Lab 1

BSS Stat 20

2022-06-12

NOTICE

If you have any collaborators, please write a sentence before Question 1 which acknowledges them. In addition, make sure that the sentence they therefore are also required to write acknowledges you.

For a template you can follow, see the course syllabus. This notice will be pasted at the beginning of every lab assignment.

Questions

Question 1

The purpose of this question is to get you familiar with using the source (or visual) editor in RMarkdown to compose written documents. The question is as follows:

> Where are you from *or* where would you like to visit?

Please answer this question in three components, each properly separated by a subheading, like as done here:

part a

If you are from the United States, give the city (and its state) that you are from. If not, give the name of a United States city (and its state) that you would like to visit.

part b

One to three paragraphs about this United States city in text, taken from the Wikipedia page for the city. As practice, before the paragraph(s), please write a sentence acknowledging where you got the information from. Ex: "I found this information at [link here]"

part c

An enumerated list of your top three favorite things about this city. (If you don't have three, you can make some up!)

Question 2

The purpose of this question is to familiarize yourself with the process of coming up with/receiving a question that could be answered with data and thinking about the exact structure of that data. Sometimes, the data you need is available to you; other times you'll have to collect it yourself. In this case you will "collect" it by constructing a part of the data yourself with made-up values.

John Arbuthnot (1667-1735) was a Scottish physician who at one point wanted to determine the true proportion of babies born female born each year. In order to collect this data, he traveled to London, England in 1710 and collected *christening records* from the parish churches there. Each christening record consisted of the child's name, parents and birth date. From these, he counted the total names by year from 1629 to 1710 that were traditionally female or male names.

part a

What do you believe the probability of a child being born female is? Provide the evidence/reasoning you used to form your opinion.

part b

Provide a sketch (upload your drawing) of what you believe his finished data set might look like. Sketch five example rows. Where appropriate, you can (and should) make up values. Make sure you label your columns.

part c

What does each row in your data set correspond to?

part d

What are the data types of each column in your data set according to the Taxonomy of Data?

part e

Now, with R code, construct a *data frame* based off the data set you have sketched. Save the final data frame into an object called `my_arbuthnot`.

Question 3

It turns out that we have access to the actual dataset Arbuthnot collected! On DataHub, it is called `arbuthnot` and can be accessed by loading the `stat20data` library and then running the line

```
data(arbuthnot)
```

Make sure you do the above steps first before continuing.

Most of this question is meant to get you practicing with some of the basic `dplyr` functions we cover during this week. Some you can do without code; if you take that route please walk us through the steps you took. Please do not use the `%>%` operator during this question because that will be for next week.

part a

Compare the values of number of boys and girls born per year in your rows of `my_arbuthnot` versus the rows of `arbuthnot`. Were you close to the actual number of children?

part b

Which year saw the *most* number of boys christened?

part c

Which year saw the *least* number of *children* christened?

part d

What was the proportion of girl names christened in 1694? Please round to three decimal places.

part e

Compute the average proportion of girl names christened per year. We will go over how to do this in detail next week; but for now template code to do this is below, depending on the name of your dataset (I am calling it here `arbuthnot_new`) and on the name of column `girl_prop` which you probably made in completing part c.

```
summarise(arbuthnot_new, mean(girl_prop))
```

Based off of your result, how is your opinion on the probability of a child being born a girl affected?

Question 4

Please complete the welcome survey. It can be accessed by clicking here. Make sure you are logged into your UC Berkeley Google account.