STAT 260 Summer 2024: R Assignment 2

Due: Friday July 5 BEFORE 6pm (PDT) on Crowdmark.

Introduction to R: Before attempting this assignment, read and work through the **Introduction to R Assignment 3** file posted on Crowdmark. This file contains a list of all the R commands needed to complete this assignment.

Submission: Follow the same submission instructions as in R Assignment 1. Make sure to copy/paste your R code and the output into a document file, and then convert it into PDF. Upload your files for submission to the assignment on Crowdmark, before Friday July 5 at 6:00pm (PDT).

Note: For each of the following, carry out your calculations **only** using R or RStudio. ***For each part, include a copy of your R code used and the output of your code.*** (i.e. Copy and paste the relevant pieces into a Word document (or other word processing document).)

Note: R can complete basic arithmetic operations such as addition (+), subtraction (-), multiplication (*), division (/), and square root (sqrt()). It should be noted that parameter values can be entered as arithmetic expressions. For example, we could enter an arithmetic expression using multiplication when writing the value of lambda in the desired Poisson function. As such, ***you must complete all of the following calculations using only R*** - no external calculator should be used.

Part 1 Background: *Although CDF tables are a useful tool for calculating probabilities by hand, they only show probability distributions for certain parameter values, and so they limit what questions we can actually solve. For example, if we want to calculate a probability using a binomial distribution with a very large n (ex. $n = 500$), or a very specific value of p (ex. $p = 0.1258$), our binomial CDF table will be of no help. In these situations, we turn to statistical software packages like R to complete quick and accurate calculations.*

Answer each of the following probability questions using R. In your document, for each of the following, define your random variable, give its distribution, and state the probability that you are attempting to calculate. Then, complete the required calculation in R, copy/pasting both your code and the output. For example:

Definition: Let X be the number of cats in a 2 km^2 city block.

Distribution: $X \sim Poisson(\lambda = 5.4)$

Probability: $P(X \geq 4)$

```
> 1-ppois(3, lambda=5.4)
[1] 0.786709
```

(a) **[3 marks]** A large glass manufacturer makes lenses for a variety of commercial applications. It is known that 0.65% of the lenses are cracked during production and are discarded before shipment.

(i) If we examine a random sample of 500 lenses, what is the probability that exactly 7 are cracked?

(ii) In a random sample of 3000 lenses, what is the probability that 20 to 25 (inclusive) lenses are cracked?

(b) **[3 marks]** In a certain region, the frequency of significant earthquakes (i.e. earthquakes that are strong enough for humans to feel) is known to follow a Poisson distribution with an average rate of 4.8 earthquakes per year. Incomplete records from a certain 10-year period indicate that there were at least 45 earthquakes during that decade. What is the probability that there were exactly 50 earthquakes during that period?

(c) **[3 marks]** A study suggests that the mean zinc content in adult human hair (in μg of zinc per g of hair) is normally distributed with a mean of 159 $\mu g/g$ and a standard deviation of 13.1 $\mu g/g$. Suppose that we test a random human hair sample, what is the probability that the sample's zinc content is between 160 $\mu g/g$ and 165 $\mu g/g$?

(d) **[3 marks]** Suppose that the in-vitro lifespan (in days) for a certain variety of Plasmodium (a unicellar eurkaryote) is known to be gamma distributed with $\alpha = 3.4$ and $\beta = 2.8$. We observe one such Plasmodium at random. What is the probability that its lifespan is no more than 7 days?

Part 2 **Background:** *In this question, we will compare the answers given by our two binomial approximations (the Poisson approximation and the Normal approximation with and without continuity correction), to the actual value given by the binomial CDF.*

Scenario: Letter mail arriving at a sorting center is fed through a machine which reads the letter's destination address. The machine's text recognition is 99.79% accurate; that is, it correctly reads and sorts the mail 99.79% of the time. Otherwise, 0.21% of the time, the machine fails to correctly read the address (i.e. it either reads the wrong address or fails to read at all). Suppose we track 4000 letters and record X, the number of letter addresses that the machine fails to correctly read. Assume that whether letters' addresses are correctly read or not are independent of one another.

Determine the probability that the machine fails to read at least 9 addresses.

(a) **[2 marks]** Calculate the probability that the machine fails to read at least 10 addresses by using the **Poisson approximation** to the binomial distribution. Include a brief sentence (and calculation(s), if necessary) justifying why it is appropriate to use the Poisson approximation here.

(b) **[2 marks]** Calculate the probability that the machine fails to read at least 10 addresses by using the **normal approximation with continuity correction** to the binomial distribution. Include a brief sentence (and calculation(s), if necessary) justifying why it is appropriate to use the normal approximation here.

(c) **[1 mark]** Repeat the calculation in (b), but this time do NOT use the continuity correction.

(d) **[2 marks]** Finally, calculate the true probability that the machine fails to read at least 10 addresses by using R's binomal CDF. Of the three approximations in (a), (b), and (c), state which was the closest to this true value.

Part 3 Background: *Many technical fields make use of simulated or training data sets. For example, to demo a new accounting application, we might simulate the wages of several hundred fake employees. The wages, although fictitious, should still be realistic and adhere to certain restrictions like its distribution and parameter values. R allows us to simulate data from many distributions including the binomial and the Poisson. Many of the examples we have seen in lecture and past assignments have been generated in R.*

Scenario: At hydropower plants, physical barriers like metal grates or fish protective screens (FPSs) are often installed over turbine inlets to prevent fish from entering the turbines. Nonetheless, some fish inevitably pass through the preventative measures, resulting in injury or death. At a certain dam during autumn, the number of fish passing through the installed FPSs and the turbines is modelled by a Poisson distribution with an average rate of 4.3 fish per hour. We will simulate a data set to represent the number of fish passing through the FPSs at the dam during a series of 24 hour periods. The first number in the data set should be the number of fish passing through the FPSs on the first day, the second number should be the number of fish passing through the FPSs on the second day, etc.

(a) **[1 mark]** Using R, simulate one week (ie. 7 days) of data. Copy and paste both your R code, and the output (i.e. the seven days of data) into your document.

(b) **[1 mark]** In R, simulate 200 days of data, and save that data in an appropriately named vector. Copy and paste only your R code into the document (do NOT paste in the 200 outcomes).

(c) **[3 marks]** Create a histogram for the simulated 200 days of data that you created in (b). Don't forget to include an appropriate title and labels for the axes. In a brief sentence comment on the shape and spread of the data in your histogram.

(d) **[2 marks]** Use R to find the mean of your data from (b). Then, determine the (theoretical) expected number fish during a 24 hour period. In a brief sentence comment on how these two values compare.