

Lab IV — Central Limit Theorem*

Peter Brinkmann

The purpose of this lab is to introduce you to the Central Limit Theorem in a hands-on fashion. I am assuming that you have worked through Lab I, available at

<http://www.math.uiuc.edu/iprom/materials.html>.

In particular, you should be able to create, modify, and run simulations. When working on this lab, please work slowly and do one step at a time. If you skip ahead or read on before finishing the task at hand, it may spoil the fun! You do not have to submit any of your work.

1 Sums of independent random variables

Let X_1, X_2, \dots, X_n be independent, identically distributed random variables, each having mean μ and variance σ^2 , and let

$$X = \sum_{i=1}^n X_i.$$

Action 1.

1. Find a mathematical expression for $E(X)$.

$$E(X) =$$

*Copyright © 2003, Peter Brinkmann (brinkman@math.uiuc.edu). Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is available at <http://www.fsf.org/copyleft/fdl.html>.

2. Find a mathematical expression for $Var(X)$. Hint: You may want to use Proposition 3.2 from Section 7.3 as well as the fact that the covariance between two independent random variables is 0.

$$Var(X) =$$

Action 2. Build an IPROM model that simulates the random variable X and displays the results in a histogram. Recall that you need to start with the model `blank.mdl` and save it under a new name. You may want to follow the following steps.

1. Pick a type of random variable from the IPROM library and copy it into your model. Exponential random variables work particularly well for the purposes of this lab, but feel free to pick your favorite random variable here.¹

If you double-click on the block representing your random variable, you can change the number of outputs, i.e., the number of random variables X_i that are being simulated. Pick some small number, say, 3. This number is the parameter n .

2. Copy a histogram block into your model and connect it to your random variable. Leave plenty of room between them because you'll be adding more blocks later on.

If you run the simulation now, you'll see a histogram that shows you the distribution of each of your independent random variables, i.e., at each number, you'll see one vertical bar for each random variable X_i . You may want to change the stop time of your simulation to a fairly large value, say 1000.

3. Now, drop a MATLAB Expression block on the line connecting the random variable and the histogram. Make sure it connects properly to the two other blocks. Double-click on the new block, change the expression to `sum(u)`, and change the number of outputs to 1.

Run the simulation a few times. If you did everything correctly, you should now see a histogram with only one vertical bar at each number. The resulting histogram gives you an idea of the distribution of the random variable $X = \sum_{i=1}^n X_i$.

¹Make sure to pick a random variable whose expectation and variance you can compute easily. You'll need this later.

We now study the distribution of X for larger and larger values of n .

Action 3. Double-click on your random variable and change the number of outputs to, say 4. Run the simulation and look at the histogram. Repeat this for $n = 5, 6, \dots$. Does the appearance of the histogram remind you of anything?

By the time you get to $n = 10$, you may have a hunch what's going on, and by the time you get to $n = 40$, this hunch should be pretty firm. Of course, I don't expect you to repeat this experiment 40-odd times, but you should run a fair number of experiments in order to get a feeling for how the distribution changes from one step to the next.

For $n = 40$, what does the histogram look like?

Answer: This histogram looks like _____.

Action 4.

1. Repeat the previous experiments with different parameters. For instance, if your random variable is exponential, you may choose $\lambda = 2$ instead of the default value $\lambda = 1$. Once again, look at the histograms for $n = 4, 5, 6, \dots$, and see what happens.
2. Repeat the previous experiments with a different random variable. Maybe you want to try a uniform random variable this time, or maybe a Gamma random variable? As before, play with the parameters as well as the number n of independent instances of your random variable and see what happens.

2 Comparing histograms

Chances are that you have observed that the distribution of X looks more and more like a bell curve as n gets larger. Let's see whether we can corroborate this observation. In order to do this, let's modify the IPROM simulation such that it allows for direct comparison of our random variable X and a normal random variable.

Action 5.

1. Delete the line connecting your MATLAB Expression and the histogram block. Copy a Multiplex block into your simulation. (A multiplex block looks like a solid vertical bar with a couple of arrowheads going in and one arrowhead going out.)

Connect the output of your Expression block to one of the inputs of the multiplex block. Connect the output of your multiplex block to the input of the histogram block.

Now, copy a normal random variable block into your model and connect its output to the free input of the multiplex block. The purpose of the multiplex block is to merge the output from two sources so that both can be displayed simultaneously.

If you run your simulation again, you'll see vertical bars in two colors, one for our random variable X , the other for the new normal random variable.²

Chances are that at this point, the histogram for X is not even close to the histogram for the normal random variable. That should not come as a surprise since the normal random variable comes with a default mean of 0 and a default variance of 1. Our next goal is to choose the parameters of the normal random variable such that they match the expectation and variance of X .

Action 6.

1. Find the expectation $E(X)$ (remember Action 1), and set the mean of your normal random variable to $E(X)$. Note that you need to evaluate $E(X)$ *numerically*, i.e., you need to enter $E(X)$ as a number, not as a formula.

If you run the simulation again, the peaks of the two histograms should agree.

2. Find the variance $Var(X)$ (Action 1 again), evaluate it numerically, and set the variance of your normal random variable accordingly.

Run the simulation again. How do the two superimposed histograms compare now?

3. Repeat this experiment using different random variables in the definition of X , i.e., try various blocks and parameters, as well as various values of n , compute $E(X)$ and $Var(X)$, adjust the parameters of the normal random variable accordingly, and see how the histograms compare.

²You may only see one vertical bar for the normal random variable. That's okay; we'll fix this in the next step.

3 Conclusion

If you've carefully followed the instructions up to this point, you've seen that the random variable X looks more and more like a normal random variable as n gets larger and larger. More precisely, $X_1 + \cdots + X_n$ minus its mean and divided by its standard deviation looks more and more like a *standard* normal random variable as n gets larger and larger. This is true regardless of the distribution of the random variables X_i . In other words, you have discovered experimental evidence of the following theorem.

Theorem 3.1 (Central Limit Theorem). *Let X_1, X_2, \dots be a sequence of independent, identically distributed random variables, each having mean μ and variance σ^2 . Then*

$$P \left\{ a \leq \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq b \right\} \rightarrow \Phi(b) - \Phi(a)$$

as $n \rightarrow \infty$.

The Central Limit Theorem is rather a remarkable result in probability theory. It explains why normally distributed random variables are abundant in the real world — regardless of what kind of random variable you look at, if you add enough copies of it, you'll get something that looks normal! It should also seem somewhat familiar because you've already seen the following special case of this theorem.

Theorem 3.2 (DeMoivre-Laplace Limit Theorem). *If S_n denotes the number of successes that occur when n independent Bernoulli trials (each resulting in a success with probability p) are performed, then for any $a < b$,*

$$P \left\{ a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b \right\} \rightarrow \Phi(b) - \Phi(a)$$

as $n \rightarrow \infty$.

In order to see why DeMoivre-Laplace is a special case of the Central Limit Theorem, just remember that if X_i is a Bernoulli random variable with parameter p , then $E(X_i) = p$ and $Var(X_i) = p(1-p)$.