PANDION: The Osprey Journal of Research & Ideas

Volume 1 | Number 1 Article 3

2020

Uniform random variate generation with the linear congruential method

Joseph Free
University of North Florida, joseph.free@unf.edu

Faculty Mentor: Ping Sa, PhD, Professor
Department of Mathematics and Statistics

Uniform random variate generation with the linear congruential method

Cover Page Footnote

The author would like to acknowledge his wife, Dani, whose support has been unwavering; his late brother-in-law, Kyle Forgy, whose presence is sorely missed; and Dr. Ping Sa, who was gracious enough to endorse this work.

Uniform random variate generation with the linear congruential method

Joseph Free

Faculty Mentor: Ping Sa, PhD
Department of Mathematics and Statistics
University of North Florida

Abstract
This report considers the issue of using a specific linear congruential generator (LCG) to create random variates from the uniform(0,1) distribution. The LCG is used to generate multiple samples of pseudo-random numbers and statistical computation techniques are used to assess whether those samples could have resulted from a uniform(0,1) distribution. Source code with annotations can be obtained by contacting the author.

1. Introduction

The Linear Congruential Generator is one of the most common methods of generating pseudo-random numbers. It is a fast and efficient tool well-suited for computer implementations. The method itself is deterministic; that is, it does *not* generate truly random numbers, but only values that give the appearance of randomness.

Definition 1 (Linear Congruential Generator [Ros13])**.** *Given an initial value x_0, called a seed, the Linear Congruential Generator recursively computes values according to the rule*

$$x_n = ax_{n-1} \mod m \tag{1}$$

where a and m are given positive integers. The quantity $\frac{x_n}{m}$ is then between 0 and 1, and is taken as an approximate value of a uniform(0,1) random variate. Alternatively, the method may also be specified as

$$x_n = ax_{n-1} + c \mod m \tag{2}$$

where c is an integer.

2. Problem

In this report, we use the LCG given by $x_n = 7^5 x_n \bmod 2^{31} - 1$ to generate a sequence of pseudo-random numbers and we analyze the possibility that they resulted from a true uniform(0,1) distribution. This is an important consideration to make as simulation of random variates from other distributions depends strongly on the the ability to generate uniformly distributed random numbers, as in the case of Inverse Transform Sampling. Random variates from more complicated distributions are the crux of realistic stochastic modeling, so it is vitally important that to ensure the quality of the values generated lest we run the risk of creating simulations that produce erroneous results. By starting with the uniform distribution we take the first step in guaranteeing the validity of random number generation from general distribution functions.

3. Method

In order to evaluate whether a given sequence of pseudo-random numbers is a realization from a theoretical distribution F, we make use two statistical tests. The first is the Kolmogorov-Smirnov test.

Theorem 1 (Kolmogorov-Smirnov)**.** *Given a sample of random numbers* $(x_1,x_2,\ldots,x_n)$*, define the sample, or empirical, distribution function* S_n *as*

$$\begin{aligned} S_n(x) &= 0 \text{ for } x < x_{(1)} \\ &= \frac{r}{n} \text{ for } x_{(r)} \le x < x_{(r+1)} \\ &= 1 \text{ for } x_{(n)} \le x \end{aligned}$$

Then the Kolmogorov-Smirnov statistic is defined to be $D = \sup_x |S_n(x) - F(x)|$*. For large* n*, if the* x_i*,* $i = 1,\ldots,n$*, are observations with true distribution* F*, then with probability* 0.99, $D < \frac{1.6276}{\sqrt{n}}$.

The second is the classic Pearson Chi-Squared Goodness-of-Fit test. For this test, we follow the recommendation of [DS86] (p. 70) and divide the sample into $k = \lfloor 1.88n^{2/5} \rfloor$ bins and calculate the statistic

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(O(i) - E(i))^2}{E(i)} \tag{3}$$

where $O(i)$ is the observed count of bin i and $E(i)$ is the expected count of bin i under the null hypothesis that the data was generated from a uniform distribution. The test statistic is then compared to the critical value $\chi^2_{1-\alpha,k-1}$, the upper $(1-\alpha)$th quantile of the chi-squared distribution with $k-1$ degrees of freedom. If $\chi^2_{k-1} > \chi^2_{1-\alpha,k-1}$, then the null hypothesis that the data is uniform is rejected.

In addition to the two tests above other heuristics are used to assess uniformity and randomness in two separate approaches. In the first approach, a single sample of 1000 pseudo-random numbers is generated using the LCG $x_n = 7^5 x_n \bmod 2^{31} - 1$ along with a random integer seed value x_0. The sample autocorrelation plot is then examined to assess randomness. The Kolmogorov-Smirnov (KS) and GOF statistics are then used to determine if the sequence could have come from the uniform(0,1) distribution. The plot of the theoretical and empirical cumulative density functions (essentially a graphical analog of the KS test) are compared; and the first, second, third, and fourth sample moments are calculated and juxtaposed against the theoretical moments of the uniform distribution. The maximum likelihood estimate (MLE) for the uniform distribution is also obtained; if the sample was generated from a uniform distribution, the MLE should be close to the parameter 1 of the uniform(0,1) distribution.

Note that the likelihood for the uniform distribution can be found by assuming that if $X_1,\ldots,X_n$ are ordered IID observations from a uniform distribution with parameter θ, then the pdf of the ith observation is $f_\theta(x_i) = 1/\theta$ for $0 \le x_i \le \theta$, and hence the likelihood function is $\mathcal{L}(\theta|\mathbf{X}) = \theta^{-n}$. By taking the derivative of the log-likelihood function, we can find that the maximum likelihood occurs at $\hat{\theta} = x_n$. That is, the ML estimator is the largest observed value of the sample.

In the second approach, 5000 samples of size 1000 are generated from the same LCG, each with a random integer seed. For each sample, the above analysis is repeated and the results saved. A bar plot of KS and GOF successes vs. failures is given, as well as the sampling distributions of the sample moments. Additionally, the empirical distribution functions $S_n^{(1)}, S_n^{(2)},\ldots,S_n^{(m)}$ for each sample is saved and used to calculate the large sample average empirical distribution function by the rule $\bar{S}(x) = \sum_{i=1}^{m} \frac{S_n^{(i)}(x)}{m}$. We expect that for large samples, $\bar{S} \to F$ as $m \to \infty$

Algorithms and source code for each approach can be found in sections 2.1 and 2.2 of the appendix, respectively.

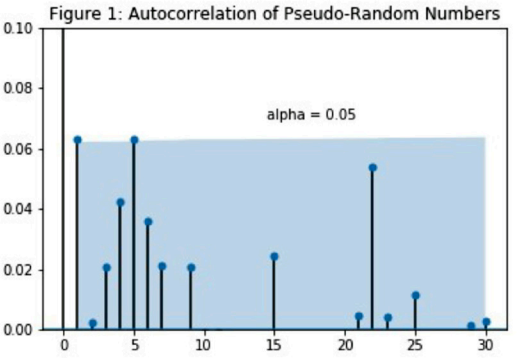
4. Results

4.1 Single Run

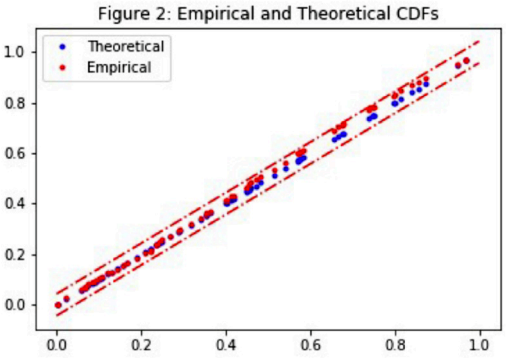
After generating the sample and calculating the KS and χ^2 statistics, it was determined that the sample "passed" both tests. The KS test generated a statistic $D = 0.388 < \frac{1.6276}{\sqrt{1000}}$, while GOF yielded χ^2 = 28.282 with p-value p = 0.4496 – resulting in a failure to reject the uniform hypothesis. Thus, we see no statistical evidence to conclude this sample resulted from any other type of distribution. In Figure 2 below a plot of the observed vs. expected CDF values is given together with a 95% confidence band. It is readily seen that all empirical CDF values fall within the band. Additionally, in Figure 1 an autocorrelation plot is given to assess independence. The plot has been scaled on the vertical axis to increase readability. Notice that all values are very small (< .1), with only two out of 30 lagged correlations plotting out of the confidence region. From this we can reasonably assume the sample is exhibiting sufficient independence. The sample moments and maximum likelihood estimator were also calculated and compared to the theoretical values. The results are summarized below:

	Expected	Observed	Δ
MLE	1	0.9997	0.0003
Mean	0.5	0.4849	0.0151
Variance	0.0833	0.0785	0.0048
Skewness	0	0.0534	0.0534
Kurtosis	-1.2	1.1021	0.0979

Taking these results into consideration, the sample has all the appearances of having been generated from a uniform(0,1) distribution, and hence the LCG did a satisfactory job of generating a sample of uniform random numbers. Output can be seen in the appendix, section 2.1.



(a) Figure 1.



(b) Figure 2.

Figures: (a) Autocorrelation plot out to lag $h = 30$ for the single sample of pseudo-random numbers. (b) A plot of values from the ECDF vs. expected theoretical CDF values. Dashed lines represent a 95% confidence band given by $b_k \pm \frac{1.36}{n^{1/2}}$, where $b_k = \frac{k - 0.5}{n}$ [JK70].

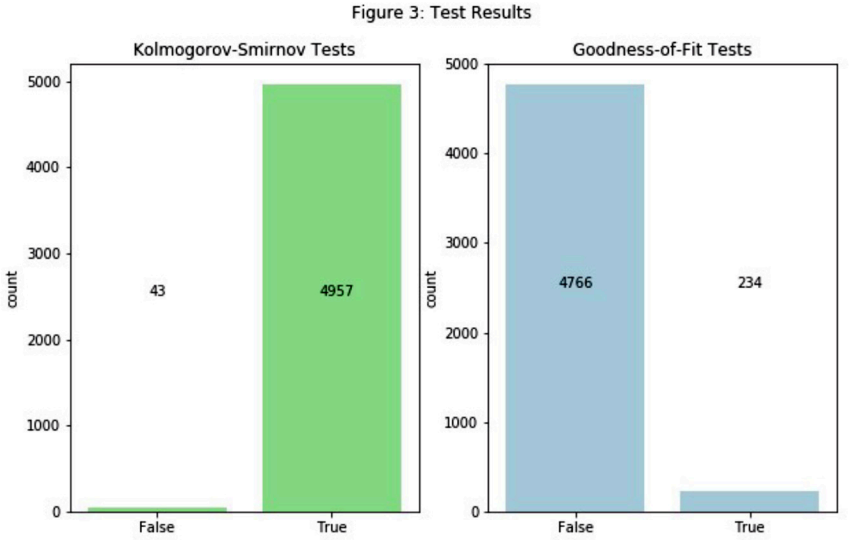


Figure 3: *left*: KS successes. True indicates the number of samples such that $D < \frac{1.6276}{\sqrt{1000}}$. *right*: GOF test results. True indicates the number of samples where the uniform hypothesis was rejected.

4.2 Multiple Runs

In order to truly assess the ability of the LCG to generate samples of uniform random numbers, we must examine the long-run behavior of the generator and assess the quality of the samples generated. To do so, the previous analysis, with the exception of the ACF plot, was replicated 5000 times. That is, 5000 new samples of length 1000 were generated and each was subjected to the KS and GOF tests, in addition to the previous heuristics.

The results of the KS and GOF tests are summarized in Figure 3 above. Recall that by theorem 1, the Kolmogorov-Smirnov statistic will be less than the specified cutoff with probability 0.99 for large n. After replicating our analysis multiple times, Figure 3 tells us that the KS statistic was less than the cutoff value for $\frac{4957}{5000} * 100 = 99.14\%$ of the samples. Similarly, the GOF test seems to indicate the samples generated could have likely come from a uniform distribution. The right panel of Figure 3 indicates that the uniform hypothesis was rejected for only $\frac{234}{5000} * 100 = 4.68\%$ of all samples, which is within a 5% margin of error.

Figure 4 shows the distribution of maximum likelihood estimators for the generated samples. For the uniform(0,1) distribution, the parameter of interest is $\theta = 1$. Hence, MLEs centered near 1 as seen here serves as valuable evidence of the uniformity of the samples. It is worth noting, however, that it is known that the ML estimators for the uniform distribution are known to underestimate the parameter.

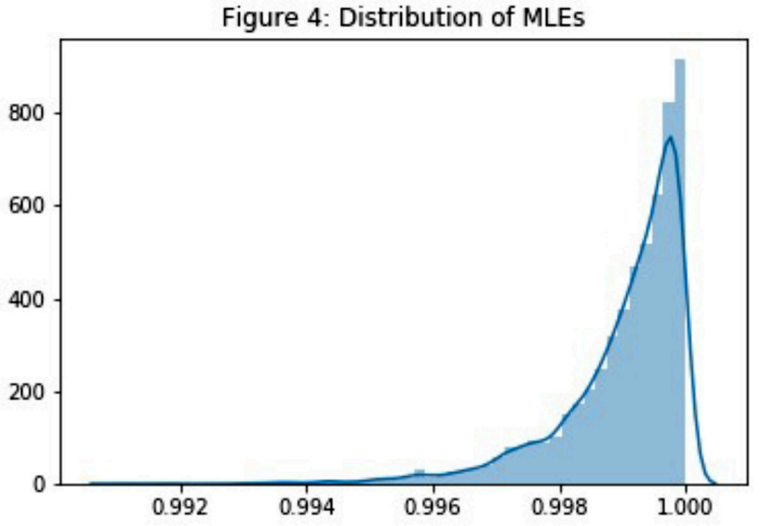


Figure 4: Distribution of MLEs for the 5000 runs. Note the distribution is skewed left with mass accumulated near 1

To assess the agreement of sample moments, the sampling distributions for the mean, variance, skewness, and kurtosis were plotted, and can be seen in Figure 5 (page 9). For each, the Shapiro-Wilk test was conducted to determine if the distributions could have resulted from a normal distribution. Additionally the average value of the sample moments was obtained and compared to the expected theoretical moments as above. These results are all present in the Figure 5. Observe that all sampling distributions are symmetric about their respective means, and all but kurtosis seem to behave normally at the $\alpha = 0.05$ level of significance. Also, as seen in the previous analysis above, the average moments all appear to fall much closer to the theoretical moments than the single sample case.

In Figure 6 (page 10), the average empirical CDF across the 5000 samples is plotted against the theoretical value. It is immediately obvious from the graph that the average ECDF appears to approach the theoretical CDF point-wise as the number of samples increases.

Finally, in Figure 7 (page 10) individual ECDFs for samples of size 20000, 10000, 5000, and 1000 are compared to assess changes that occur due to the number of observations generated. Not surprising, the larger the individual sample, the closer the sample distribution gets to the theoretical distribution. In the plot, 95% confidence bands are given, and it can be seen that for smaller sample sizes, the ECDF deviates slightly from the expected CDF, but those deviations smooth out and the bands become tighter for increasing n.

5. Discussion

The preceding results seem to strongly indicate that the LCG will truly generate random variates and samples from a uniform(0,1) distribution. However, these results are strictly empirical rather than mathematical. So the LCG will truly generate uniform variates *in a practical sense*. In general, though, the LCG ultimately fails at generating truly uniform random variables for a few reasons. The first and foremost reason is that the LCG is obviously not random: it is a deterministic algorithm based on modular arithmetic with a large prime modulus. In this case, the LCG used is actually the MINSTD LCG implemented in some C and C++ standards, and the constant $a = 7^5$ is a multiplicative cyclic generator modulo $2^{31}-1$. This guarantees that after a finite but large number of iterations, the LCG will repeat itself. So at the surface this generator is fundamentally not random. It is also fundamentally incapable of

truly generating uniform(0,1) random numbers. This is because a truly uniform(0,1) random number could be any point in the interval (0,1), whereas the LCG is only capable of producing numbers in the finite set $\{0,1/(m-1),\ 2/(m-1),\ldots,(m-1)/m\}$, where $m = 2^{31}-1$. Thus, it would be theoretically possible, for instance, to observe the value or from a true uniform(0,1) distribution, but this is impossible using the LCG. More generally speaking, the problem is that the LCG is only capable of producing values from a finite, countable subset of an uncountable support.

6. Conclusion

Based on the analysis and discussion conducted in this report, we may conclude that for practical purposes the Linear Congruential Generator $x_n = 7^5 x_n \bmod 2^{31} - 1$ behaves near identically to the uniform(0,1) distribution, but this is an empirical distinction only. Theoretically the generator fails to be precisely uniform(0,1) because 1) it is not truly random, and 2) the support of the generator is only a small (comparatively speaking) finite subset of the uniform(0,1) support. However it should be noted that the theoretical considerations do not preclude the empirical ones. The LCG's ability to mimic uniform random variables is a valuable resource and should not be discounted for simulation purposes.

References

[DS86] Ralph B. D'Agostino and Michael A. Stephens. *Goodness-of-Fit Techniques*. Marcel Dekker, Inc., USA, 1986.

[JK70] Norman L. Johnson and Samuel Kotz. *Distributions in statistics. Continuous univariate distributions. 2*. Houghton Mifflin Co., Boston, Mass., 1970.

[Ros13] Sheldon M. Ross. *Simulation, Fifth Edition*. Academic Press, Inc., USA, 2013.

Figure 5: Sampling Distribution of Moments

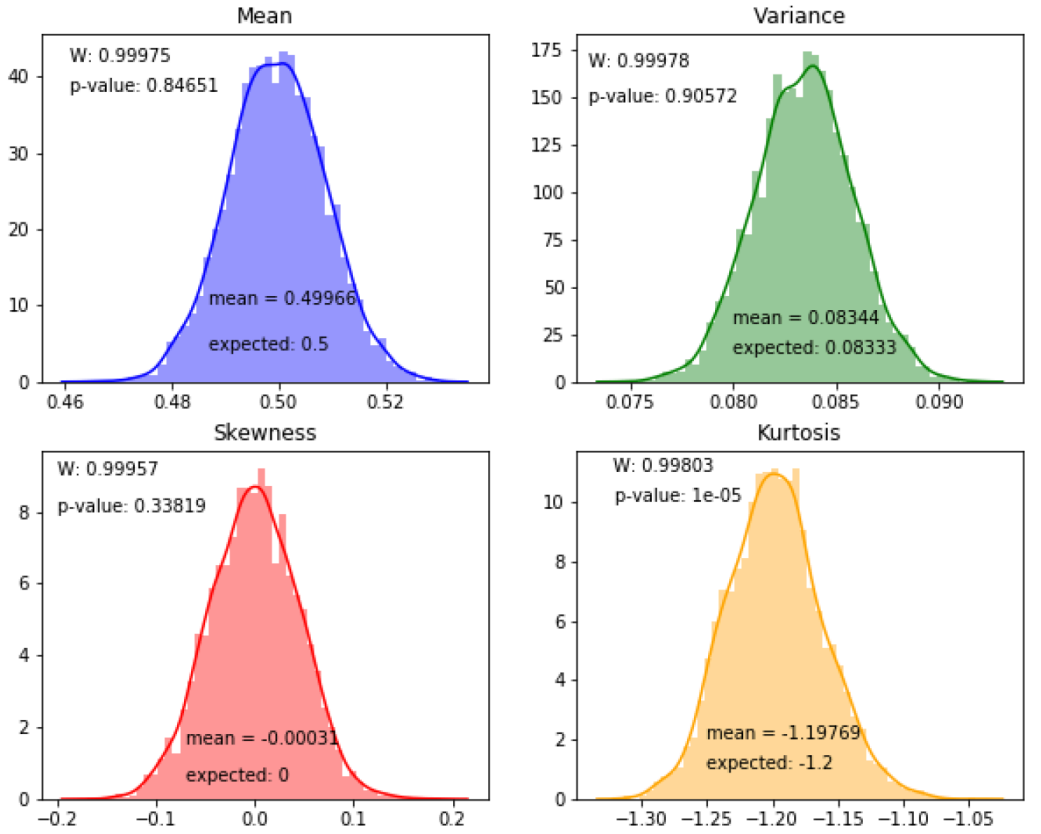


Figure 5: Sampling distributions for the first four moments. Shapiro-Wilk test results are given in the upper left of each subplot, and the average vs. theoretical moments are given in the center.

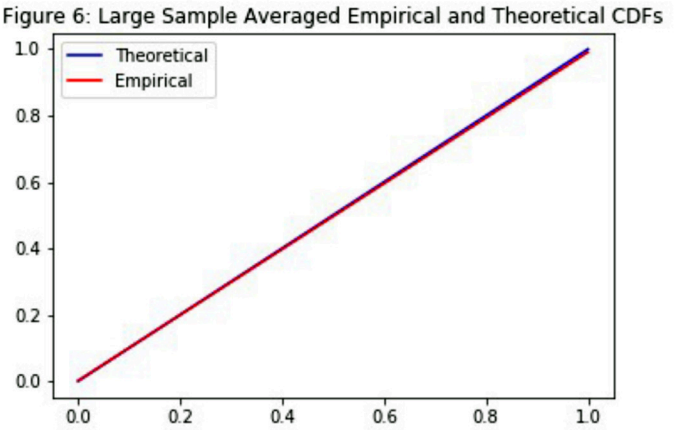


Figure 6: This is the plot of the averaged empirical CDF. By the LLN, we expect that the averaged ECDF approaches the theoretical CDF as the number of samples increases.

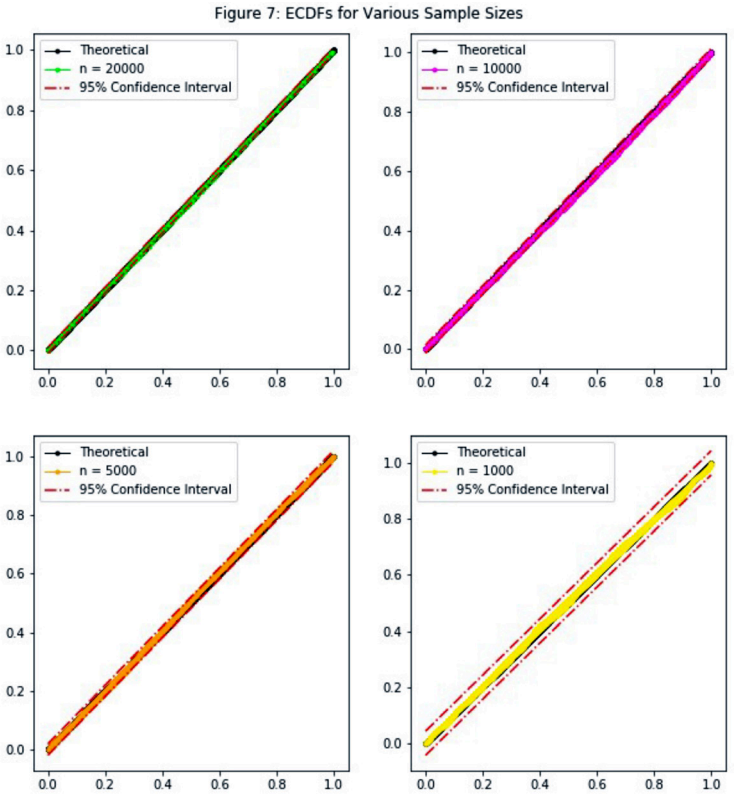


Figure 7: Individual ECDFs vs. theoretical for differing sample size. *Top left:* ECDF with n = 20000 observations. *Top right:* ECDF with n = 10000 observations. *Bottom left:* ECDF with n = 5000 observations. *Bottom Right:* ECDF with n = 1000 observations.