CWRU DSCI351-451: Homework 2

Prof.:Roger French, TA: JiQi Liu

04 September, 2018

Contents

2.1.0.0.1 HW2, 5 points, 5 questions.

- Due Tuesday September 11th
  - Before Class
- Answers to these problems may be in the back of the book (OIStats-v3),
  - so you can check your work.

- The grading is done on how you show your thinking,
  - explain yourself and
  - show your Rcode and
  - the output you got from your code.
- Code style is important
  - Follow Rstudio code diagnostics notices
  - And the Google R Style Guide
  - Also available in your class repo, cheat sheets

To be done as an Rmd file,

- where you turn in
  - the Rmd file and
  - the compiled pdf showing your work.
  - and the R script of IntroR.R

You will want to produce a report type format

- (html and pdf type document) to turn in.
- And not an ioslides or beamer (slide type) compiled output.
  - These are presentation formats, and can be fussy

Also are you backing up your git repo

- in a second and third location,
- to avoid corruption problems?

2.1.0.1 1. (1/2 pt.) R Calculator:

In the 1-Assignments/hw/hw2 folder in your repo,

- You will find an R script with some basic R variable problems.

- Complete these using proper R commands in your R script file
  - and submit the solution.

- Don't forget attribution, versioning and licensing.

2.1.0.2 2. (1/2 pt.) Data Basics: OpenIntroStats Exercise 1.7 1n Chapter 1, pg. 57.

Answer this in this Rmd file and

- explain what you are doing,
- i.e. show your R code and work.

2.1.0.2.1 OIS Exercise 1.7 Fisher's irises.

Sir Ronald Aylmer Fisher was

- An English statistician, evolutionary biologist, and geneticist
- Who worked on a data set that contained
  - sepal length and width, and petal length and width
  - from three species of iris flowers
    - (setosa, versicolor and virginica).
- There were 50 flowers from each species in the data set.

2.1.0.2.2 (2a) How many cases were included in the data?

Show you R code!

2.1.0.2.3 (2b) How many numerical variables are included in the data?

- Indicate what they are, and
- if they are continuous or discrete.

2.1.0.2.4 (2c) How many categorical variables are included in the data,

- and what are they?
- List the corresponding levels (categories).

2.1.0.3 3. (1 pt.) Examining Numerical Data:. Factory defective rate.

A factory quality control manager decides

- to investigate the percentage of defective items produced each day.
- Within a given work week (Monday through Friday)
  - the percentage of defective items produced was
    - 2%, 1.4%, 4%, 3%, 2.2%.

2.1.0.3.1 (3a) Calculate the mean for these data.

- Show your R code!

2.1.0.3.2 (3b) Calculate the standard deviation for these data,

- showing each step in detail.

2.1.0.4 4. (1 pt.) Examining Numerical Data: OpenIntroStats Exercise 1.47 in Chapter 1, pg 66.

2.1.0.4.1 Exercise 1.47 Means and SDs.

For each part, compare distributions (1) and (2)

- based on their means and standard deviations.

You do not need to calculate these statistics;

- simply state how the means and the standard deviations compare.

Make sure to explain your reasoning.

- Hint: It may be useful to sketch dot plots of the distributions.

2.1.0.5 5. (1/2 pt.) For Loops

Using a for loop

- complete the problem below in the given code space
- Create a data frame of
  - the average temperature (Temp) and
  - wind speeds (Wind) for each month
- The data frame must have 3 columns -
  - average temperature,
  - average wind speed, and
  - month number (5, 6, etc.),
- colnames are up to you

You may only use one for loop - You may not hard code (i.e. type in manually) - the number of each month -Hint: you may find the unique() function useful

```r
data("airquality")
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

```r
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

2.1.0.6 6. (1/2 pt.) Normal Distribution:

2.1.0.6.1 OpenIntroStats Exercise 3.2 in Chapter 3, pg 158.

Area under the curve, II:

What percent of a standard normal distribution N (mu = 0, sigma = 1)

- is found in each region?
- Be sure to draw a graph.

Four parts of this problem.

- (a) For Z > -1.13
- (b) For Z < 0.18
- (c) For Z > 8
- (d) For |Z| < 0.5

2.1.0.7 Links

http://www.r-project.org

http://rmarkdown.rstudio.com/