Project Report

Project – Sampling methods

Joanikij Chulev

University College Roosevelt

SCIMATH202 - Theory of Statistics and Data Analysis

Prof. Dr. Ir. Richard van den Doel

April, 2023

Abstract

In this report, we investigate various methods for obtaining samples from discrete and continuous probability distributions. For discrete distributions, we explore the Binomial, Geometric, Negative Binomial and Poisson distributions and provide examples of how to obtain samples from each. Additionally, we investigate methods for obtaining samples from other discrete distributions through research such as Bernoulli. For continuous distributions, we use the various sampling methods to obtain samples from various distributions, including the Uniform, Normal, Gamma, and Beta distributions. Furthermore, we investigate alternative sampling methods. Through research and experimentation, we conclude that these methods can be used to obtain samples from almost all probability density functions.

Introduction

Sampling from probability distributions is a fundamental tool in many areas of science and engineering. Obtaining samples from probability distributions is useful for simulating real-world scenarios, performing Monte Carlo simulations, and testing statistical hypotheses. In this report, we investigate various methods for obtaining samples from both discrete and continuous probability distributions.

Discrete probability distributions, such as the Binomial, Geometric, Negative Binomial, and Poisson distributions, have important applications in fields such as biology, physics, and finance. For instance, the Poisson distribution may be used to simulate the accident frequency in a certain location, while the Geometric distribution can simulate the distribution of claim amounts. The Binomial distribution can also be used to model the proportion of defective products in a sample. In this report, we provide examples of how to obtain samples from each distribution using the methods of the probability theory.

For continuous probability distributions, we use the inverse sampling method to obtain samples from various distributions, including the Normal, Gamma, and Beta distributions. The inverse sampling method has a long history in scientific computing, and it is widely used in scientific research. For example, the Normal distribution can be used to model the distribution of weights of a product or problem, whether that weight is real or a logical construct Additionally, we investigate the accept-reject or Monte-Carlo sampling method, Central Limit Theorem and the Box-Muller transform as alternative sampling methods. These methods have been used in many fields, including physics, computer science, and engineering, and we provide examples of how they can be used to obtain samples from the Normal Distribution.

Mathematica, a potent computational software program that is frequently used in engineering and scientific research, was used for the study and testing that went into this report. The investigation of the methods described above is best done using Mathematica, which has a number of built-in functions for producing random numbers from various probability distributions. Furthermore, Mathematica offers robust visualization tools that let us investigate the characteristics of probability distributions and the outcomes of our simulations. We can get precise and trustworthy results from our research and tests by using Mathematica, as well as explore a variety of probability distributions and sampling techniques.

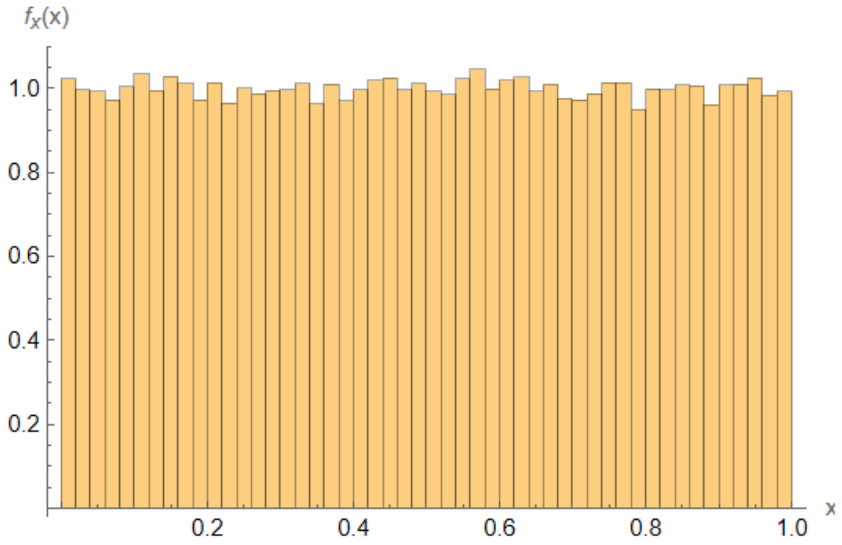
Through research and experimentation, we conclude that the methods presented in this report can be used to obtain samples from almost all probability density functions. The methods presented in this report can also be extended and applied to other probability distributions and fields of research, providing a useful framework for scientific inquiry.

Discrete probability distributions

Uniform distribution (Discrete and Continuous)

The discrete/continuous uniform distribution is a probability distribution that assigns equal probability to all values in a given range. It is commonly used as a baseline distribution for testing and comparing other distributions and sampling methods. We also followed this logic and it is used the most for transforming samples into other distributions in these experiments due to its simple nature. The probability density function of the uniform distribution is quite simple having only one or two parameters. N for discrete, a and b for continuous. where a and b are the lower and upper bounds of the range, respectively.

The RandomReal function in Mathematica is used to generate the sample, and it takes two arguments: the range of values for the sample and the number of samples to generate. Once the sample is generated, various statistical properties of the sample can be analyzed, such as the mean, variance, and higher moments.



Geometric distribution-transforming the uniform sample to geometric

The geometric distribution is a discrete probability distribution that models the number of independent Bernoulli trials needed to achieve the first success. The probability mass function of the geometric distribution is:

$$f(x) = p(1-p)^{x-1} for\ x = 1,2,3,\ldots$$

where p is the probability of success in each trial.

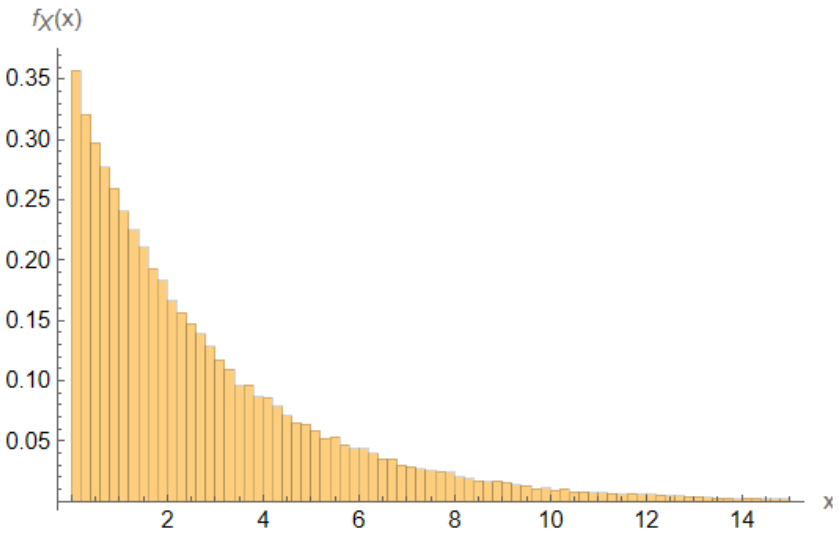
To generate a sample from the geometric distribution, we can use the inverse transform method, which involves using the inverse of the cumulative distribution function (CDF) to transform a sample of uniform random numbers on the interval [0,1] to a sample of random numbers from the desired distribution. In the case of the geometric distribution, the CDF is:

$$f(x) = 1 - (1-p)^x$$

and its inverse can be expressed as:

$$f^{-1}(y) = \left[\frac{Log(1-y)}{Log(1-p)}\right]$$

where y is a random number generated from the uniform distribution on [0,1].

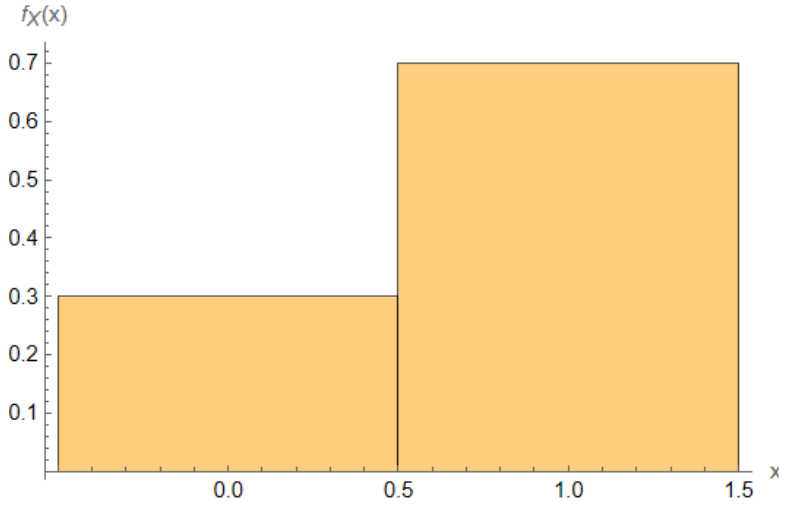


Bernoulli distribution-transforming the uniform sample to Bernoulli

The Bernoulli distribution is a discrete probability distribution that represents a random variable with two possible outcomes: success (with probability p) and failure (with probability 1-p). Through trying various test, we came to the conclusion that the best approach is for a sample to be generated from the uniform distribution, and then converted to a sample of the Bernoulli distribution by assigning a value of 1 if the generated value is less than or equal to the probability of success p, and 0 otherwise. The mathematical formula for the Bernoulli distribution is:

$$f(x) = p^x(1-p)^{1-x}, for\ x = 0\ or\ 1$$

where x is the outcome (either 0 or 1), and p is the probability of success.



Binomial distribution-transforming the Bernoulli sample to binomial

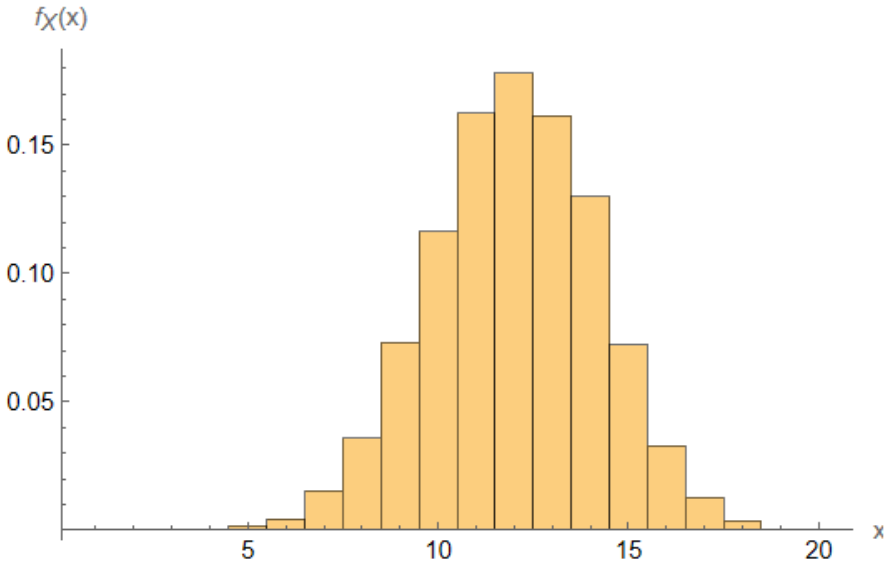
We generated a sample from the binomial distribution by using a previously generated sample from the Bernoulli distribution. The binomial distribution is the probability distribution of the number of successes in a fixed number of independent Bernoulli trials. The parameter n denotes the number of trials and p denotes the probability of success for each trial.

The solution we thought of is to divide the Bernoulli sample into blocks using:

$$\{n * (i - 1) + 1, n * i\}$$

which gives us the indices for the i-th block of n trials in the sample x.

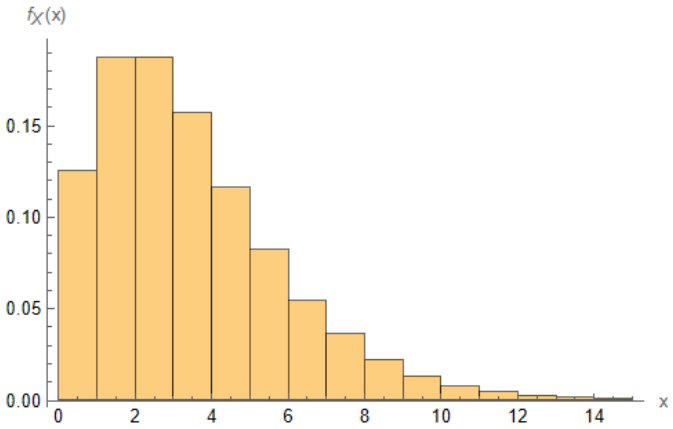
The method we used creates a sequence of subsets of n elements from the Bernoulli sample, where each subset corresponds to a single binomial trial. The number of successes in each trial is obtained by taking the sum of the elements in each subset. These sums are then collected into a list, which represents a sample from the binomial distribution.



Negative binomial distribution-transforming the Bernoulli sample to negative binomial

We generated a set of samples from the negative binomial distribution. The negative binomial distribution is a discrete probability distribution that describes the number of failures before a fixed number of successes occurs in a sequence of Bernoulli trials. In this method we used, the probability of success in each Bernoulli trial is set to p, and the fixed number of successes is set to r. For each sample, a sequence of Bernoulli trials is generated and counts the number of failures until r successes are reached. The sequence of trials is generated by repeatedly generating a random number between 0 and 1 and comparing it to the probability of success p. If the generated number is less than or equal to p, it is considered a success, thus we move on to the next trial. If the generated number is greater than p, it is considered a failure, and we increment the number of failures by 1. Once the sequence of trials reaches r successes, the number of failures up to that point is recorded as a sample from the negative binomial distribution. The loop continues until our desired

number of samples have been generated. This method as the previous methods for the Bernoulli and Binomial distributions relies on logic.

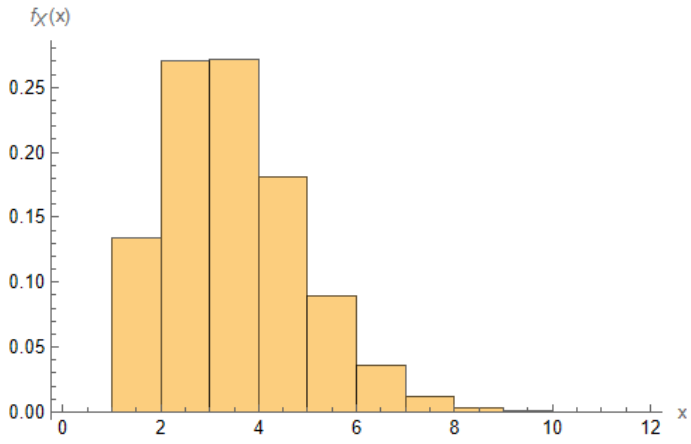


Poisson distribution-transforming the exponential sample to Poisson

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, given that the events occur independently and at a constant rate. It is often used to model the number of occurrences of events in a fixed period of time, such as the number of phone calls received per hour or the number of customers arriving at a store per day. The Poisson distribution has a single parameter, λ, which represents the average number of events that occur in the fixed interval.

To convert an exponential sample to a Poisson sample, we need to count the number of events that occur in a fixed interval of time or space, given that the events occur independently and at a constant rate. We can do this by summing the entries in the exponential sample until the sum reaches 1, and recording the number of entries it took to reach 1 as a Poisson sample (We expect the sum of exponential random variables with rate parameter λ to equal 1 on average after $\frac{1}{\lambda}$ random variables). This process is repeated until our desired number of samples is generated.

The resulting list sample contains the number of entries it took for the sum to reach a value of 1, which follows a Poisson distribution.



Continuous probability distributions

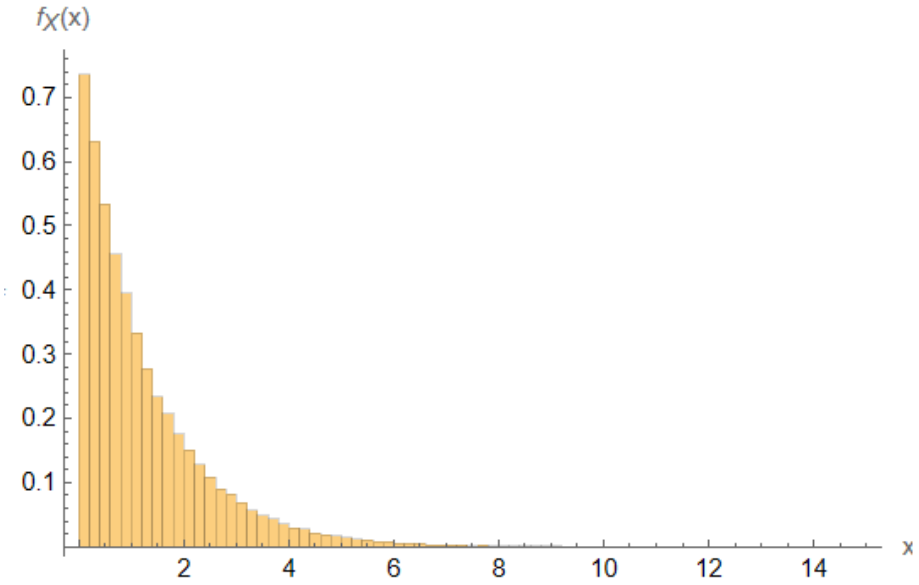
Exponential distribution-transforming the uniform sample to exponential

The exponential distribution is a continuous probability distribution that describes the time between events in a Poisson process, where events occur independently and at a constant average rate. It is widely used in various fields, such as physics, engineering, and finance, to model processes that involve waiting times or lifetimes of systems. The inverse transform method is a popular way to generate random samples from the exponential distribution because it is simple and efficient, and it relies on only basic mathematical functions.

We also utilized this method. Firstly, we generated a sample from the uniform distribution. Then, wrote a solution to transform the uniform sample into a sample of the exponential distribution using the inverse transform method. Specifically, it applies the formula for the inverse CDF of the exponential distribution, which is given by:

$$f^{-1}(y) = -\frac{Log(1 - y)}{\lambda}$$

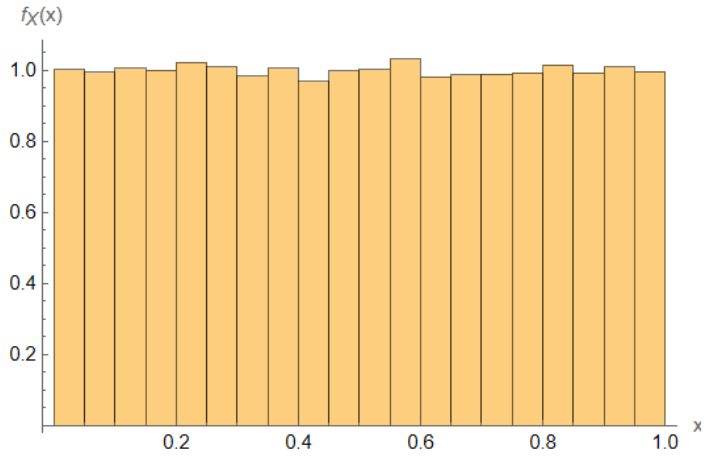
where y is a value between 0 and 1 sampled from the uniform distribution, and λ is the rate parameter of the exponential distribution.



Uniform distribution-transforming the exponential sample to uniform

Using the transformation method, I was able to change a sample from an exponential distribution into one from a uniform distribution. When given random numbers from one distribution, the transformation method is a popular method for producing random numbers from the other distribution. Specifically, if X is a random variable with cumulative distribution function f (x), and Y is a uniform random variable on the interval [0, 1], then the inverse of the cumulative distribution function $f^{-1}(Y)$ has the same distribution as X. The cumulative distribution function of the exponential distribution is given by:

$$f(x) = 1 - e^{-\lambda x}$$



Beta distribution-transforming the uniform samples to beta

The beta distribution is a continuous probability distribution with two shape parameters, denoted by α and β. The probability density function (PDF) of the beta distribution is given by:

$$f(x;\ \alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

We used a method that generates a sample from the beta distribution using the acceptance-rejection method. The acceptance-rejection method is a general algorithm for generating random variables from any probability distribution, provided that a candidate distribution is available from which it is easy to generate random variables. In this case, the candidate distribution is the uniform distribution over [0, 1], and the target distribution is the beta distribution with parameters α and β. The Metropolis-Hastings algorithm generates a Markov chain whose stationary distribution is the target distribution, by accepting or rejecting candidate samples based on an acceptance probability.

The algorithm works as follows: (in this case)

-Generate a random value x from the candidate distribution (i.e., the uniform distribution over [0, 1]).

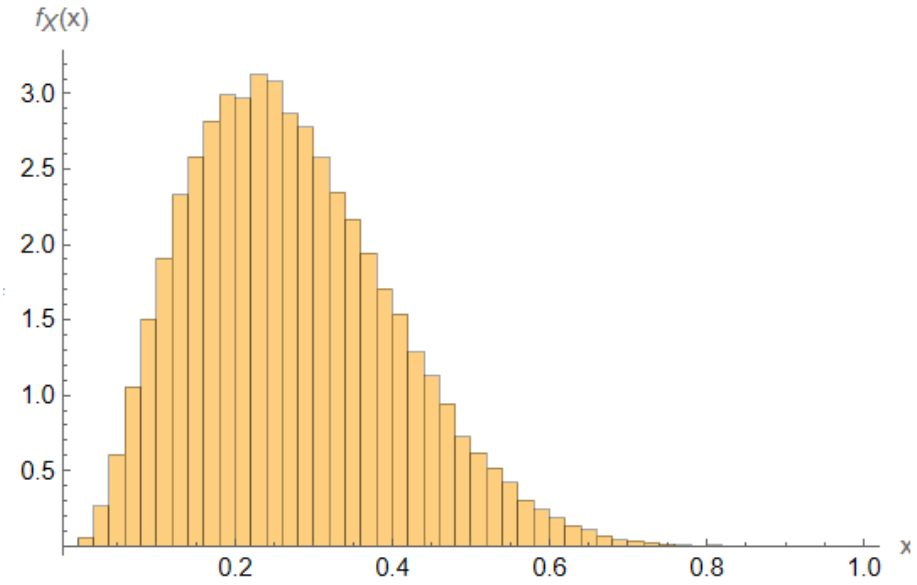
-Transform x to a sample from the candidate distribution using an inverse transformation method. In this case, we use the transformation $x\ =\ u1^{\frac{1}{\alpha}}$, where u1 is the random value generated in step 1.

-Compute the acceptance probability $y\ = \frac{f(x;\alpha,\beta)}{g(x),}$, where $g(x)$ is the PDF of the candidate distribution. In this case, $g(x)\ =\ 1$.

-Generate a second random value u2 from the candidate distribution.

-If $u2\ <\ y$, accept the sample x; otherwise, reject it and go back to step 1.

-Repeat steps 1-5 until the desired number of samples is obtained. (In our case 100000).



Gamma distribution-transforming the beta sample to gamma

The transformation is based on the fact that if X is a beta-distributed random variable with parameters α and β, then $Y = \frac{X}{1-X}$ follows a distribution known as the "beta-prime" distribution with parameters α and β. The beta-prime distribution is also known as the "inverted beta" or "beta distribution of the second kind". Note that if we let $Y = \frac{X}{1-X}$, then we can solve for X in terms of Y:

$$X = \frac{Y}{1 + Y}$$

Where X is a sample variable from the beta distribution.

Now we can sample using the following steps:

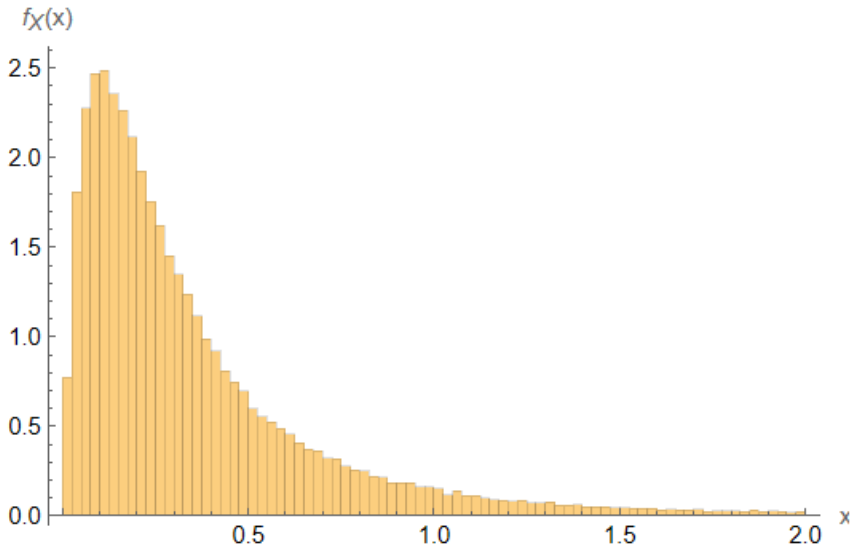
-Generate a random sample x from a beta distribution with parameters α and β.

-Calculate $Y = \frac{X}{1-X}$. get the corresponding values from the beta prime distribution with the same shape parameters.

-Calculate k = α.

-Calculate $\theta = \frac{1}{\alpha + \beta}$. The parameter $\theta = \frac{1}{\alpha + \beta}$ comes from the definition of the Beta distribution.

-The transformed values are then scaled and shifted to obtain the corresponding values from the gamma distribution with shape parameter alpha and scale parameter.

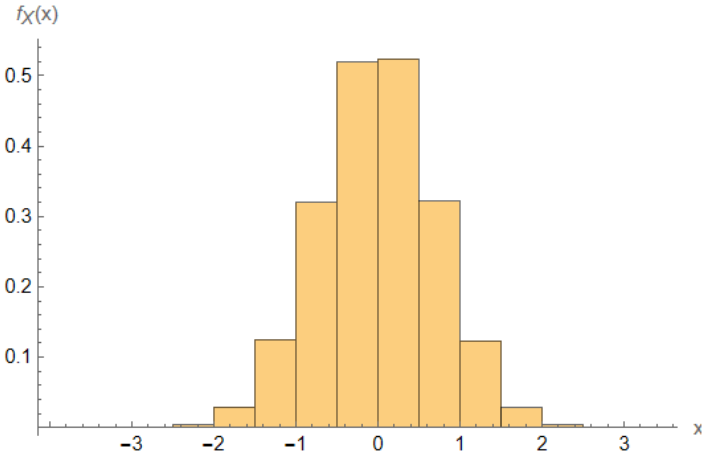


Normal distribution-transforming the uniform sample to normal using the Accept-Reject Method

The Accept-Reject method is a Monte Carlo technique that generates random samples from a given probability distribution by using a proposal distribution that is easy to sample from and has a density that bounds the target distribution. The basic idea behind this method is to repeatedly generate samples from the proposal distribution and accept or reject each sample based on a comparison of the proposal density and the target density.

In this specific implementation, we want to generate random samples from a standard normal distribution (mean 0, variance 1). We use a uniform distribution as the proposal distribution because it is easy to sample from and its density is constant over the interval [0,1].

The reason why this method works is because the uniform distribution is a bounding distribution for the normal distribution, meaning that the ratio of their densities is always less than or equal to 1. This means that when the ratio is greater than or equal to 1, the sample x can always be accepted. When the ratio is less than 1, the probability of accepting a sample x is proportional to the ratio.



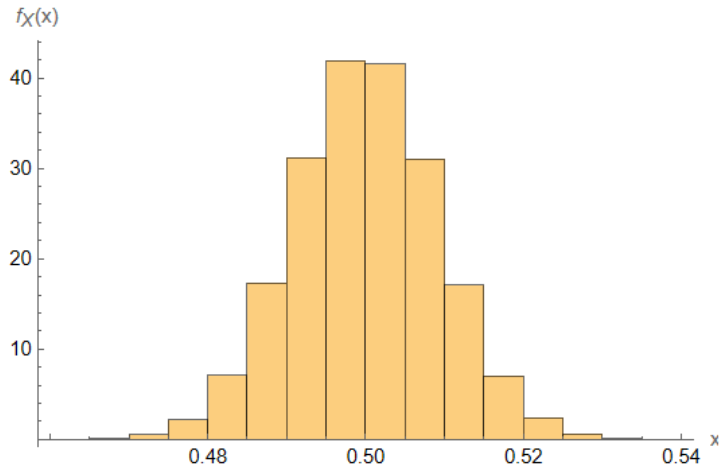
Normal distribution-transforming the uniform sample to normal using Central Limit Theorem

The Central Limit Theorem states that if we take a large number of independent and identically distributed (i.i.d.) random variables and calculate the average of these variables, then the distribution of the averages will be approximately normal, regardless of the underlying distribution of the individual variables.

In our case random variables are generated from a uniform distribution between 0 and 1. Next, we calculate the standard deviation of the sample means, which is equal to the standard deviation of the original uniform distribution divided by the square root of the sample size n. This gives us an estimate of how much the sample means vary around the true mean of the distribution.

Then, we calculate the standard error of the sample means, which is equal to the standard deviation of the sample means divided by the square root of the sample size n. This gives us an estimate of how much the sample means vary from sample to sample.

Finally, we create a normal distribution with mean equal to the average of the sample means and standard deviation equal to the standard error. This approximates the distribution of the sample means, which is expected to be approximately normal due to the Central Limit Theorem. (Through testing after 50000 numDist the normal bell curve started to appear).



Normal distribution-transforming the uniform sample to normal using Box-Muller transform

The Box–Muller transform, by George Edward Pelham Box and Mervin Edgar Muller transform is a mathematical method used to generate a sample of normally distributed random numbers from a sample of uniformly distributed random numbers. The method works as follows:

Generate two independent random variables U1 and U2 from a uniform distribution on the interval (0,1).

Compute the variables X and Y by transforming U1 and U2 as follows

$$X = \sqrt{-2Log(U_1)}cos(2\pi U_2)$$

$$Y = \sqrt{-2Log(U_1)}sin(2\pi U_2)$$

The variables X and Y are independent and identically distributed normally distributed random variables with mean 0 and variance 1.

To obtain a sample from a normal distribution with mean μ and standard deviation σ, we can simply transform X as follows:

$Z = \mu + \sigma * X$

where Z is the desired normal random variable.

The Box-Muller transform works because of a fundamental result in probability theory, the central limit theorem as we defined it before. The Box-Muller transform takes advantage of this fact by generating two independent random variables from a uniform distribution, and then transforming them into two independent and identically distributed normally distributed random variables.

References

Bain, L. J., & Engelhardt, M. (2013). Introduction to probability and mathematical statistics. Cengage Learning.

Devroye, L. (1986). Non-uniform random variate generation. Springer-Verlag.

Gentle, J. E. (2003). Random number generation and Monte Carlo methods. Springer Science & Business Media.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). Continuous univariate distributions (Vol. 1). John Wiley & Sons.

Ross, S. M. (2010). Simulation. Academic Press.

Wolfram Research, Inc. (2021). Mathematica (Version 12.3) [Computer software]. Champaign, IL: Wolfram Research, Inc.

Exponential distribution. (2021, April 16). In Wikipedia. Retrieved April 25, 2023, from https://en.wikipedia.org/wiki/Exponential_distribution

Generating Poisson distribution from exponential distribution. (n.d.). In Statistics How To. Retrieved April 25, 2023, from https://www.statisticshowto.com/poisson-distribution-from-exponential/

Monte Carlo method. (2022, April 1). In Wikipedia. Retrieved April 25, 2023, from https://en.wikipedia.org/wiki/Monte_Carlo_method

Negative binomial distribution. (2022, February 22). In Wikipedia. Retrieved April 25, 2023, from https://en.wikipedia.org/wiki/Negative_binomial_distribution

Normal distribution. (2022, April 12). In Wikipedia. Retrieved April 25, 2023, from https://en.wikipedia.org/wiki/Normal_distribution

Poisson distribution. (2022, April 1). In Wikipedia. Retrieved April 25, 2023, from https://en.wikipedia.org/wiki/Poisson_distribution

Rejection sampling. (2022, March 23). In Wikipedia. Retrieved April 25, 2023, from https://en.wikipedia.org/wiki/Rejection_sampling

Central limit theorem. (2022, April 20). In Wikipedia. Retrieved April 25, 2023, from https://en.wikipedia.org/wiki/Central_limit_theorem

Wolfram, E. (n.d.). Box-Muller Transformation. In Wolfram MathWorld. Retrieved April 26, 2023, from https://mathworld.wolfram.com/Box-MullerTransformation.html