PDAT 610: Module 4 Homework

Introduction

In this assignment, you'll have an opportunity to explore some visualization and simulation based on the R commands you've seen in Module 4. We'll be looking at the published results of a survey conducted by the Gapminder Foundation. They are an organization dedicated to using data and statistics to promote "a fact-based worldview everyone can understand." Over the years, they have served as an important source of effectively-communicated data on the state of the world.

As an optional activity, I encourage you to

- Take their quiz yourself at http://forms.gapminder.org/s3/test-2018 before you read about the results.

  [Note: They may ask for your personal information after showing you your scores, but you don't need to do that. I suspect it's either for putting you on their mailing list or required if you want you scores actually added to their database.]

- Read about the results of the survey at https://www.gapminder.org/ignorance/gms/.
- Check out the main Gapminder site at https://www.gapminder.org if you're interested. They've got some interesting TED talks and data visualizations.

Submission Method

- Create an R Markdown document containing both your code and any written responses or explanations.
- Make sure to number your answers so that the grader can easily see which question you were answering.
- The code for any question part should be contained in a single code block
- When you're done, knit your code to a PDF file, and submit that file through Blackboard.
- Hint: Sometimes students only test their code by running the individual chunks interactively in R Studio. I'd suggest checking often to make sure your code will knit to a final document. It's easier to diagnose any knitting problems if you know exactly what you just typed that caused the knitting to break (as opposed to waiting until the end and troubleshooting the whole document).

Assignment

1. Gapminder claims that the percentage of correct answers out of 12 questions is distributed as follows:[1]

Number Correct	0	1	2	3	4	5	6	7	8	9	10	11	12
Percentage of Respondents	14%	25%	24%	17%	10%	6%	3%	1%	0%	0%	0%	0%	0%

   In R, create a numeric vector containing these percentages, and then use the `barplot` command to create a graph of the distribution of "number correct." Make sure the graph axes are labeled correctly, and add an explanatory title. [Note: You should be able to use `?barplot` to find all the necessary labeling options.]

2. Describe the shape of the graph you obtained. [In other words, skewed (which way?), symmetric, bell-shaped, bimodal, etc.?]

3. Gapminder observes that this relatively poor performance on their quiz shows a perpetuation of what they would argue are harmful misconceptions about world development. To bring this point home, they compare human performance on the quiz to what would have been achieved by a group of chimpanzees choosing randomly.

   a. If there are 12 questions, each with three options, and the chimpanzees really do choose randomly, the number of correct responses X should be described by a binomial probability distribution. Indicate the values of n and p for this binomial distribution.

   b. Use the `dbinom` function to create a new vector of probabilities reflecting the chimp's expected performance on the quiz. As in Question 1, plot the probability that a chimp will get everywhere from 0 to 12 correct. [Note: If you use R's vectorized calculations, you need only use the `dbinom` function once.]

   c. Describe how humans compare to chimps by comparing distribution center, spread and shape in your answer.

4. Overall, the average number of correct answers from humans was 2.22 out of 12, which gives a probability of 0.185 that a human would get any specific question correct.

   a. Use the `dbinom` command again to create a barplot of the expected distribution of correct human answers, if human responses were also described by a binomial distribution.

   b. Write code that uses the `pbinom()` function to calculate $P(X \geq 5)$ for the binomial description of the number of correct human responses.

   c. How closely does your graph from Question 4a resemble the actual distribution of number of correct answers from Question 1? In other words, for which number of correct answers were the actual human probabilities higher or lower than the binomial distribution?

[1] Based on free survey results from Gapminder: gapm.io/gms17, licensed under Creative Common License CC BY 4.0.

5. You should have seen that there were some differences between the actual human performance, and that predicted by the binomial distribution. But perhaps that difference can be explained by random variation—the fact that the humans were merely one possible random sample.

    a. Gapminder states that they surveyed 12,000 people. Write code to create a simulated set of results using the `rbinom` command. [Remember that $p = 0.185$.] You should get a vector whose length is 12,000, with each entry containing the simulated number of correct answers from one respondent. After generating this simulated data, graph it using a combination of `barplot` and `table` commands to obtain a *relative frequency distribution*.

    [Note: You could use the `hist` command with this raw data, but it's formatting is a bit finicky.]

    b. This part is something you can do "by hand" in R Studio, and the results don't have to be included in this document. We want to see if random sampling from our simulated human binomial distribution seems likely to produce anything that looks like the actual data. You could write fancier code to do this, but I'm not asking you to do that here.

    What I would like you to do is to push the "green triangle" button in R Studio to run your code chunk from Question 5a many times. You should get a different random sample every time, provided you haven't included a `set.seed()` command. Does the graph of your random sample *ever* look like the actual human results (i.e., bars with heights that match the actual human results)? In other words, does it seem likely that the real human results could have been one random results from a binomial distribution with the appropriate success probability?

6. Finally, let's think about what's going on here. The binomial distribution makes several assumptions in order to calculate the probability of a certain number of successes:

    - There are two possible outcomes: success or failure. (In this case correct or incorrect answers.)
    - There are a fixed number of trials. (In this case, a trial is one person answering one question, and our random variable is the number of correct that a person got out of that total number of trials.)
    - The probability of success is constant over all trials.
    - Results of one trial are independent of other trials. (In other words, knowing the success/failure result for one trial or some trials won't help you predict results of other trials.)

    One explanation for any differences between the real human results and your simulated results could lie in the fact that one or more of these assumptions is not actually satisfied. Thinking about this real-world situation, briefly explain why or why not these assumptions might be satisfied. Your explanation should be sure to explicitly connect the general assumption to the specifics of this real-world example.