Data Science Module 5 Exercises

Brady Lamson

2/28/2022

9: Statistical Foundations

9.3: Simulations

Exercise 1:

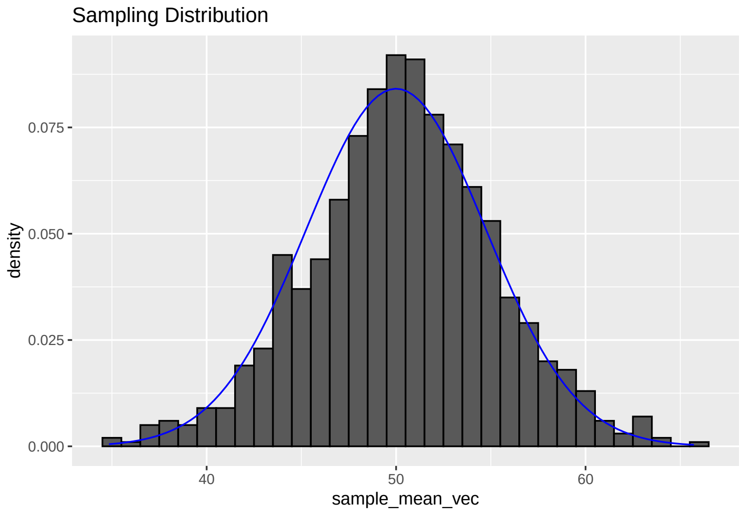
```r
# A
sample_mean_vec <- c()

for(i in 1:1000) {
    sim_sample <- rnorm(n = 10, mean = 50, sd = 15)
    sample_mean_vec <- c(sample_mean_vec, mean(sim_sample))
}
```

```r
#B
sample_mean <- mean(sample_mean_vec) %>% round(digits = 3)
sample_standard_error <- sd(sample_mean_vec) %>% round(digits = 3)

glue::glue("
    The mean of the sample mean vector is approximately {sample_mean},
    and the standard error of the vector is approximately {sample_standard_error}.
")
```

```
## The mean of the sample mean vector is approximately 50.311,
## and the standard error of the vector is approximately 4.892.
```

c) We can see that both values we got are very close to their theoretical values. The theoretical mean would be 50 and we can calculate the theoretical standard error, $\sigma/\sqrt{n} = 15/\sqrt{10} = 4.74342$.

```r
# D

ggplot(data = data.frame(sample_mean_vec)) +
    geom_histogram(
        mapping = aes(x = sample_mean_vec, y = stat(density)),
        binwidth = 1,
        color = "black") +
    geom_function(
        fun = dnorm,
        args = list(mean = 50, sd = 15/sqrt(10)),
        color = "blue") +
    labs(title = "Sampling Distribution")
```



The blue line represents an idealized normal distribution. What we can see is that our simulation comes incredibly close. The center is right around 50, the shape follows the same curve and the density matches as well.

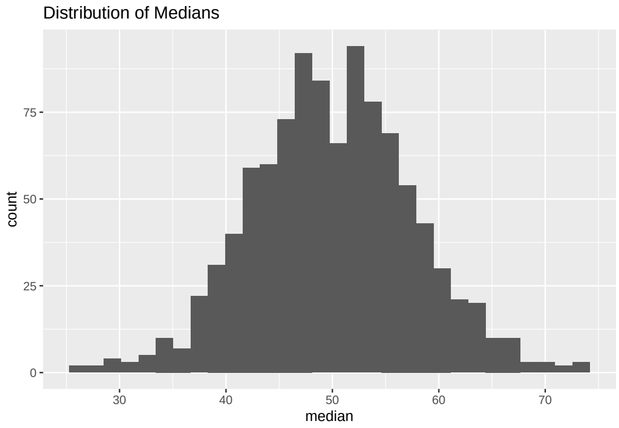
Exercise 2

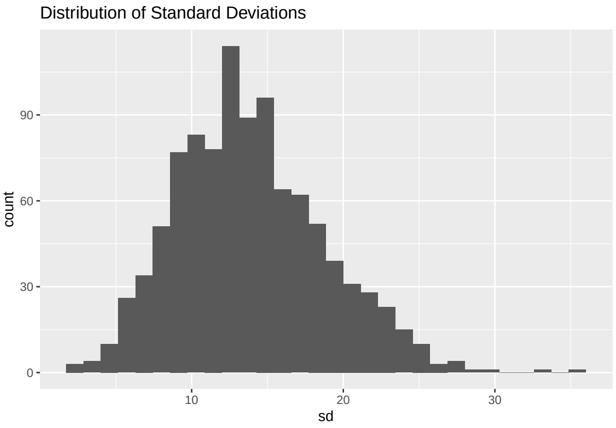
```r
sample_mean_vec <- c()
sample_median_vec <- c()
sample_sd_vec <- c()
sample_min_vec <- c()
sample_max_vec <- c()
sim_vec <- c()

for(i in 1:1000) {
    sim_sample <- rnorm(n = 5, mean = 50, sd = 15)
    sim_vec <- c(sim_vec, sim_sample)
    # simulated statistic vectors
    sample_mean_vec <- c(sample_mean_vec, mean(sim_sample, na.rm = TRUE))
    sample_median_vec <- c(sample_median_vec, median(sim_sample, na.rm = TRUE))
    sample_sd_vec <- c(sample_sd_vec, sd(sim_sample, na.rm = TRUE))
    sample_min_vec <- c(sample_min_vec, min(sim_sample, na.rm = TRUE))
    sample_max_vec <- c(sample_max_vec, max(sim_sample, na.rm = TRUE))
}
```

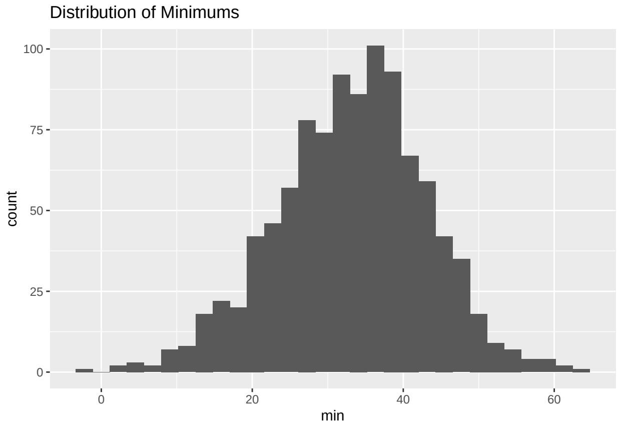
```r
simulation_stats <- tibble(
    median = sample_median_vec,
    sd = sample_sd_vec,
    min = sample_min_vec,
    max = sample_max_vec
)
```

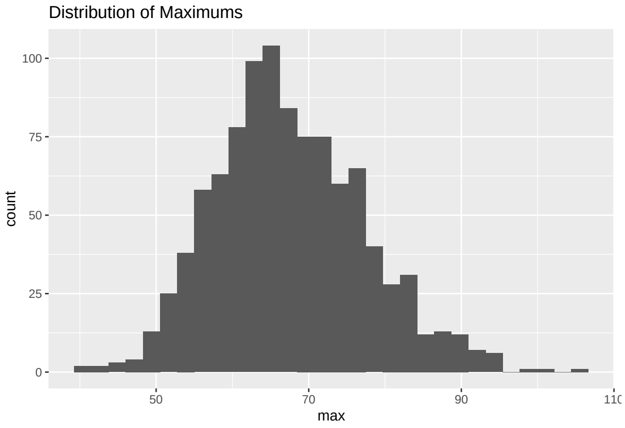
```r
data.frame(
    average = sapply(simulation_stats, FUN = mean),
    standard_error = sapply(simulation_stats, FUN = sd)
) %>%
    kbl()
```

	average	standard_error
median	50.01884	7.602022
sd	13.88497	4.879357
min	33.15363	9.889900
max	67.56669	9.821515

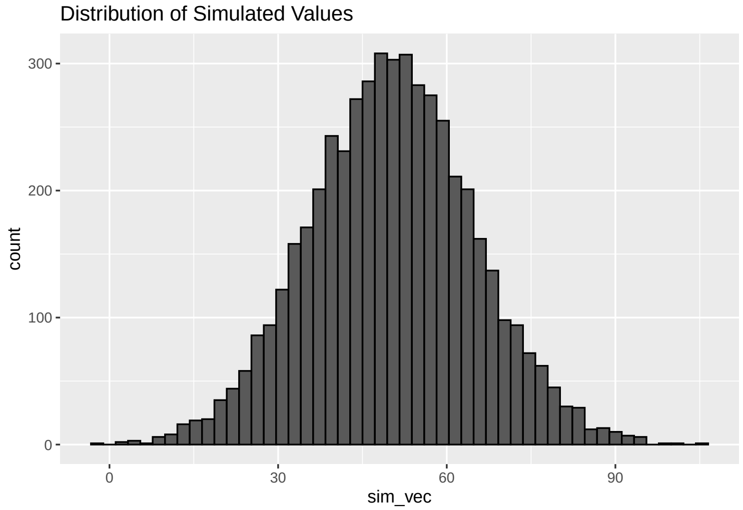








```
ggplot() +
    aes(x = sim_vec) +
    geom_histogram(bins = 50, color = "black") +
    labs(title = "Distribution of Simulated Values")
```



The distribution of the simulated values seems very much normal. The center is right around 50 and the count dips as the values get further and further away from the mean.

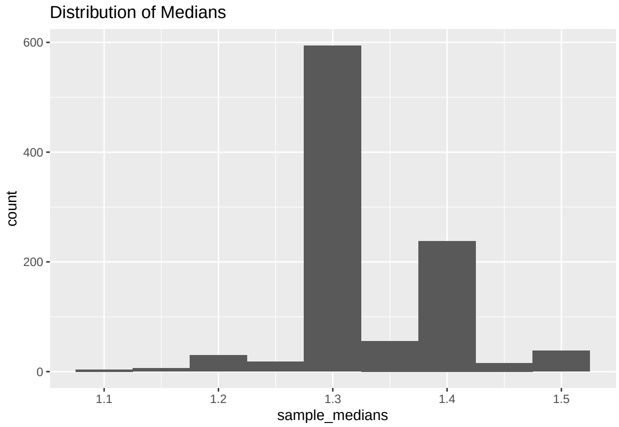
9.4: The Bootstrap

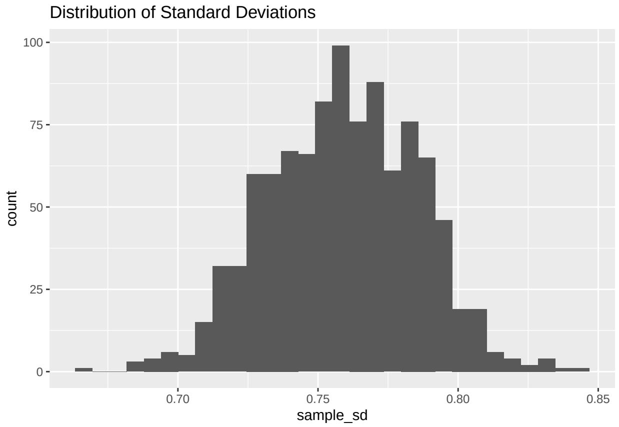
Exercise 3

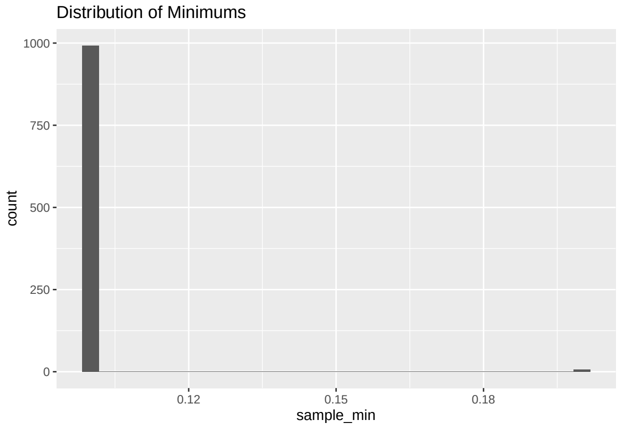
```
B <- 1000
sample_df <- tibble(
    sample_medians = rep(NA, B),
    sample_sd = rep(NA, B),
    sample_min = rep(NA, B),
    sample_max = rep(NA, B),
)

sim_vec <- c()

for(i in 1:B) {
    resamp <- slice_sample(.data = iris,
                           n = 150,
                           replace = TRUE)

    # Simulated values
    sim_vec <- c(sim_vec, resamp$Petal.Width)

    # Simulated statistics
    sample_df$sample_medians[i] <- median(resamp$Petal.Width)
    sample_df$sample_sd[i] <- sd(resamp$Petal.Width)
    sample_df$sample_min[i] <- min(resamp$Petal.Width)
    sample_df$sample_max[i] <- max(resamp$Petal.Width)
}
```

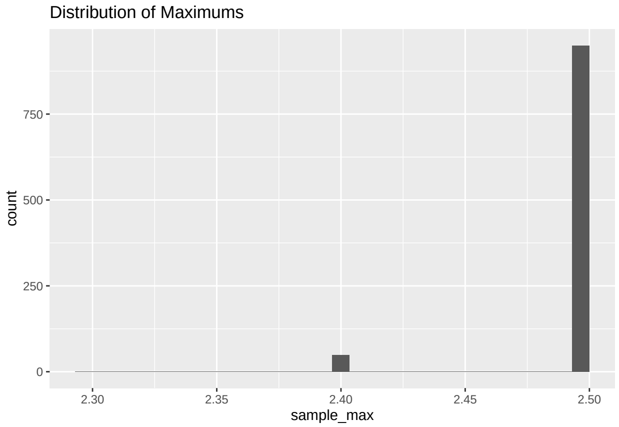
```
data.frame(
    average = sapply(sample_df, FUN = mean),
    standard_error = sapply(sample_df, FUN = sd)
) %>%
    kbl()
```

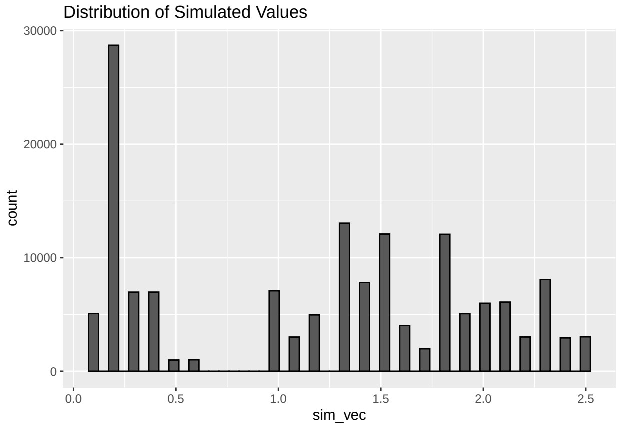
	average	standard_error
sample_medians	1.3310000	0.0638990
sample_sd	0.7589661	0.0266265
sample_min	0.1007000	0.0083414
sample_max	2.4949000	0.0224610











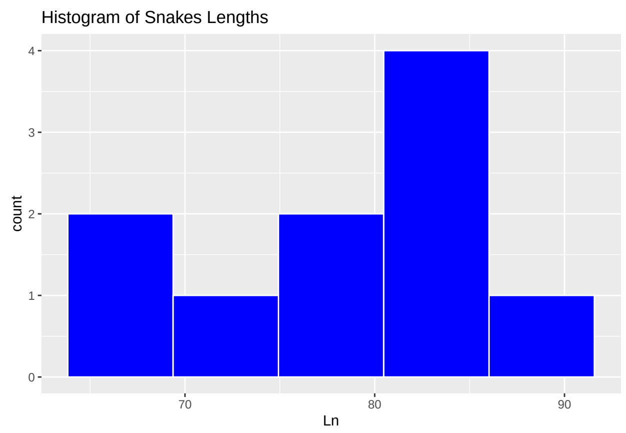
All of the plots here are very much **not** normal. The medians have two values that are far more frequent than any others, whereas the minimums and maximums only really have 1 value that occurs at all. The standard deviations seem to have a relatively normal distribution though, with a slight left skew. As for the general simulated values, There's a very non normal distribution with a peak close to 0 and a fair frequency of values between 1.3 and 1.8.

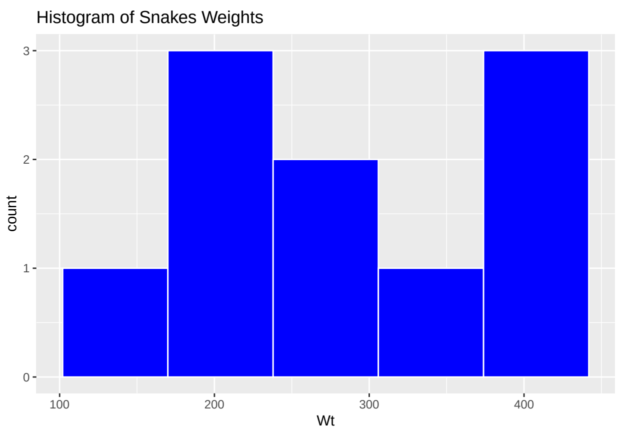
9.5: Outliers

Exercise 4

```
SnakeID <- 1:10
Ln <- c(85.7, 64.5, 84.1, 82.5, 78.0, 65.9, 81.3, 71.0, 86.7, 78.7)
Wt <- c(331.9, 121.5, 382.2, 287.3, 224.3, 380.4, 245.2, 208.2, 393.4, 228.3)
Snakes <- data.frame(SnakeID, Ln, Wt)
```

```
ggplot(data = Snakes) +
    geom_histogram(mapping = aes(x = Ln),
                   fill = "blue",
                   color = "white",
                   bins = 5) +
    ggtitle("Histogram of Snakes Lengths")

ggplot(data = Snakes) +
    geom_histogram(mapping = aes(x = Wt),
                   fill = "blue",
                   color = "white",
                   bins = 5) +
    ggtitle("Histogram of Snakes Weights")
```



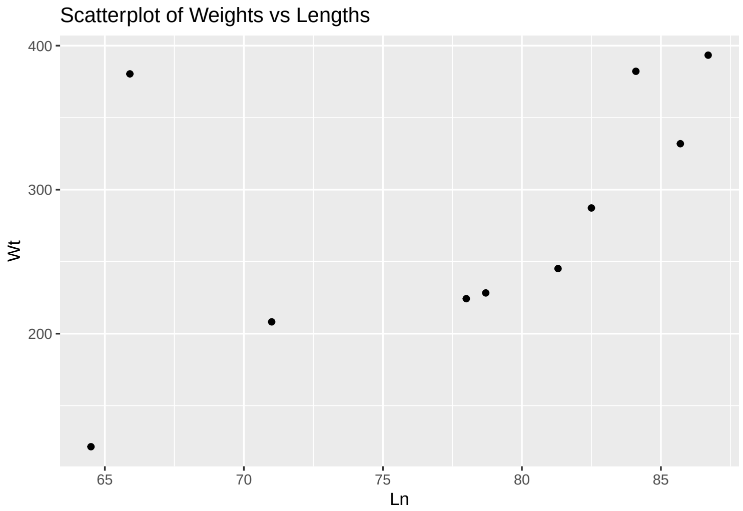


a)

From these two histograms I absolutely cannot tell what the outlier is. There's too few bins to properly differentiate here.

b)

```
ggplot(data = Snakes) +
    geom_point(mapping = aes(x = Ln, y = Wt)) +
    ggtitle("Scatterplot of Weights vs Lengths")
```



Here it is a lot easier to notice the outlier, it's the value in the top left.