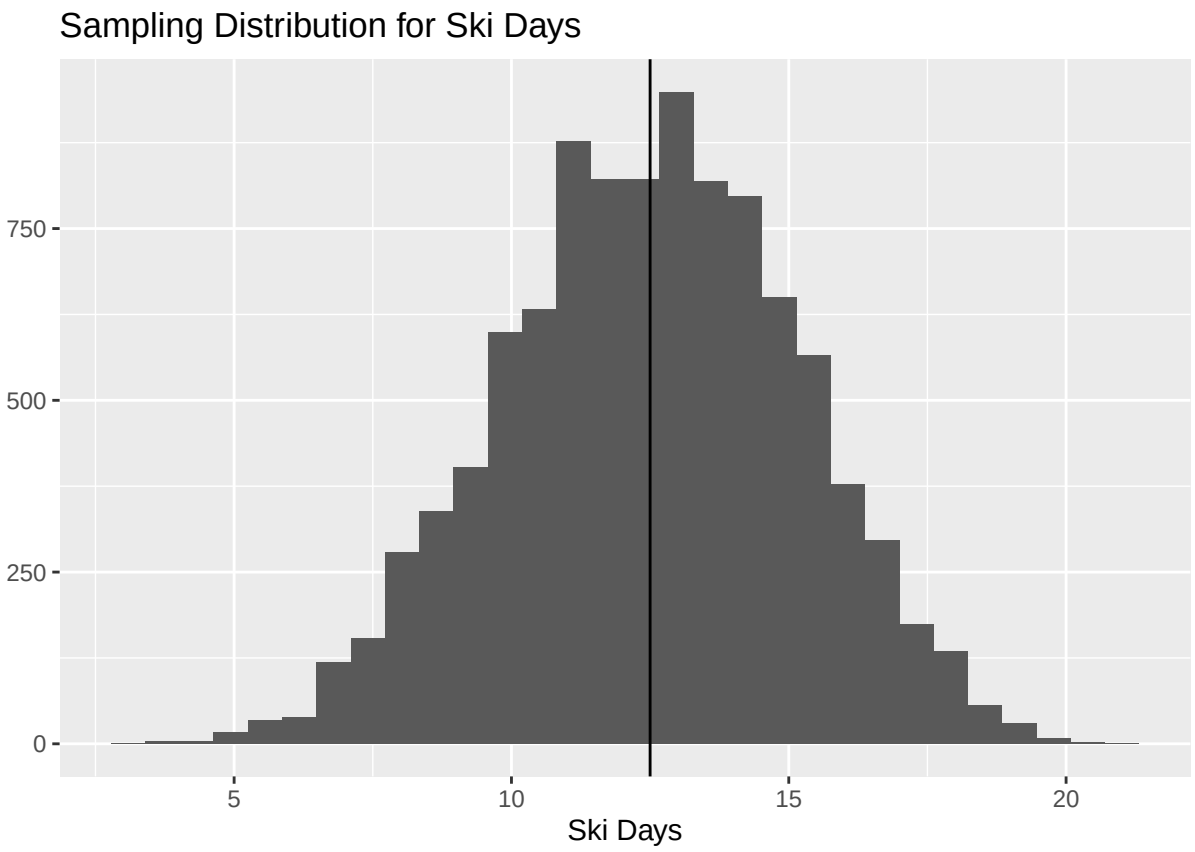
```
thousand_samples <- replicate(10000, sample(sampling_frame$snow_days, 12, replace = F))
all_point_estimates <- colMeans(thousand_samples)
```

a. (4 points)

Construct a histogram of these point estimates computed from these samples and include a summary caption/comment.

```
library(ggplot2)
tibble(val = all_point_estimates, var = 'estimate') %>% ggplot(aes(val)) + geom_histogram() +
  geom_vline(xintercept = truth) + ggtitle('Sampling Distribution for Ski Days') + ylab('') +
  xlab('Ski Days')
```

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



The vertical line represents the truth and the histogram shows the possible values resulting from the sampling procedure.

b. (4 points)

Describe the bias and variance of the *point estimator* (the process that generated the 1000 samples)

```
bias <- mean(truth - all_point_estimates)
variance <- var(all_point_estimates)
```

The bias of the point estimator is 0.0117 and the variance is 6.9629691.

c. (4 points)

Suppose another *point estimator* used the sample median as the point estimate, how would you (computationally) compare the two procedures?

We could use the same procedure, but then instead of taking the mean of the samples, we would use the median. Then the variance and bias of the two procedures could be compared.