Activity 4: Diving Penguins

10/10/2022

Terminology review

In today's activity, we will use simulation-based methods for conducting a hypothesis test for a linear regression slope or correlation. Some terms covered in this activity are:

- Correlation
- Slope
- Regression line

Diving Penguins

Emperor penguins are the most accomplished divers among birds, making routine dives of 5–12 minutes, with the longest recorded dive over 27 minutes. These birds can also dive to depths of over 500 meters! Since air-breathing animals like penguins must hold their breath while submerged, the duration of any given dive depends on how much oxygen is in the bird's body at the beginning of the dive, how quickly that oxygen gets used, and the lowest level of oxygen the bird can tolerate. The rate of oxygen depletion is primarily determined by the penguin's heart rate. Consequently, studies of heart rates during dives can help us understand how these animals regulate their oxygen consumption in order to make such impressive dives.

In this study, the researchers equipped emperor penguins with devices that record their heart rates during dives. The data set reports Dive Heart Rate (beats per minute), the Duration (minutes) of dives, and other related variables. Is there an association between dive heart rate and the duration of the dive?

```
diving <- read_csv("Diving_Penguins.csv")
glimpse(diving)
```

```
Rows: 125
Columns: 4
$ Dive_HeartRate <dbl> 88.8, 103.4, 97.4, 85.3, 60.6, 77.6, 44.3, 32.8, 94.2, ~
$ Depth          <dbl> 5.0, 9.0, 22.0, 25.5, 30.5, 32.5, 38.0, 32.0, 6.0, 10.5~
$ Duration       <dbl> 1.050000, 1.183333, 1.916667, 3.466667, 7.083333, 4.766~
$ Bird           <chr> "EP19", "EP19", "EP19", "EP19", "EP19", "EP19", "EP19",~
```

Vocabulary review

1. In this study, which variable will be on the x-axis? Is it the response variable or the explanatory variable? Is it quantitative or categorical?

The penguin's heart rate - Dive_HeartRate - will be on the x-axis as the explanatory variable for the study. This is quantitative (beats per minute).

2. Which variable is on the y-axis? Is it the response variable or the explanatory variable? Is it quantitative or categorical?

The duration of the penguin's dive - Duration - will be on the y-axis as the response variable for the study. This is quantitative (minutes).

3. How is this study different from other studies we have considered? *(Hint: consider the number and types of variables.)*

We now have two quantiative variables rather than one quantative variable.

Exploratory Data Analysis

Visualizing the Relationship - Scatterplot

A *scatterplot* is a graph showing a dot for each observational unit, where the location of the dot indicates the values of the observational unit for both the explanatory and response variables.

Typically, the explanatory variable is placed on the x-axis and the response variable is placed on the y-axis.

When describing a relationship or association between two quantitative variables as seen through a scatterplot, we look for three aspects of association: form, direction, and strength.

- The **form** of association between two quantitative variables is described by indicating whether a line would do a reasonable job summarizing the overall pattern in the data or if a curve would be better. It is important to note that, especially when the sample size is small, you don't want to let one or two points on the scatterplot change your interpretation of whether or not the form of association is linear. In general, assume that the form is linear unless there is compelling (strong) evidence in the scatterplot that the form is not linear.

- The **direction** of association between two quantitative variables is either positive or negative, depending on whether or not the response variable (`Duration`) tends to increase (positive association) or decrease (negative association) as the explanatory variable (`Dive_HeartRate`) increases.

- In describing the **strength** of association revealed in a scatterplot, we see how closely the points follow a straight line or curve. If all the points fall pretty close to this straight line or curve, we say the association is strong. Weak associations will show little pattern in the scatterplot, and moderate associations will be somewhere in the middle.

- When exploring the relationship and describing a scatterplot which visualizes two quantitative variables, we look for **unusual observations** or apparent **outliers**. Points which fall far from the trend of other points have the potential to be high leverage or high influential points.
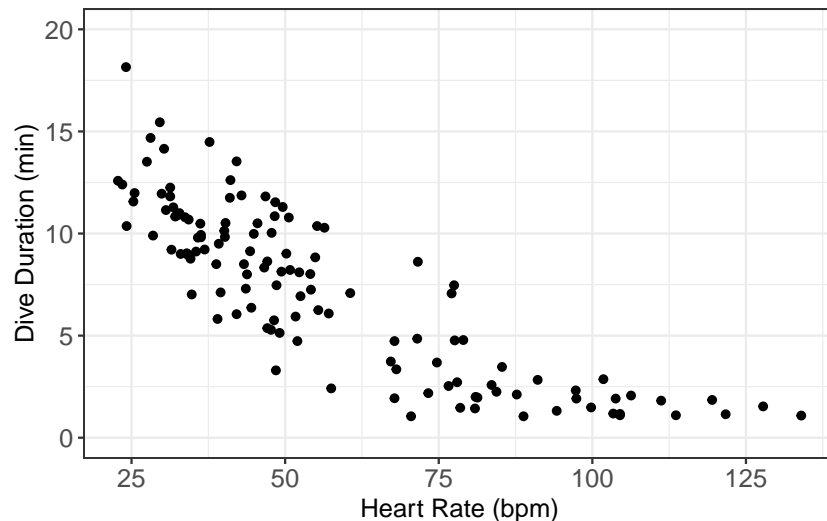
4. Describe the association between the variables as revealed in the scatterplot above. *(Hint: Remember to comment on direction, strength, and form of association as well as unusual observations. Also, make sure you use the context of the study)*

- Direction: Negative association (as Heart Rate increases, the Duration of the dive decreases).

- Strength: There is a strong relationship between heart rate and duration as indicated by the obvious trend in the data.

- Form: The association between heart rate and duration is close to linearly related (can reasonably fit a straight line through the set of points).

- Unusual Observations/Outliers: Most points fall within the general trend, but there are a few penguin's with heart rates around 75 bpm that have a little higher dive durations than the rest. We have only a couple observations where the penguin's heart rate was above 125 bpm.

Summarizing the Relationship - Correlation

Describing the direction, form, and strength of association based on a scatterplot, along with investigating unusual observations, is an important first step in summarizing the relationship between two quantitative variables. Another approach is to use a summary statistic. One of the statistics most commonly used for this purpose is the **correlation coefficient**. When the relationship has a roughly linear form, it's strength and direction can be quantified by the correlation.

The sample correlation coefficient, denoted r, is a single number that takes a value between -1 and 1, inclusive. Negative values of r indicate a negative association, whereas positive values of r indicate a positive association. It is important to note that the correlation coefficient within a population is denoted ρ and r is an estimate of ρ.

Key Idea

The correlation coefficient is only applicable for data which has a linear form; non-linear data is not summarized well by the correlation coefficient. In fact, we could say that the correlation coefficient is a numerical summary of the **strength** and **direction** of a linear association between two quantitative variables.

Some key ideas to remember

- Correlation measures the relationship between a pair of variables; the correlation is the same regardless of which one is explanatory and which is response. *(Be careful, the same is not true for regression coefficients!)*

- Correlation is a number without units. It is not a percent!
- The stronger the linear association is between the two variables, the closer the value of the correlation coefficient will be to either -1 or 1, whereas weaker linear associations will have correlation coefficient values closer to 0. Moderate linear associations will typically have correlation coefficients in the range of 0.3 to 0.7, or -0.3 to -0.7.
- Correlation can be sensitive to outliers and extreme values of either variable.

Calculating the Correlation

The correlation coefficient uses a rather complex formula that is rarely computed by hand; instead, people almost always use a calculator or computer to calculate the value of the correlation coefficient. We will use the **moderndive** R package to obtain the correlation between two numerical variables. Specifically, we will use the `get_correlation()` function to obtain our sample correlation coefficient.

The code will always look something like this:

```
get_correlation(data = <NAME OF DATASET>,
                <Y-VARIABLE> ~ <X-VARIABLE>)
```

You will be responsible for (1) filling in the name of the data set and (2) filling in the names of the x-variable and y-variable.

I've filled in the code for you to obtain our sample correlation

```
get_correlation(data = diving,
                Duration ~ Dive_HeartRate)
```

```
# A tibble: 1 x 1
     cor
   <dbl>
1 -0.846
```

5. Interpret the correlation obtained. Is it positive or negative? What does this imply about the relationship between the variables? Is it strong, moderate, or weak? How does this connect to what you saw in the scatterplot?

A correlation coefficient of $r = -0.84$ tells us we have a strong negative association between a penguin's heart rate (bpm) and the duration of their dive (mins). This connects to our scatterplot because we can see the points follow a fairly apparent downward trend.

6. Say you had another observation at (100, 15), as in a penguin with a heart rate of 100 beats per minute who dove for 15 minutes. How do you think this would change the correlation coefficient?

The absolute value of the correlation coefficient would become smaller (i.e. closer to zero) because the point does not follow the strong trend.

Least Squares Regression

7. If you knew the heart rate of a penguin, what might be a way to determine how long you would expect for them to dive for based on the data?

With tools we previously learned, we could take the average of dive time for penguins with that heart rate from the data set. But now that we are considering a line, we could input the penguin's heart rate into our equation to find what dive duration we'd expect.

Correlation measures strength of the association between two quantitative variables when the sample data points tend to follow a straight line. A natural question is then: what line do the points tend to follow?

8. Based on the scatterplot, would you say that a straight line could summarize the relationship between dive duration and heart rate reasonably well?

Yes, the trend is reasonably linear/straight.

9. Using the scatterplot below, draw the line you believe fits the data the best. How did you decide where to draw your line? Is your line the same as your group members?

You just visually fit a linear regression by eye. There are some interesting perceptual properties of fitting lines through a set of points that we will not get into. Your line is likely not the same as your group member's but they might all have similarities.

Finding the "Best" Regression Line

Your regression line will be different from my regression line and from other people in our class. There may be similarities, but no one will draw the exact same line! So, how then do we decide what line is the "best"?
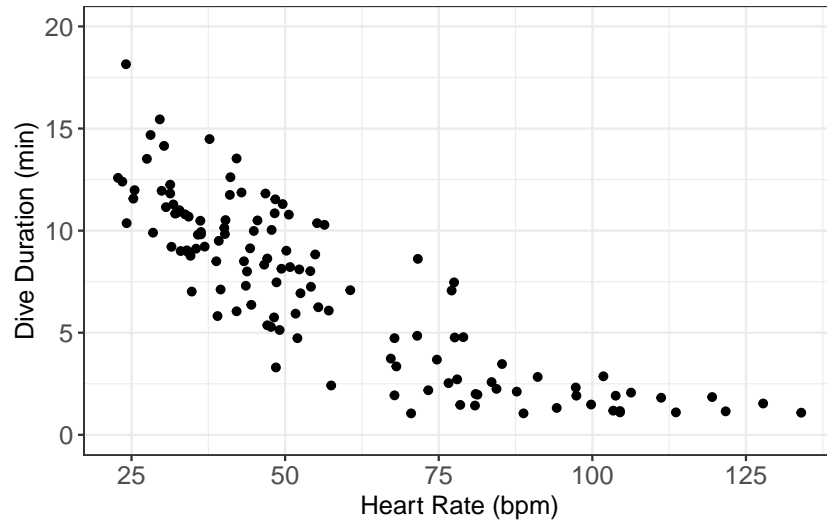
In statistics we use a method called "least squares." The idea is that we minimize the sum of the squared distances between each point and the line. That's a mouthful! Let's visualize what this means.

10. On your plot, draw the vertical distance between some of the points and the line you drew.

These vertical lines between the observed dive duration (points) and the predicted dive duration (line) for that heart rate are called residuals.

These vertical distances are how far off your estimated duration is from what was actually seen in the data. These values are called **residuals**. The least squares method finds the line that minimizes the *square* of these residuals.

11. When you square a residual what happens? *Hint: What happens if you square positive number? A negative number?*

Squaring a number makes it positive, so all of the squared residuals will be positive.

12. When you square a large residual what happens?

Squaring a large residual will blow the resulting value up (e.g. $8^2 = 64$).

13. When you square a number between 0 and 1 what happens?

Squaring a small residual (between 0 and 1) will result in a value smaller than your residual. Squaring a smaller value larger than 1 still results in a relatively small number (e.g. $2^2 = 4$).

Obtaining Coefficient Estimates from `R`

We will always use `R` to find the equation of the "best" regression line. Specifically, we will use the `lm()` function. The `lm` stands for *linear model* - the method we believe best models the relationship between our two variables.

The code will always look something like:

```
name_of_lm <- lm(<Y-VARIABLE> ~ <X-VARIABLE>,
                 data = <NAME OF DATASET>)
```

In the context of these data, here is the code I used to find the regression line:

*Note: We named our linear regression model **diving_lm**.*

```
diving_lm <- lm(Duration ~ Dive_HeartRate,
                data = diving)
```

Once we've told `R` to find the regression line, we need to obtain the estimated coefficients that go with the line! To do this, we will use the `get_regression_table()` function from the moderndive package.

Note: We can use this function on the linear regression model we named above.

```
get_regression_table(diving_lm)
```

```
# A tibble: 2 x 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept           14.8     0.466      31.6       0     13.8     15.7
2 Dive_HeartRate    -0.131     0.007     -17.6       0   -0.146   -0.116
```

14. Using the output from the `R` code above, write the equation of the regression line. Note that we've used variable names in the equation, not generic x and y. And put a carat ("hat") over the y-variable name to emphasize that the line gives predicted values of the y (response) variable.

$$\widehat{\text{Dive Duration}} = 14.8 + -0.131 \times (\text{Heart Rate})$$

Notation: The equation of the best fit line is written as $\hat{y} = b_0 + b_1 \times (\text{x})$ where

- b_0 is the intercept
- b_1 is the slope
- x is a value of the explanatory variable
- $\hat{y}$ is the predicted value for the response variable

15. Is the slope positive or negative? Explain how the sign of the slope tells you about whether your data display a positive or a negative association.

The observed slope $b_1 = -0.131$ is negative which tells us the data display a negative association because the line will be decreasing from left to right.

Key Idea

For a given data set, the signs (positive or negative) for the correlation coefficient and the slope of the regression line must be the same.

Interpreting the Coefficients

Let's investigate what the slope means in the context of heart rate and dive duration.

16. Use the least squares regression line to predict the diving duration for penguins with a heart rate of 75 beats per minute.

$\text{Dive } \widehat{\text{Duration}} = 14.8 + -0.131 \times 75 = 4.975$

17. Use the least squares regression line to predict the diving duration for penguins with a heart rate of 76 beats per minute.

$\text{Dive } \widehat{\text{Duration}} = 14.8 + -0.131 \times 76 = 4.844$

18. By how much do your predictions in #8 and #9 differ? Does this number look familiar? Explain.

$4.975 - 4.844 = 0.131$, notice this is the same value as our estimated slope (b_1).

19. These questions above were designed to help you interpret the slope. Interpret the slope in context:

The slope of the regression line predicting dive duration based on heart rate is 0.131, meaning that for every 1 beat per minute increase in heart rate, the predicted dive duration decreases by 0.131 minutes.

> **Key Idea**
> The slope coefficient of a least squares regression model is interpreted as the predicted change in the mean response (y) variable for a one-unit change in the explanatory (x) variable.

Let's investigate the meaning of the y-intercept in the context of dive duration and heart rate.

20. Use the least squares regression line to predict the diving duration for a penguin with a heart rate of 0 beats per minute.

$\text{Dive } \widehat{\text{Duration}} = 14.78 + -0.131 \times 0 = 14.8$

{A penguin with a heart rate of 0 bpm could theoretically dive for a durration of 14.758 minutes. Seems strange, huh? See the notes about extrapolation below!}

21. Your answer to #12 should look familiar. What is this value?

14.78, notice this is the same value as our estimated intercept (b_0).

> **Key Idea**
> The y-intercept of a regression line is interpreted as the predicted value of the response variable when the explanatory variable has a value of zero.

Be cautious!

While we can make predictions using our least squares regression line, we should always be wary of extrapolation in interpreting the intercept or other values outside the original data range.

> **Key Idea**
> Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is known as ***extrapolation*** and can give very misleading predictions.

22. What was the lowest value of heart rate observed in these data?

The lowest heart rate shown on the graph is about 20 (22.8 to be exact).

23. What heart rates do you believe would be an extrapolation?

Any heart rate below the minimum of about 20 and above a heart rate of about 135 would be extrapolation because we do not have data on the duration of a dive outside of this range.

MY HOBBY: EXTRAPOLATING
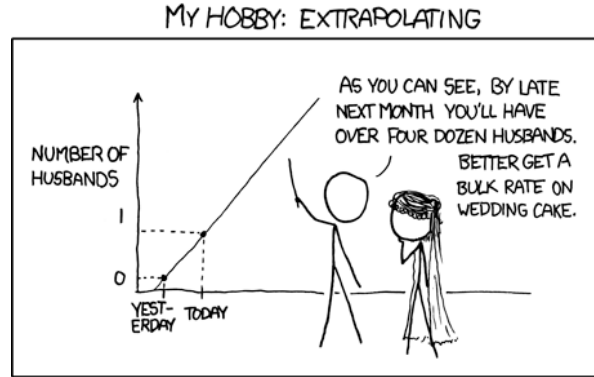
Coefficient of Determination

A quantity related to the correlation coefficient (r) is called the coefficient of determination or R-squared (R^2). The coefficient of determination (R^2) is the percentage of total observed variation in the response variable that is accounted for by changes (variability) in the explanatory variable.

Keep in mind that R^2, like correlation, requires that the relationship between the explanatory and response variables is linear!

R^2 values are reported as proportions, but can also be thought of as a percent. Values close to 1 or 100% indicate that the explanatory variable is able to explain a large portion of the variability in the response.

We calculate R^2 by squaring the correlation (r).

24. Calculate the coefficient of determination (R^2) for the relationship between heart rate and dive duration. *Hint: Look back at question 5.*

$R^2 = (-0.84)^2 = 0.7056$.

25. Complete the following statement to interpret what this value means in the context of the data:

The coefficient of determination is 70.56%, this means that 70.56% of the variation in penguin's dive duration is attributable to changes in their heart rate.

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?*

26. Write out the null and alternative hypothesis in words.

Null: There is no association between a penguin's dive heart rate and the duration of their drive.

Alternative: There is an association between a penguin's dive heart rate and the duration of their drive.

27. Using the research question, rewrite the hypotheses using notation. Use the slope as the summary measure. *Hint: What is the symbol for the population slope?*

$H_O : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

Inference for the Slope Coefficient

28. Based on the slope coefficient you reported in question 14, do you think there is convincing evidence of an association between a penguin's heart rate and the duration of their dive?

The estimated slope from the observed data is, $b_1 = -0.131$. This value is not zero and indicates a negative trend, but it is not too far away from 0, so it is hard to say.

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 - (x,y) pairs.

- Step 2: Assume the null hypothesis is true - rip apart the pairs and make a pile of heart rate values and a pile of dive duration values.

- Step 3: Generate a new sample - match a randomly drawn heart rate with a randomly drawn dive duration until all cards have a random match (this is called permuting).

- Step 4: Calculate the statistic of interest - calculate the slope from the linear regression line fit on the randomly matched pairs of heart rate and dive duration.

- Step 5: Repeat and plot the statistic of interest from each random permutation on a dot-plot to create your null distribution.

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 = " _____")
```

29. 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`): Duration

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

- Number of repetitions: n permutations - we will do 5000.

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

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

30. 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?

Change the Summary measure input to "correlation".

Here is a preview of what the output of this code looks like for 5000 permutations:

Note: We name the values in our null distribution (plotted in the next histogram), `null_dist`

14

```
head(null_dist) %>% knitr::kable(digits = 3)
```

replicate	stat
1	-0.003
2	-0.008
3	0.002
4	0.004
5	0.003
6	0.012

31. What does the `replicate` column represent?

The replicate column represents one permutation of the observed data (replicate 1 = permutation 1, replicate 2 = permutation 2, etc.).

32. What does the `stat` column represent?

The stat column is the calculated slope from the linear regression fit on that replicate / permutation of data.

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.
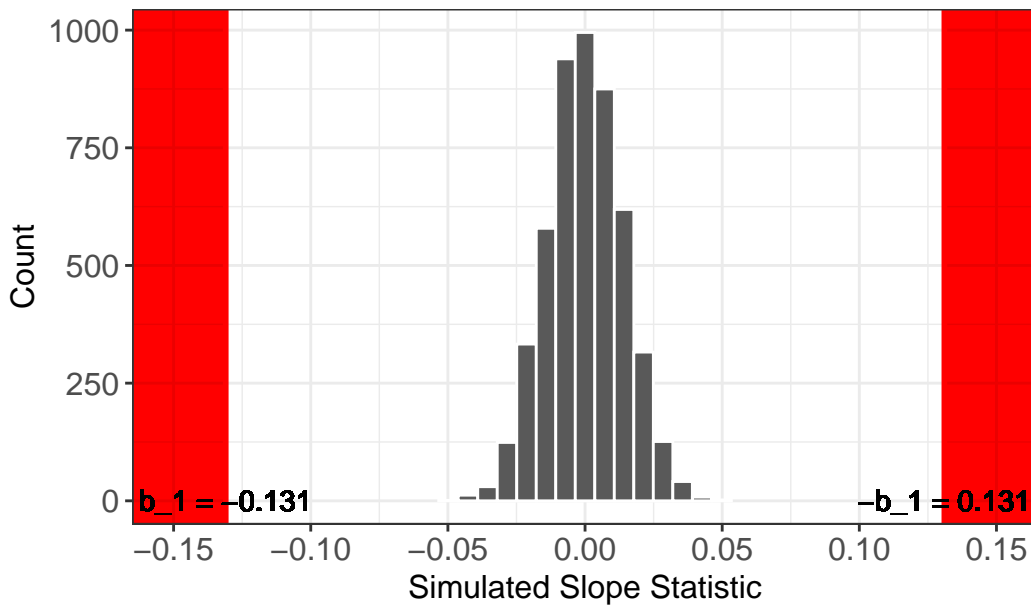
33. Why is the distribution centered at 0?

The null distribution is centered at 0 because it is simulated under the assumption there is no association between the heart rate and dive duration of a penguin (that the slope is 0).

34. Mark where the observed slope lies on the distribution.
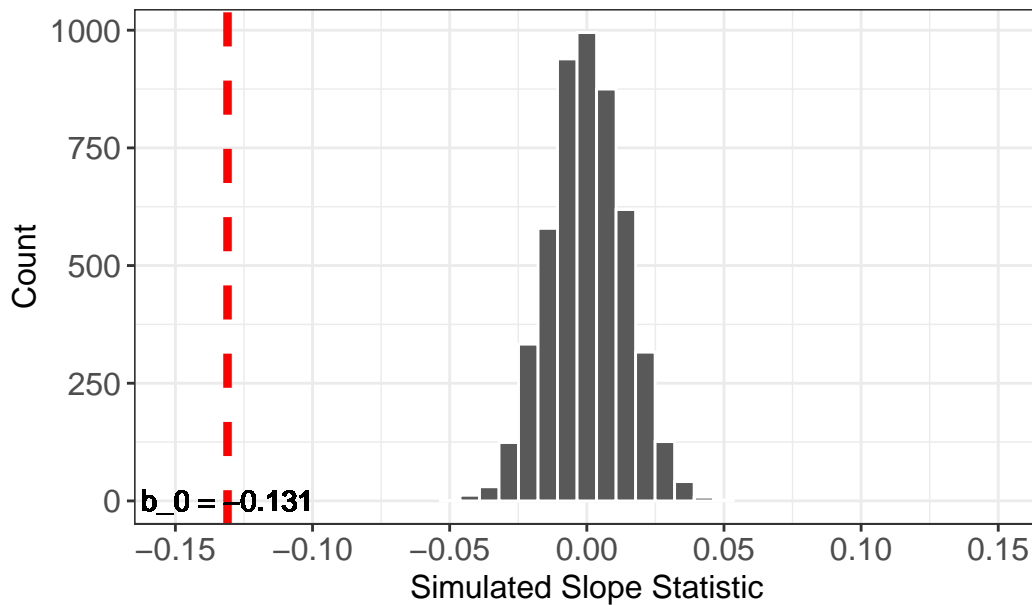
35. Shade or circle what area you will use to calculate the p-value.

36. Estimate what p-value we will obtain from R.

Since none of our permutation samples resulted in an estimated slope as extreme as |-0.131|, I expect the p-value to be 0.

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` data set
- **specifying** what variables to use
- **calclating** the statistic we are interested in

*Note: we name this value in R, **obs_slope**.*

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

obs_slope
```

```
Response: Duration (numeric)
Explanatory: Dive_HeartRate (numeric)
# A tibble: 1 x 1
    stat
   <dbl>
1 -0.131
```

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

```r
get_p_value(null_dist,
            obs_stat = _____,
            direction = "_____")
```

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

- `obs_stat` = obs_slope

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

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

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.

The warning message from R is telling us to be cautious of reporting a p-value of 0 because our result is based on permutations. If we don't have a sufficient number of permutations, saying there are no statistics as extreme as -0.131 may be inaccurate. So, rather than reporting a p-value of 0, we say our p-value < 0.0001.

Communicate the results and answer the research question

39. Based on the p-value and an $\alpha = 0.05$, write a conclusion in context of the problem.

p-value $< 0.0001 < 0.05 \implies$ Reject the null hypothesis.

We have evidence to conclude there is an association between a penguin's heart rate and their dive duration. Specifically, we have evidence to conclude the population slope between the heart rate and dive duration for all pegnuins is different from 0.

40. 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!*

I would expect to see a p-value < 0.0001 since testing the slope is equivalent to testing the correlation. Recall, the correlation coefficient is used to calculate the slope.

Connection to Confidence Intervals

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

No, since we rejected our null hypothesis based on a small p-value, our null value of 0 would not fall in the confidence interval and we would also reject based on the confidence interval.

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.