STA 2260

HOMEWORK 1 - R PORTION

Kevin Bailey

Due 9/4/2020

Instructions

- Create a new `R` Script
  - When saving this file, use the format `LastName_FirstName_hw1`, this should be a `.R` file when saved.
- When doing separate parts, create clear breaks in the code separated by a comment about which part will follow. Something like:

```
code is here

# part (b)
more code is here
```

The goal of this assignment is to create three histograms in `R` from simulated data. One set of data will be discrete, one will be continuous, and one will be categorical (ordinal, specifically). I will start you off:

```
set.seed(1) # This controls random results. Run this BEFORE simulating the data.

# ?sample and ?seq for documentation
# seq(...) generates a sequence of numbers "from" some number "to" another, inclusive,
    "by" is the increment argument, so "by=2" will have a sequence of numbers
    incrementing by 2.
# replace=T means once a number is picked, it can be used again. replace=F is "without
    replacement", so a number won't appear more than once.
discrete_data <- sample(x=seq(from=0, to=75, by=1), size=420, replace=T)

# rnorm(...) generates n random numbers from a normal distribution with mean and
    standard deviation as specified. Don't worry to much about this, it's just to get
    truly continuous data.
continuous_data <- rnorm(n=400, mean=300, sd=5)

# LETTERS is just a built in set of the capital letters, "A", "B", "C", etc. Individual
    letters are accessed by array notation, such as LETTERS[1] returning "A".
# prob=... is just a list of probabilities corresponding to the letters, so LETTERS[1]
    has a 50% chance of being selected, for example.
ordinal_data <- sample(LETTERS[c(1:4, 6)], size=420, replace=T, prob=c(0.5, 0.3, 0.15,
    0.1, 0.05))
```

Part (a)

A common rule for determining the number of bins is to do

$$\text{Number of bins} \approx \sqrt{n}$$

where n is the size of your data set (and usually we want between 5-20 bins inclusively). If this number is a decimal, it is typically rounded up (so we take the ceiling). Construct separate histograms (remember that we use left-closed, right-open histograms) for the discrete and continuous data which have a number of bins corresponding to the rule above. What type of shape does the *continuous* data have? What are the features that made you come to your conclusion?

Then, create a barplot for the ordinal data (there is something you need to do to the data for `barplot(...)` to work). This type of barplot has a particular name, it is a *Pareto Chart*.

Part (b)

Now, copy and paste the ordinal data's line to a new line and change the `prob=c(...)` argument around to cause the barplot to no longer be a pareto chart. ~~Note that the sums of all the arguments in `prob=c(...)` need to equal 1!~~ Not actually true, this is just the probability of selecting the corresponding letter. Set another seed before you do this (any number can be inside) so your results are reproducible! You should have *two* lines assigning something to `ordinal_data`. Now, create a barplot again for the *new* ordinal data you just created.

Part (c)

Create a function that takes an argument for the size of the data set. This function should:

- Simulate a random *sample* of data from 10 to 40 in increments of 0.25.
  - The size of the data is determined by the argument you pass to the function.
- Draw a histogram with the appropriate number of bins (that is left-closed, right-open).
  - You may want to remember how `breaks=seq(...)` interacts with `length.out` regarding number of bins. :)
  - The command `sqrt(...)` may be useful.

After your function is created, call it a few times by passing in $n = 25$ and $n = 300$, comment on some of the results you see. The end-ish of the second Week 1 video (lecture 2) should help here. What kind of variation is there between the two aside from number of bins?

FEEL FREE TO CONTACT ME ON ANY OF THESE PARTS AT ANY TIME! I want to help you understand! Feel free to reference the second video and the R script in Week 1 as well. Also, communicate with each other as well as me! Peers are always a helpful resource too. :)