Assignment 4

Steven Spiegel

2023-09-17

Assignment 4 responses

Below are my responses for Assignment 4.

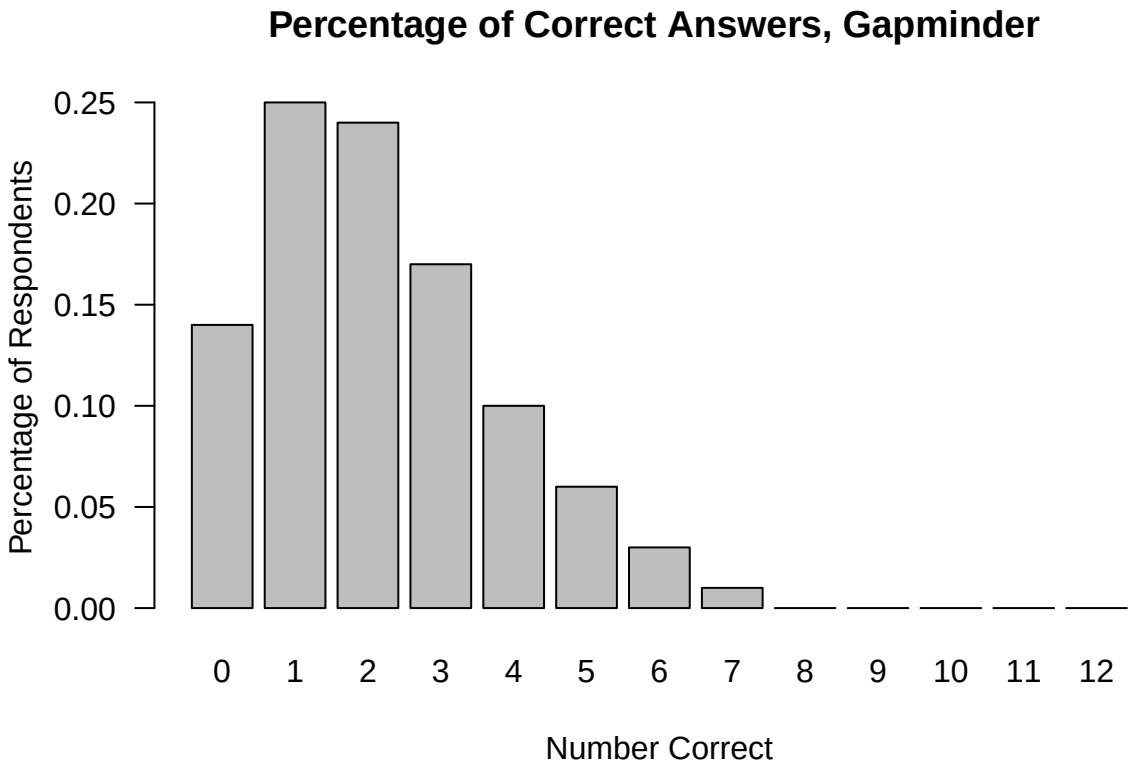
Question 1: **Gapminder claims that the percentage of correct answers out of 12 questions is distributed as follows:**

Number Correct	0	1	2	3	4	5	6	7	8	9	10	11	12
Percentage of Respondents	.14	.25	.24	.17	.10	.06	.03	.01	0	0	0	0	0

In R, create a numeric vector containing these percentages, and then use the barplot command to cre- ate 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.]

Response: **See code below**

```
gap <- c(14,25,24,17,10,6,3,1,0,0,0,0,0) * 0.01
names(gap) <- seq(0,12)
barplot(gap, xlab = "Number Correct",
        ylab = "Percentage of Respondents",
        main = "Percentage of Correct Answers, Gapminder ",
        las = 1)
```



Question 2: **Describe the shape of the graph you obtained. [In other words, skewed (which way?), symmetric, bellshaped, bimodal, etc.?**

Response:

The graph appears to be skewed right, with a *p-value* appearing to be around $\frac{3}{24}$. The majority of the answers are less than 5 correct, appearing to by approximately 95 percent.

Question 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.**

Part 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

Part 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.]

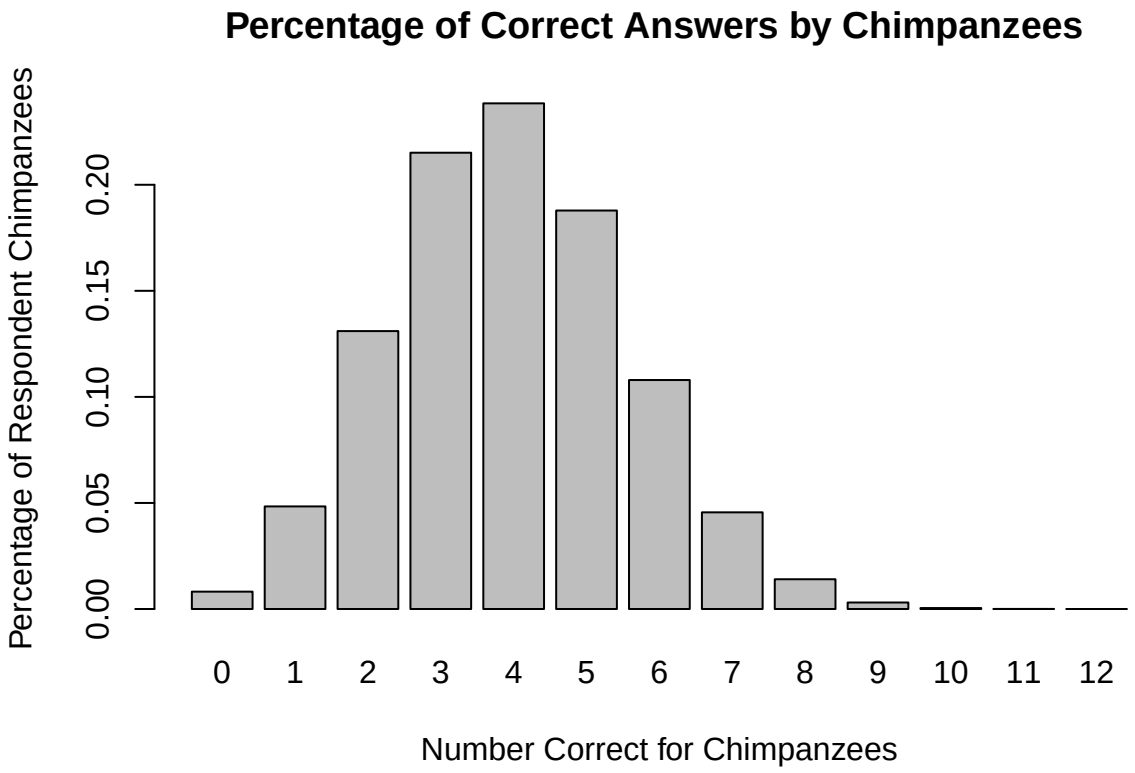
Part C: Describe how humans compare to chimps by comparing distribution center, spread and shape in your answer.

Response:

Part A: The number of correct responses can be described as a binomial distribution with $n = 12$ and probability $p = \frac{1}{3}$.

Part B: **See code below**

```r
x <- seq(0,12)
db <- dbinom(x= x, size = 12, prob = 0.33)
names(db) <- x
barplot(db, xlab = "Number Correct for Chimpanzees",
        ylab = "Percentage of Respondent Chimpanzees",
        main = "Percentage of Correct Answers by Chimpanzees ")
```



Part C: The "Chimp" distribution is much more symmetrical with the mean value at 4. The human distribution is shifted left and skewed right, appearing to have a mean of approximately .21.

Question 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.**

Part 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**

Part 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.**

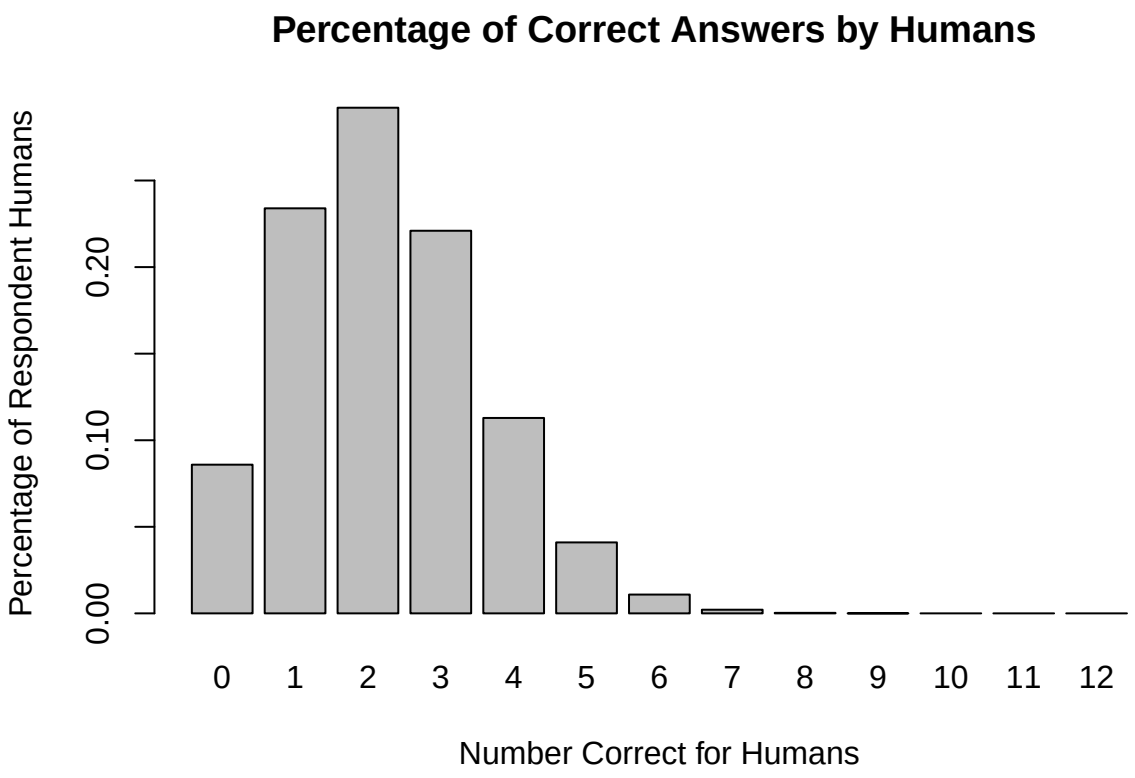
Part C: **How closely does your graph from question 4a resemble the actual distribution of number of correct answers from Question 1? In other words, for which number of correct answers were the actual human probabilities higher or lower than the binomial distribution?**

Responses:

Part A: **see code below**

```r
x <- seq(0,12)
db <- dbinom(x= x, size = 12, prob = 0.185)
names(db) <- x
```

```
barplot(db, xlab = "Number Correct for Humans",
        ylab = "Percentage of Respondent Humans",
        main = "Percentage of Correct Answers by Humans ")
```



Part B: **see code below**

```
prob <- 1 - sum(pbinom(5, 12, 0.185)) #or
protail <- sum(pbinom(5, 12, 0.185,lower.tail = FALSE))
protail
```

```
## [1] 0.01329958
```

```
prob
```

```
## [1] 0.01329958
```

Part C: The distribution produced from part A is more symmetric than the results given by the survey. More respondents only got 1 correct than predicted by the binomial distribution. More respondents got 6-7 questions correct than the binomial, thus demonstrating the skewed nature of the distribution.

Question 5: **You should have seen that there were some differences between the actual human performance, and hat 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.**

Part 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.

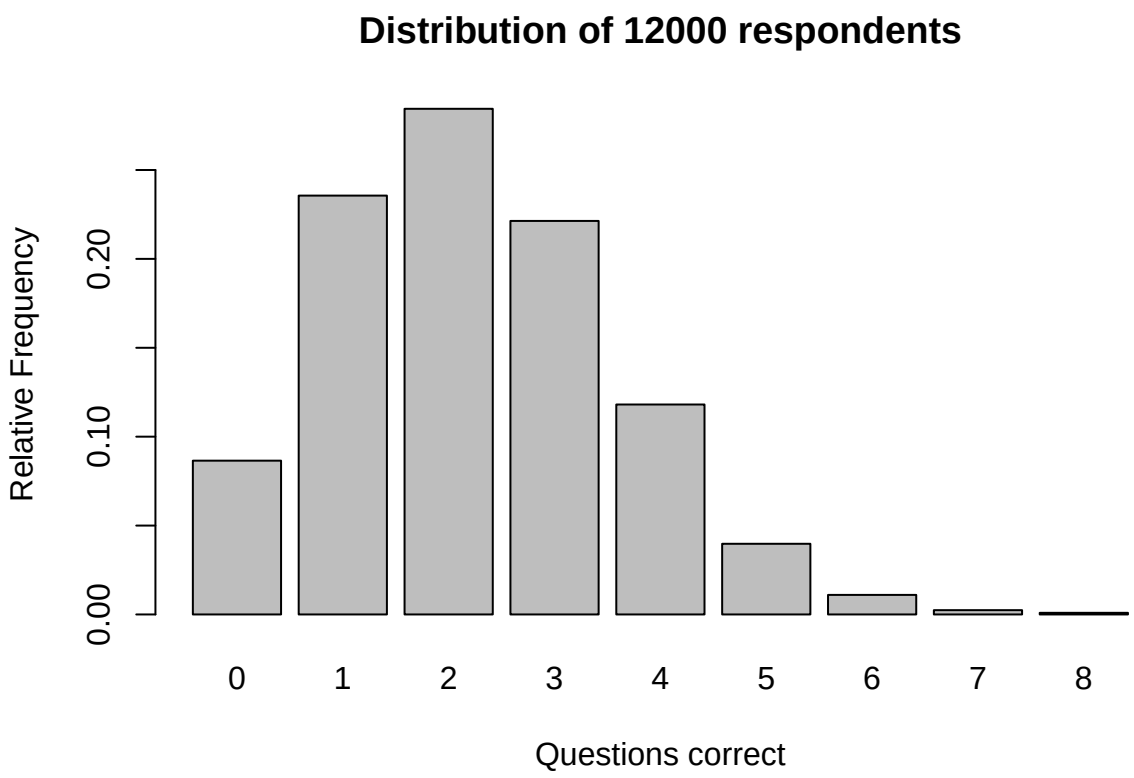
Part B: **his part is something you can do "by hand" in R Studio, and the results don't have to be included in this document. We want to see if random sampling from our simulated human binomial distribution seems likely to produce anything that looks like the actual data. You could write fancier code 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?

Response:

Part A: **see code below**

```
sim <- rbinom(12000, 12, 0.185)
barplot(table(sim)/ length(sim), xlab = "Questions correct",
        ylab = "Relative Frequency",
        main = "Distribution of 12000 respondents")
```



Part B: **see code below**

Most of these plots look mostly symmetrical. Even after many attempts, they all look mostly like the binomial distribution with very few having any skew whatsoever. The third plot in Figure 1 is the only one that looks somewhat skewed, however many of these plots only have slight variation in their distribution.

Question 6: **Finally, let's think about what's going on here. The binomial distribution makes**

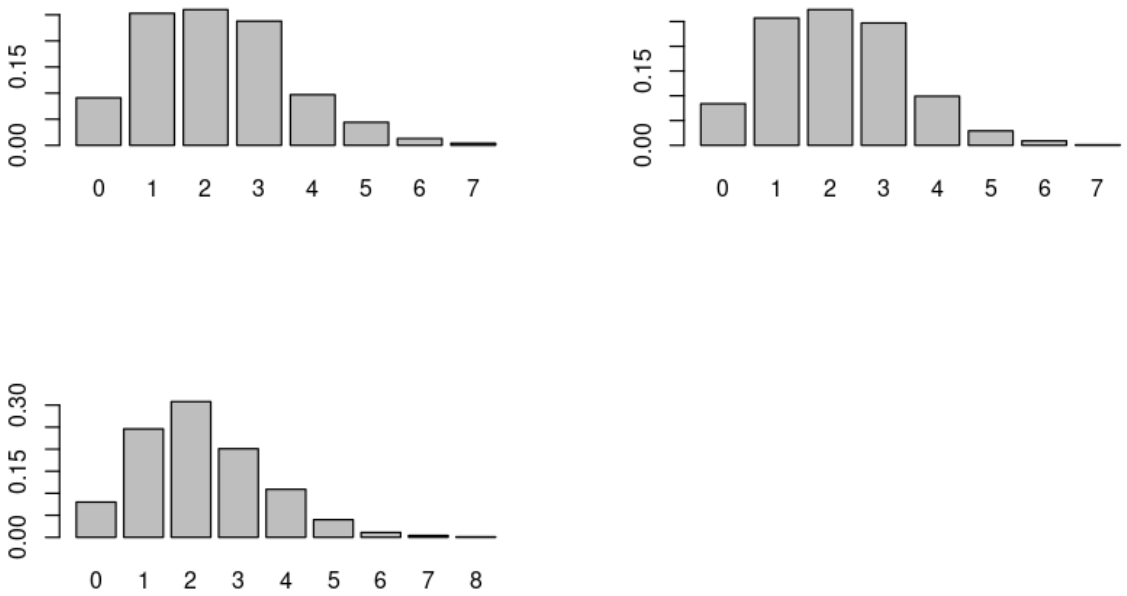


Figure 1: First trial

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.

Response:

There are two assumptions that don't necessarily seem applicable in this situation, particularly that the probability of success is constant over all trials and the independence of trial. It certainly seems plausible that certain questions are easier than others, or subject-specific surveys can vary in difficulty. Some of the questions in my test were certainly easier than other questions. The quiz I took were facts about the US so several of the questions were more familiar.

The other assumption that is likely not true is trial independence. The survey I took allowed me to predict

the answer of another. I can't remember the specifics, but I do remember one question informing my answer of a separate question. This may not be true of all the surveys but some of the questions and their phrasing can sometimes assist in answering a different question.