Data project 1: Birds, birds, birds

Cynthia Li (student ID 3034939311)

Problem 1: Critical Thinking

1. What are two likely sources of error/noise/uncertainty in this dataset (particularly in the max_individuals measurement)?

(1) Citizen scientists may fail to count all the individuals of the species within their count site. In this case, the number of individuals of the species they record is less than the real data.
(2) Citizen scientists may include some individuals of other similar species when counting the number of individuals of a specific species they focus. In this case, the number they record is greater than the real data.

2. In the description in the introduction, we mentioned that the birder's report "the maximum number of individuals of each species seen at one time". Why do you think that FeederWatch asks for the maximum number of individuals seen at one time (i.e., the flock size), rather than individual bird reports, or the total number of birds of each species that visited the feeder throughout the day?

The fock size of a species seen at one time can imply habits or life styles of this bird species. The fock size is also a stronger indicator of the distribution of birds' habitats compared to individual bird reports. It is more likely that a species of birds like to live in a certain place if a considerably big flock of this species - rather than an individual - is observed in this place.

3. The number of citizen scientists in the Bay area contributing to the FeederWatch project has increased from 2011 to 2021. If we are interested in exploring changes in flock size over time, how will this increase in data contributions influence our estimates?

The increase in data contributors increases the sample size, makes the sample data more general and thus reduces bias. The sample mean of flock size is one useful statistic we care about when we explore changed in flock sizeover time. The increase in data contributors increases sample size and makes the sample data more representative of the population.

Problem 2: Working with the data

1. How many unique feeder locations were there in the provided data in 2011 and 2021? Howmany feeder locations provide data in both 2011 and 2021?

There are 73 unique feeder locations in the provided data in 2011.

There are 119 unique feeder locations in the provided data in 2021.

```
## Joining, by = "loc_id"
```

There are 18 feeder locations in the provided data in both 2011 and 2021.

2. What were the five species that visited the feeders in the largest flocks in 2011 and in 2021 (based on the average flock size in these years)? Have the species changed? For the species that remained in the list in 2021, have the sizes of the flocks changed?

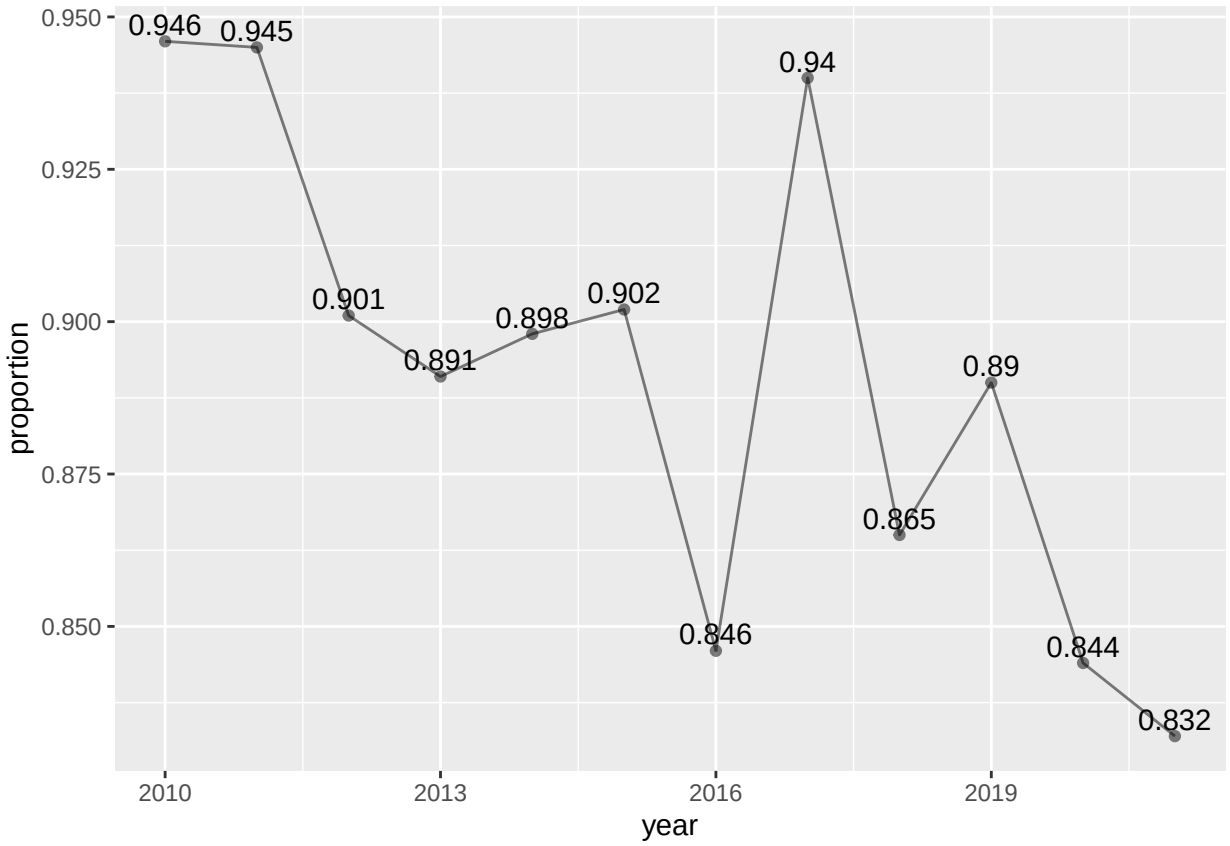
Red-winged Blackbird, Spinus sp.(goldfish sp.), Lawrence's Goldfinch, Wild Turkey, and Cedar Waxwing are the five species that visited the feeders in the largest flocks in 2011.

Cedar Waxwing, Rock Pigeon(Feral Pigeon), Wild Turkey, Pine Siskin, and Lesser Goldfinch are the five species that visited the feeders in the largest flocks in 2021.

The species that visited the feeders in the largest flocks have have changed from the data in 2011 to the data in 2021.

Cedar Waxwing and Wild Turkey are the two species that on the top-5 list in both 2011 and 2021. The average flock size of Cedar Waxwing has increased by approximately one from 13.78 to 14.70. The average flock size of Wild Turkey has decreased from 13.96 to 7.32.

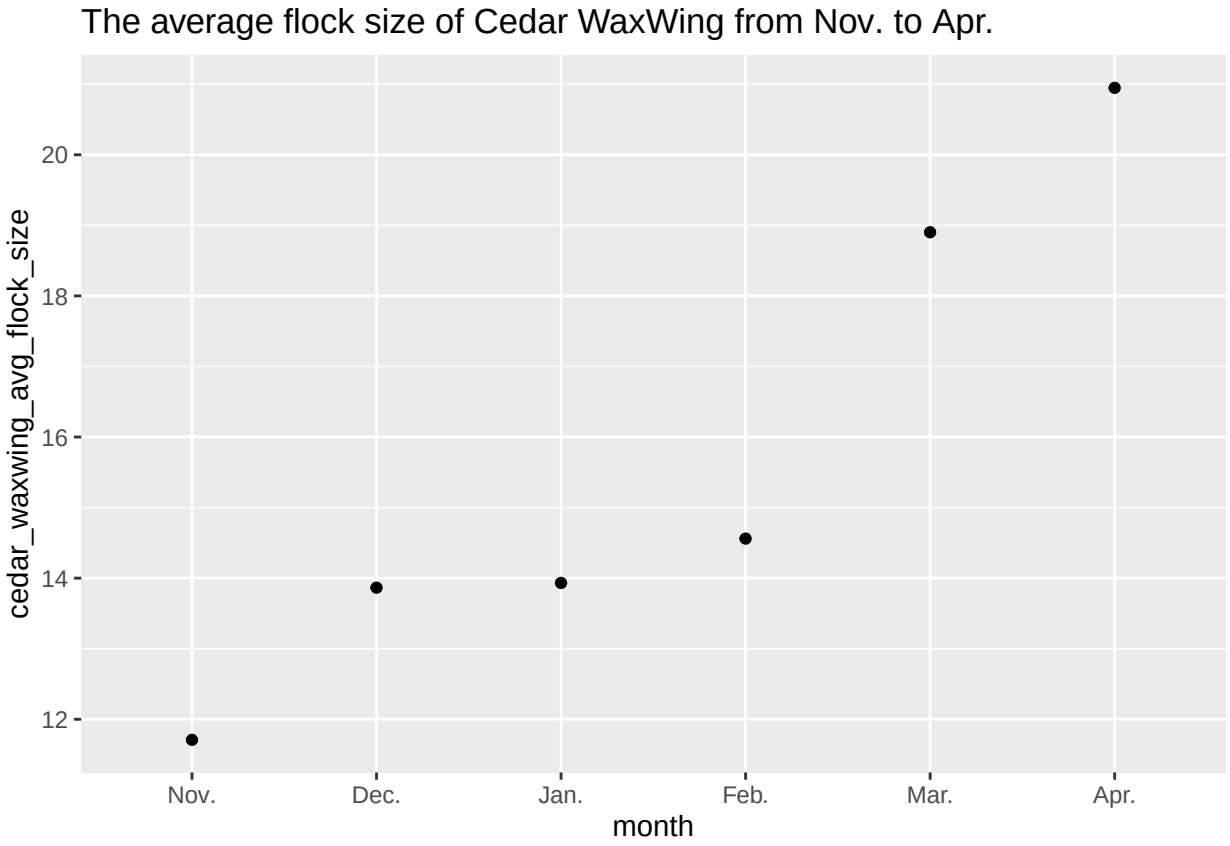
3. Have the proportion of feeders reporting at least one Dark-eyed Juncos visit per year increased or decreased over time? Provide some evidence of your answer.



The proportions of feeders reporting at least one Dark-eyed Juncos visit per year have generally decreased from 2010 to 2021. A noticeable exception of this decreasing tread is that the proportion increased from 0.846 to 0.94 from 2016 to 2017.

Problem 3: EDA

```
##   mon cedar_waxwing_avg_flock_size month
## 1  11                     11.70833  Nov.
## 2  12                     13.86486  Dec.
## 3  01                     13.93197  Jan.
## 4  02                     14.56028  Feb.
## 5  03                     18.90196  Mar.
## 6  04                     20.94828  Apr.
```



The flock size of Cedar Waxwings increases from winter to spring.

Problem 4: Parameter estimation

1. What is the population that this data is designed to capture?

The population is the birds in the Bay Area from early November to early April.

2. Do you think this data is a biased or unbiased representation of the population you identified in the previous question? Explain why or why not.

This data is biased representation of all the birds in the Bay Area from early November to April because the locations of the citizen scientists' home bird feeders may disproportionately cover all the possible places birds appear. For example, citizen scientists and bird feeder may be fewer in wild places like high mountain and deep forest.

Regardless of your answer to the previous question, let's make the almost certainly unrealistic assumption that the observations in the data are a representative IID sample of whatever population you indicated in part 1 of this question.

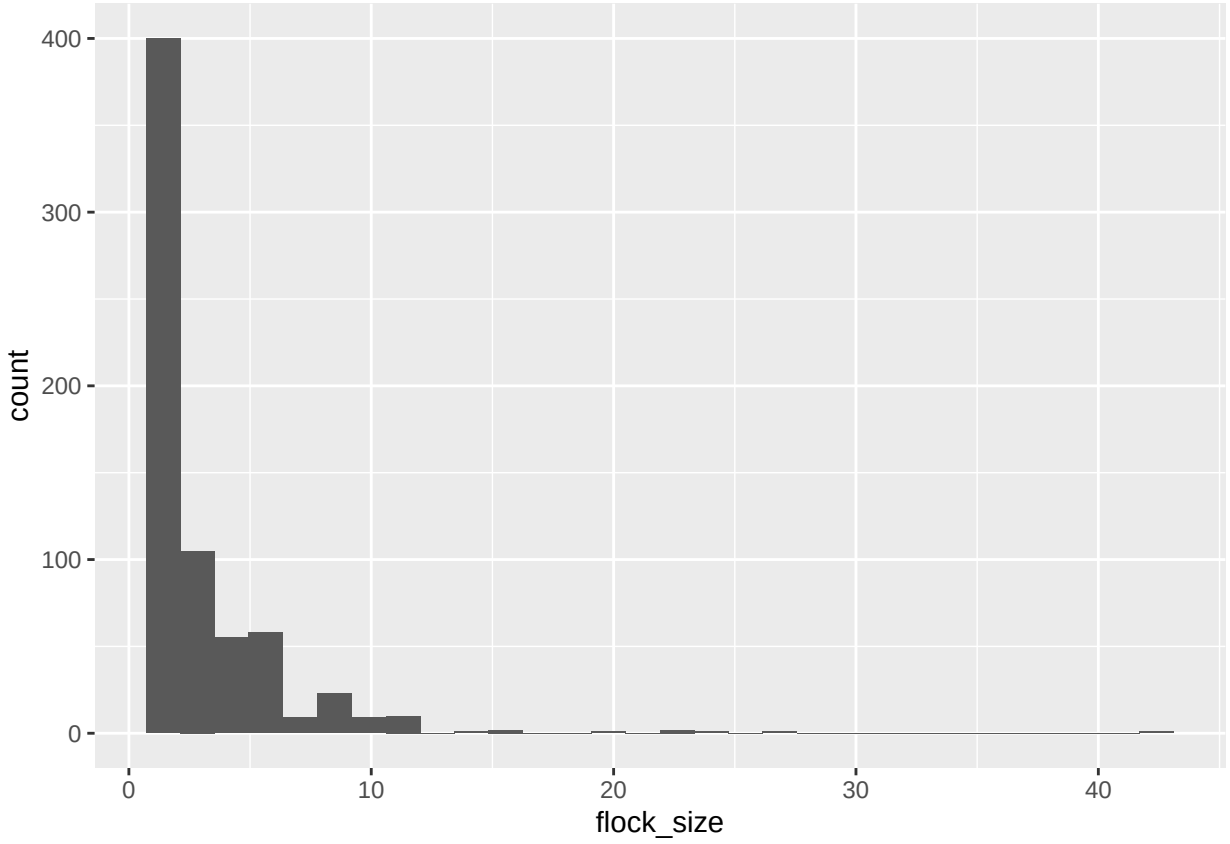
3. Compute an estimate of the average flock size of Dark-eyed Juncos feeding at bird feeders in 2021 in the Bay Area based on the sample mean.

```
mean(filter(obs, year == 2021 & species_name == 'Dark-eyed Junco')$max_individuals)
```

```
## [1] 3.116519
```

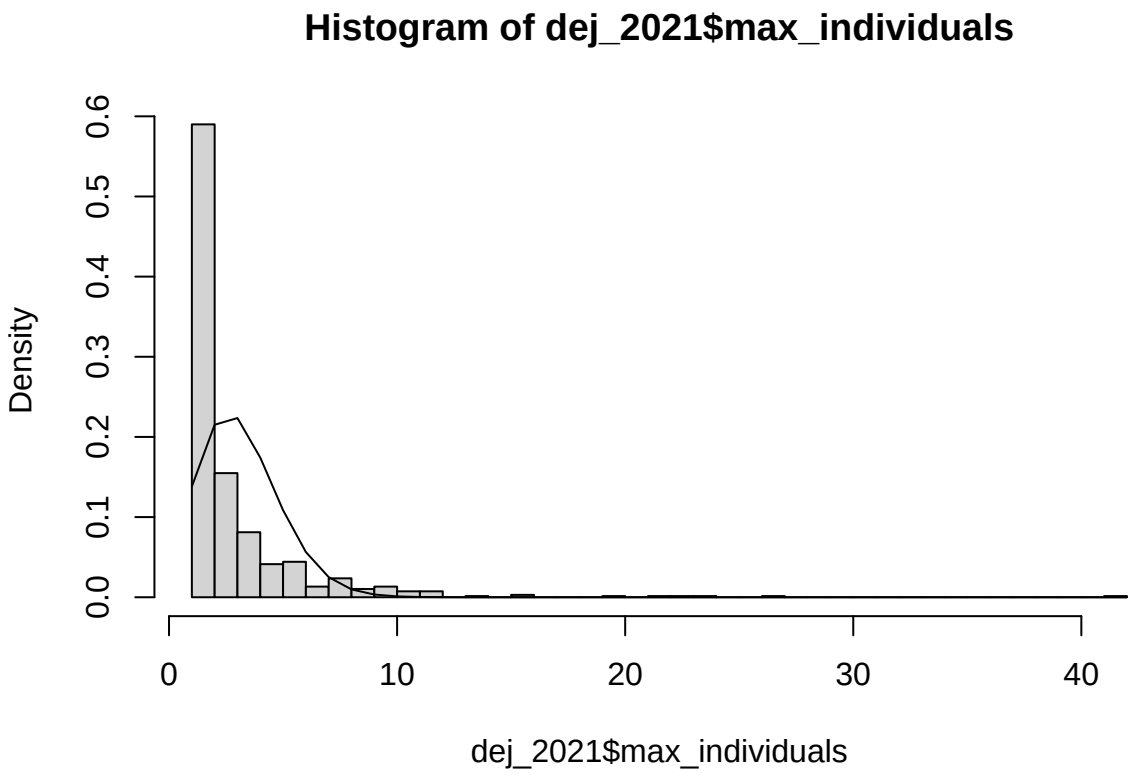
4. Plot a histogram of the distribution of Dark-eyed Junco flock size in 2021.

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



5. Assume that the flock size data comes from a Poisson distribution. Use the MLE estimate to estimate the parameter lambda to fit a Poisson distribution to the data using your MLE estimate lambda^MLE. Report your lambda^MLE parameter estimate computed from the data, and plot the Poisson distribution that you fit on top of your histogram of flock size from the previous question. Based on your plot, do you think that the Poisson assumption is reasonable?

The MLE estimate lambda^MLE to the parameter lambda of Poisson distribution is the sample mean 3.117 from Part4.3.



Based on the plot above, the Poisson assumption is not reasonable. Because of the highest density at flock_size = 1, exponential distribution may be a better assumption.

6. Use your estimated Poisson distribution to generate 1,000 parametric bootstrap sampled versions of your average flock size estimate and plot a histogram of your bootstrapped estimates. Report an estimate of the bias and variance of your average flock size sample mean estimate. Show your code for this question.

Reference: Professor's bootstrap_mean.R on bcourse.

```
# Read the csv file and assign the data frame to the variable name "birds"
birds <- read_csv("bay_area_birds.csv")
```

```
## Rows: 148317 Columns: 7

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): species_name, loc_id
## dbl  (4): year, max_individuals, latitude, longitude
## date (1): date

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

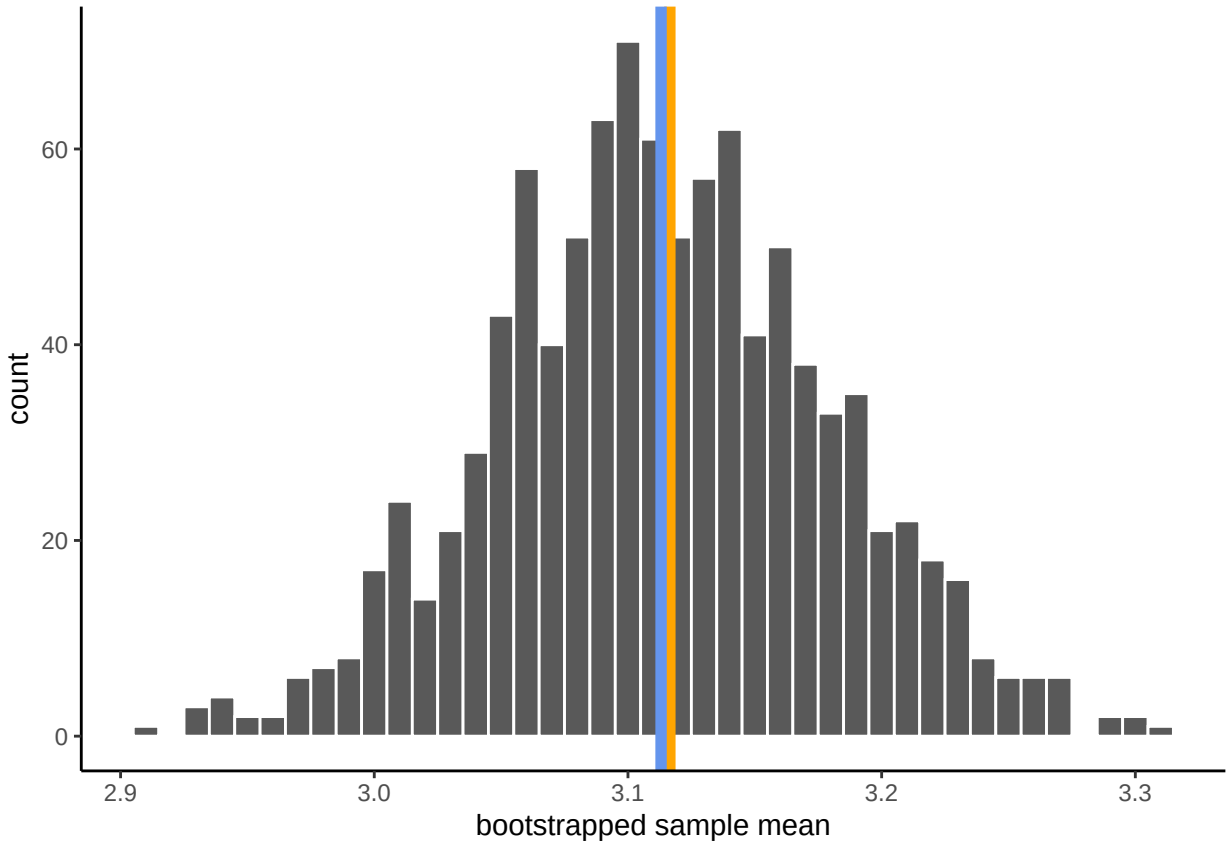
```r
# Get the sample data
sample_data <- filter(birds, year == 2021 & species_name == "Dark-eyed Junco")$max_individuals
sample_mean <- mean(sample_data)
# Pseudo-random seed
set.seed(1234567)

# draw 1000 parametric bootstrap samples
p_boot_mean_df <- map_df(1:1000, function(i) {
  # sample from the estimated Poisson distribution with lambda mle sample mean
  bootstrap_data <- rpois(length(sample_data), mean(sample_data))
  # compute the sample mean of the parametric bootstrap sample
  data.frame(boot_mean = mean(bootstrap_data))
})

# parametric bootstrap estimate of the sample mean bias:
p_bias <- mean(p_boot_mean_df$boot_mean) - mean(sample_data)
p_bias
```

```
## [1] -0.003464602
```

```r
# parametric bootstrap estimate of the sample mean SD
p_var <- var(p_boot_mean_df$boot_mean)
p_var
```

```
## [1] 0.004312383
```

```r
# Plot a histogram of the bootstrapped sample means
p_boot_mean_df %>%
  ggplot() +
  geom_histogram(aes(x = boot_mean), color = "white",
                 binwidth = 0.01) +
  # add a line for the sample estimate of the mean
  geom_vline(xintercept = mean(sample_data),
             color = "orange", size = 2) +
  # add a line for the bootstrapped estimate of the mean
  geom_vline(xintercept = mean(p_boot_mean_df$boot_mean),
             color = "cornflowerblue", size = 2) +
  labs(x = "bootstrapped sample mean") +
  theme_classic()
```



7. Generate 1,000 non-parametric bootstrapped versions of your estimate and plot a histogram of the distribution of your bootstrapped estimates. Report an estimate of the bias and variance of your average flock size sample mean estimate. Show your code for this question.

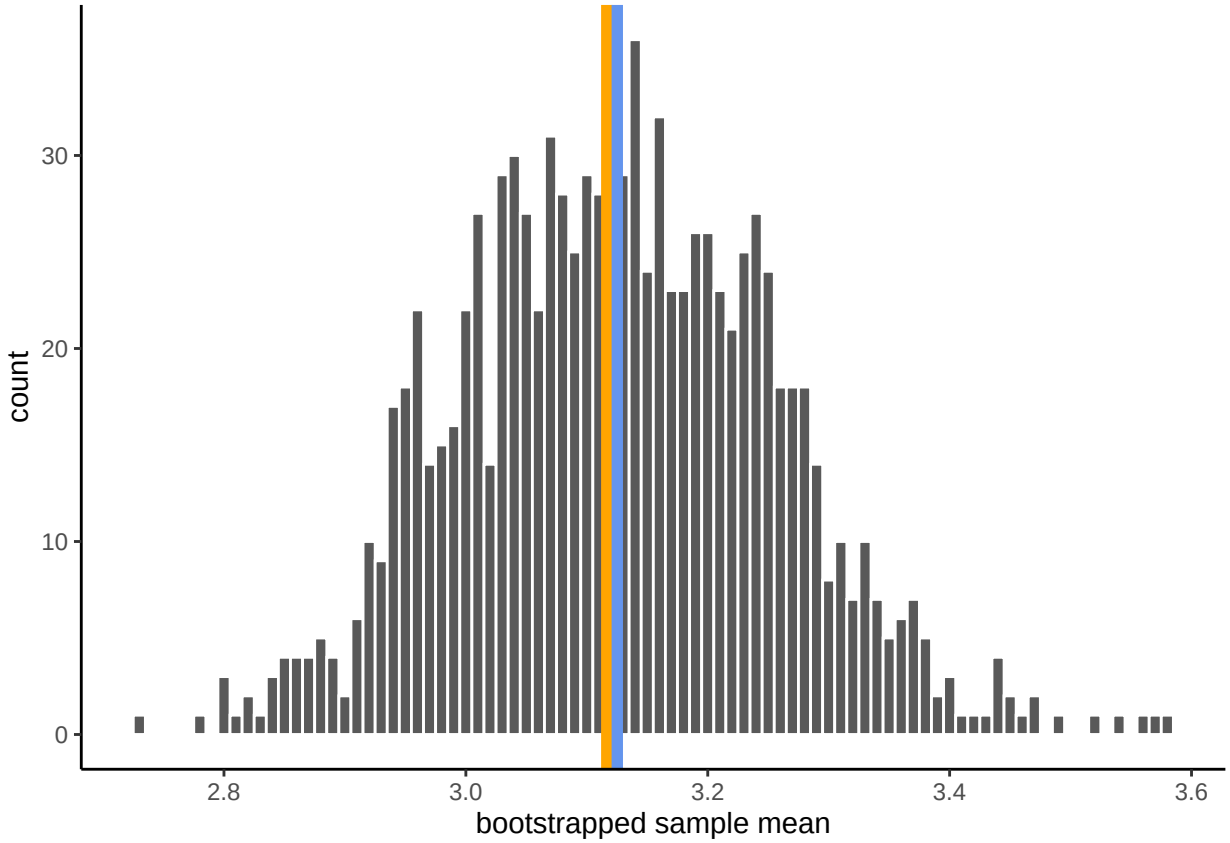
```
# draw some non-parametric bootstrap samples
np_boot_mean_df <- map_df(1:1000, function(i) {
  # sample from the data with replacement
  bootstrap_data <- sample(sample_data, length(sample_data), replace = TRUE)
  # compute the sample mean of the bootstrap sample
  data.frame(boot_mean = mean(bootstrap_data))
})

# np bootstrap estimate of the sample mean bias:
np_bias <- mean(np_boot_mean_df$boot_mean) - mean(sample_data)
np_bias
```

```
## [1] 0.008665192
```

```
# np bootstrap estimate of the sample mean SD
np_var <- var(np_boot_mean_df$boot_mean)
np_var
```

```
## [1] 0.01702949
```

```r
# Plot a histogram of the bootstrapped sample means
np_boot_mean_df %>%
  ggplot() +
  geom_histogram(aes(x = boot_mean), color = "white",
                 binwidth = 0.01) +
  # add a line for the sample estimate of the mean
  geom_vline(xintercept = mean(sample_data),
             color = "orange", size = 2) +
  # add a line for the bootstrapped estimate of the mean
  geom_vline(xintercept = mean(np_boot_mean_df$boot_mean),
             color = "cornflowerblue", size = 2) +
  labs(x = "bootstrapped sample mean") +
  theme_classic()
```



Part 5: Become a citizen scientist!