Module 2, Assigment 1

Ellen Bledsoe

2022-09-19

Purpose

The goal of this assignment is to explore computational reproducibility and apply the base R coding skills we've learned and practiced in class and lab.

Task

Write R code to successfully answer each question below and/or write text to successfully answer the question.

Criteria for Success

- Code is within the provided code chunks
- Code chunks run without errors
- Code produces the correct result
  - Code attempts will get half points
  - Code that produces the correct answer will receive full points
- Text answers correctly address the question asked

Due Date

Sept 19 at 1pm MST

Assignment Questions

This assignment is worth 20 points total. Each question is worth 1 point.

Definitions (1 point each)

In your own words, define/describe the following terms. These don't need to be technical descriptions but rather how you are thinking about them.

Grading note: I've copied these definitions from my slides. If their definitions are the same, they have copied directly from the slides and not used their own words, so take 0.5 points off.

1. *Reproducibility*: ability to repeat the original study using the same data, materials, and methods

2. *Open science*: Scientific research conducted and communicated in an honest, accessible, and transparent way, such that–at a minimum–a study can be reproduced and/or replicated.

3. *R*: R refers to both the programming language and the software that interprets scripts written in the language.

4. *RStudio*: RStudio is an integrated development environment (IDE) that helps us interact with R more easily.

Vectors (1 point each)

5. Run the first code chunk below to create the object heights. Use the `mean()` function to calculate the mean value of height.

Hint: remember to check the answer key!

```
heights <- c(63, 69, 60, 65, NA, 68, 61, 70, 61, 59, 64, 69, 63, 63, NA, 72, 65, 64, 70, 63, 65)
```

```
mean(heights)
```

```
## [1] NA
```

6. This should yield an odd result caused by the `NA` values. To resolve this, use the help function to remind yourself about the argument `na.rm = TRUE` that applies to many R functions

```
help(mean) # same as ?mean
```

Issue a revised command to calculate the mean value in `heights` and send the result to the console.

```
mean(heights, na.rm = TRUE)
```

```
## [1] 64.94737
```

7. Write a line of code that selects the 6-10th height values.

```
heights[6:10]
```

```
## [1] 68 61 70 61 59
```

8. Create a character vector called `rhymes` that contains the following values: cat, rat, bat, hat

```
rhymes <- c("cat", "rat", "bat", "hat")
```

Data Frames

IMPORTANT! Run the following code chunk to read in the cactus pad dataset. *You don't need to understand what is happening, though I've added some comments in case you are interested!*

We will be using that data for the remainder of the assignment.

Once you run the code chunk, the data frame will be saved in your environment as the object `cactus`.

9. How many rows does this data frame have? How many columns? You can either use code to figure this out or look at the object in the environment.

   Rows: 50

   Columns: 13

```
# optional space for code to answer the question above
# str(pads)
# dim(pads)
```

10. Look at the first 6 rows of data. You can either do this by using a function or by using index sub-setting.

```
head(pads)
```

```
##   group_id                                                    group_members
## 1        1 Jaelyn Dennis, Katharine Cole, Jose Rodriguez, Rheanna Fernandez
## 2        1 Jaelyn Dennis, Katharine Cole, Jose Rodriguez, Rheanna Fernandez
## 3        1 Jaelyn Dennis, Katharine Cole, Jose Rodriguez, Rheanna Fernandez
## 4        1 Jaelyn Dennis, Katharine Cole, Jose Rodriguez, Rheanna Fernandez
## 5        1 Jaelyn Dennis, Katharine Cole, Jose Rodriguez, Rheanna Fernandez
## 6        2                                       Zach, Britsy, Riley, Woods
##   temp_C              species  size paddle_id length_in width_in depth_in
## 1     40 Opuntia ficus-indica Large         1      9.00     2.50     0.25
## 2     40 Opuntia ficus-indica Large         2     10.00     6.00     0.25
## 3     40 Opuntia ficus-indica Large         3      8.00     4.00     0.25
## 4     40 Opuntia ficus-indica Large         4     11.00     6.00     0.25
## 5     40 Opuntia ficus-indica Large         5      7.00     4.50     0.50
## 6     40   Opuntia santa-rita Large         1      3.75     3.75     0.50
##   spines insects damage location
## 1      N       N   None  Seventh
## 2      N       N   None    Sixth
## 3      N       N   None    Third
## 4      N       N   Some    Fifth
## 5      N       N   Some   Fourth
## 6      N       N   None    Eigth
```

11. Calculate the mean of depth of all the cactus pads that were measured.

```
mean(pads$depth_in)
```

```
## [1] 0.4149
```

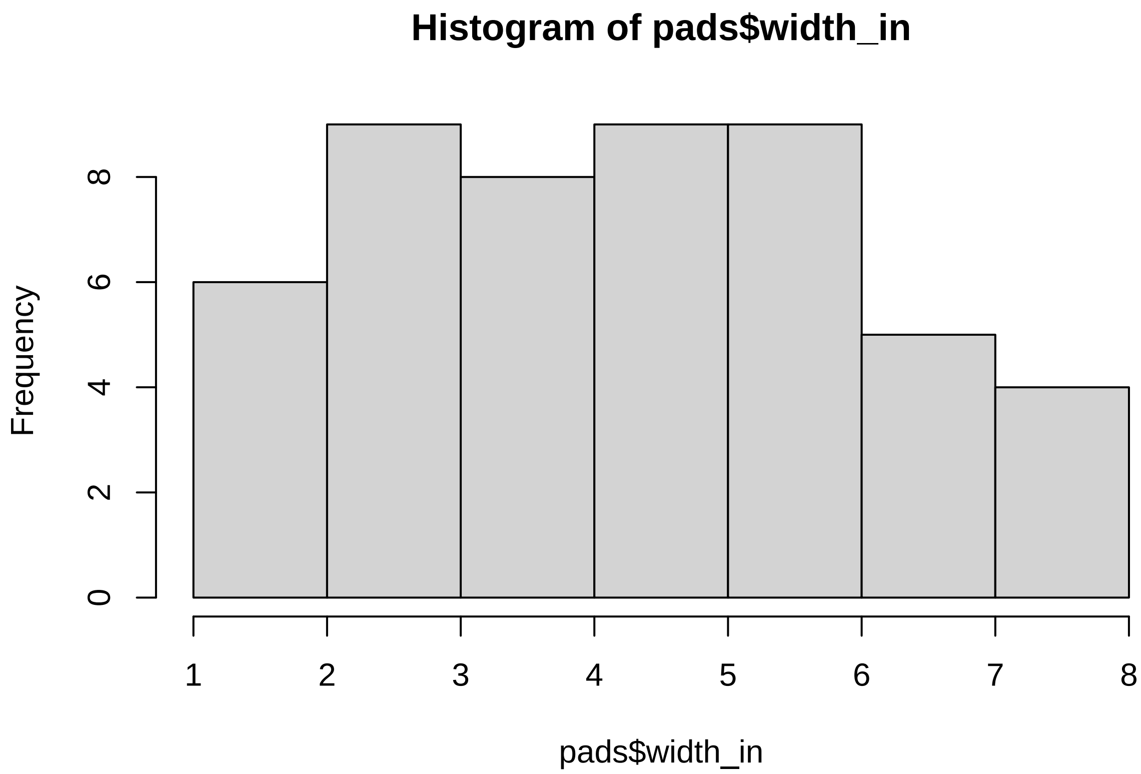
12. Use the `round()` function to around the mean of the cactus pad depths (Q11) to only 1 place after the decimal.

```r
round(mean(pads$depth_in), digits = 1)
```

```
## [1] 0.4
```

13. Create a histogram with the width of all the cactus pads that were measured.

```r
hist(pads$width_in)
```



14. In 1-2 sentences, describe what each axis on this histogram represents.

Answer: The x-axis indicates groupings/bins of width measurements for each cactus pad. The y-axis indicates the frequency with which measurements are in each group.

15. Describe what the following line of code is doing. Make sure to include each part of the code (i.e., before the arrow, the arrow, after the arrow…). (As a point of reference, this took me ~6 sentences in the answer key).

```r
pad_depth <- pads[pads$depth_in > 1, ]
```

Answer: ***pad_depth****is the name of the object we are creating. the assignment arrow (‘<-‘) says that we are creating the object pad_depth to include whatever the code on the right says. On the right, we first tell R that we want to select something from the* ***pads*** *data frame. The [] indicate that we are subsetting. We set a condition in the rows section that indicates we only want rows in which the value in the depth_in column is greater than 1. Leaving the space after the comma blank indicates that we want all columns.*

Calculating the Coefficient of Variance

One type of measure of variability that we did not get a change to cover in class is called the "coefficient of variance." I recommend taking a look at that slide in the lecture. The coefficient of variance is a way of making the amount of variation in the data relative to the data values themselves. This way, we can calculate standard deviation for the weight of squirrels and elephants and be able to compare them. Without standardization, the values for the elephant weights will be much bigger than the squirrel weights, but that doesn't actually mean that they have more variation. We want to know if they have more variation once we've accounted for the fact that elephant weights will naturally be much larger. We do this by dividing the standard deviation by the mean. We can then multiply by 100 to get a percentage.

The remainder of the questions in this assignment lead you through the steps to calculate the coefficient of variance for pad length for two of the three cactus species that we measured. This will give us relative amounts of variation in between the two species.

16. Create 2 different data frames: one called `ficus` with only rows that have *Opuntia ficus-indica* in the species column and one called `santa_rita` with only rows which have *Opuntia santa-rita* in the species column. (Hint: take a look at Q15 for inspiration on how to do that)

```r
ficus <- pads[pads$species == "Opuntia ficus-indica", ]
santa_rita <- pads[pads$species == "Opuntia santa-rita", ]
```

17. Calculate the mean pad length for each species.

```r
mean_ficus <- mean(ficus$length_in)
mean_ficus
```

```
## [1] 8.42875
```

```r
mean_santa <- mean(santa_rita$length_in)
mean_santa
```

```
## [1] 5.9
```

18. In a similar fashion, calculate the standard deviation (`sd()`) for the weight of both species.

```r
sd_ficus <- sd(ficus$length_in)
sd_ficus
```

```
## [1] 2.192934
```

```r
sd_santa <- sd(santa_rita$length_in)
sd_santa
```

```
## [1] 2.138276
```

19. Using the objects (not just the numbers) you just created in questions 22 and 23, calculate the coefficient of variance (CV) for each species. Remember to multiply by 100 to convert into a percentage.

```
cv_ficus <- (sd_ficus / mean_ficus) * 100
cv_ficus
```

```
## [1] 26.01731
```

```
cv_santa <- (sd_santa / mean_santa) * 100
cv_santa
```

```
## [1] 36.24196
```

20. In 2-3 sentences, interpret the results of your calculations from Q17-19.

Answer: Both the mean and sd for ficus are nearly twice that of santa rita. The larger sd would suggest more spread in the data in ficus than santa rita. However, once we account for the means, the percent of variation in both species is ~30%, so they actually have similar amounts of variation.

Turning in Your Assignment

Follow these steps to successfully turn in your assignment on D2L.

1. Click the `Knit` button up near the top of this document. This should produce a PDF file that shows up in the `Files` panel on the bottom-right of your screen.
2. Click the empty box to the left of the PDF file.
3. Click on the blue gear near the top of the `Files` panel and choose Export.
4. Put your last name at the front of the file name when prompted, then click the Download button. The PDF file of your assignment is now in your “Downloads” folder on your device.
5. Head over to D2L and navigate to Module 1 Assignment 2. Submit the PDF file that you just downloaded.