Exam 1 (Practice)

Data Science for Studying Language & the Mind

Estimated time: 60 minutes

You may need more time if programming is completely new to you, or less if you have some experience already.

Instructions

- The quiz is closed book/note/computer/phone
- If you need to use the restroom, leave your exam and phone with the TA
- You have 90 minutes to complete the exam. If you finish early, you may turn in your exam and leave early

Part 1: Data science with R

Name: _____
PennKey: _____

Lab section TA: _____

R basics: general

(a) Suppose you run the following code. Which command could you run to remove the `y` variable from the current environment? Choose all that apply.

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

- ☐ `ls(y)`
- ☐ `rm(list = ls())`
- ☐ `rm(y)`
- ☐ `remove(y)`

(b) Write an expression that would assign the value `10` to the variable name `my_var`?

(c) Which of the following would install the `praise` package? Choose all that apply.

- ☐ `library(praise)`
- ☐ `install(praise)`
- ☐ `install.packages("praise")`
- ☐ `install.packages(praise)`

(d) Suppose you run the following code. Which functions would return the structure of the object you defined? Choose one.

```
x <- c("bus", "stop")
```

- ☐ `length(x)`
- ☐ `print(x)`
- ☐ `str(x)`
- ☐ `attributes(x)`
- ☐ None of the above

R basics: vectors, operations, subsetting

(a) Suppose you run the following code. What will length(x) return? Write your answer in the box below.

```
x <- 1:5
```

(b) Suppose you run the following code. What will `y > 4` return? Write your answer in the box below and **show your work**.

```
x <- seq(2, 8, by = 2)
y <- x[c(-4)]
```

(c) Suppose you run the following code. What will `typeof(x)` return?

```
x <- c("true", "false", "true", "false")
```

- ☐ `double`
- ☐ `integer`
- ☐ `character`
- ☐ `logical`

(d) Suppose you create the following data frame and assign it to the `df` variable. What will `sum(df$antique)` return?

```
  antique age show
1       1   2    a
2       2   3    b
3       3   4    c
```

- ☐ NULL
- ☐ Error: no columns include value "antique"
- ☐ 6
- ☐ 9
- ☐ 15

Data importing

(a) Which of the following will load the `readr` package into the current environment? Choose all that apply.

- ☐ `library(tidyverse)`
- ☐ `library(readr)`
- ☐ `install.packages("tidyverse")`
- ☐ `install.packages("readr")`
- ☐ `import(tidyverse)`
- ☐ `import(readr)`

(b) Suppose `print(x)` returns the following. What will `is.data.frame(x)` return? Write your answer in the box below.

```
# A tibble: 4 x 3
      x     y     z
  <int> <int> <int>
1     1     5     9
2     2     6    10
3     3     7    11
4     4     8    12
```

(c) Suppose you import "junesales.csv", shown below, with the following code. What would `data$Sale` return? Choose one.

```
Year, Month, Day, Sale
2023, June, 1, 0
2023, June, 2, 1
2023, June, 3, 0
2023, June, 4, 1
```

```
data <- read_csv("junesales.csv",
    col_types = list(Sale = col_logical())
)
```

- ☐ A `double` vector with values `0 1 0 1`
- ☐ A `logical` vector with values `FALSE TRUE FALSE TRUE`
- ☐ A double vector with values `NA NA NA NA`
- ☐ `NULL`

(d) Suppose you import a dataset with `readr`, but when you `print(data)` you notice that the `age` column was identified as `character` when you were expecting `double`. Given the resulting tibble, which of the following arguments could you include in blank in the code below to solve this problem?

```
# A tibble: 4 x 3
  age   graduated   gpa
  <chr> <lgl>     <dbl>
1 18    FALSE      NA
2 na    FALSE       3.8
3 25    TRUE        2.9
4 21    TRUE        3.1

data <- read_csv("data.csv", ________)
```

- ☐ `.drop = NA`
- ☐ `skip = 1`
- ☐ `guess_max = Inf`
- ☐ `na = c("na")`
- ☐ `col_names = FALSE`

Data visualization: basics

Section 5 makes use of the `durationsGe` dataset and plots A, B, and C in the appendix.

(a) Which of the plots above (A and B) did the code blocks below generate?

```
# code 1
ggplot(durationsGe, aes(x = DurationOfPrefix, fill = Sex)) +
    geom_density(fill = "lightgray") +
    theme_classic(base_size = 12) +
    labs(y = "") +
    scale_fill_manual(values = c("white", "gray", "black"))
# code 2
ggplot(durationsGe, aes(x = DurationOfPrefix)) +
    geom_density(fill = "lightgray") +
    theme_classic(base_size = 12) +
    labs(y = "")
```

- ☐ Code 1 generates plot A, code 2 generates plot B
- ☐ Code 2 generates plot A, code 1 generates plot B
- ☐ Code 1 and 2 both generate plot A
- ☐ Code 1 and 2 both generate plot B
- ☐ Code 1 and 2 generate neither plot A nor plot B

(b) Which geoms could be depicted in plots A and B? Choose all that apply.

- ☐ `geom_histogram()`
- ☐ `geom_smooth()`
- ☐ `geom_line()`
- ☐ `geom_density()`
- ☐ `geom_bar()`

(c) True or false, the following code blocks generate the same figure. Write your answer in the following box and **explain why**.

```
# code block 1
ggplot(
    data=durationsGe,
    mapping = aes(y = DurationOfPrefix, x = Sex)) +
        geom_bar(stat = "identity")

# code block 2
  ggplot(
    aes(y = DurationOfPrefix, x = Sex),
    durationsGe) +
        geom_bar(stat = "identity")
```

(d) The code below makes use of a new geom, `geom_rug()`, to generate plot C, in which each individual data point is plotted along the x-axis like a "rug". In the box below, rewrite the code such the color of the rug is mapped to the `Sex` variable and the bars of the histogram are filled in with the color "lightblue".

```
ggplot(durationsGe, aes(x = DurationOfPrefix)) +
    geom_rug() +
    geom_histogram() +
    theme_classic(base_size = 12)
```

Data visualization: layers

Section 5 makes use of the `durationsGe` dataset and plots D, E, and F in the appendix.

(a) Which of the following would add a small amount of random noise around each point in plot D? Choose all that apply.

- ☐ add the argument `position = "jitter"` to `geom_point()`
- ☐ add the argument `position = "random"` to `geom_point()`
- ☐ add the argument `rand_noise = TRUE` to `geom_point()`
- ☐ replace `geom_point()` with `geom_jitter()`
- ☐ replace `geom_point()` with `geom_noise()`

(b) Which of the following could change plot D to plot E? Choose all that apply

- ☐ add `facet_wrap(~Sex)`
- ☐ add `facet_wrap(~Sex, ncol = 2)`
- ☐ add `facet_wrap(~Sex, ncol = 3)`
- ☐ add `facet_grid(Sex~.)`
- ☐ add `facet_grid(.~Sex)`
- ☐ add `facet(.by = c(Sex))`

(c) Which of the following arguments to `geom_histogram()` could be present in the code that returned plot F? Choose all that apply.

- ☐ `bins=12`
- ☐ `bins=11`
- ☐ `binwidth=1`
- ☐ `binwidth=3`
- ☐ `stat="identity"`

(d) Which of the following layers are required to produce plot F? Note that the plot uses the complete theme `theme_minimal()` and the font is 20pt Palatino. Choose all that apply.

- ☐ `theme_minimal(use=TRUE)`
- ☐ `theme_minimal(base_size = 20, base_family = "Palatino")`
- ☐ `labs(title = "Histogram of speech rate")`
- ☐ `font(size=20, family="Palatino")`
- ☐ `scale_fill_manual(values = c("gray"))`

Data wrangling

Section 6 makes use of the `durationsGe` dataset in the appendix.

(a) The `Sex` variable in the `durationsGe` dataset has the following distinct values: `"male"` `"female"` `NA`. How many rows would be in the object returned by the following code block? Write your answer in the box below.

```
durationsGe %>%
    filter(Sex %in% c("female")) %>%
    summarise(minBirthYear=min(YearOfBirth, na.rm=TRUE), .by=c(Sex))
```

(b) True or false, the following code options are equivalent.

```
# 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

(c) Fill in the blank in the code below such that it returns a new column called "count", which counts of the number of rows in the `durationsGe` dataset per `Sex`

```
ratings %>% group_by(Sex) %>% summarise(_________)
```

(d) True or false, the following code options are equivalent.

```r
# option 1
durationsGe %>%
    select(Freq=Frequency, Speaker:DurationOfPrefix) %>%
    mutate(AgeInYears = 2023 - YearOfBirth, .before = Freq)

# option 2
durationsGe %>%
    select(Frequency:DurationOfPrefix) %>%
    rename(Freq = Frequency) %>%
    mutate(AgeInYears = 2023 - YearOfBirth, .before = 1)
```

- ☐ True
- ☐ False

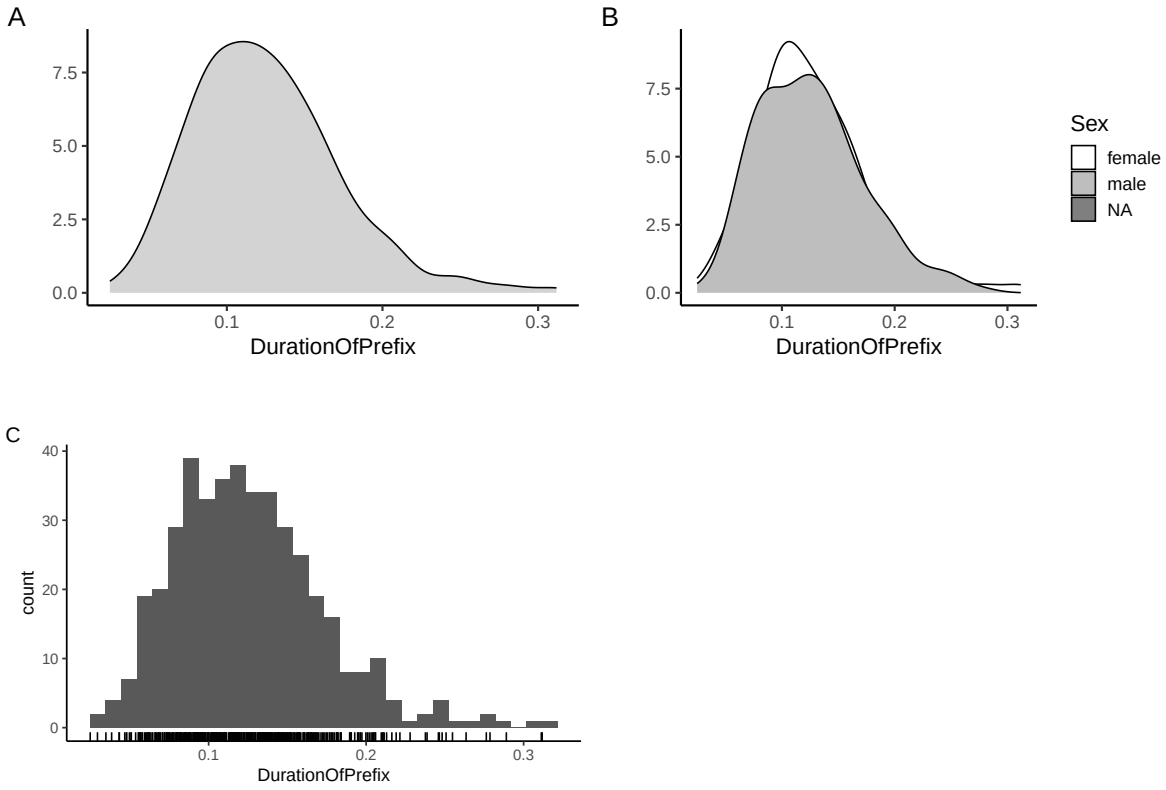
Appendix A: Data

Sections 4-6 make use of `durationsGe` data in the `languageR` package. The dataset includes the duration of the prefix ge- in Dutch by various speakers from the Spoken Dutch Corpus.

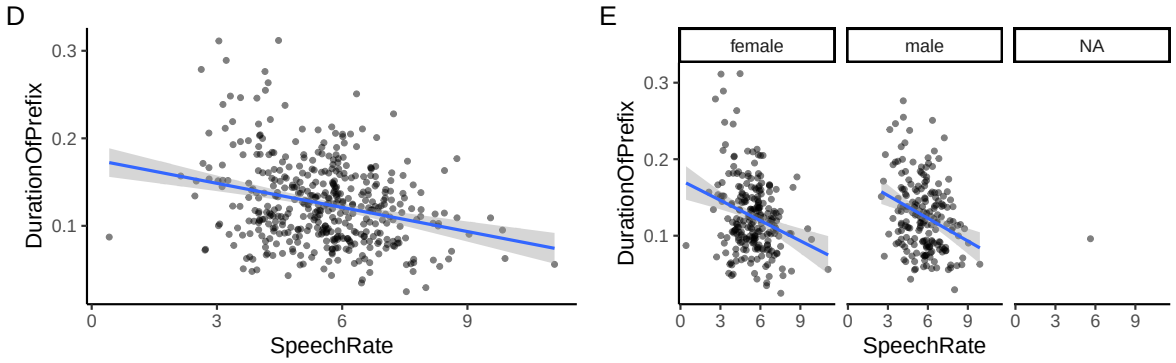
```
library(languageR)
glimpse(durationsGe)
```

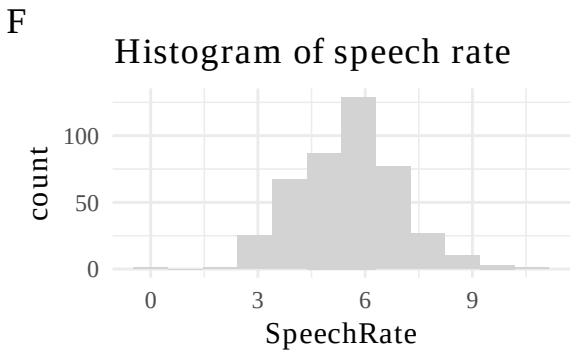
```
Rows: 428
Columns: 8
$ Word                <fct> geprikt, gepresteerd, gevolgd, geprikkeld, gestaak~
$ Frequency           <int> 13, 25, 309, 16, 40, 42, 1301, 10, 73, 19, 39, 6, ~
$ Speaker             <fct> N01159, N01077, N01032, N01128, N01204, N01151, N0~
$ Sex                 <fct> male, male, 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,~
```

Appendix B: Plots section 4



Appendix C: Plots section 5





Part 2

The data

This quiz refers to data simulated from Johnson & Newport (1989), who studied the English language proficiency of 46 native Korean or Chinese speakers who arrived in the US between the ages of 3 and 39. The researchers were interested in whether the participants' age of arrival to the United States played a role in their English language proficiency.

The simulated data are stored in the tibble `johnson_newport_1989`. Here is a `glimpse()` at the tibble for your reference:

```
glimpse(johnson_newport_1989)
```

```
Rows: 69
Columns: 4
$ score     <dbl> 270.8899, 270.2497, 267.1322, 268.3546, 263.7737, 263.8069, ~
```

```
$ age       <dbl> 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, ~
$ ageGroup  <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", ~
$ langGroup <chr> "native", "native", "native", "native", "native", "native", ~
```

Sampling distribution

Johnson and Newport (1989) reported the mean and standard deviation of participants' scores on the English proficiency test, grouped by an `ageGroup` variable, which divides age into 5 groups. Below we computed the **median** and **IQR** as the descriptive statistics on our simulated data. Then, we used `infer` to generate the sampling distribution for the **17-39 year old age group**, visualize the distribution, and shade the confidence interval.

```
# A. compute descriptive statistics by group
johnson_newport_1989 %>% group_by(ageGroup) %>%
    summarise(n = n(), median = median(score),
        lower = quantile(score, 0.25), upper = quantile(score, 0.75))
```

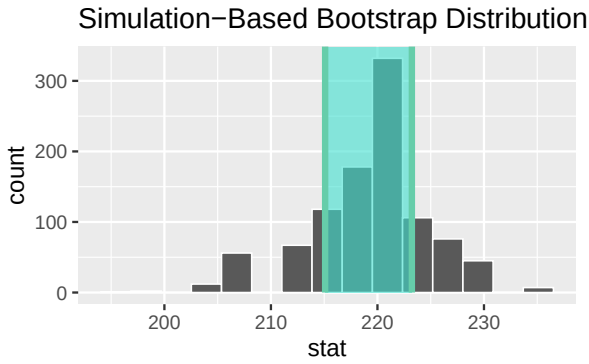
```
# A tibble: 5 x 5
  ageGroup     n median lower upper
  <chr>    <int>  <dbl> <dbl> <dbl>
1 0           23   268.  268.  270.
2 11-15        8   233.  228.  237.
3 17-39       23   220.  201.  233.
4 3-7          7   271.  269.  273.
5 8-10         8   263.  256.  266.
```

```
# B. generate the sampling distribution 17-39 group
samp_distribution <- johnson_newport_1989 %>%
    filter(ageGroup == "17-39") %>%
    specify(response = score) %>%
    generate(reps = 1000, type = "bootstrap") %>%
    calculate(stat = "median")

# C. get confidence interval
ci <- samp_distribution %>%
    get_confidence_interval(type = "percentile", level = 0.68)

# D. visualize sampling distribution and confidence interval
samp_distribution %>%
    visualize() +
    shade_ci(endpoints = ci)
```

(a) True or false, the descriptive statistics reported above are parametric.



☐ True
☐ False

(b) The sampling distribution of the median looks approximately Gaussian. The probability densitiy function for the Gaussian distribution is given by which of the following equations?

☐ $\frac{\sum_{i=i}^{n} x_i}{n}$
☐ $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
☐ $\frac{1}{max-min}$
☐ $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$

(c) Fill in the blanks in the sentence below to describe what happens *on each repeat* in code B above, in which we constructed the sampling distribution.

Draw _____ data points _____ replacement, compute the _____.

(d) The shaded area of the figure shows a 68% confidence interval. If we were to increase the `level` of confidence to 95%, the confidence interval would become:

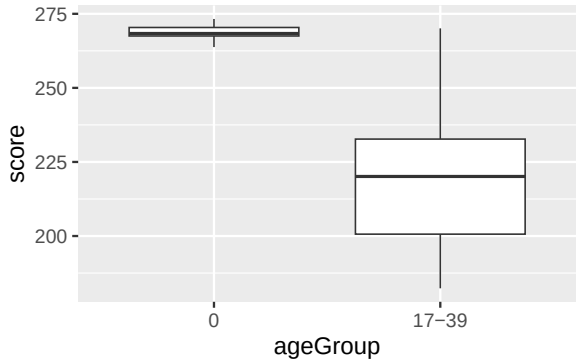
☐ Narrower
☐ Broader
☐ Stay the same
☐ There's insufficient information to determine this

Hypothesis testing

Suppose we want to know whether the participants who arrived as adults (17-39 age group) acheived native performance. We decide to address this question via the 3-step hypothesis test-

ing framework in which we investigate the difference in **medians** between the native English speakers (0 age group) and the 17-39 age group.

```
# A. visualize difference with a boxplot
johnson_newport_1989 %>%
    filter(ageGroup %in% c("0", "17-39")) %>%
    ggplot(aes(y = score, x = ageGroup)) +
    geom_boxplot()
```

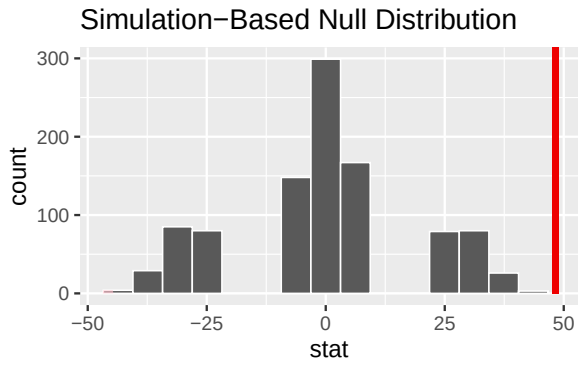


```r
# B. compute observed difference in means
diff_medians <- johnson_newport_1989 %>%
    filter(ageGroup %in% c("0", "17-39")) %>%
    specify(response = score, explanatory = ageGroup) %>%
    calculate(stat = "diff in medians", order = c("0", "17-39"))

# C. construct the null distribution with infer
null_distribution <- johnson_newport_1989 %>%
    filter(ageGroup %in% c("0", "17-39")) %>%
    specify(response = score, explanatory = ageGroup) %>%
    hypothesize(null = "independence") %>%
    generate(reps = 1000, type = "permute") %>%
    calculate(stat = "diff in medians", order = c("0", "17-39"))

# D. visualize the null and shade p-value
null_distribution %>%
    visualize()  +
    shade_p_value(obs_stat = diff_medians, direction = "both")
```

```
Warning in (function (mapping = NULL, data = NULL, stat = "identity", position = "identity",
i Did you mean to use `annotate()`?
```



(a) Step 1 is to pose the null hypothesis. True or false, the null hypothesis here is that the observed difference in medians is due age group (age of arrival in the US).

☐ True
☐ False

(b) Step 2 is to ask, if the null hypothesis is true, how likely is our observed pattern of results? We quantify this likelihood with:

- ☐ diff in medians
- ☐ correlation
- ☐ liklihood estimation
- ☐ p-value

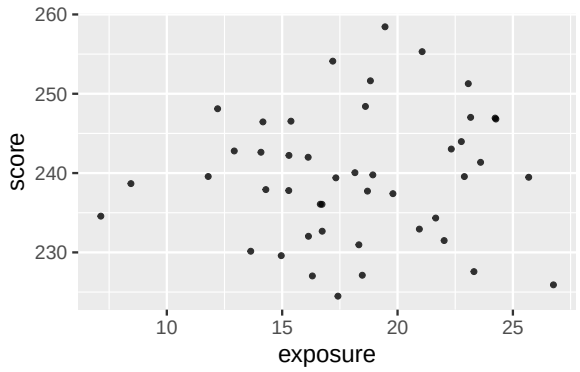
(c) Step 3 is to decide whether to reject the null hypothesis. Johnson and Newport concluded that the two groups were significantly different from each other, suggesting that participants who arrived to the US after age 17 did not acheive native proficiency. This implies that they:

- ☐ Reject the null hypothesis
- ☐ Fail to reject the null hypothesis
- ☐ Prove the research hypothesis to be true
- ☐ Prove the null hypothesis to be true

(d) When we calculate the p-value from the simulated null distribution using the `get_p_value()` function, we get $\mathbf{p = 0}$. Is this a problem? Why or why not? Explain what a p-value of 0 means in this context.

Correlation

Johnson and Newport (1989) also wanted to ask whether years of exposure to English predicted score on the English proficiency task. To address this, they computed the correlation between score and exposure.



(a) Given the scatterplot of these data, which of the following could be their observed correlation?

☐ -0.88
☐ 0.88
☐ 0.16
☐ 0.5

(b) True or false, the correlations computed on these data were subject to sampling variability.

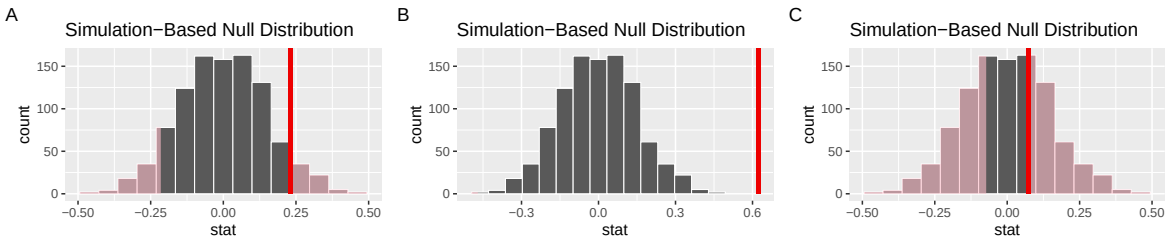
☐ True
☐ False

(c) Johnson and Newport used hypothesis testing to determine whether the correlation they observed was significantly different from zero. We computed a p-value of `0.624` on the correlation we observed in our simulated data. Which figure could represent this p-value visualized on a null distribution generated nonparametrically from 1000 repetitions?

```
Warning in (function (mapping = NULL, data = NULL, stat = "identity", position = "identity",
i Did you mean to use `annotate()`?

Warning in (function (mapping = NULL, data = NULL, stat = "identity", position = "identity",
i Did you mean to use `annotate()`?

Warning in (function (mapping = NULL, data = NULL, stat = "identity", position = "identity",
i Did you mean to use `annotate()`?
```



(d) What type of relationship does the correlation between years of exposure and score suggest?

☐ Linear
☐ Nonlinear
☐ Independence
☐ Permute