Senior Thesis
Week 1 R Assignment

Background:
Modern statistical problems require a mix of analytical, computational, and simulation techniques for implementing and understanding statistical inferential methods. We will use the statistical software package `R` as our method of implementing numerical simulation studies to investigate the performance of our statistical methods.

The term "Monte Carlo" (MC) is often applied to numerical simulation methods, essentially referring to any use of random simulation. The details of generating random numbers is, in general, beyond the scope of our project. We will make use of `R`'s built-in number generation whenever possible.

The strength of numerical simulation studies relies on convergence concepts studied in probability and statistics. In particular, the Weak Law of Large Numbers and the Central Limit Theorem. Recall that if $X_1, X_2, \ldots, X_n$ are independently and identically distributed (IID) random variables, then the following statements are true:

- $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i \xrightarrow{p} E(X)$ (WLLN)
- $\frac{\bar{X} - E(X)}{S/\sqrt{n}} \xrightarrow{d} N(0,1)$ (CLT)

The first result basically states that the sample mean will be close to the underlying population mean, and the second says that the way that the sample mean behaves from one sample to the next (the sampling distribution), once standardized, behaves like a standard normal random variable. The CLT provided here is actually an alternative version than the one typically presented in a probability or statistics course, but the result still holds. It turns out that these results also apply to pseudo-random variables (those generated by a computer). For our purposes, the above results basically state that we can estimate any quantity we want arbitrarily well by simply simulating the process enough times in a computer.

Assignment:
In this first assignment, we will simply convince ourselves of some of these properties and see how simulation can be used in various contexts. In future assignments, we will build toward conducting a full simulation study. Over the next week, complete the tasks listed below. You may use the `R` script provided as a starting point filling in the necessary details.

1. For the first problem, we will use numerical simulation to simply estimate an unknown quantity. Consider the following integral:

$$\int_0^{1.5} e^{e^x} dx$$

This integral does not have an antiderivative and must therefore be evaluated using numerical methods. One option is estimating the integral using simulation techniques by recognizing the following:

$$\int_0^{1.5} \frac{1.5e^{e^x}}{1.5} dx = E\left(1.5e^{e^X}\right)$$

where $X \sim Unif(0, 1.5)$. That is, the integral can be expressed as an expected value (or mean) of a function of X, where X is a random variable with density function

$$f_X(x) = 2/3 \qquad 0 < x < 3/2$$

Therefore, according to the WLLN, we can estimate the integral by doing the following:

- Generate a large sample of independent random variables from a $Unif(0, 1.5)$ distribution.
- Consider the sample mean of the function: $n^{-1} \sum_{i=1}^{n} 1.5e^{e^{X_i}}$.

Estimate the integral for $n = 100$, $n = 1000$, and $n = 10000$. Report your three estimates.

2. Now, we want to move toward examining an unknown distribution. Let $X_1, X_2, \ldots, X_{40} \overset{IID}{\sim} Exp(1/5)$; that is, let each X be distributed according to an exponential distribution with rate parameter equal to $1/5$ (so that $E(X_i) = 5$). Note that there are two parameterizations of the exponential distribution; this one is consistent with that given in `R`. Now, I am interested in the distribution of $\bar{X} = n^{-1} \sum_{i=1}^{40} X_i$, the sample mean. It turns out that there is theory available that says that $\bar{X} \sim Gamma(40, 5/40)$, where 40 is the "shape" parameter and $5/40$ is the "scale" parameter. However, this result is not straight-forward to those without a strong probability background. Your task is to simulate this result by performing the following steps:

- Generate a large number ($m = 1000$) samples, each of size $n = 40$ from an exponential distribution with rate parameter $1/5$.
- For each of the m samples, compute the sample mean.
- Take the m sample means and use them to construct a histogram, and overlay the theoretical density (Gamma distribution described above) on the plot to establish the result. This plot shows the sampling distribution of the sample mean.

Submit the plot.

3. One characteristic of an estimator is its bias. Let θ be a parameter of interest (like the mean), and $\hat{\theta}$ an estimator (like the sample mean). The bias is given by

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Thus, this quantity can be estimated using a simulation study by comparing the sample mean of the estimator in repeated samples with the true parameter.

Suppose $X_1, X_2, \ldots, X_n \overset{IID}{\sim} Exp(\lambda)$. Now, we know that $E(X_i) = 1/\lambda$ and further $Var(X_i) = 1/\lambda^2$. Typically, we estimate the variance of a population using the sample variance. But, based on the above fact that $Var(X) = [E(X)]^2$ for the exponential distribution suggests that we could estimate the variance by taking the square of the mean. Specifically, let parameter of interest θ be the population variance, and consider estimating it by $\hat{\theta} = \bar{X}^2$. Determine the bias in this estimator when $n = 30$ using a simulation with $m = 10000$ replications.

Submit the estimate of the bias (which should be in the ball-park of 0.83).