Bonnie Turek

Eco 634 – Lab 8: Modeling Continuous Data 1

10/27/21

*I worked with Matt, John, and Mandy on this assignment.

Penguin Boot 1

Q1. 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, na.rm = TRUE)
```

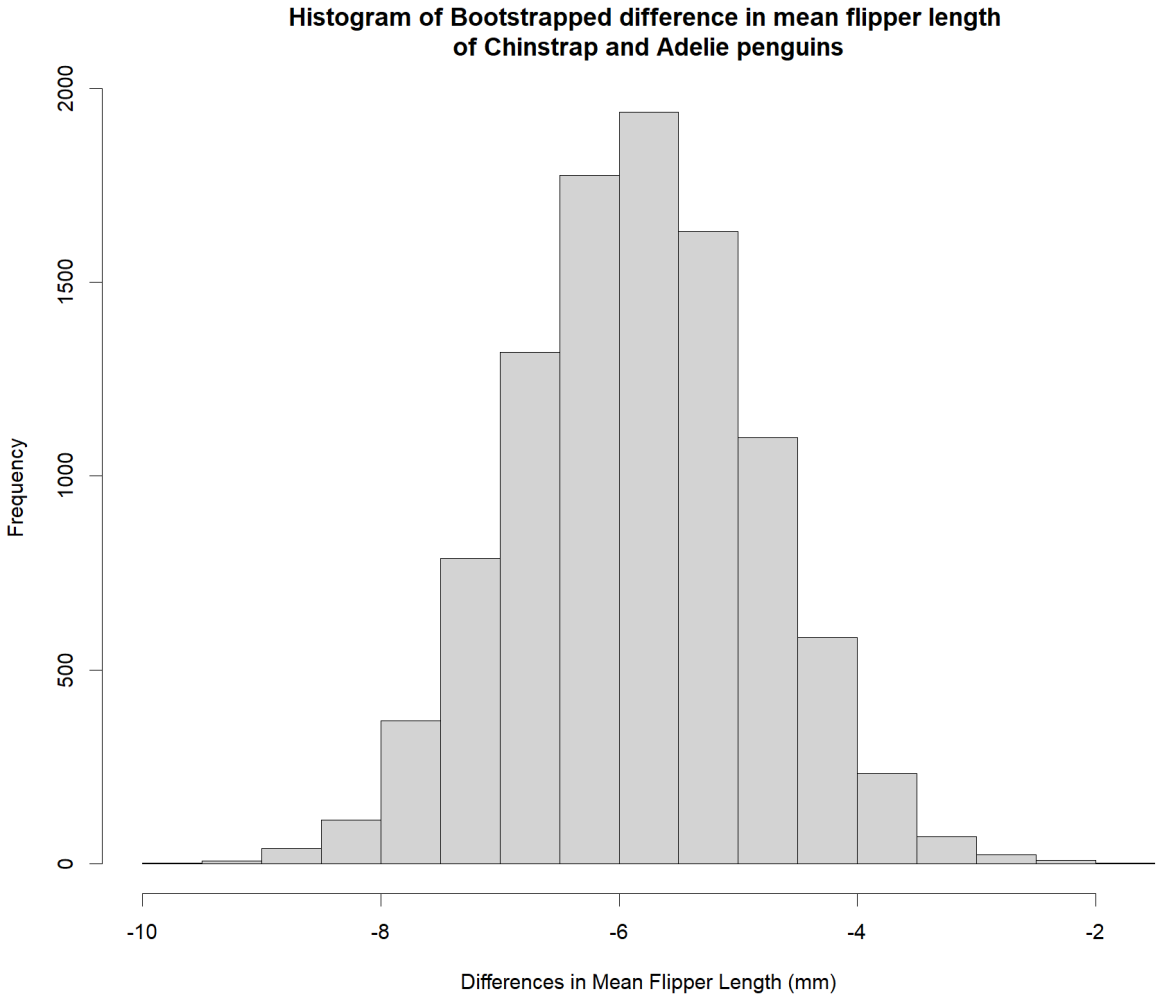
#t is the differences in mean flipper length of the two boot output, that's why we subset t from pen_boot

Std. Dev = **1.016693**

R Code:

```
install.packages("boot")

require(boot)

install.packages("simpleboot")

require(simpleboot)

#subsetting penguin data so we only have Adelie and Chinstrap species with only Flipper length

pen_dat = penguin_dat[,0:5]

pen_dat2 = pen_dat[, -c(2,3,4)]

adelie_dat = droplevels(subset(pen_dat2, species != "Chinstrap"))

chinstrap_dat = droplevels(subset(pen_dat2, species != "Adelie"))

#run actual two bootstrap

pen_boot = two.boot(sample1 = adelie_dat$flipper_length_mm,

                    sample2 = chinstrap_dat$flipper_length_mm,

                    FUN = mean, R = 10000, na.rm = TRUE)

print(pen_boot)
```

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



R Code:

```
hist(pen_boot$t, main = "Histogram of Bootstrap difference in mean flipper length

of Chinstrap and Adelie penguins",

    xlab = "Differences in Mean Flipper Length (mm)")
```

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

quantile(pen_boot$t,c(0.025, 0.975), na.rm = TRUE)

2.5% 97.5%

-7.817353 -3.889847

Q4. 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.

sd(pen_boot$t, na.rm = TRUE) = 1.016693

mean(pen_boot$t, na.rm = TRUE) = -5.852683

median(pen_boot$t, na.rm = TRUE) = -5.853476

No, I don't think the resampled differences in means follow a skewed distribution. Just looking at the histogram, they actually appear to follow a normal distribution. You can also see where the mean and median values of -5.85 fall, right in the center of the normally-distributed histogram of diff in means. If the resampled differences in means followed a skewed distribution, the histogram bins would appear to be heavier on either the left or right side. This is not the case. Most of the frequent differences in means fall around the -5.8mm value. The mean and median values being so close is also an indicator that the distribution is not skewed.

Penguin ECDF

Q5. Show the R-code you used to create `pen_ecdf()`

pen_ecdf = ecdf(pen_boot$t)

Q6. 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.

#mean difference of -4.5 or greater

1 - pen_ecdf(-4.5)

0.0836 = 8.4%

This makes sense when looking at the histogram of the differences in means too! There is only a small portion of the distribution in this range.

Q7. 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.

#mean difference of -8 or smaller?

pen_ecdf(-8)

> 0.0159 = 1.6%

This also makes sense in looking at the histogram of differences in means. The majority of values of differences in means are centered around -5.8, but -8 and smaller are towards the far left of the histogram x-axis, where there are much less frequent occurrences of differences in means this small.

Hypotheses

Q8. 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.

Null hypothesis:

The two penguin species, Adelie and Chinstrap, DO NOT differ in mean flipper lengths.

Alternative hypothesis:

The two penguin species, Adelie and Chinstrap, DO differ in mean flipper lengths.

Pines – Non-parametric test

Q9. What was the p-value? Show the R-code you used to find out.

#wilcoxon test on two treatments on pines

wilcox.test(pine ~ treatment, data = dat_tree)

Wilcoxon rank sum test with continuity correction

>W = 48, **p-value = 0.1005**

alternative hypothesis: true location shift is not equal to 0

The p value of 0.1 is not below the 0.05 criteria to reject the null. So based on the resulting p-value we assume the difference in the mean number of pine seedlings in between the treatments of control and clipped DO NOT differ significantly. We cannot reject the null.

Pines – Bootstrap

1. Use `two.boot()` to create a bootstrapped data set of the differences in mean tree count between the clipped and control treatments.
    - Save your results as `tree_boot`.
2. Use `quantile()` to find a 95% CI.

Q10. What were the endpoints of your bootstrap CI? Show the R-code you used to find out.

The endpoints of my 95% Bootstrap CI were **4.12 and 29.63.**

```
#pines - bootstrap

tree_boot =

  two.boot(

    subset(dat_tree, treatment == "clipped")$pine,

    subset(dat_tree, treatment == "control")$pine,

    FUN = mean,

    R = 10000,

    na.rm = TRUE

  )

# sum(tree_boot$t >= 0)

# sum(tree_boot$t < 0)

boot.ci(tree_boot)

hist(tree_boot$t, main = "Bootstrap sampling distribution")

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

Q11. What is the observed difference in mean tree counts and does it fall within the 95% bootstrap CI?

The observed difference in mean pine tree counts between the clipped and control treatments is **16.**

This DOES fall within the 95% Bootstrap CI of 4.12-29.63. The difference in observed means between clipped and control in our sample actually follows along nicely with the mean of the bootstrapped difference in means (mean(tree_boot$t) = 16.063).

```
#R code to calc orig data diff in means

dat_control = droplevels(subset(dat_tree, treatment == "control"))

dat_clipped = droplevels(subset(dat_tree, treatment == "clipped"))

control_mean = mean(dat_control$pine)

clipped_mean = mean(dat_clipped$pine)

dif_means = clipped_mean - control_mean

print(dif_means)
```

>dif_means = 16

Resampling Model Coefficients

Q12. Briefly describe the Simpson diversity index and explain what it quantifies.

According to the provided metadata, the Simpson diversity index is a stand-based diversity index (proportion), where each stand is a disjunct patch based on floristic community, seral stage and canopy closure. In other words, it's a measure of diversity which takes into account the number of species present, as well as the relative abundance of each species based on vegetation cover types. Specifically vegetation cover types relates to the diversity in landscape composition as defined largely by vegetation seral (intermediate stages of succession) stage.

Q13. Show the code you used to z-standardize the s.sidi column.

```
#standardize s.sidi now

# Calculate the sample mean and sd:

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

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

# Use the subset-by-name symbol ($) to create a new column of z-standardized values.

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

Q14. Show the code for your completed loop.

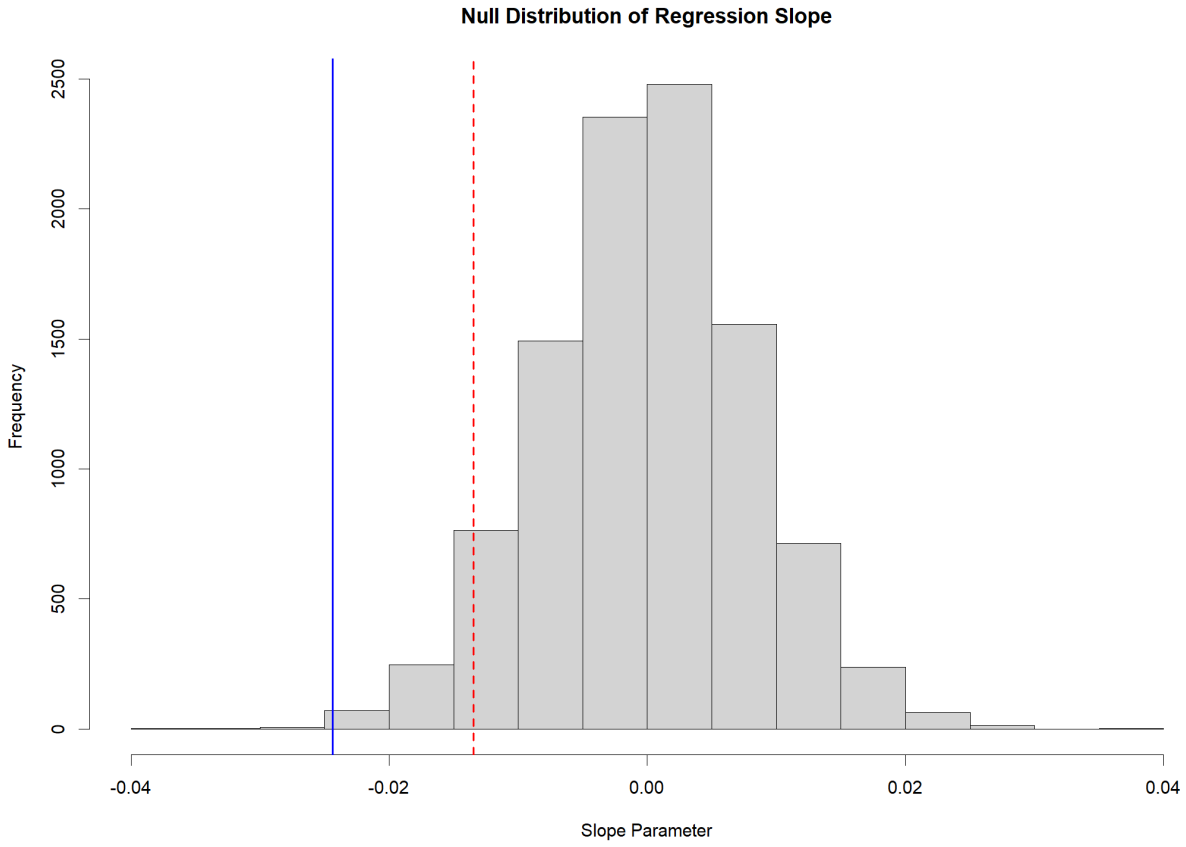
```
#use a loop to resample many times to create null dist
m = 10000
result = numeric(m)
for(i in 1:m)
{
  index_1 = sample(nrow(dat_1), replace = TRUE)
  index_2 = sample(nrow(dat_1), replace = TRUE)
  dat_resampled_i =
    data.frame(
      b.sidi = dat_1$b.sidi[index_1],
      s.sidi = dat_1$s.sidi[index_2]
    )
  fit_resampled_i = lm(b.sidi ~ s.sidi, data = dat_resampled_i)
  result[i] = coef(fit_resampled_i)[2]
}
```

Q15. In your report, include a plot of your histogram with vertical lines showing the observed slope and the critical value from the resampled MC slopes.

hist(result, main = "Null Distribution of Regression Slope", xlab = "Slope Parameter")

abline(v = slope_observed, lty = 1, col = "blue", lwd = 2)

abline(v = slope_montecarlo, lty = 2, col = "red", lwd = 2)



Q16. What was your critical value? Was the observed slope less than the critical value?

slope_montecarlo = quantile(result, c(.05))

print(slope_montecarlo)

5%

-0.01344603 = Critical value

Yes, the observed slope (blue solid line: -0.024) was less than the critical value (red dotted line: -0.013)

Q17. 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 analysis.

Here we used Monte Carlo randomization to assess the significance of regression parameters, specifically the slope. We standardized both Simpson's diversity index for breeding birds: (`b.sidi`) and Simpson's diversity index for vegetation cover types (s.sidi) so that these two model variables were comparable / on the same range.

At first, we observed a negative relationship between bird diversity and vegetation diversity; specifically, bird diversity declines as the vegetation diversity increases. However, this seems counter-intuitive since we would expect greater vegetation density would result in greater habitat availability and therefore greater bird diversity.

Our resampling technique (Monte Carlo) breaks the associations among the data so to assume that there is no real relationship between bird diversity and vegetation diversity. Then we ran a loop to resample and create a null distribution. We can then compare how our critical value slope under the null distribution compares to the observed slope parameter from the sample.

In looking at the resulting observed slope versus monte carlo resample critical slope, as well as observing the null distribution histogram and red and blue lines:

We can conclude that the original observed slope parameter value (-0.024) falls outside of the threshold critical value of (-0.013), so we have strong evidence to reject the null that there is no real relationship between bird diversity and vegetation diversity. In fact, it is likely then that there IS an observed negative relationship between bird diversity and vegetation diversity. This relationship did not occur by chance. If it did occur by chance, we would expect the observed slope value to fall within the larger "hump" of values in the null distribution.