Exam 1

Data Science for Studying Language & the Mind

Instructions

The exam is worth **73 points**. You have **1 hour and 30 minutes** to complete the exam.

- The exam is closed book/note/computer/phone except for the provided reference sheets
- If you need to use the restroom, leave your exam and phone with the TAs
- If you finish early, you may turn in your exam and leave early

(5 points) Preliminary questions

Please complete these questions *before* the exam begins.

(a) **(1 point)** What is your full name?

(b) **(1 point)** What is your penn ID number?

(c) **(1 point)** What is your lab section TA's name?

(d) **(1 point)** Who is sitting to your left?

(e) **(1 point)** Who is sitting to your right?

Please do not turn the page until the exam begins

1. (16 points) R basics

(a) **(2 points)** Suppose you run the following code. What would `ls()` return?

```
x <- 1 + 2
y <- 3 + 4
z <- 0
```

- ☐ `x y z`
- ☐ `rm(y)`
- ☐ `3 7 0`
- ☐ 3

(b) **(2 points)** Which of the following occur in the code block below? Choose all that apply

```
# add 5 and 3 and store as x
x <- sum(c(5, 3))
```

- ☐ a message
- ☐ a comment
- ☐ a function
- ☐ a vector
- ☐ the assignment operator

(c) **(2 points)** Suppose you run the code below in Google Colab and receive the following error. What is the issue?

```
x <- 1+2
```

NameError
Traceback (most recent call last) /tmp/ipython-input-4199144622.py in <cell line: 0>()
—-> 1 x <- 1+2

NameError: name 'x' is not defined

- ☐ `x` is a protected name in R.
- ☐ You must use `=`, not `<-`
- ☐ You need to change the runtime type to R.
- ☐ You have another variable named `x`

(d) **(3 points)** Suppose you run the following code. What will `length(x)` return?

“‘

- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ an error

(e) **(2 points)** Suppose you run the following code. What will `typeof(x)` return? Choose one.

```
x <- c("one", "two", "three", "four", 5)
```

- ☐ `double`
- ☐ `character`
- ☐ `logical`
- ☐ `Error: vectors must be atomic`

(f) **(2 points)** What will the following code block return? Choose one.

```
x <- matrix(10:30, nrow = 2, ncol = 11)
typeof(x)
```

- ☐ `integer`
- ☐ `double`
- ☐ `matrix`
- ☐ `vector`

(g) **(3 points)** Suppose you run the following code. What will be returned? Choose one.

```
c(1, 2, 3) + c(1, 2, 3)
```

- ☐ `2 4 6`
- ☐ `1 4 9`
- ☐ `1 2 3 2 4 6 3 6 9`
- ☐ `2 3 4 3 4 5 3 5 6`
- ☐ Error: non-numeric argument to binary operator‘

3. (12 points) Data visualization

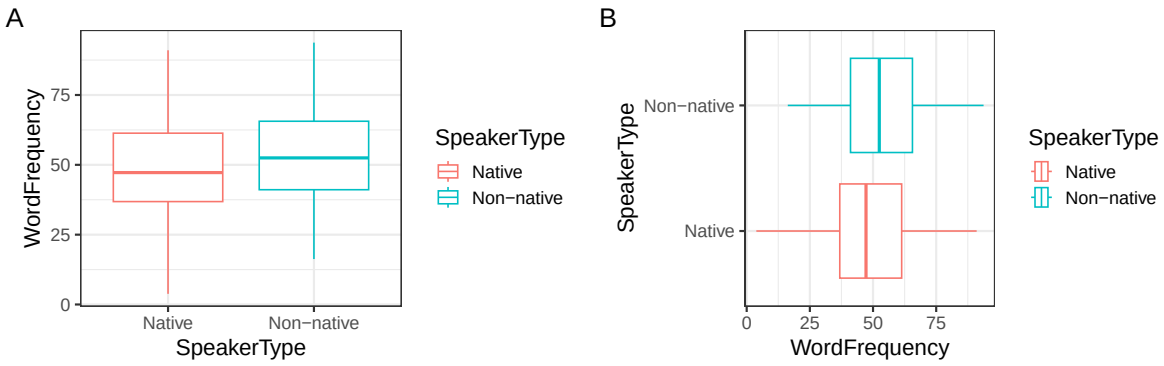
Suppose we measure the reaction times (in milliseconds) of both native and non-native speakers as they process words of varying frequency (measured as occurrences per million words). We store these data in a tibble called `rt_by_speaker`. The first 6 rows of this tibble are printed below for your reference.

```
# A tibble: 6 x 3
  WordFrequency ReactionTime SpeakerType
          <dbl>        <dbl> <chr>
1          38.8         773. Non-native
2          45.4         754. Non-native
3          81.2         711. Non-native
4          51.4         495. Native
5          52.6         851. Non-native
6          84.3         719. Non-native
```

Suppose we run the following code to visualize the effect of SpeakerType on WordFrequency.

```
ggplot(
    rt_by_speaker,
    aes(x = WordFrequency, y = SpeakerType, color = SpeakerType)
) +
    geom_boxplot()
```

(a) **(2 points)** Which of the following plots will be returned? Choose one.

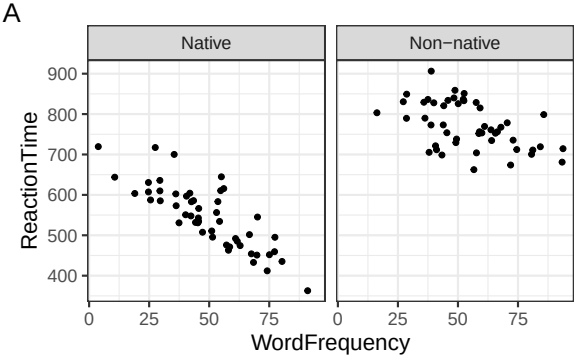


☐ A
☐ B
☐ There is not enough information to distinguish

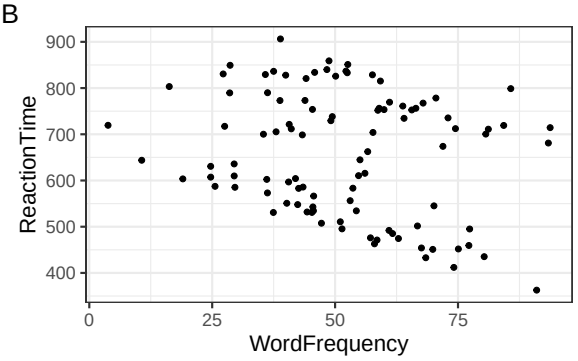
(b) **(2 points)** Suppose we run the folowing code to see the effect of WordFrequency on ReactionTime. Which of the following plots will be returned? Choose one.

```
  ggplot(
    rt_by_speaker,
    aes(y = ReactionTime, x = WordFrequency)
) +
facet_grid(.~SpeakerType) +
geom_point()
```

A



B



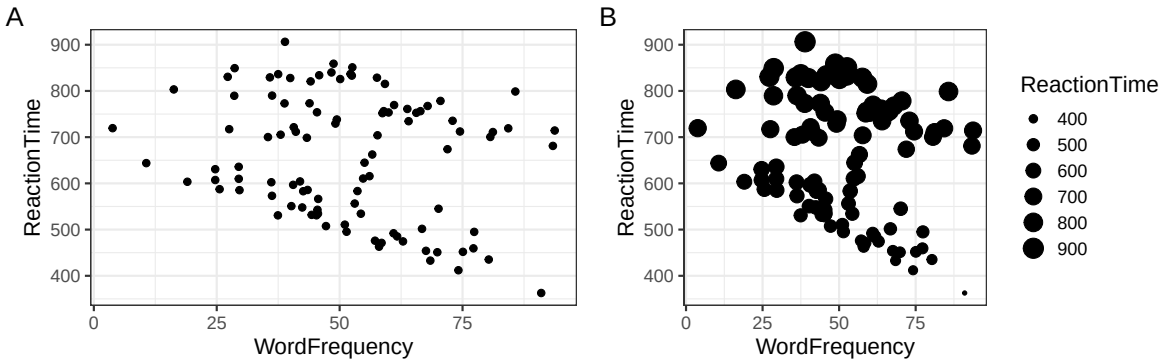
☐ A
☐ B
☐ There is not enough information to distinguish

(c) **(2 points)** In plot B above, which of the following layers could be added to change the y-axis label to `RT`?

☐ `labs(y = "RT")`
☐ `labs(x = "RT")`
☐ `annotate(geom = "text", y = ReactionTime, text = "RT")`
☐ `y-axis(ReactionTime = "RT")`

(d) **(2 points)** Suppose we run the folowing code to see the effect of WordFrequency on ReactionTime. Which of the following plots will be returned? Choose one.

```
ggplot(
    rt_by_speaker,
    aes(y = ReactionTime, x = WordFrequency, size = ReactionTime)
) +
geom_point(size = 2 )
```



☐ A
☐ B
☐ There is not enough information to distinguish

(e) **(2 points)** When `ggplot2` maps a categorical variable to an aesthetic, it automatically assigns a unique value of the aesthetic to each level of the variable. This process is called:

☐ Faceting
☐ Scaling
☐ Mutating
☐ Filtering

(f) **(2 points)** In a `geom_point()` which of the following aesthetics should we set to make the points more translucent?

☐ translucence = 0.5
☐ alpha = 0.5
☐ fill = "none"
☐ color = "translucent"

3. (18 points) Data wrangling

Suppose your are working with the `durationsGe` dataset. Here's a glimpse at the data again to refresh your memory.

```
Rows: 428
Columns: 8
$ Word                <fct> geprikt, gepresteerd, gevolgd, geprikkeld, gestaak~
$ Frequency           <int> 13, 25, 309, 16, 40, 42, 1301, 10, 73, 19, 39, 6, ~
$ Speaker             <fct> N01159, N01077, N01032, N01128, N01204, N01151, N0~
$ Sex                 <fct> male, male, female, female, female, female, male, ~
$ YearOfBirth         <int> 1944, 1980, 1939, 1979, 1963, 1956, 1979, 1944, 19~
$ DurationOfPrefix    <dbl> 0.238703, 0.082057, 0.120832, 0.106897, 0.133441, ~
$ SpeechRate          <dbl> 3.144654, 6.882591, 6.870229, 7.217848, 5.866667, ~
$ NumberSegmentsOnset <int> 2, 2, 1, 2, 2, 1, 2, 2, 1, 3, 1, 2, 1, 2, 3, 1, 2,~
```

(a) **(3 points)** How many rows would be in the object returned by the following code block? Choose one.

```
durationsGe %>%
    group_by(Sex)
```

- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 428

(b) **(6 points)** Fill in the blanks in the partially completed code below such that it returns the following tibble, including the `mean` and `median` of the `DurationOfPrefix` variable.

```
durationsGe %>% group_by(__a__) %>% summarise(____b_____, _____c_______)
```

```
# A tibble: 3 x 3
  Sex      mean median
  <fct>   <dbl>  <dbl>
1 female 0.125  0.120
2 male   0.126  0.120
3 <NA>   0.0960 0.0960
```

(i) **(2 points)** Fill in blank a.

(ii) **(2 points)** Fill in blank b.

(i) **(2 points)** Fill in blank c.

(c) **(3 points)** True or false, the following code options would return identical tibbles.

```
# option 1
durationsGe %>%
    select(Frequency) %>%
    filter(Frequency > 40) %>%
    distinct()

# option 2
just_freq <- select(durationsGe, Frequency)
freq_under_40 <- filter(just_freq, Frequency > 40)
distinct(freq_under_40)
```

☐ True
☐ False

(d) **(2 points)** Suppose we run the following code block. What will `n()` do? Choose one.

```
rt_by_speaker %>%
    summarise(n = n())
```

- ☐ remove all `NA`s from the dataset
- ☐ count of the number of rows in the dataset
- ☐ add the string `n` before each value in `SpeakerType`
- ☐ Nothing, because it requires a grouping
- ☐ Throw `Error: Missing arguments to n()`

(e) **(2 points)** True or false, the following code blocks are equivalent

```
# option 1
rt_by_speaker %>% select() %>% glimpse()

# option 2
select() %>% rt_by_speaker %>% glimpse()
```

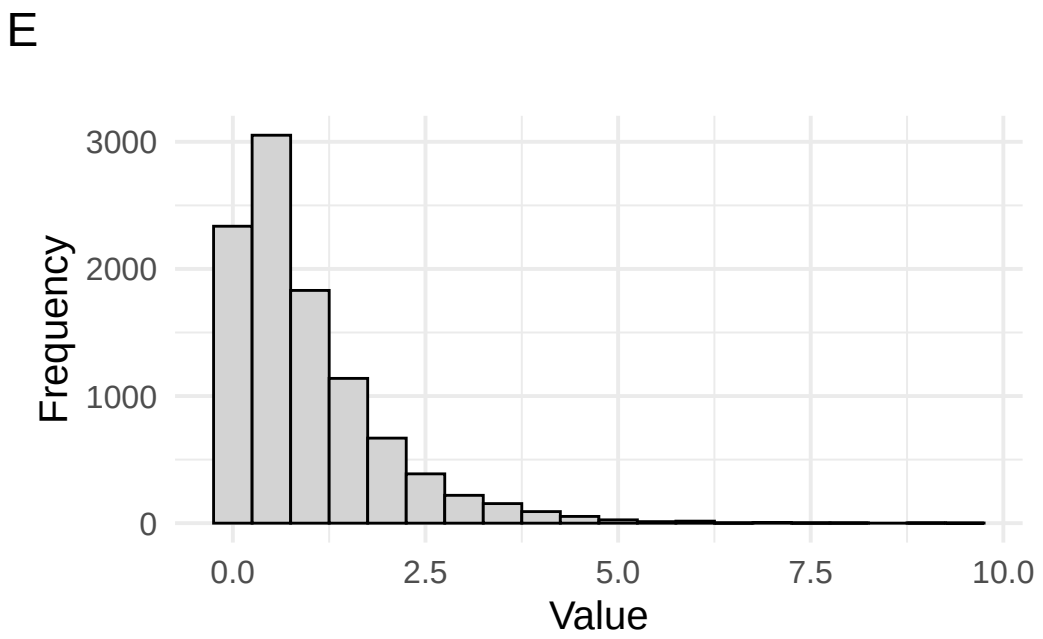
- ☐ True
- ☐ False

(f) **(2 points)** Suppose we want to create a new column `ReactionTimeSeconds` which converts every value in the `ReactionTime` column from milliseconds to seconds. Which of the following `dplyr` functions could accomplish this? Choose one.

- ☐ `group_by` with `summarise()`
- ☐ `filter()`
- ☐ `select()`
- ☐ `mutate()`
- ☐ `rename()`

4. (16 points) Sampling distribution

Suppose we visualize the frequency distribution of a value in a dataset, shown below.



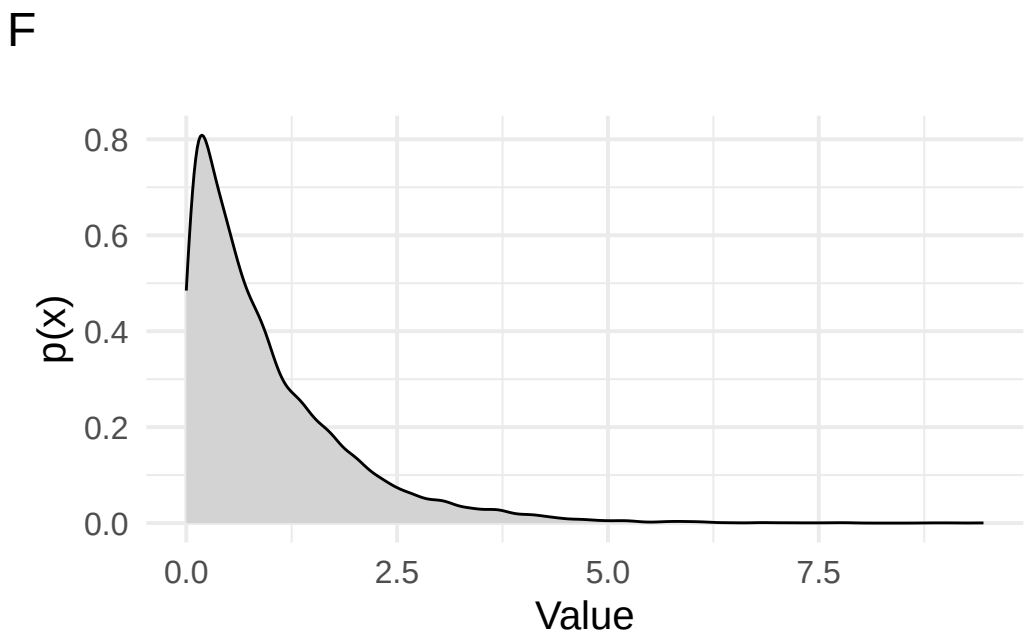
(a) **(2 points)** Which of the following would best summarise the *spread* of these data? Choose one.

- ☐ mean
- ☐ median
- ☐ sd
- ☐ interquartile range
- ☐ p-value

(b) **(2 points)** Which of the following would best summarise the *central tendency* of these data? Choose one.

- ☐ mean
- ☐ median
- ☐ sd
- ☐ interquartile range
- ☐ p-value

(c) **(2 points)** Suppose we visualize the probability density function of the distribuition that generated these data (below) What could be the height of the probability density function at a value of 8? Choose one.



☐ 0.8
☐ 0.2
☐ 0.001
☐ 1

(d) **(2 points)** Which of the following parameters define the uniform probability density function? Choose all that apply.

☐ max
☐ min
☐ mean
☐ sd
☐ none of the above, uniform distributions are nonparametric

(e) **(6 points)** Suppose we want to generate the bootstrap sampling distribution for the *median* of `value` in our dataset. Fill in the blanks to accomplish this with the `infer` package, generating 1000 bootstrapped samples. Note that `data` has 500 rows.

```
library(____a_____)

data %>%
    specify(response = value) %>%
    generate(reps = _____b_____, type = _____c_____) %>%
    calculate(stat = _____d_____)
```

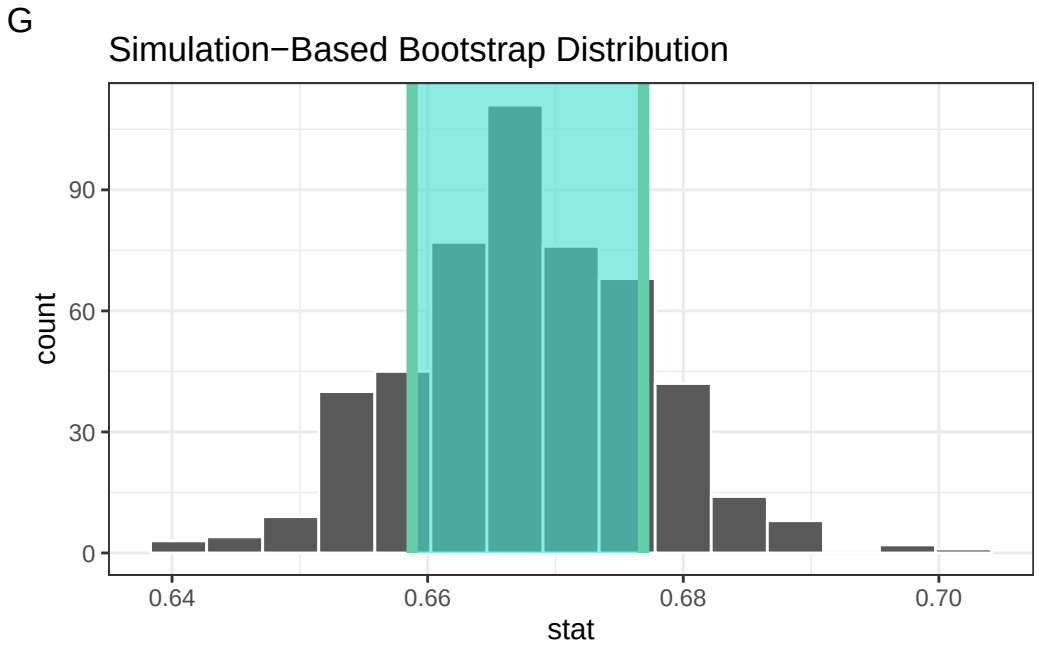
(i) **(1 point)** Fill in blank a.

(ii) **(2 points)** Fill in blank b.

(iii) **(2 points)** Fill in blank c.

(iii) **(1 points)** Fill in blank d.

(f) **(2 points)** The shaded area of the figure shows the 68% confidence interval. If we were to increase the `level` of confidence to 95%, the confidence interval would become (choose one):

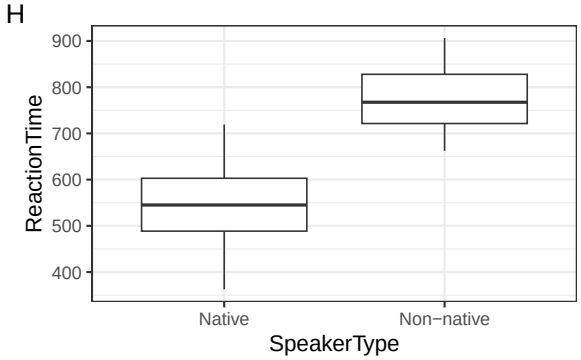


☐ Narrower
☐ Wider
☐ Unchanged
☐ Not enough information to determine this

5. (11 points) Hypothesis testing

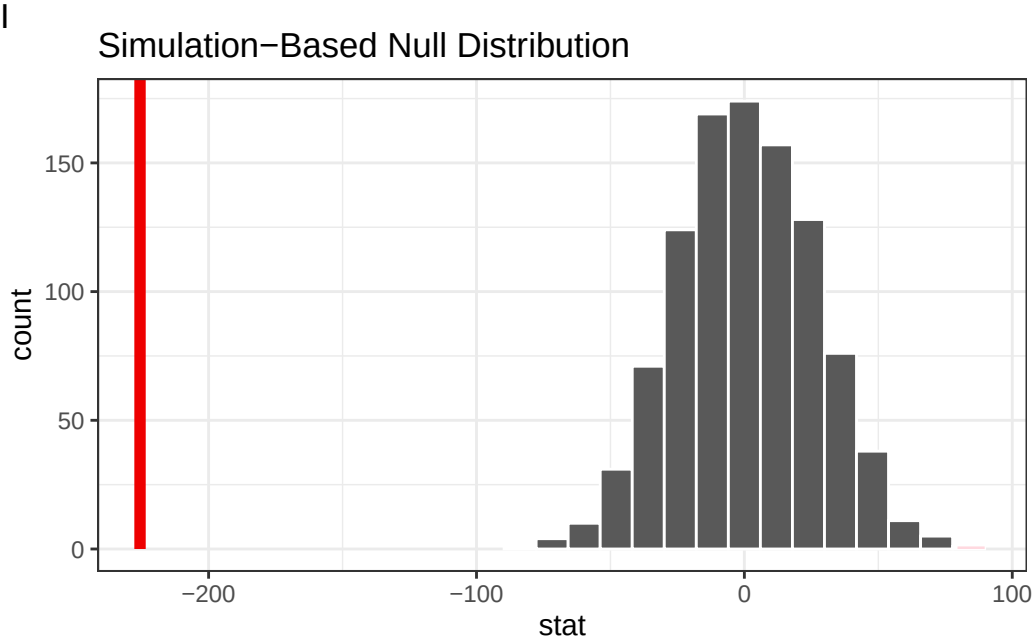
Suppose we want to determine whether there is a difference in mean `ReactionTime` between our Native and Non-native speakers in the `rt_by_speaker` dataset. We can visualize these data with a boxplot.

H



Then we use `infer` to generate the sampling distribution for the difference in mean `ReactionTime` between the Native and Non-native speakers. We've visualized this distribution here and called `shade_p_value()` to generate the vertical line.

I



(a) **(3 points)** Step 1 of the 3-step hypothesis testing framework is to pose the null hypothesis. Which of the folloing could be the null hypothesis? Choose one.

- ☐ Native speakers have faster Reaction Times than Non-native speakers.
- ☐ Non-native speakers have faster Reaction Times than Native speakers.
- ☐ Any difference in the average Reaction Time between Native and Non-native speakers is due to random chance.
- ☐ Native and Non-native speakers always have exactly the same Reaction Times on every trial.

(b) **(2 points)** Step 2 is to ask, if the null hypothesis is true, how likely is our observed pattern of results? Given the figures above, what will the returned p-value be?

(c) **(2 points)** Step 3 is to decide whether to reject the null hypothesis. True or false, we should reject the null hypothesis. Assume our threshold for rejection is $p < 0.05$.

- ☐ True
- ☐ False

(d) **(2 points)** Why do we pose a null hypothesis? Choose one.

- ☐ It is the hypothesis most likely to be true.
- ☐ It allows us to generate predictions based on prior beliefs.
- ☐ It is the hypothesis for which we can simulate data.
- ☐ It ensures that the alternative hypothesis is proven false.

(e) **(2 points)** True or false, we can compute a p-value by counting the number values in our null distribution that are more extreme than the actual observed value and then dividing by the total number of simulations that we generated.

- ☐ True
- ☐ False