Practice Problems 1

Solutions

Instructions

To complete your assignment, please follow these steps:

0. Download R and RStudio (if you haven't already). See the *Useful Resources* section on Moodle for installation links and videos with step-by-step installation instructions. **Get started on this part as soon as possible.** If you get stuck, you'll need to reach out to your instructor and/or come to office hours, so plan accordingly!

1. Download `practice1.Rmd` from Moodle and save it some place on your computer that you can easily find again (i.e., NOT in your Downloads folder). **I strongly encourage you to create a new folder dedicated to work for this class.** See the *File Structure and Organization* video in the Useful Resources section on Moodle for tips on how to do this.

2. Make sure that the file you downloaded is called `practice1.Rmd` and not `practice1.Rmd.txt`. The latter often happens when you use Safari on a Mac—try downloading the file using a different browser (e.g., Chrome) instead, or edit the file name (as explained in the *File Structure and Organization* video).

3. Open `practice1.Rmd` in RStudio. See the *Intro to R and RStudio*, *R Data Types*, and *R Error Messages and Troubleshooting* videos.

4. Add your section number, name, and the due date to the third, fourth, and fifth lines of the file.

5. Make sure you've installed the `readr` and `NHANES` packages — you'll need both of those for this assignment. Depending on how far you've made it in the lecture videos and in-class activities, you might have installed some of these packages already. Open the `Packages` panel (usually in the bottom right corner) to see the list of all packages that are already installed. Look to see if `readr` and `NHANES` are listed here. If they're on this list, that means they're already installed and you're good to go. If any packages are missing from this list, type `install.packages('packagename')` in the *Console* (usually in the bottom left corner) and hit enter. See the *R Packages* video for more details.

6. Try *Knitting* your document: click the `Knit` button at the top of this screen (look for yarn and needle). See the *Intro to RMarkdown* video for more details, and get in touch with your instructor or the preceptors if you encounter an error messages.

7. Answer the questions in Parts 1–4. Click `Knit` occasionally along the way to make sure everything looks okay.

8. Once you're done with Parts 1–4, click `Knit` one final time. This will turn your R Markdown document into a nicely formatted HTML file. Look at the file to make sure it looks like you want it to (no R code errors, no data print out that lasts 50 pages, etc.).

9. Once your HTML file pops up and you've checked that it looks like what you want, click `Open in Browser` and then Print and select "Save/Export as pdf". You will upload the pdf to Moodle.

10. Submit both this `.Rmd` file and the knitted `.pdf` version you generated in Step 9 to Moodle.

Part 1. Background Survey

I want to know a little more about you! Please answer the following questions. Note that there are no "correct" answers to the questions: you'll get credit as long as you put effort into answering each question.

> If you're curious, here are Kelsey's answers to the Background Survey questions…

a. What is your preferred name (and, if you'd like to share, what are your pronouns)?

> Kelsey or Professor Grinde (she/her/hers)

b. Please write a short pronunciation guide for you name: spell it out phonetically, tell me what it rhymes with, etc.

> KELL-see GRIN-dee

c. What is your hometown?

> I grew up in Plymouth, Minnesota (a suburb of Minneapolis, about 30 minutes from Mac), and now I live in St. Paul

d. What is your class year at Macalester?

> This is my third year teaching at Mac

e. What is (are) your major(s) and minor(s)? If you haven't declared yet, let me know what your passions are right now!

> I graduated from St. Olaf College with a major in mathematics and a concentration in statistics. I was also one course shy of a Spanish major!

f. What are your career and/or post-graduation ideas or goals?

> After graduating from St. Olaf, I headed to Seattle to get my Phd in Biostatistics at the University of Washington. I graduated from UW in August 2019, started a teaching postdoc at Macalester in September 2019, and then started a new position at Mac as an Assistant Professor in September 2020.

g. Have you taken any previous statistics courses? If so, which classes have you taken and when did you take them?

I've taken somewhere around 26 stats courses throughout my lifetime. My first stats class was AP Stats in high school, and to be honest I thought it was incredibly boring. But, during my sophomore year at St. Olaf I took "Statistical Modeling" (a course very similar to this one) and, much to my surprise, I absolutely loved it! I've been studying statistics ever since.

h. Do you have any other prior experience with statistics (e.g., a class in another department, job, research)?

I've been working as a researcher in the field of statistical genetics ever since my junior year of college, and also have some experience as a statistical consultant.

i. Do you have any previous computing experience (e.g., Excel, R, Stata, Matlab, C, Python)? If so, in what context?

I first learned R in my statistics courses at St. Olaf. I now use it every day for my research and consulting work (and when I need a calculator!). Besides R, I often use shell scripting and occassionally Python.

j. Why are you taking this course? What are you hoping to learn/gain from this class?

I love teaching this course: it was a course like this that first got me interested in being a statistician! Learning how to use statistics to answer important scientific questions and getting hands-on experience with data analysis showed me how powerful and exciting statistics can be. I hope you all are able to gain the same insight from this course!

k. What do you do when you're not in class? (e.g., are you an athlete? in a musical ensemble? involved in any campus organizations? board game enthusiast?)

I love all things related to soccer (playing, watching, and coaching, although I'm not doing any coaching at the moment). I'm also a big fan of biking, hiking, and trying new restaurants. Although I grew up in Minnesota, I'm still relatively new to St. Paul: if you have any favorite hiking trails or restaurants around campus, or elsewhere in the Twin Cities, I'd love to hear your recommendations!

l. Looking through the Project Description on Moodle, which of the project datasets sounds most interesting to you? Is there another dataset, not on this list, that you'd be more interested in working with? (This is not a commitment to using a particular dataset for your project, but note that you will need to pick a dataset by the end of next week.)

> If you're thinking about using a dataset that's not on the list of recommended datasets, I'll be in touch!

m. Is there anything else you would like me to know about you?

> I am always happy to chat about questions you have about course material, ways I can make this class better for you, career/post-grad planning, or anything else! Catch me after class, come to office hours, or email me to set up a time to talk. I'm looking forward to getting to know all of you throughout the semester.

Part 2. Load Data in R

NHANES Data

Enter code into the code chunk below to load the `NHANES` dataset into R, and then print out the names of all the variables in the dataset. Don't forget to load any packages you might need first! (*Hint: see Activity 2*)

```
# load package first
library(NHANES)

# then load dataset
data(NHANES)

# then print out variable names
names(NHANES)
```

```
##  [1] "ID"               "SurveyYr"         "Gender"           "Age"
##  [5] "AgeDecade"        "AgeMonths"        "Race1"            "Race3"
##  [9] "Education"        "MaritalStatus"    "HHIncome"         "HHIncomeMid"
## [13] "Poverty"          "HomeRooms"        "HomeOwn"          "Work"
## [17] "Weight"           "Length"           "HeadCirc"         "Height"
## [21] "BMI"              "BMICatUnder20yrs" "BMI_WHO"          "Pulse"
## [25] "BPSysAve"         "BPDiaAve"         "BPSys1"           "BPDia1"
## [29] "BPSys2"           "BPDia2"           "BPSys3"           "BPDia3"
## [33] "Testosterone"     "DirectChol"       "TotChol"          "UrineVol1"
## [37] "UrineFlow1"       "UrineVol2"        "UrineFlow2"       "Diabetes"
## [41] "DiabetesAge"      "HealthGen"        "DaysPhysHlthBad"  "DaysMentHlthBad"
## [45] "LittleInterest"   "Depressed"        "nPregnancies"     "nBabies"
## [49] "Age1stBaby"       "SleepHrsNight"    "SleepTrouble"     "PhysActive"
## [53] "PhysActiveDays"   "TVHrsDay"         "CompHrsDay"       "TVHrsDayChild"
## [57] "CompHrsDayChild"  "Alcohol12PlusYr"  "AlcoholDay"       "AlcoholYear"
## [61] "SmokeNow"         "Smoke100"         "Smoke100n"        "SmokeAge"
## [65] "Marijuana"        "AgeFirstMarij"    "RegularMarij"     "AgeRegMarij"
## [69] "HardDrugs"        "SexEver"          "SexAge"           "SexNumPartnLife"
## [73] "SexNumPartYear"   "SameSex"          "SexOrientation"   "PregnantNow"
```

Intro Survey Data

Next, enter code in the code chunk below to load the `introsurvey.csv` dataset into R, and then print out the first six rows of the dataset. Don't forget to load any packages that you might need, and make sure that you've downloaded the `.csv` file from Moodle and it's saved in the same folder as this RMD. (*Hint: see Activity 3*)

```
# load package first
library(readr)

# then read in dataset
results <- read_csv("../activities/introsurvey.csv")
```

```
## Rows: 51 Columns: 10
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (4): Timestamp, What is your declared or potential major? (If you are a ...
## dbl (6): How many hours of sleep did you get last night?, How many cups of c...
```

```
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# then print out first six rows
head(results)
```

```
## # A tibble: 6 × 10
##   Timestamp  `How many hours… `How many cups … `What is your d… `What is your a…
##   <chr>                 <dbl>            <dbl> <chr>                       <dbl>
## 1 8/26/2021…              6.5                3 math                         2014
## 2 9/4/2021 …              7                  0 Computer Science             2025
## 3 9/6/2021 …              8                  0 econ                         2025
## 4 9/6/2021 …              8                  0 math                         2024
## 5 9/7/2021 …              7                  0 Math                         2022
## 6 9/8/2021 …              7                  0 Biology                      2023
## # … with 5 more variables:
## #   How many stats courses have you taken in the past? <dbl>,
## #   Have you used R or RStudio before? <chr>,
## #   On a scale of 1 to 10, how excited are you about this class? <dbl>,
## #   When is your birthday? <chr>,
## #   How many unread emails do you have in your inbox right now? <dbl>
```

NOTE: depending on when you downloaded the "introsurvey.csv" file from Moodle, you may see slightly different output here (because you're working with a different version of the dataset). That's okay! The code will be the same either way.

Also note that I used slightly different syntax for reading in the CSV file than we are used to seeing in class. I did this because of how I organize files on my computer. In this case, my RMD for these solutions is saved in a "practice" folder within my Stat 155 folder, whereas the introsurvey.csv file is saved in my "activities" folder. In order to tell R to look for the intro survey dataset in that actitivities folder, I specified the *path* to get to the file ("../activities/introsurvey.csv") rather than just giving it the file name ("introsurvey.csv"). This is useful to keep in mind if you want to save your RMD and CSV files in different folders.

Part 3. Data Context

We'll consider the NHANES dataset for the following questions. (*Hint: see Activity 2*)

Who

What does each row of this data set represent (i.e., what are the cases)?

Each row in the NHANES dataset represents a person who participated in the study. There are 10,000 of them (see below).

How many cases are there?

```
# use this code chunk to find the number of cases
nrow(NHANES)
```

```
## [1] 10000
```

What

What do the columns of this data set represent (i.e., what are the variables)?

Each of the columns in this dataset represents a question that was asked or something that was measured on those study participants. There are a few different types of information we have on each person: demographic info, physical measurements, health variables, and lifestyle variables. There are 76 total variables (see below).

How many variables are there?

```
# use this code chunk to find the number of variables
ncol(NHANES)
```

```
## [1] 76
```

Pick four variables that you're interested in learning more about, and open the help file for the NHANES package data to read more about them. List those variables, and their type (quantitative or categorical), here.

> If a dataset is stored inside an R package, then you can open up a help file for that dataset to learn more about it. Do this by typing `?NHANES` in the Console.

> Your answer to this question depends on which variables you listed. In general, quantitative variables are variables that have numerical values with units, and categorical variables are variables whose values are words or numbers without units.

> Paying attention to units is important: you can have something with numerical values that doesn't have units (e.g., zipcode), and this would be considerd a *categorical* variable.

> Another thing to look out for is variables that cut a quantitative variable into ranges or categories; for example, `HHIncome` in the NHANES data takes income in US dollars and cuts it into categories (0 - 4999, 5000-9999, etc) — this is now a categorical variable. There are a few variables like this in the NHANES dataset.

Part 4. Data Limiations

We'll consider the Intro Survey dataset for the following questions. (*Hint: see Activity 3*)

Bias

The Stat 155 Intro Survey asked about your declared or intended major. Suppose a researcher wants to use your responses to this question to learn about the academic interests of the entire Macalester student body (a larger population of interest). Would you caution them against it? Why/why not? What if the researcher were interested in the responses to one of the other questions on the survey — would your suggestion differ?

> Yes, I would caution them against it. Students who volunteer (choose) to be in Stat 155 may have different major interests than the general Macalester student body. (In other words, the students in Stat 155 are not *representative* of the Macalester student body when it comes to majors.)

> The birthday question, for example, might be fine to generalize to the larger population as I don't believe that when an individual's birthday falls impacts their choice of courses.

What type of bias is this question asking about?

> This is an example of *sampling bias*: the people in our sample (Stat 155 students) may not be representative of the larger population of interest (all Macalester students).

More bias

The Stat 155 Intro Survey asked about your enthusiasm for this course on a scale from 1 to 10. Suppose the instructor wanted to use these data to gauge the enthusiasm of students in this class. Would you caution them against it? Why/why not? What if responses to the survey had not been collected anonymously — would your answer change?

> Since it was collected anonymously, students may have given honest responses but it may still be hard to provide a 1 since they knew the instructor would look at the data. Having your name attached to your response might make you even less likely to report a 1, even if that's truly how you feel.

What type of bias is this question asking about?

> This is an example of *information bias* (specifically response/social desirability bias).

Other limitations

Considering the context under which these data were collected, do you have any other concerns about the data quality (do the values represent the true values) or data representation (who/what might be excluded from the data)? In other words, can you think of any other limitations that this dataset may have?

There are many possible answers to this question! Here are a few ideas (not an exhaustive list):

- rounding of sleep hours
- inconsistency of “cups” of coffee
- recall bias (can you remember how many cups you drank or how many unread emails you have)
- the categories provided may not have been specific enough to accurately represent someone’s experience
- unread email count wasn’t taken at the same time by all people (that fluctuates throughout the day and everyone took the survey at different points in their day)
- even though data were collected anonymously, we can still identify which respondent was the instructor (based on answers to questions like graduation year and number of previous stats classes)

Part 5. Knit

Once you’ve answered all the questions above, you’ll need to `Knit` this document and then save that knitted document as a PDF. See the assignment *Instructions* at the top of this document and on Moodle for more details.