Matthew Jusino

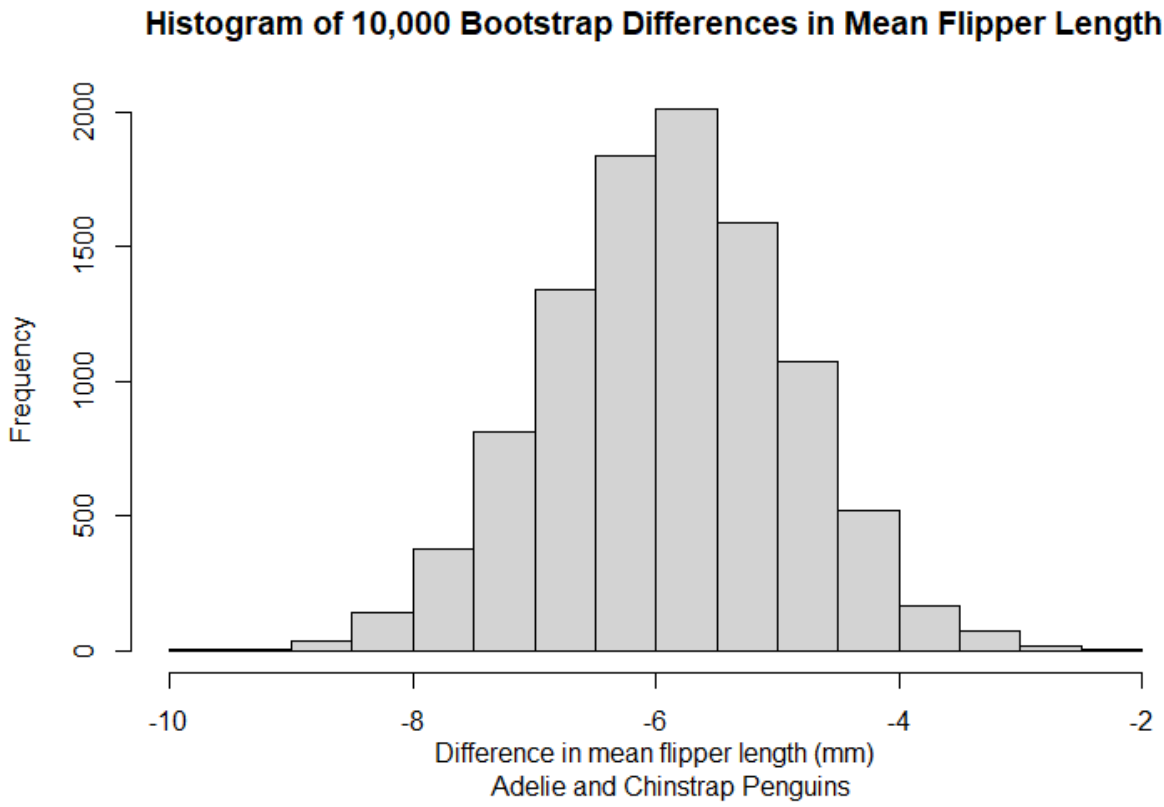
Lab 8

- **Q1 (1 pt.):** Calculate the standard deviation of the differences in mean flipper length from your bootstrap simulation. Show the R-code you used to find do the calculation.

```
sd(pen_boot$t)
```

sd = 0.9971849

- **Q2 (2 pts.):** Include your histogram of bootstrapped differences in your lab report (you don't need to show the R-code but make sure your plot includes appropriate title, axes, etc.).



- **Q3 (2 pts.):** What was the 95% bootstrap CI you calculated using `quantile()`? Show the R-code you used to answer the question.

```
quantile(pen_boot$t, probs = c(0.025, 0.975))
```

2.5% 97.5%

-7.862110 -3.989321

- **Q4 (4 pts.):** Do you think the resampled differences in means follow a skewed distribution? Your answer should make reference to the *mean*, *median*, **and** *histogram* of the differences in means.

The histogram of the distribution shows a nearly normal distribution centered around the mean of -5.898, determined using the summary function. The median, -5.884, is almost the same, the difference could be considered negligible, meaning that is doesn't really follow a skewed distribution. For all intents and purposes, it can be treated as a normal distribution.

- **Q5 (2 pts.):** Show the R-code you used to create `pen_ecdf()`

```
pen_ecdf = ecdf(pen_boot$t)
```

- **Q6 (2 pts.):** What is the probability, according to the empirical distribution function, of observing a mean difference of -4.5 *or greater*? Show the R code you used to perform the calculation.

```
1 - pen_ecdf(-4.5)
```

0.0781 or a 7.81% chance

- **Q7 (2 pts.):** What is the probability, according to the empirical distribution function, of observing a mean difference of -8 *or smaller*? Show the R code you used to perform the calculation.

```
pen_ecdf(-8)
```

0.0185 or a 1.85% chance

- **Q8 (3 pts.):** State the null and alternative hypotheses of a *two-sample*, *two-tailed* test for the difference in mean flipper lengths between the two penguin species.

The null hypothesis for the test would be that the mean flipper lengths are both equal. And if either species is greater or less than the other species, that would be the alternative hypothesis, that their mean flipper length are not equal.

- **Q9 (2 pts.):** What was the p-value? Show the R-code you used to find out.

```
pine_cont = droplevels(subset(dat_veg, treatment == "control"))

pine_clip = droplevels(subset(dat_veg, treatment == "clipped"))

wilcox.test(pine_cont$pine, pine_clip$pine)
```

p-value = 0.1005

- **Q10 (1 pt.):** What were the endpoints of your bootstrap CI? Show the R-code you used to find out.

```
quantile(tree_boot$t, probs = c(0.025, 0.975))
```

```
2.5%    97.5%
4.25    30.00
```

- **Q11 (1 pt.):** What is the observed difference in mean tree counts and does it fall within the 95% bootstrap CI?

The observed difference in mean tree counts is 16, which falls within the 95% bootstrap CI of 4.25 to 30.

- **Q12 (2 pts.):** Briefly describe the Simpson diversity index and explain what it quantifies.

The Simpson diversity index is a measure of the richness and evenness of different species in a population. It quantifies the diversity of species along a scale of 0 to 1, with greater values having greater sample diversity.

- **Q13 (2 pts.):** Show the code you used to z-standardize the `s.sidi` column.

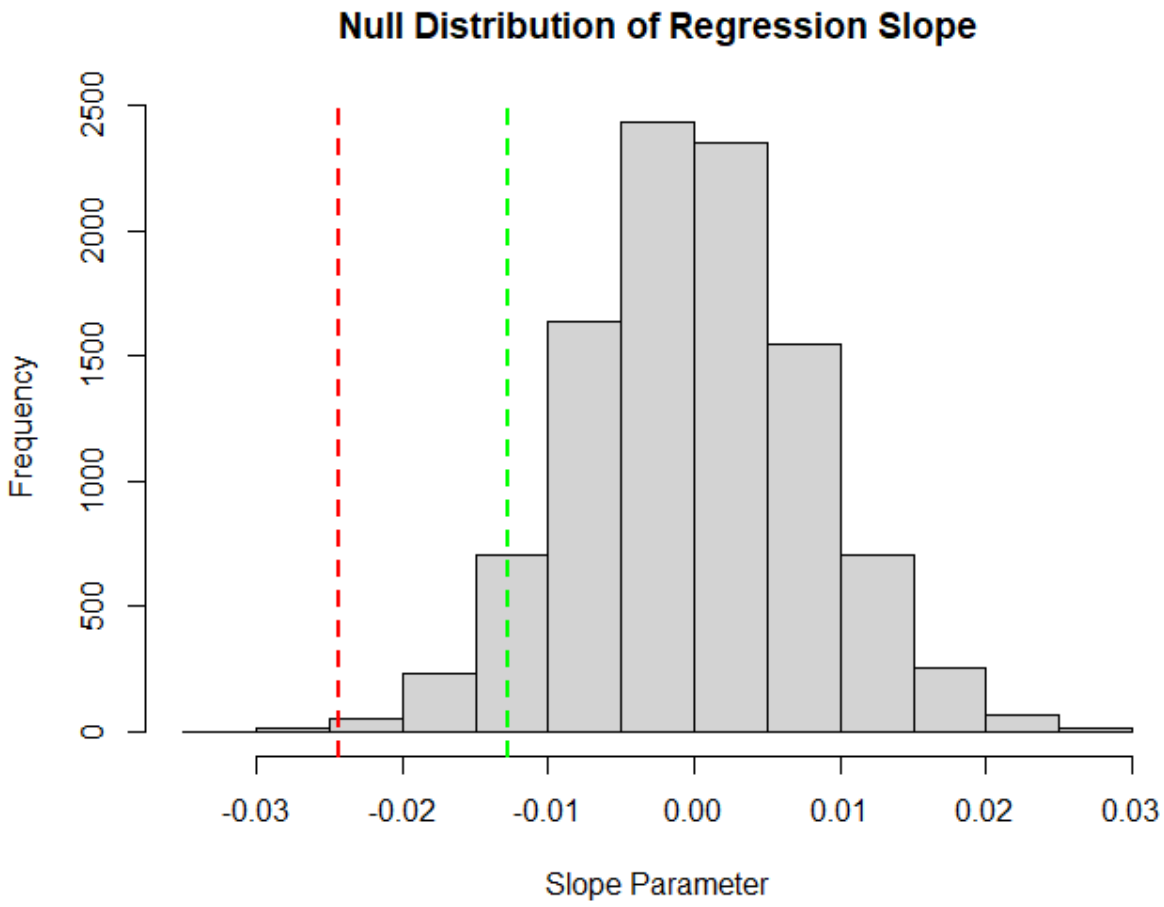
```
s_sidi_mean = mean(dat_all$s.sidi, na.rm = TRUE)

s_sidi_sd = sd(dat_all$s.sidi, na.rm = TRUE)

dat_all$s.sidi.standardized = (dat_all$s.sidi - s_sidi_mean)/s_sidi_sd

mean(dat_all$s.sidi.standardized)
```

- **Q14 (6 pts.):** Show the code for your completed Monte Carlo simulation loop.

```
m = 10000

result_mc = numeric(m)

for(i in 1:m)

{

 index_1 = sample(nrow(dat_1), replace = TRUE)

 index_2 = sample(nrow(dat_1), replace = TRUE)

 dat_resampled_i = data.frame(b.sidi = dat_1$b.sidi[index_1], s.sidi = dat_1$s.sidi[index_2])

 fit_resampled_i = lm(b.sidi ~ s.sidi, data = dat_resampled_i)

 result_mc[i] = coef(fit_resampled_i)[2]
```

}

- **Q15 (2 pts.):** In your report, include a plot of your histogram of Monte Carlo resampled slopes. Include vertical lines showing the observed slope and the critical value from the resampled MC slopes.



- **Q16 (2 pts.):** What was your critical value? Was the observed slope less than the critical value?

The critical value was -0.01278298. The observed slope was less than the critical value, with a value of -0.02437131

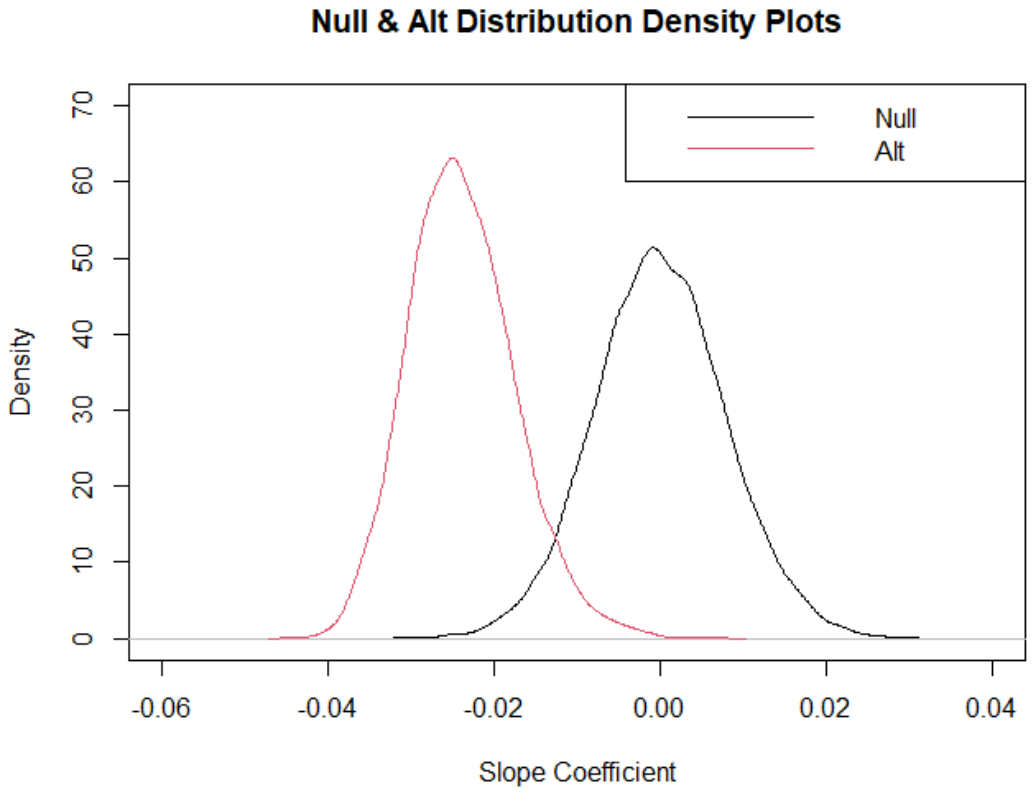
- **Q17 (3 pts.):** What is your conclusion regarding the evidence of a negative relationship between vegetation cover diversity and bird diversity? Make sure to justify your conclusions using the results of your simulation analysis.

It would appear that there is indeed a negative relationship between vegetation cover diversity and bird diversity, as the observed slope of the negative trendline in the actual data set was approximately double the magnitude of the slope of the negative trendline in a Monte Carlo resampled data set, showing that there is a significant difference between the data observed and how it would look with no association between the two diversities.

- **Q18 (2 pts.):** Show the code you used in your bootstrap loop.

```
m = 10000

result_bs = numeric(m)

for(i in 1:m)

{

 bs_index_1 = sample(nrow(dat_1), replace = TRUE)

 dat_boot = dat_1[bs_index_1, ]

 fit_bs1_i = lm(b.sidi ~ s.sidi, data = dat_boot)

 result_bs[i] = coef(fit_bs1_i)[2]

}
```

- **Q19 (4 pts.):** Include your double density plot. For full credit your plot must include:
  - a legend
  - the two density curves, in different colors
  - appropriate axis labels and title



- **Q20 (2 pts.):** Recall that the bootstrap curve shows the distribution of plausible values for the slope coefficient if we could resample the original data. The Monte Carlo curve shows the distribution of plausible values for the slope coefficient if the null hypothesis were true. How can you interpret the region that falls under both curves?

I would interpret the region that falls under both curves as a region of uncertainty. Values that fall in this region could support either the null or the alternative hypothesis, and as such these values cannot be used to make any meaningful conclusions about the data.