BIOS 6611 Homework 3

The Central Limit Theorem and Functions of Random Variables

A. The Central Limit Theorem

1a. Assume we are interested in the binary outcome (i.e., 0 or 1) of an experiment where the probability of "success" is 0.15. Generate and save a vector with 500 sample means (i.e., the mean of five-hundred simulated "experiments"), where each sample mean is from a sample size of 10 simulated from `rbinom()` in R (recall, the R syntax for `rbinom()` designates n=10 and size=1 to sample 10 cases with a binary outcome in one "experiment").

1b. Repeat for sample sizes of n = 20, n = 30, n = 40, and n = 50.

1c. Calculate the mean and standard deviation associated with each of the five sets of $\bar{x}$ values.

1d. Create histograms of the sampling distribution of the mean, for each sample size *n*. Provide meaningful labeling (i.e., include a title and label the relevant axes).

1e. Is there a value of the sample size *n* (i.e., 10, 20, 30, 40, or 50) where the distributions begin to look normal?

B. The CLT and the Cauchy Distribution

The Cauchy distribution is famous in part because its mean and variance are undefined. As a result, the CLT does not apply.

Repeat the simulation exercises from Problem A, using sample size values of n = 10, n = 50, n = 100, and n = 1000, substituting the random function `rcauchy()` and no other argument besides the sample size `n` parameter. (Hint: If you run `?rcauchy` to pull up the documentation on this function, you'll see that `rcauchy` has the `location = 0` and `scale = 1` arguments by default. That is, these are the defaults that will be used for our simulation, and any other simulation, if we don't specify anything.)

How do the trends you observe differ relative to Problem A?

C. Estimating Hospital Budget

After receiving one of two medical procedures (coded “1” for standard, “2” for new), patients admitted to a hospital were followed for one month. The total medical costs per patient incurred by the hospital over this month were tabulated in the column “Cost” of **ProcedureCost.csv**, in units of $1000. Note that costs may be zero if no additional medical care was provided to a patient.

Part 1

Provide R code that reproducibly creates a table in the following format. The item in each cell of the table should be the count of patients that match each category. By “reproducible”, we mean that your R code should be able to remake the table if a new “ProcedureCost.csv” file is provided in the same format (but with different data).

Hint: Refer to R Lab 1 Exercise 4: Study Design Sprints.

		Cost	
		Zero	Non-Zero
Procedure Group	1		
	2		

Part 2

For samples in each procedure group, reproducibly calculate the **proportion of non-zero costs** (p_1 and p_2), the **mean non-zero cost** (m_1 and m_2, i.e., cost among subjects that have some positive cost), and the **variance in the non-zero costs** (v_1 and v_2).

Part 3

Both the (i) estimated total patient cost as well as (ii) the frequency of non-zero costs are important for hospital planning. We can model the cost (Y) for a given patient in a given procedure group by considering Y as the product of two random variables (i.e., Y = RZ) where:

- R = a Bernoulli random variable (binomial with n=1) that takes values of 0 or 1 (for non-zero cost)
- Z = a random variable that takes values between 0 and infinity (for cost)

Assuming that Z and R are independent, **derive mathematical expressions for the expected values and variance of the cost Y for a given subject**, in terms of p = Pr(R = 1), m = E(Z), and v = Var(Z).

Hints for Part 3:

- When two random variables are independent, the expected value of the product is the product of their expectations: i.e., $E(XY) = E(X)E(Y)$. Note that this does not hold true for the variance.
- Alternatively, those with more statistical background might consider the formal definition of the expectation for a function of two variables: $E(XY) = \sum_{X \in x} \int_y xy\, f(x,y) dy$, where $f(x,y)$ is the joint distribution of X and Y.
- The variance of a random variable is $Var(X) = E(X^2) - (EX)^2$.

Part 4 – Extra Credit

The hospital expects different distributions for Y (cost) between the two procedure groups. Using your estimates for p_1, m_1, v_1 and p_2, m_2, v_2 based on the sample data (i.e., the values you calculated from Part 2), the hospital is interested in estimating how much they should budget for next year, if: (1) they anticipate treating $n_1 = 120$ and $n_2 = 200$ subjects for each respective group and (2) they want a less than 20% chance of this total expenditure exceeding their budget, i.e. for procedure groups 1 and 2 summed together?

Formulate this question in terms of random variables using correct notation. Assume that the expenditures across the two procedure groups are independent of each other. Finally, give a numerical recommendation for the budget.

Hint: Could the Central Limit Theorem apply here? Can we use the results from Part 3? What R function(s) is (are) needed to answer this: dnorm? pnorm? qnorm?

Part 5

Carry out the budget calculation in 4 using simulation. Assume that the Z value (the cost per patient) for each procedure group is a *gamma* distributed random variable with E(Z) = m and Var(Z) = v, specific to each procedure group. The random variable R is as above – Bernoulli with p = Pr(R = 1) as estimated per group in Part 2.

Write R code to simulate 10,000 sets of data to answer the budget question in Part 4, and give a numerical answer for the budget.

Hint: *n* gamma distributed observations can be simulated in R using the following code.

```
shape <- m^2/v
scale <- v/m
Z <- rgamma(n, shape = shape, scale = scale)
```

Part 6 – Extra Credit

Compare your results in Part 4 and Part 5.

a. In general, under what circumstances will these values be similar?
b. What level of accuracy do you estimate can be achieved for each of the groups using an approximation via simulation instead of deriving the theoretical properties of the random variable?