Week 4 Day 2: Inferences for Diving Penguins

Recall the Context of Today's Exploration

Let's remember what the purpose of this study was. For the penguin study, researchers equipped emperor penguins with devices that record their heart rates during dives. The data we analyzed contained Dive Heart Rate (beats per minute), the Duration (minutes) of dives, and other related variables. These researchers were interested in studying if there was evidence of an association between a penguin's dive heart rate and the duration of their dive?

1. Write out the null hypothesis **in words**.

2. Using the research question, rewrite the hypotheses using notation. Use the slope as the summary measure.

Inference for the Slope Coefficient

3. Based on the slope coefficient you found yesterday, do you think there is convincing evidence of an association between a penguin's heart rate and the duration of their dive?

Hypothesis Testing

In a hypothesis test we are interested in comparing two things:

- what we observed in our data (our observed slope)
- what would have happened if the null hypothesis was true

In order to compare what we saw in this study to what would have happened if the null hypothesis was true, we need to generate a distribution of slopes that we would have expected to see if the null was true. This is called our **null distribution**. We then see where our observed slope falls on that distribution.

If our observed statistic falls in the middle of the distribution, then it is fairly likely to happen if the null is true. However, if our observed statistic falls in the tails of the distribution, then it is unlikely to happen if the null is true.

Simulation-Based Methods

Similar to what we talked about with confidence intervals and bootstrapping, a null distribution is also a *sampling distribution*. It is, however, a special type of sampling distribution. It is a distribution of sample statistics that could have been observed **if the null hypothesis was true**.

Much like we used a bootstrap or a *t*-distribution to approximate the true sampling distribution, there are two ways to approximate the true null distribution, (1) using simulation or (2) using a *t*-distribution.

In this activity we will focus on option 1, but in the lab we will focus on option 2.

Simulating what could have happened under the null hypothesis

Let's start by thinking about how one simulation would be created on the null distribution using cards.

Step 1: Write each penguin's heart rate and dive duration on 125 cards.

Step 2: (assume the null hypothesis is true)

Step 3: (generate a new sample)

Step 4: (calculate the statistic of interest)

How do we do this in `R`?

We will use the **infer** package to help us simulate what could have happened if the null was true. The layout for the code looks similar to what you saw in last week's activity, so let's try and fill in what is missing.

```
diving %>%
  specify(response = ________,
          explanatory = ________) %>%
  hypothesise(null = "independence") %>%
  generate(reps = ________,
           type = ________) %>%
  calculate(stat = "________")
```

4. What inputs should be entered for each of the following to create the simulation to test regression slope?

- Response variable (choose `Duration` or `Dive_HeartRate`):

- Explanatory variable (choose `Duration` or `Dive_HeartRate`):

- Number of repetitions:

- Type of method to use when generating new samples (choose `"permute"` or `"bootstrap"`):

- Summary measure (choose `"slope"` or `"correlation"`):

5. Suppose we wanted to complete the simulation test using correlation as the summary measure, instead of slope. Which input(s) in #4 would need to be changed to test for correlation? What input(s) should you use instead?

Here is a preview of what the output of this code looks like:

```
glimpse(null_dist)
```

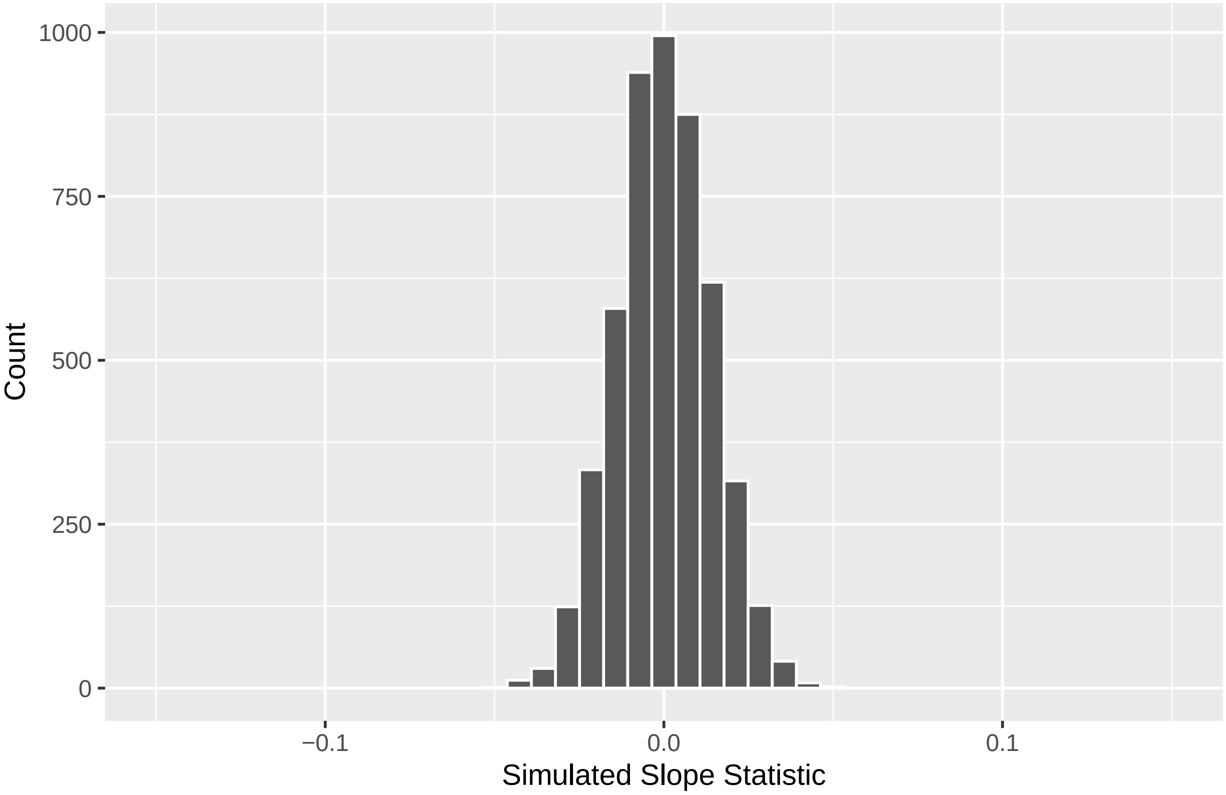
```
## Rows: 5,000
## Columns: 2
## $ replicate <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ stat      <dbl> -0.0032294944, -0.0082793790, 0.0022390702, 0.0038737611, 0.~
```

6. What does the `replicate` column represent?

7. What does the `stat` column represent?

Visualizing the Null Distribution

Below is a null distribution of slope statistics, generated using the code you completed above. I chose to use 5000 reps, which might be slightly different from what you chose.



8. Why is the distribution centered at 0?

9. Mark where the observed slope on the distribution.

10. Shade or circle what area you will use to calculate the p-value.

11. Estimate what p-value we will obtain from `R`.

Calculating the p-value

Since our distribution consists of 5000 slope statistics, we will need to use `R` to find our p-value. To find a p-value, `R` needs to first know what our observed statistic is. We can actually do this using some of the code from before!

What I'm doing in the code below is:

- making a new object named `obs_stat` that contains the number for the observed slope
- using the `diving` dataset
- `specifying` what variables to use
- `calclating` the statistic we are interested in

```
obs_slope <- diving %>%
  specify(response = Duration,
          explanatory = Dive_HeartRate) %>%
  calculate(stat = "slope")
```

The next part is to compare this `obs_stat` to the distribution of shuffled slope statistics (stored in `null_dist`).

```
get_p_value(null_dist,
            obs_stat = ________,
            direction = "________")
```

12. What inputs should be entered for each of the following to calculate the p-value?

- `obs_stat =`

- `direction =` (`"greater"`, `"less"`, or `"two-sided"`):

Using your inputs, I obtained the p-value for our observed slope statistic. The p-value is:

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of ‘reps‘ chosen in the ‘generate()‘ step. See
## ‘?get_p_value()‘ for more information.

## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

13. Why did I get a warning message from `R`? What does the warning tell me?

Communicate the results and answer the research question

12. Based on the p-value, write a conclusion in context of the problem.

13. Based on the p-value you obtained for testing the slope, what p-value do you think you would get if you tested the correlation instead? *Hint: think about the relationship between slope and correlation!*

Connection to Confidence Intervals

14. If you were to make a 95% confidence interval for the true slope (β_1‘), would the interval contain 0? Why or why not?

Take-home messages

1. To create one simulated sample on the null distribution when testing for a relationship between two quantitative variables, we separate the x-values from the y-values. We then shuffle the y-values and pair them with a new x-value. Once we have a new dataset, we find the regression line for the shuffled data and plot the slope or the correlation for the shuffled data.
2. To obtain a p-value for the observed slope we need to carry out the following steps:

- obtain a distribution of statistics that could have happened if the null was true (null distribution)
- locate the observed slope on the null distribution
- count how many shuffled slopes are as large or larger than the observed slope
- divide the number of slopes by the total number of reps used (e.g., $\frac{4}{1000}$)
- multiply this by two, since we almost always have a two-sided alternative

3. The p-value for a test for correlation should be approximately the same as the p-value for the test of slope. In the simulation test, we just change the statistic type from slope to correlation and use the appropriate sample statistic value.