R Project 1

February 6, 2021

Please use the statistical software R in order to answer the following questions. When doing so, please use R-notebooks to segment your code and output according to the problem. Further please leave ample comments in your R code explaining what you are doing.

You will work in groups of two (sometimes three) to complete the project. You should write and submit your own project report (the html file generated using the notebook) to Canvas. You will be assigned a grade out of 25 (the grade out of 100 will be re-scaled)

Please refer to the tentative schedule for submission deadlines.

1 Some Advice Before You Start

1. Spend some time getting used to the syntax of R. Play around with assigning values to variables, handling vectors and iterating through them etc. Attempting the questions without getting a handle on the basics will likely end in frustration.

2. Functions are your friend. If you have a computation that needs to be repeated often consider creating a function that handles the computation and calling it from inside a loop.

3. Make sure you know what version of R you have installed. It is not uncommon to spend time trying to get code to work only to realize that the code you are using is not compatible with the version at hand.

4. The official documentation can be useful. It is worthwhile to take a look before using a new command for the first time.

5. The notebooks allow you to try out little pieces of code separately. This can be used to great effect when you are trying to debug your code.

6. Sometimes it's good to take a break and revisit the code later. Starting from scratch may also prove useful when debugging isn't getting you anywhere.

2 Groups

1. Group 1: Applbaum, Ori and Chiang, Philip

2. Group 2: Choi, Joanna and Desai, Mihir

3. Group 3: Duffy, Stephen and Ferrell, Markus
4. Group 4: Frech, James and Ilangovan, Diwakaran
5. Group 5: Marquardt, John and Martinez Garcia, Elvis
6. Group 6: McPhillips, Dylan and Mustafa, Farzad
7. Group 7: O'Meara, Connor and Ogochukwu, Munachimso
8. Group 8: Roberts, Noah and Sorkin, Neil
9. Group 9: Sun, Diane and Walter, Timothy

3 Problem 1: Plotting Probability Densities (30 points)

1. What are the commands 'dbinom', 'pbinom' and 'barplot' used for?
2. Draw barplots for the following values of (n, p)
   (a) $(5, .1), (10, .1), (20, .1), (30, .1), (50, .1)$.
   (b) $(5, .6), (10, .6), (15, .6), (30, .6)$.
   (c) $(5, .9), (10, .9), (20, .9), (30, .9), (50, .9)$.

   Notice that n increases, the binomial plot looks increasing like a normal distribution.
3. For the Poisson distribution, use the commands 'dpois' and 'barplot' to draw barplots for the following values of λ $\lambda = 2, 4, 5, 10, 20, 30, 50$.
4. Do you notice anything about the Poisson distribution as λ increases?

4 Problem 2: Simulating Experiments (35 points)

Suppose a vendor stocks four types of pens, each sold for \$2, \$3.5, \$6, and \$10. Suppose X keeps track of revenue from a single sale, and we know that X has the following distribution:

x	2	4	6	10
$P(X = x)$	.3	.35	.25	.1

1. Use the formula $\mu = E(X) = \Sigma xp(x)$ to compute the expected revenue of the vendor from the sale of these four pens.
2. Compute the standard deviation of X, call it σ.
3. Assuming that every sale of a pen is independent from any other sale, and also assuming that the probability distribution for every sale is given by X, simulate an experiment (with $B = 8000$ replicates) to compute the average earnings of the vendor on a given day assuming that he sells exactly n pens, where $n = 2, 6, 12, 20, 100, 150$. Hint: The command replicate may be useful in this context.
4. For each n, suppose $\overline{X}_n$ is the sample mean when working with sample of size n. Using the simulation:

(a) compute the distribution of $\overline{X}_n$.

(b) compute the expected value of sample mean $E(\overline{X}_n)$ and the sample standard deviation $s_{\overline{X}_n}$.

(c) compute the histogram for the distribution of $\overline{X}_n$ (for the 8000 replicates).

(d) Experiment with other values for the number of replicates and sample sizes.

(e) Write a short summary of your observations from this experiment. You could consider commenting on:

   i. The effect of increasing the sample size on the sample mean and it's distribution

   ii. The effect of increasing the number of replicates on the accuracy distribution of $\overline{X}_n$ approximated by the simulation.

5 Problem 3: Random Samples from Continuous Distributions (35 points)

For each of the following distributions, do $B = 10000$ replications of random samples of size $n = 5, 10, 15, 25, 30, 50, 100$. Use these to compute the expected values of the sample means $\overline{X}_n$ and standard deviations $s_{\overline{X}_n}$ for each experiment. Further compute the expected sample means and standard deviations suggested by the Central Limit Theorem. Compare the results (for sample mean and standard deviation) from your simulations to the theoretical values postulated by the CLT (See Chapter 5.4 for a statement of the Central Limit Theorem. Also print, paste functions may be useful here).

1. Normal distribution with $\mu = 10$ and $\sigma = 2$.

2. Exponential distribution with $\lambda = 2$.

3. Gamma distibution with $\alpha = 2$ and $\beta = 1/3$.

4. Chi-Squared distribution with $\nu = 5$.