Data Project: Birds, birds, birds

Leah Hong

Part 1: Critical Thinking

- Possible sources of error in the data set includes miscounting the birds, misidentifying a species, or entering observations incorrectly. For example, a participant may have a difficult time distinguishing between bird species if they are visually similar to one another.
- FeederWatch asks for the maximum number of individuals seen at one time (flock size) instead of individual bird reports because counting the total number of individual species per day may lead to double counting individuals if they visit the same feeder multiple times throughout a day. The total bird counts may be inaccurately reflecting the actual total number of birds. Therefore, reports on the maximum flock size may be a more accurate reflection of thea ctual maximum flock size of the birds visiting the feeders.
- Since the maximum flock size visiting the feeders is reported, the estimates will be a more accurate reflection of the maximum flock size of each bird species.

Part 2: Working with the data

First, we will filter the data for 2011 and 2021 separately. We are finding the number of unique feeder locations in 2011 as well as 2021.

```
# data in 2011 only
birds_2011 <- birds %>% filter(year == 2011)
uniq_loc_2011 <- length(unique(birds_2011[["loc_id"]]))

# data in 2021 only
birds_2021 <- birds %>% filter(year == 2021)
uniq_loc_2021 <- length(unique(birds_2021[["loc_id"]]))
```

Then, we will find the number of unique feeder locations that are in both 2011 and 2021.

```
# the number of non-unique values in both 2011 and 2021
uniq_2011 <- unique(birds_2011[["loc_id"]])
uniq_2021 <- unique(birds_2021[["loc_id"]])
uniq_2011_2021 <- c(uniq_2011, uniq_2021)
uniq_loc_2011_2021 <- length(uniq_2011_2021) - length(unique(uniq_2011_2021))
```

- In 2011, there are 73 unique feeder locations provided.
- In 2021, there are 119 unique feeder locations provided.
- In both 2011 and 2021, there are 18 unique feeder locations provided.

Now, we will group the birds by species then compute the mean for 2011 and 2021. We will output the five species with the largest flocks.

```
# 2011: grouped by species then computed the mean for each group
mean_by_group_2011 <- arrange(aggregate(birds_2011[, 4], list(birds_2011$species_name),
mean), desc(max_individuals))

# list the five species with the largest flocks
five_species_2011 <- tail(mean_by_group_2011, 5)$Group.1
```

```
# 2021: grouped by species then computed the mean for each group
mean_by_group_2021 <- arrange(aggregate(birds_2021[, 4], list(birds_2021$species_name),
mean), desc(max_individuals))

# list the five species with the largest flocks
five_species_2021 <- tail(mean_by_group_2021, 5)$Group.1
```

```
tail(mean_by_group_2011)
```

	Group.1 <chr>	max_individuals <dbl>
88	Red-shouldered Hawk (elegans)	1
89	Red-tailed Hawk (calurus/alascensis)	1
90	Rose-breasted/Black-headed Grosbeak	1
91	Rufous-crowned Sparrow	1
92	Spotted Towhee (oregonus Group)	1
93	White-tailed Kite	1

6 rows

```
tail(mean_by_group_2021)
```

	Group.1 <chr>	max_individuals <dbl>
97	Sharp-shinned Hawk	1
98	Townsend's Solitaire	1
99	Varied Thrush	1
100	Wilson's Warbler	1
101	Wrentit	1
102	Yellow-billed Magpie	1

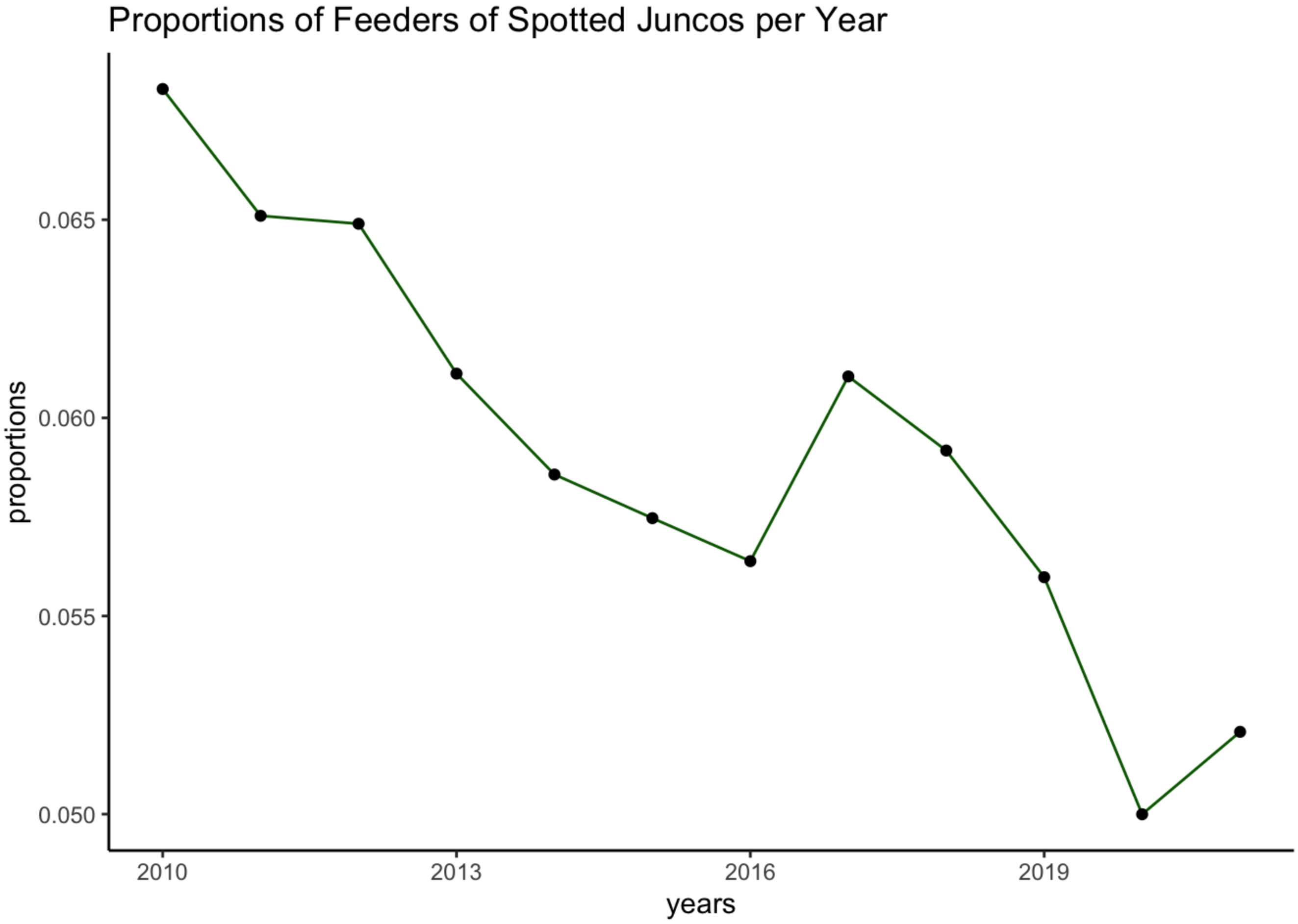
6 rows

- The five species that visited the feeders in the largest flocks in 2011 are Cedar Waxwing, Wild Turkey, Lawrence’s Goldfinch, Spinus sp. (goldfinch sp.), and Red-winged Blackbird.
- The five species that visited the feeders in the largest flocks in 2011 are Lesser Goldfinch, Pine Siskin, Wild Turkey, Rock Pigeon (Feral Pigeon), and Cedar Waxwing.
- All of the species have changed except Cedar Waxwing and Wild Turkey. For the species that remained in the list in 2021, the sizes have changed. For the Cedar Waxwing, the average flock size has gone from 13.78 to 14.65. For the Wild Turkey, the average flock size has gone from 13.96 to 7.32.

Next, we will create a line plot that will depict the trend in Juncos visiting feeders.

```
# df of year and counts for each year
counts_of_year <- birds %>% group_by(year) %>% summarise(counts = n())

# filter to get rows with species "Dark-eyed Junco"
dark_eyed_juncos <- birds %>% filter(species_name == "Dark-eyed Junco")
counts_of_juncos <- dark_eyed_juncos %>% group_by(year) %>% summarise(counts = n())

juncos_val <- counts_of_juncos$counts
year_val <- counts_of_year$counts

# proportions are in order of increasing year
proportions <- juncos_val / year_val

# table with year and proportions
years <- counts_of_juncos$year
props_table <- data.frame(years, proportions)
```

```
# line plot showing the years vs proportions to see an increase or decrease
ggplot(data = props_table, aes(x = years, y = proportions)) +
  geom_line(color = "darkgreen") +
  geom_point(color = "black") +
  labs(title = "Proportions of Feeders of Spotted Juncos per Year") +
  theme_classic()
```

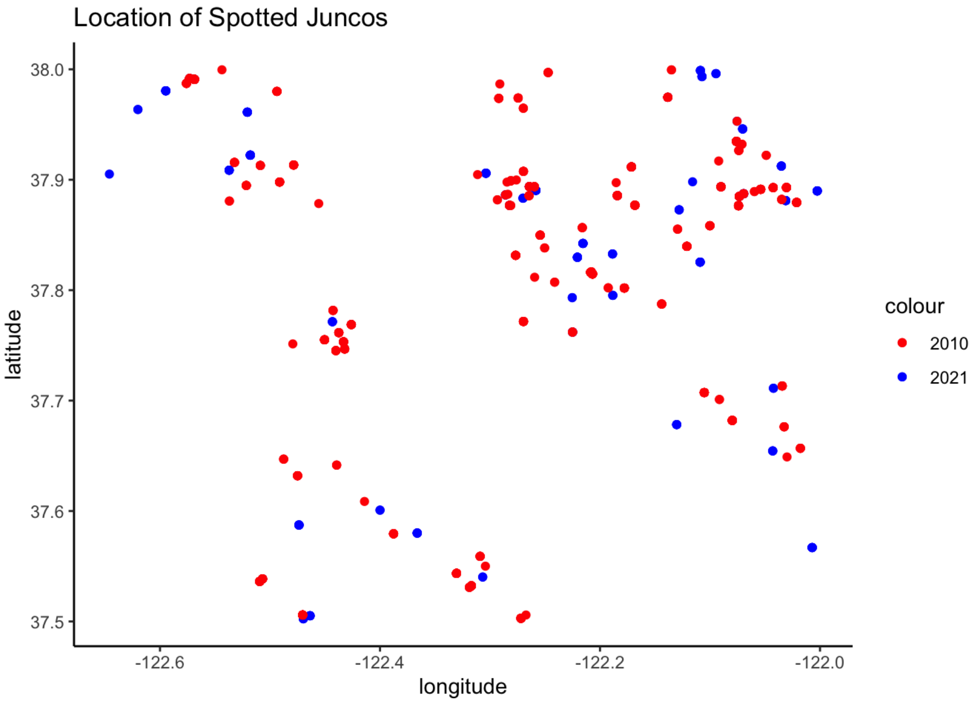


- Looking at the graph above, we can see that over time, the proportions tend to decrease over time except 2016 to 2017 and 2020 to 2021 where there are slight increases.
- There is an overall decrease, but it is not a monotonic decrease.

Part 3: EDA

The graph I chose to portray is the latitude vs longitude of the bird species Dark-eyed Juncos in the years 2010 and 2021. Since 2010 is the earliest year in the data set and 2021 is the latest year in the data set, I wanted to see if there was a large change in where the Dark-eyed Juncos were spotted. For instance, I wanted to see if there were new locations where Juncos were spotted or areas that do not have spotted Juncos anymore.**

```
edited_2010 <- birds %>%
  filter(year == 2010) %>%
  filter(species_name == "Dark-eyed Junco") %>%
  group_by(latitude, longitude)

edited_2021 <- birds %>%
  filter(year == 2021) %>%
  filter(species_name == "Dark-eyed Junco") %>%
  group_by(latitude, longitude)

ggplot() +
  geom_point(data = edited_2010, aes(x = longitude, y = latitude, color = "red")) +
  geom_point(data = edited_2021, aes(x = longitude, y = latitude, color = "blue")) +
  scale_color_manual(labels = c("2010", "2021"), values = c("red", "blue")) +
  labs(y = "latitude",
       x = "longitude",
       title = "Location of Spotted Juncos") +
  theme_classic()
```



A scatterplot showing the areas where the flock Dark eyed Juncos were spotted in 2010 vs 2021

- Looking at the graph, I see that there seems to be a couple of new spots where Juncos are located. One particular area that stood out to me was at (-122.45, 37.75). In the year 2010, there were so many more spotted observations of Juncos compared to 2021. This could be due to the fact that Juncos did not like the area they were living in and decided to populate elsewhere. This could also be due to the fact that there are less Junco birds now than eleven years ago.
- Another point is that the locations of volunteers could be different in 2010 vs 2021. There could be new people who are contributing to the dataset, which could account for the new sightings of Junco in a certain area.
- Nonetheless, looking at the graph, it seems as though the Dark-eyed Juncos relatively stayed in the same area over the years.

Part 4: Parameter estimation

The population that this data is designed to capture is bay area birds between November and April.**

- I believe the data is a biased representation of the population I identified in the previous part because the data does not involve data for the months May to September. Since the data is not representative of all birds in the Bay Area since some species may travel to the Bay Area during the months not included, the data is biased. In addition, the volunteers are observing and counting birds in flocks when they prefer to, which does not allow for all birds to have an equal probability of being observed.

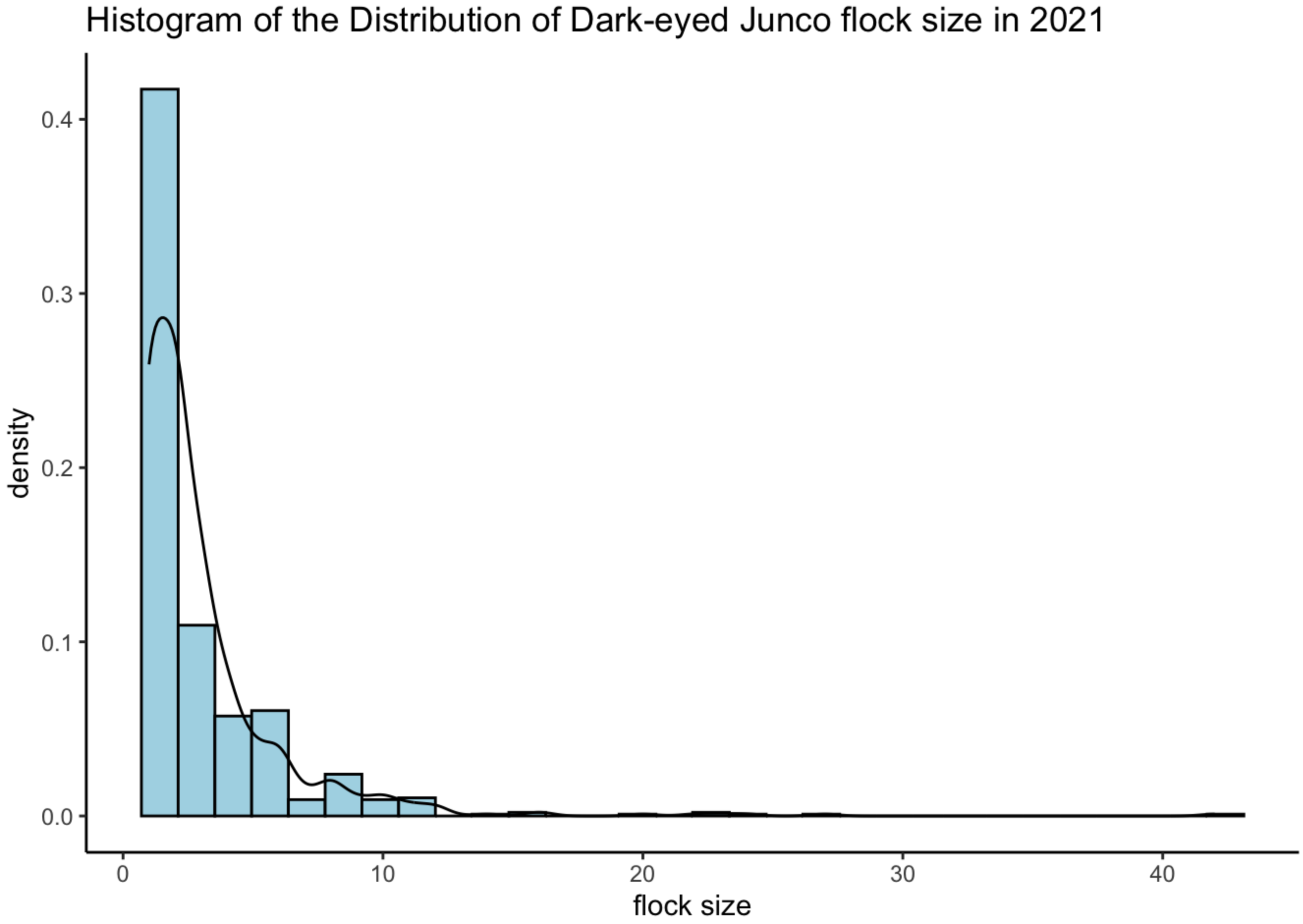
```
# only Dark-eyed Junco
dark_eyed_juncos <- birds %>% filter(species_name == "Dark-eyed Junco", year == 2021)

# table of averages of flock size per year
averages_by_year <- aggregate(dark_eyed_juncos$max_individuals, list(dark_eyed_juncos$ye
ar), FUN = mean)

# average of all the average flock sizes per year to get sample mean
sample_mean <- mean(averages_by_year$x)
```

- 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 is about 3.12.

Below, we will plot a histogram of the distribution of Dark-eyed Junco flock size in 2021.

```
# filter to get values only with Dark-eyed Junco in 2021
dark_eyed_juncos_2021 <- birds %>% filter(species_name == "Dark-eyed Junco", year == 202
1)

ggplot(dark_eyed_juncos_2021, aes(x = max_individuals)) +
  geom_histogram(aes(y = after_stat(density)), colour = 1, fill = "lightblue") +
  geom_density() +
  labs(x = "flock size",
       title = "Histogram of the Distribution of Dark-eyed Junco flock size in 2021") +
  theme_classic()
```

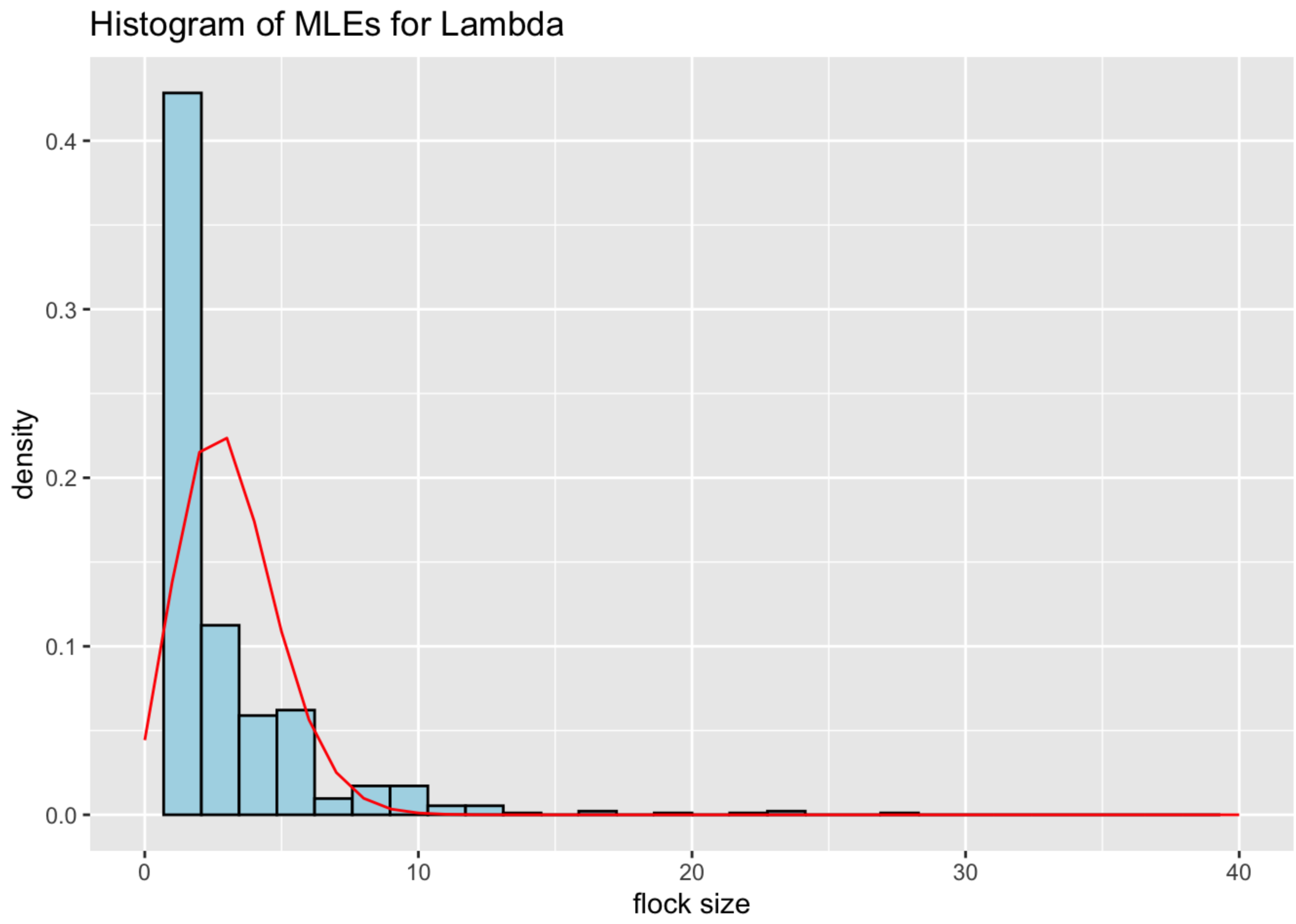
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- I found that the MLE is equivalent to the sample mean of the n observations in the sample. In other words, the MLE estimate to estimate the parameter lambda is equivalent to the sample mean of 3.116519 computed in a previous part.

```
lambda <- 3.116519
dark_eyed_juncos_2021 <- birds %>% filter(species_name == "Dark-eyed Junco", year == 202
1)
xvals <- seq(0, 50, seq = 1)
density_vals <- dpois(xvals, lambda)
dens_df <- data.frame(x = xvals, y = density_vals)

ggplot() +
  geom_histogram(data = dark_eyed_juncos_2021, aes(x = max_individuals, y = after_stat(d
ensity)), colour = 1, fill = "lightblue") +
  geom_line(data = dens_df, aes(x = x, y = y), color = "red") +
  xlim(0, 40) +
  ggtitle(paste0("Histogram of MLEs for Lambda")) +
  labs(x = "flock size")
```

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



- Based on my plot, I think the Poisson assumption is reasonable since it is equal to the sample mean.

```
lambda <- 3.116519
dark_eyed_juncos_2021 <- birds %>% filter(species_name == "Dark-eyed Junco", year == 202
1)
num_samples <- nrow(dark_eyed_juncos_2021)

p_boot_mean_df <- map_df(1:num_samples, function(i) {
  # sample from the approximated distribution
  bootstrap_data <- rpois(num_samples, lambda)
  # compute the sample mean of the parametric bootstrap sample
  data.frame(boot_mean = mean(bootstrap_data))
})
```

Let us compute the bias estimate, which is the average of the bootstrapped estimates minus the original sample mean estimate.

```
p_bias <- mean(p_boot_mean_df$boot_mean) - mean(dark_eyed_juncos_2021$max_individuals)
p_bias
```

```
## [1] -0.001068125
```

Let us compute the variance estimate, which is the variance of the bootstrapped estimates.

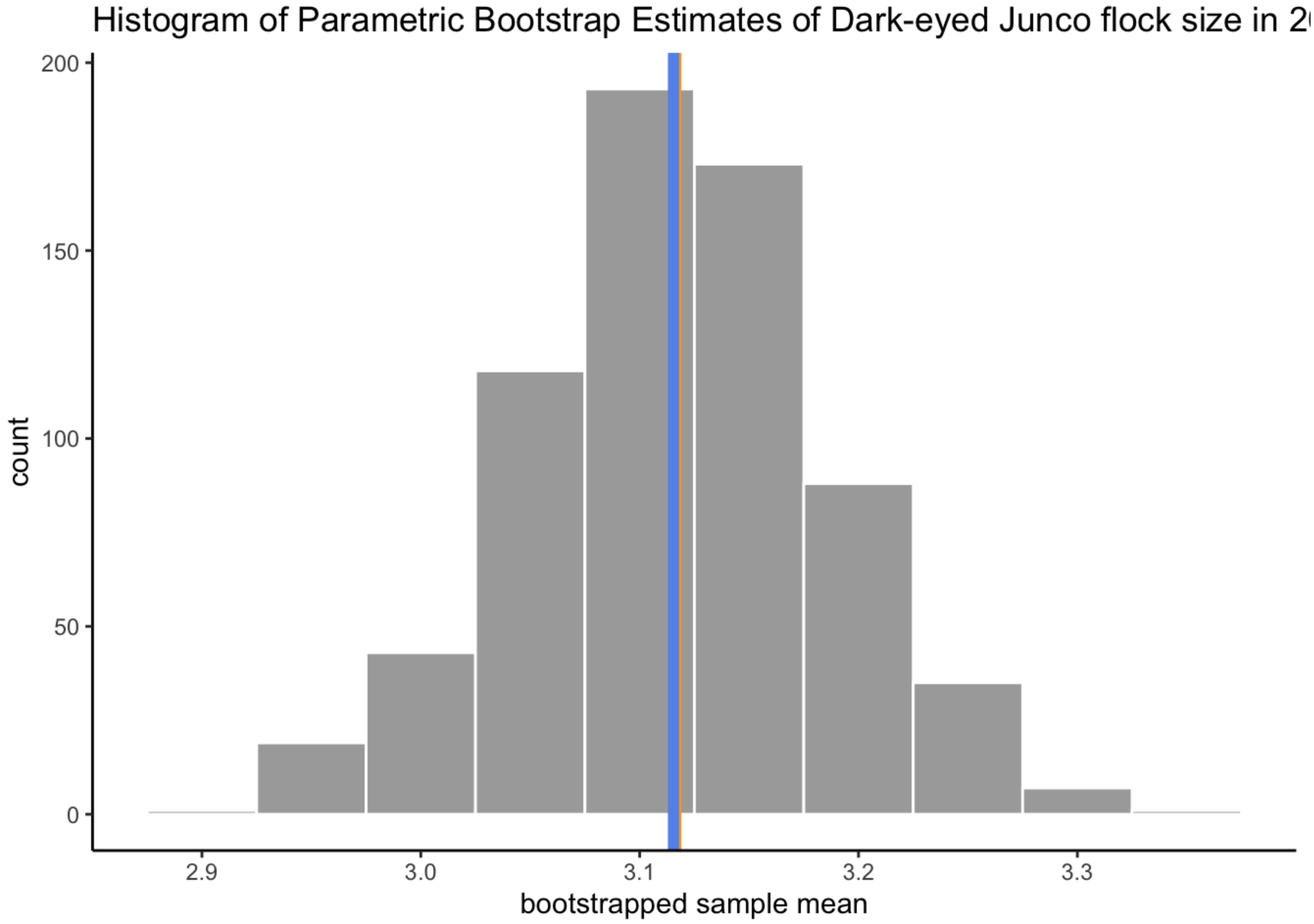
```
# an estimate of the variance of the average flock size sample mean estimate
p_var <- var(p_boot_mean_df$boot_mean)
p_var
```

```
## [1] 0.004994714
```

Let us compute a histogram of the parametric bootstrapped sample means below.

```
p_boot_mean_df %>%
  ggplot() +
  geom_histogram(aes(x = boot_mean), color = "white", fill = "darkgray",
                 binwidth = 0.05) +
  # line for the sample estimate of the mean
  geom_vline(xintercept = mean(dark_eyed_juncos_2021$max_individuals),
             color = "orange", size = 2) +
  # line for the bootstrapped estimate of the mean
  geom_vline(xintercept = mean(p_boot_mean_df$boot_mean),
             color = "cornflowerblue", size = 2) +
  labs(x = "bootstrapped sample mean",
       title = "Histogram of Parametric Bootstrap Estimates of Dark-eyed Junco flock siz
e in 2021") +
  theme_classic()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
N <- 1000
dark_eyed_juncos_2021 <- birds %>% filter(species_name == "Dark-eyed Junco", year == 202
1)

# draw some non-parametric bootstrap samples
np_boot_mean_df <- map_df(1:N, function(i) {
  # sample from the data with replacement
  bootstrap_data <- sample(dark_eyed_juncos_2021$max_individuals, length(dark_eyed_junco
s_2021$max_individuals), replace = TRUE)
  # compute the sample mean of the bootstrap sample
  data.frame(boot_mean = mean(bootstrap_data))
})
```

Let us compute the non-parametric bootstrap estimate of the sample mean bias.

```
np_bias <- mean(np_boot_mean_df$boot_mean) - mean(dark_eyed_juncos_2021$max_individuals)
np_bias
```

```
## [1] 0.001626844
```

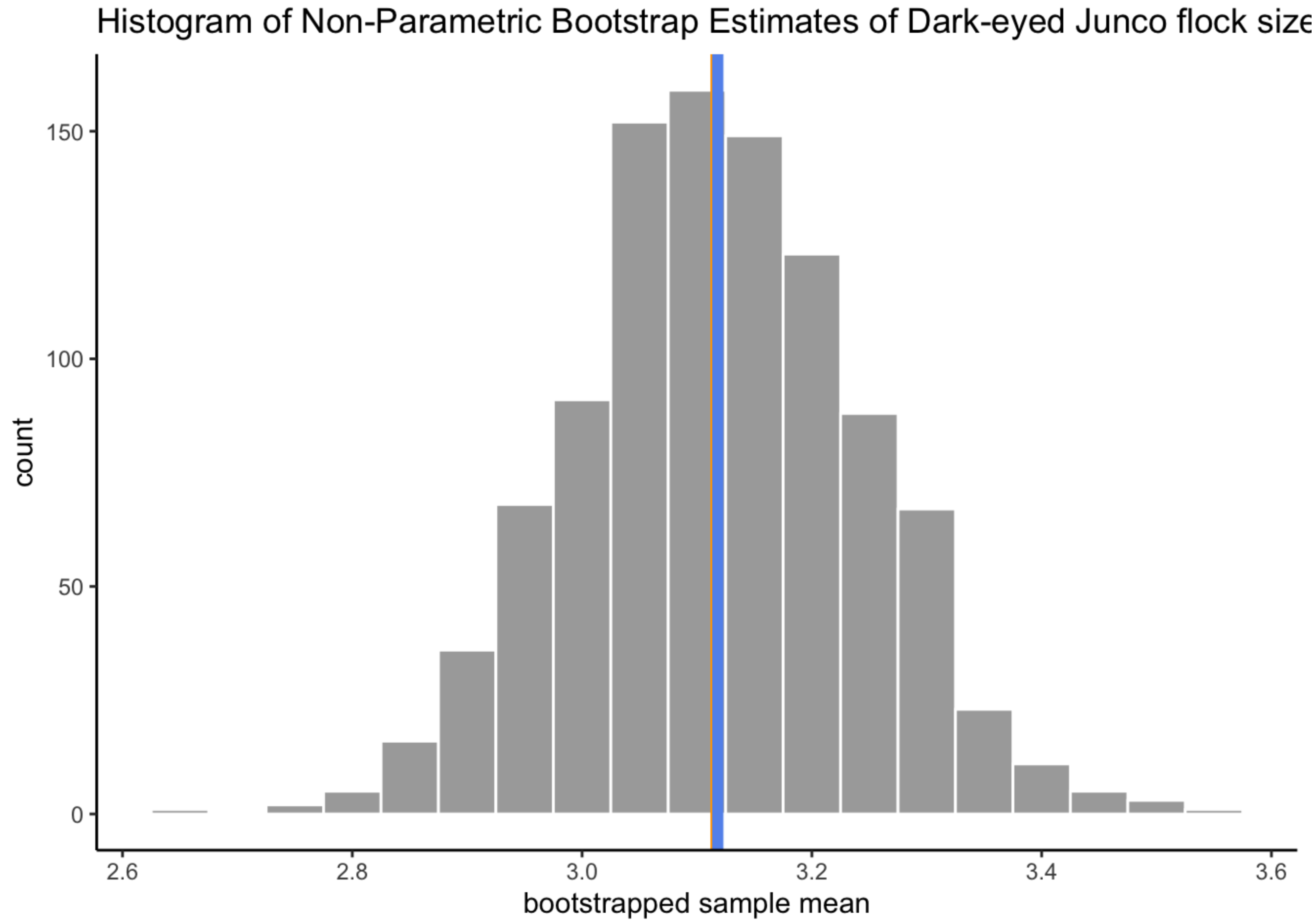
Let us compute the non-parametric bootstrap estimate of the sample mean standard deviation.

```
np_var <- var(np_boot_mean_df$boot_mean)
np_var
```

```
## [1] 0.01561724
```

Finally, let us compute a histogram of the bootstrapped sample means.

```
# a histogram of the bootstrapped sample means
np_boot_mean_df %>%
  ggplot() +
  geom_histogram(aes(x = boot_mean), color = "white", fill = "darkgray",
                 binwidth = 0.05) +
  # line for the sample estimate of the mean
  geom_vline(xintercept = mean(dark_eyed_juncos_2021$max_individuals),
             color = "orange", size = 2) +
  # line for the bootstrapped estimate of the mean
  geom_vline(xintercept = mean(np_boot_mean_df$boot_mean),
             color = "cornflowerblue", size = 2) +
  labs(x = "bootstrapped sample mean",
       title = "Histogram of Non-Parametric Bootstrap Estimates of Dark-eyed Junco flock
size in 2021") +
  theme_classic()
```


Histogram of Non-Parametric Bootstrap Estimates of Dark-eyed Junco flock size
count
bootstrapped sample mean