STAT 471: Introduction to R Programming Lecture

Lecture 10: Simulations in R for Normal, Poisson, Central Limit Theorem, and Law of Large Numbers

Cory Suzuki

Department of Mathematics & Statistics
California State University, Long Beach

20 October 2025

Overview

1. Background of Fundamental Mathematical Statistics Theorems

2. Simulations in R

Central Limit Theorem

Recall from STAT 381 that the CLT in essence, is saying that the normalized sample mean of an arbitrary distribution (the random variables have to be sampled from such a distribution) will converge in distribution to a Standard Normal Distribution. From elementary statistics, the watered down version of this is "The distribution of sample means will be approximately normally distributed regardless of the distribution of the population provided that the sample is large enough". The formal theorem is given below.

Central Limit Theorem (CLT)

If $X_1, ..., X_n$ is a random sample from a distribution with mean μ and finite variance $\sigma^2 > 0$, then $Z_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$

Proof. Skip, just know what the theorem is saying here. □

Law of Large Numbers

There are two versions of the Law of Large Numbers: the Strong and Weak Laws. However we'll only consider the Weak Law of Large Numbers. This states that as the sample size gets asymptotically large, the sample mean will converge in probability to the theoretical population mean.

Weak Law of Large Numbers

If $X_1, ..., X_n$ is a random sample from a distribution with finite mean $E[X_i] = \mu$, then $\bar{X} \xrightarrow{P} \mu$. This means that for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0$$

Proof. Skip, just know what the theorem is saying here. □

Stochastic Processes

Stochastic processes can be useful to represent random variables and their evolution over time or space.

Definition: Stochastic Process

A stochastic process $\{X(t) : t \in T\}$ is a family of random variables where t denotes the time step and T is a nonempty set. A stochastic process is discrete if T is finite, otherwise if T is infinite, then it is continuous.

Some examples we'll explore are Random Walks, Martingales (good for financial applications, and Autoregressive (AR) processes (we'll see more of this when we explore time series.
If you are interested in the theoretical aspects of stochastic processes, the book by Sheldon Ross is an excellent reference. For now we'll just see how these are coded in R.

Why Simulations?

- Simulations help us visualize and analyze data from distributions
- Simulations help us validate statistical models and are the backbone to resampling methods and model validation (Cross Validation)
- Simulations help us to generate synthetic data to represent unknown distributions empirically (this means by rough estimation)
- Simulations provide insights on randomness and variability!

The Bare Bones Basics

For creating simulations of random variables, we can use built-in functions in R (which we have done before) or develop our own algorithms.

- If the built-in function starts with "r", it generates random sample of random variables from that distribution (rbinom() for example)
- If it starts with "d", it calculates particular values from it's probability density/mass function (dnorm())
- "p" will return the cumulative density function/area of the curve to the left of a particular input value (pnorm())
- "q" will return the quantile (qbinom())
- For some distributions like the χ^2 distribution, we can use transformation methods to generate simulated data as well.

Simulations for Distributions: Normal and Poisson

To simulate distributions, we first randomly sample from the distribution in R. However, here we want to do this for different parameter values to observe any changes in the density of the distribution. usually when we perform simulations, we like to plot both:

- the density histogram
- the superimposed density curve

Simulations of CLT and LLN

- For simulating the CLT, we can use our simulations and increase the size of our sample (in this case the number of random variables we choose to generate)
- For simulating the LLN, we can check how close our sample mean is close to the true mean as we increase the number of simulations we carry out.

Simulating Mixture Distributions: Using What We Know to Model What We Don't Know

- Mixture distributions are self-explanatory in the name, distributions that are mixtures of already existing statistical distributions. We actually already know of one, the Chi-Squared distribution.
- To generate discrete and continuous mixture distributions respectively, we can sample from a uniform distribution and transform it with an inverse transform of the CDF. If you would like more information on the underlying theory of this technique, consult *Statistical Computing with R* by Maria Rizzo
- Let's pretend that we completely forgot how to generate Chi-squared simulated data and we don't have access to the built-in function in R. We can use the fact that the Chi-Squared distribution is the squared sum of squares of independent random variables that are standard normal. Precisely, if $Z_1, ..., Z_n \sim N(0, 1)$ and are i.i.d., then we know that $X \sim \chi^2(n)$ with n degrees of freedom where $X = Z_1 + ... + Z_n$. We'll see some examples of this type of technique in R.

Monte Carlo Simulations

To preface some of the Monte Carlo-based techniques we're about to see in the next coming weeks, we need to know about what Monte Carlo is.

- Monte Carlo simulations can estimate probabilities based on the idea of repeatedly sampling enough times (based on LLN) despite our beliefs of uncertainty
- Often used to model uncertainty in pricing options, futures, and portfolios (financial analysts/quants)
- Used for statistical and scientific computing for estimating the values of integrals
- Approximates unknown distributions and can be used to either supplement or compare against traditional hypothesis tests

Next Time...

Linear and Logistic Regression Analysis

Announcements

Midterm Part B is due this Saturday at 11:59pm!