DEPARTMENT of COMPUTER SCIENCE AND ELECTRICAL ENGINEERING
University of Maryland, Baltimore County
ITE325 Suite
1000 Hilltop Circle, Baltimore, MD 21250

chuck.laberge@umbc.edu // p: 410.455.5684
csee.umbc.edu

MEMO Number CMPE320-S21-103
DATE: 27 March 2021
TO: CMPE320 Students
FROM: EFC LaBerge
SUBJECT: The Central Limit Theorem: The Magic of Large Numbers

1 INTRODUCTION

One of the most important theorems in the study of probability and random processes is the Central Limit Theorem (CLT). Let us define the *sample sum* of N *independent* random variables, $Y = \sum_{k=1}^{N} X_k$,where each random variable, X_k, has a finite mean, m_k, and a finite variance, σ_k^2. The Central Limit Theorem states that in this case, Y has a probability density function (pdf) that approaches a Gaussian pdf with mean $m = \sum_{k=1}^{N} m_k$ and a variance $\sigma^2 = \sum_{k=1}^{N} \sigma_k^2$ as N grows large (but not infinite!).

Put it this way: if we add a bunch of well-behaved independent random variables together, the pdf of the sum is approximately Gaussian, and the approximation gets better as we add more terms (i.e., more independent random variables) to the sum. The approach to Gaussian in the limit as N gets large does not depend on the shape of the pdfs, or even on the X_k random variables being identically distributed. All that is required is that they are independent!

We will study the CLT during the course of this project, but you *do not* need to wait for us to cover it in class. You can do all of the project without covering the CLT.

Warning #1: Thinking is required!

Warning #2: Follow directions!

Warning #3: It isn't on the web, so don't bother looking.

2 PROJECT TASKS

Perform the following tasks, then document your results and submit them in written form in accordance with the instructions in Section 3, below. You may use this document as a format.

2.1 Sum of Independent, Identically Distributed (iid) Random Variables from *U*(0,1)

Generate the sum of N random variables distributed $X_k \sim U(0,1), k = 1,2...,N$ for $N = 2,\ N = 6,\ N = 12$. MATLAB users will use the function `rand`.

Generate a large number of such sums, say 100,000 or more, for each value of N . Plot a histogram of the results for each N , scaling the histogram appropriately to be a probability density function. In each case, compute the mean and standard deviation of the samples and compare it to the theoretical expected value and variance for an infinitely large ensemble of such sums.

On the same plot as the histogram, plot an appropriately scaled Gaussian curve with the theoretical mean and variance.

Discuss what you did and what you observed and why it does or does not make sense.

One plot required for each value of *N* for a total of three plots.

2.2 Sum of Independent, Identically Distributed (iid) Discrete Random Variables

Repeat all the sections of 2.1 where the random variables are generated using `randi` simulate the rolling of the N fair eight-sided dice, followed by the sum of the values from each roll. Repeat this experiment a large number of times to create the histogram of the sum (I'm *not* interested in the values of the individual rolls!). Repeat the whole process for $N = 2, N = 10, N = 50$.

For each value of *N*, compute the mean and standard deviation of the samples and compare it to the theoretical expected value and variance for an infinitely large ensemble of such sums.

On the same plot as the histogram, plot an appropriately scaled Gaussian curve with the theoretical mean and variance.

Discuss what you did and what you observed and why it does or does not make sense.

One plot required for each value of *N* for a total of two plots

2.3 Sum of Independent, Identically Distributed (iid) Random Variables from $p_X(x) = 0.5e^{-0.5x}$

Repeat all the sections of 2.1 where the random variables are generated using the function `randx` provided with Project 1. Use $N = 2, N = 10, N = 100$. Note that this pdf has a sharp discontinuity at $x = 0$, but that eventually the histogram *does* approach the Gaussian! The CLT is a *powerful* theorem!

In each case, compute the mean and standard deviation of the samples and compare it to the theoretical expected value and variance for an infinitely large ensemble of such sums.

On the same plot as the histogram, plot an appropriately scaled Gaussian curve with the theoretical mean and variance.

Discuss what you did and what you observed and why it does or does not make sense.

One plot required for each value of *N* for a total of three plots.

2.4 Sum of Independent, Identically Distributed (iid) Bernoulli Trials

Let a single Bernoulli trial result in either a one (1) or a zero (0), with $\Pr[X = 1] = 0.5$. Perform N independent trials. What is the form of the pmf of the random variable K = number of 1's in N trials ? (I'm looking for a specific name here, go review the standard pmfs.) Because the random variable K is a sum of independent random variables, each of which has finite mean $m = 0.5$ and finite variance, $\sigma^2 = 0.5$, the CLT should hold. Use values of $N = 4, N = 8,$ and N as large as you can without causing a MATLAB overflow. Note that N is the number of Bernoulli/Binary random variables in the *sum*. The sum itself produces *one* value of the random variable K. You need to do this process many times to generate your histogram.

Each figure should consist of two subplots. On the first, plot the (theoretical) probability density function of the sum of N independent Bernoulli trials. Plot the CLT Gaussian approximation on the same plot and compare the results.

On the second, generate and plot the scaled histogram for the sum of a large number of random trials of the sum of N iid Bernoulli experiments and compare it to the theoretical. Plot the CLT Gaussian approximation on the same histogram plot and compare the results.

There are three figures, one for each value of N and each figure has two subplots.

3 INSTRUCTIONS FOR PROJECT REPORT

3.1 Report Format

The project report shall be in the same form as this document, with an introduction, simulation and discussion section, and a "what I learned" section. Each section shall contain the content identified in Section 2 and the appropriate Section 3 subsection below. The report shall be in Times New Roman 11 point font. MATLAB pictures shall be pasted in-line in the report (this is a useful skill to know!); shall be numbered consecutively; shall be appropriately titled; the axes shall be appropriately labeled; the curves shall be appropriately identified by an appropriate legend.

3.2 Section 2 Content

Section 2 of the report shall be titled "Simulation and Discussion" and shall contain the simulation plots and a discussion of each plot. The discussion shall address the points identified in Section 2, and any other interesting observations that occur to you. Remember, I know this stuff: you don't. So take a look at the plots and tell me what you see and what it means to you.

3.3 Section 3 Content

Section 3 of the report shall be titled "What I learned" and shall contain a summary of what information you observed, what insights you gained, etc. Section 3 shall also contain a subsection critiquing the project and suggesting improvements that I could institute for next spring. Finally, Section 3 shall contain an estimate of how much time you spent on the project, including reading, research, programming, writing, and final preparation.

3.4 Questions

I will accept questions regarding the project through the Ask the Professor discussion forum through 6 PM on Tuesday April 13, 2021. Please plan to check the Ask the Professor discussion forum frequently to learn of clarifications and hints (if I give any!). In my opinion, Project 3 is actually *easier* than Project 2, so I'm less likely to give direct assistance.

3.5 Project Grading

The project will be graded in the following way:

75% of the project score shall depend on the technical, theoretical, and graphical presentations of the tasks set out in Section 2 of this document.

25% of the project score shall be based on an evaluation of the technical writing against the Rubric on Technical Writing, posted on Blackboard, including grammar, clarity, organization, etc. For the purpose of this document, you can assume that the intended audience consists of your CMPE320 classmates.

3.6 Project Delivery

The project shall be delivered by 11:59 PM on Wednesday, April 14, 2021. I am *not* inclined to give any extensions because of the adverse effect on Project 4 and Project 5.

Delivery shall be by submission of a PDF file as a Blackboard assignment. This is an individual assignment

3.7 Academic Integrity

The academic integrity provisions you signed at the beginning of class are in effect. You may discuss the interpretation of the assignment and approaches to solve the various problems amongst yourselves. You MAY NOT share MATLAB code, plots, text, etc. Do your own work.

I *will* be looking at source files for similarities, so please do not even attempt to copy work.

You ***may not*** ask for assistance on the project from the TA/grader staff, although you may ask for help with the various concepts. So, for example, if you don't understand a Gaussian pdf you may ask for help understanding pdfs in general and Gaussian pdfs in particular. You may not ask for help completing the tasks assigned in this document. You *may* ask me for assistance via the Ask The Professor forum. You may ask for help on the concepts during my open office hours.