Simfit

Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
https://simfit.org.uk
https://simfit.silverfrost.com

Given two numbers A and B with $B > A$, then a random variable that can take all values in the interval between A and B with equal probability is referred to as having a uniform distribution, $U(A, B)$, or alternatively a rectangular distribution. This distribution is of immense value in simulation studies as will be explained subsequently. Two frequently encountered special cases are when only integer values are allowed, and also when $A = 0$ and $B = 1$.

1 Definitions

A random variable Y distributed as $U(A, B)$ has probability density function $g(y)$, cumulative distribution function $G(y)$, expectation $E(Y)$ and variance $V(Y)$ given by

$$\begin{aligned}
g(y) &= \frac{1}{B-A}\\
G(y) &= \frac{y-A}{B-A}\\
E(Y) &= \frac{A+B}{2}\\
V(Y) &= \frac{(A+B)^2}{12}.
\end{aligned}$$

It is interesting to note two important facts used by SimFIT concerning any arbitrary continuous random variable X with distribution function $F(x)$, and a random variable Y which follows a continuous uniform distribution on (0,1), say with distribution $G(y)$, i.e. with $A = 0$ and $B = 1$, so that $G(y) = y$.

1. If $U(0, 1)$ random numbers $y_1, y_2, \ldots, y_n$ are available, then random numbers $x_1, x_2, \ldots, x_n$ with distribution function $F(x)$ can be generated from them using

   $$x_i = F^{-1}(y_i).$$

   This can be appreciated from a graph of the cumulative distribution $F(x)$ as a function of x but taking as vertical axis $Y = F(x)$ so that

   $$P(X \le x) = P(Y \le F(x)) = G(F(x)) = F(x).$$

2. Conversely, given random numbers $x_1, x_2, \ldots, x_n$, then uniformly distributed random numbers $y_1, y_2, \ldots, y_n$ can be generated from them using

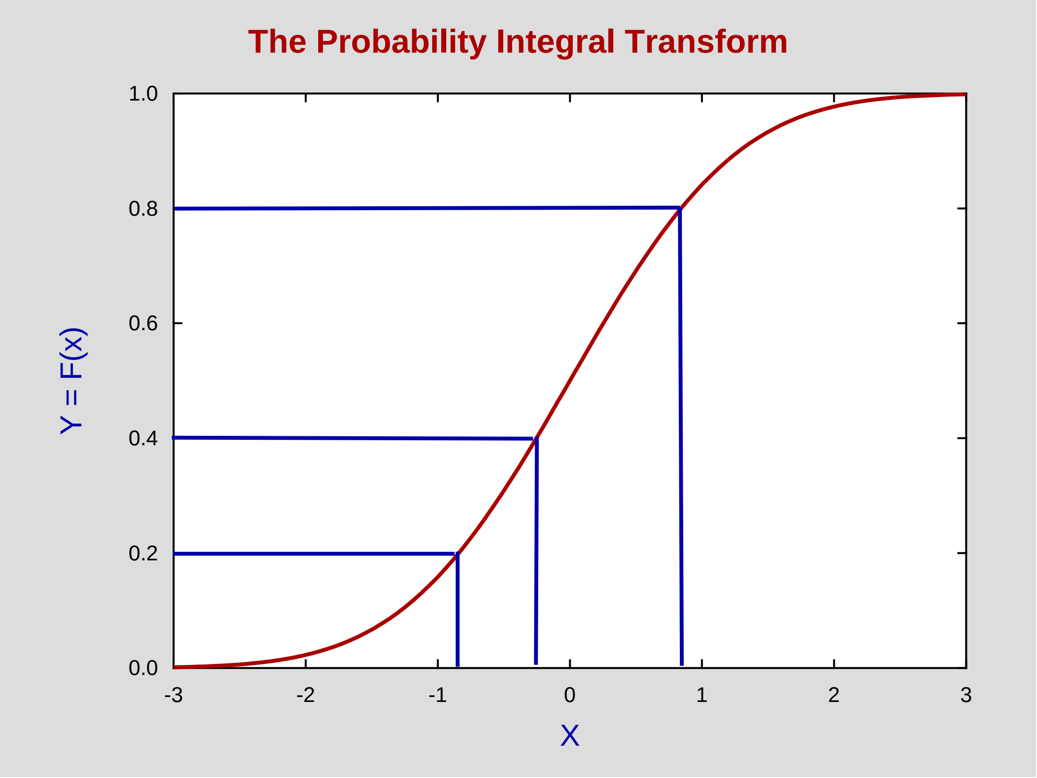
   $$y_i = F(x_i).$$

   This follows since

   $$P(Y \le y) = P(X \le F^{-1}(y)) = F(F^{-1}(y)) = y = G(y).$$

2 The Probability Integral transform

Not surprisingly there are technical details to consider before accepting the previous results, known as the probability integral transform. However, the following diagram and table illustrating the uniform distribution on 0,1 with distribution function $G(y)$ and the standard normal distribution with distribution function $F(x)$ should make it clear.



y	$G(y)$	x	$F(x)$
0.2	0.2	-0.8416	0.2
0.4	0.4	-0.2533	0.4
0.8	0.8	0.8416	0.8

The point is that equally spaced divisions on the Y axis correspond to unequal divisions on the X axis, but the probabilities in the intervals are identical. In other words, in terms of the inverse standard normal distribution function,

$$F^{-1}(0.2) = -0.8416$$
$$F^{-1}(0.4) = -0.2533$$
$$F^{-1}(0.8) = 0.8416,$$

so that

$$P(-0.8416 \le X \le -0.2533) = P(0.2 \le Y \le 0.4) = 0.2.$$

3 Pseudo random numbers

Computers cannot generate true random numbers, but they can generate extremely long deterministic sequences of numbers that do have properties closely similar to random numbers. As all such schemes are cyclic, the starting point in the sequence can be determined arbitrarily, usually by using the system clock, or from a fixed starting point using a seed or array of seeds. Generation of such pseudo random numbers begins by obtaining a sample of n such $U(0,1)$ numbers that are then transformed into numbers from a selected distribution. As evaluation of $F(.)$ and $F^{-1}(.)$ for standard distributions requires numerical methods, SimFIT does not use the scheme $x = F^{-1}(y)$ outlined in 1. above, as there are more convenient techniques. However SimFIT does use the scheme $y = F(x)$ outlined in 2. above to transform numbers into $U(0,1)$ numbers, because this is valuable when testing if numbers do arise from an assumed distribution, and it is particularly useful when visually inspecting values generated in experiments if these can be transformed so as to be collected into bins with equal probability, as will be illustrated.

Unfortunately, pseudo random numbers do have appreciable autocorrelation and other deficiencies, particularly if long sequences are required for simulation, and much ingenuity has been expended to surmount such obstacles. Accordingly, methods to test the performance of particular random number generators have been developed, and SimFIT provides the option to test the random number generator provided. This is a Marsaglia-Zaman type using subtract-with-borrow and has a cycle length of 2^{1376}, compared to the value of 2^{57} available with some standard linear congruential generators.

4 Testing the Simfit U(0,1) generator

Choose `[A/Z]` from the main SimFIT menu and open program **rannum** when the following options will be available.

```
Generate sequences of random numbers
Generate random matrices
Generate random permutations
Generate and plot random walks
Test the current U(0,1) generator
Set the seed type
```

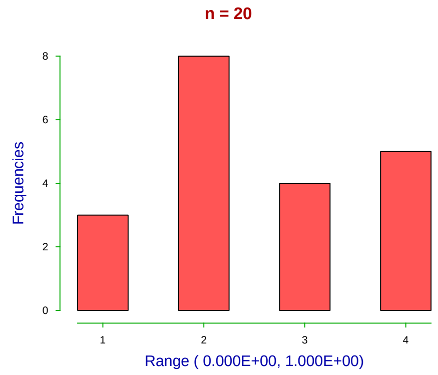
After selecting the option to test the current $U(0,1)$ generator these further options become available.

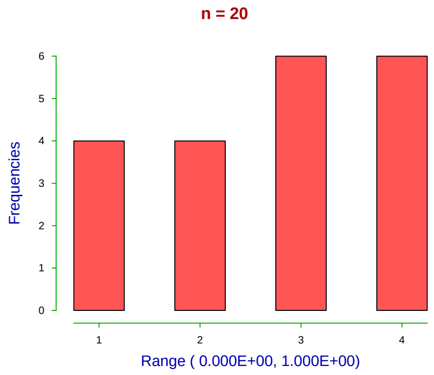
```
Runs up (or down) test
Bar chart plot
Chi-square test
Kolmogorov-Smirnov test
```

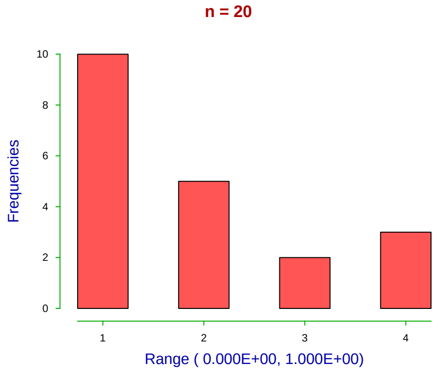
The runs up (or down) test requires a very large sample and tests for significant autocorrelations, the bar chart plot simply displays a histogram, the chi-square test measures departure of the histogram from a $U(0, 1)$ distribution, while the Kolmogorov-Smirnov test examines the maximum deviation of the sample cumulative distribution from the expected straight line.

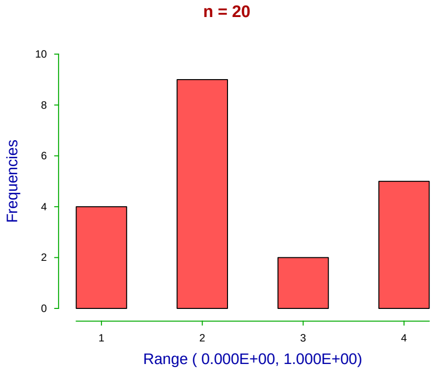
It is often advised that a minimum sample size of $n = 20$ is required to test if a sample is consistent with an assumed distribution and, although statistical tests like the above can be employed, decisions are more often made by visual inspection of a histogram. Now in the limit of very large samples with many bins histograms do converge in shape to the population distribution. However the next examples are intended to demonstrate that, in reality, sample sizes much greater than $n = 20$ are required to carry conviction. The $U(0, 1)$ distribution is particularly suited for this purpose as the histogram should have every bin frequency of approximately the same size, since the probability density function is a horizontal line.

The following histograms display four consecutive simulations using program **rannum**, and note that the usual advice is to have an expected value of at least 5, and preferably an observed value of the same order, for each bin. Of course, a major failing of analysis based on histograms is that the visual appearance and results from statistical analysis depend on the number of bins chosen.

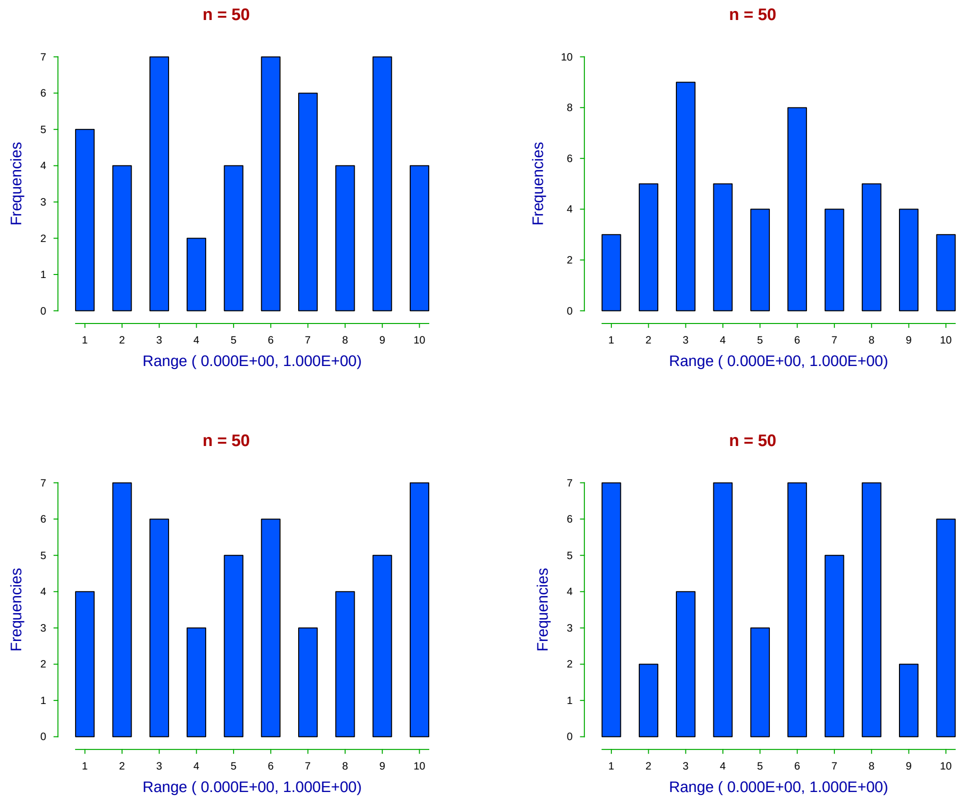




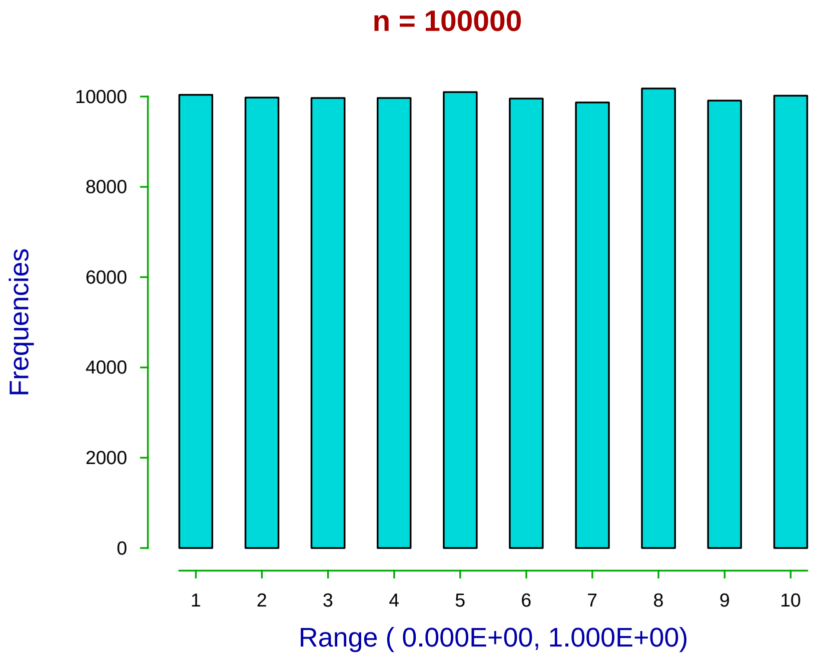




It will be clear from these results that a sample size of $n = 20$ is insufficient and could easily lead to false conclusions, as the histogram can suggest almost any shape for the population distribution. Increasing the sample size to $n = 50$ can still appear to be rather low as will be clear from the next four successive simulations.



Actually numerical results concerning the sample size required can be obtained from the SimFIT section on power as a function of sample size. Meanwhile here is the sort of convincing result obtained with large samples, in the next case $n = 10000$.


n = 100000
Frequencies
Range (0.000E+00, 1.000E+00)