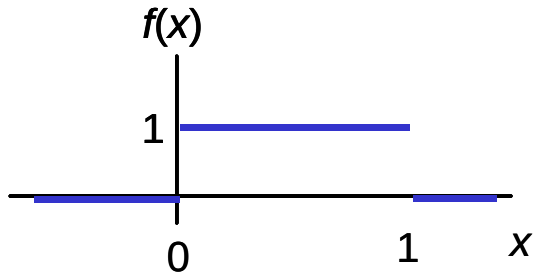
7. RANDOM NUMBER - GENERATION

7.1 Properties of Random Numbers

A sequence of random numbers $R_1, R_n, \ldots,$ must have two important statistical properties: Uniformity and Independence

Each R_i should be an independent realization (or sample) from a $R \sim U[0,1]$ variable:



$$E(R) = \int_0^1 x dx = \tfrac{1}{2}x^2|_0^1 = \tfrac{1}{2}$$

$$Var(R) = \int_0^1 x^2 - [E(R)]^2 = \tfrac{1}{3}x^3|_0^1 - [\tfrac{1}{2}]^2 = \tfrac{1}{3} - \tfrac{1}{4} = \tfrac{1}{12}$$

Consequences of Uniformity and Independence Properties:

1. If the interval $[0, 1]$ is divided into n subinterval of equal length, the expected number of observations in each interval is N/n, where N is the total number of observations

2. The probability of observing a value in a particular interval is independent of the previous values drawn.

7.2 Generation of Pseudo-Random Numbers

"Pseudo" means false, so false random numbers are being generated. If the method is known the numbers cannot be truly random. The goal of any generation scheme is to "imitate" the real properties of a uniform distribution and independence as closely as possible.

Problems that can occur with Pseudo Random Numbers:

1. Non-Uniformity

2. Numbers are discrete valued and not continuous on $[0, 1]$

3. The mean of generated numbers $\neq E(R) = \frac{1}{2}$

4. The variance of generated numbers $\neq Var(R) = \frac{1}{12}$

5. There may be dependence. For Example:

    a. Autocorrelation between numbers
    b. Numbers Successively higher or lower than adjacent numbers
    c. Several Numbers above the mean are followed by several number below the mean

Usually, pseudo-random numbers are generated by a computer as part of the simulation. Numerous methods are available. In selecting a routine, there are a number of important considerations.

1. The routine should be fast

2. The routine should be platform independent and portable between different programming languages.

3. The routine should have a sufficiently long cycle (much longer than the required number of samples).

4. The random numbers should be replicable. Usefull for debugging and variance reduction techniques.

5. Most importantly, the routine should closely approximate the ideal statistical properties.

> Inventing techniques that seem to generate random number is easy; inventing techniques that really produce sequences that appear to independent, uniformly distributed is incredibly difficult.

7.3. Techniques for Generating Random Numbers

The linear congruential method is most widely known. Will also report an extension that yield sequences with a longer period.

7.3.1. Linear Congruential Method

Generates a sequence $X_1,\ X_2, X_3, \ldots$ using

$$X_{i+1} = (aX_i + c)\, mod\, m, \tag{7.1}$$

where X_0 is called the seed variable, a is the constant multiplier, c is the increment, and m is the modulus.

When $c \neq 0$ $(c = 0)$, (7.1) is referred to as the **mixed congruential method** (**multiplicative congruential method**).

The selection of the valeus for a, c, m and X_0 drastically affect the statistical properties and the cycle length. Variations of (7.1) are quite common in computer packages.

EXAMPLE 7.1: $X_0 = 27$, $a = 17$, $c = 43$ and $m = 100$.
Here integer values will be between 0 and 99 because of the modulus m. Note that random integers are being generated rather than random numbers. These integers should be uniformly distributed. Convert to numbers in $[0, 1]$ by normalizing with modulus m:

$$R_i = \frac{X_i}{m} \qquad (7.2)$$

$X_0 = 27$

$X_1 = (17 \cdot 27 + 43)\, mod\, 100 = 502\, mod\, 100 = 2$

$R_1 = \frac{2}{100} = 0.02$

$X_2 = (17 \cdot 2 + 43)\, mod\, 100 = 77\, mod\, 100 = 77$

$R_2 = \frac{7}{100} = 0.77$

$\vdots$

etc.

Of primary importance is uniformity and statistical independence. Of secondary importance is maximum density and maximum period within the sequence R_i,$i = 1, 2, \ldots$.

Note that, the sequence can only take values in:

$$\{0, 1/m, 2/m, \ldots, (m-1)/m, 1\}$$

Thus R_i is discrete rather than continuous.

This is easy to fix by choosing large modulus m. Values such as $m = 2^{31} - 1$ and $m = 2^{48}$ are in common use in generators appearing in many simulation languages). Maximum Density and Maximum period can be achieved by the proper choice of a, c, m and X_0.

- For m a power of 2, say $m = 2^b$, and $c \neq 0$, the longest period $P = m = 2^b$ is achieved provided c is a relative prime to m (that is the greates common factor of c and m is 1), and $a = 1 + 4 \cdot k$ where k is an integer.

- For m a power of 2, say $m = 2^b$, and $c = 0$, the longest period $P = m/4 = 2^{b-2}$ is achieved provided the seed X_0 is odd and the multiplier $a = 3 + 8 \cdot k$, for some $k = 0, 1, \ldots$.

- For m is a prime number and $c = 0$, the longest period $P = m - 1$ is achieved provided the multiplier a has the property that the smallest integer k such that $a^k - 1$ is divisible by m is $k = m - 1$.

EXAMPLE 7.2: $a = 11,\ m = 2^6 = 64$, $c = 0$ and $X_0 = 1, 2, 3$ and 4.

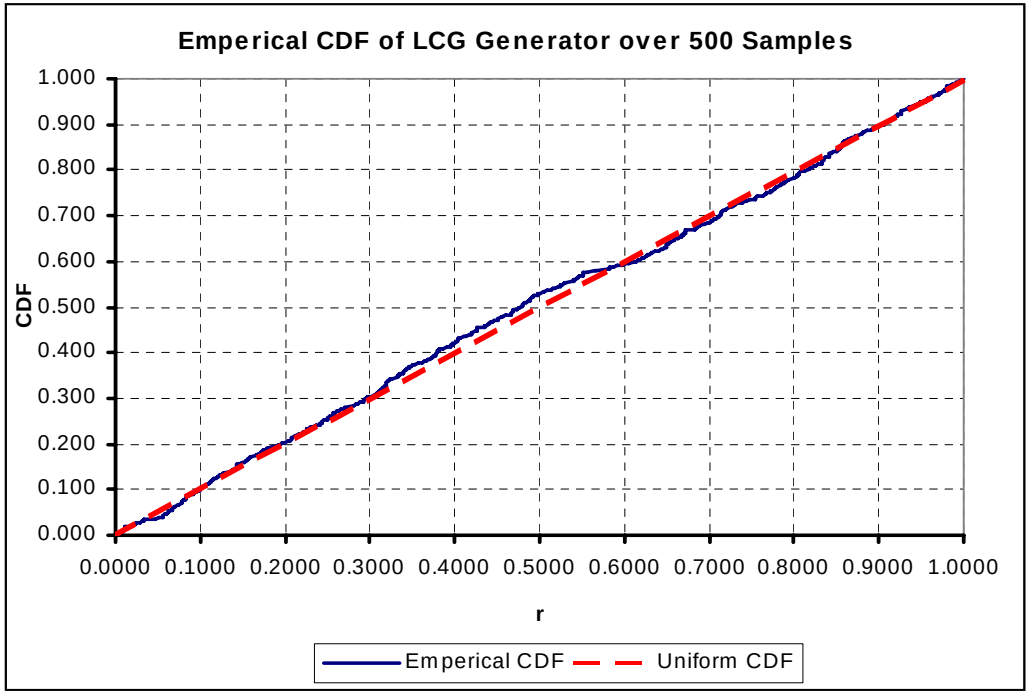
EXAMPLE 7.3: $a = 19,\ m = 10^2 = 100$, $c = 0$ and $X_0 = 63$.

EXAMPLE 7.4: $a = 7^5 = 16807,\ m = 2^{31} - 1$, $c = 0$.

This one is in actual use and extensively tested. Conditions ensure a period $P = m - 1$. By changing X_0 the user can control repeatibility of sequence

k=0	X_k	R_k	Gap
0	1234567		
1	1422014746	0.0006	0.6616
2	456328559	0.6622	0.4497
3	849987676	0.2125	0.1833
4	681650688	0.3958	0.0784
5	1825340118	0.3174	0.5326
6	1687465831	0.8500	0.0642
7	1569179335	0.7858	0.0551
8	2097898185	0.7307	0.2462
9	1988278849	0.9769	0.0510
10	9584176	0.9259	0.9214

Over 500 Samples	
Average	0.499
Variance	0.086
Average Gap	0.002



7.3.2. Combined Linear Congruential Method

Random Number Generator of Example 7.4. with a period of

$$2^{31} - 1 \approx 2 \cdot 10^9$$

is no longer adequate with current advancements of computing powers. Area of new research is deriving generators with longer periods (See, L'Ecuyer [1998]). This research uses the following important result.

If $W_{i,1}, W_{i,2}, \ldots W_{i,k}$ are independent discrete valued random variables (not necesarrily identical distributed), but one of them, say $W_{i,1}$, is uniformly distributed on the integers $0, \ldots, m_1 - 2$, then

$$W_i = \left(\sum_{j=1}^{k} W_{i,j} \right) mod\, (m_1 - 1)$$

is uniformly distributed on the integers $0, \ldots, m_1 - 1$.

1. Let $X_{i,1}, X_{i,2}, \ldots X_{i,k}$ be the i-th output from k different multiplicative $(c = 0)$ congruential generators, where first generator has modulus m_1 and the multiplier a_1 chosen so that the period is $m_1 - 1$.
2. The first generator then produces $X_{i,1}$ that are approximately uniform distributed on $1, \ldots, m_1 - 1$.
3. Hence, $W_{i,1} = X_{i,1} - 1$ are approximately uniform distributed on $0, \ldots, m_1 - 2$

L'Ecuyer suggest combined generators of the form

$$X_i = \left(\sum_{j=1}^{k} (-1)^{j-1} X_{i,j} \right) mod\, (m_1 - 1)$$

with

$$R_i = \begin{cases} \frac{X_i}{m_1} & X_i > 0 \\ \frac{m_1 - 1}{m_1} & X_i = 0. \end{cases}$$

Note that the $(-1)^{j-1}$ coefficient implicitly perform the subtraction $X_{i,1}-1$. For example, if $k=2$,

$$(-1)^0(X_{i,1}-1)+(-1)^1(X_{i,2}-1)=\sum_{j=1}^{2}(-1)^{j-1}X_{i,j}$$

Maximum possible period for such a generator is:

$$P=\frac{\prod_{i=1}^{k}(m_k-1)}{2^{k-1}}$$

which is achieved by the following generator.

EXAMPLE 7.5:

1. Selected seed $X_{1,0} \in [1, 2147483562]$, Selected seed $X_{2,0} \in [1, 21474833398]$.

2. Evaluate each individual generator

$$X_{1,j+1} = (40014 \cdot X_{1,j}) \, mod \, 2147483563$$

$$X_{2,j+1} = (40692 \cdot X_{2,j}) \, mod \, 21474833399$$

3. Set

$$X_{j+1} = (X_{1,j+1} - X_{2,j+1}) \, mod \, 2147483562$$

4. Return

$$R_{j+1} = \begin{cases} \frac{X_{j+1}}{2147483563} & X_j > 0 \\ \frac{2147483562}{2147483563} & X_j = 0. \end{cases}$$

5. Set $j = j + 1$ and go to Step 2.

ARENA RANDOM NUMBER GENERATOR:

$$A_n = (1403580A_{n-2} - 810728A_{n-3})\, mod\, 4294967087$$

$$B_n = (527612B_{n-1} - 1370589B_{n-3})\, mod\, 4294944443$$

$$Z_n = (A_n - B_n)\, mod\, 4294967087$$

$$U_n = \begin{cases} \frac{Zn}{4294967088} & Z_n > 0 \\ \frac{4294967087}{4294967088} & Z_n = 0. \end{cases}$$

Seed = a six-vector of first three A_n's, B_n's

7.4. Tests for Random Numbers

Desirable Properties: Uniformity and Independence

1. Frequency Test: Uses the Kolmogorov-Smirnic or the Chi-Square Test to compare the distribution of the set of numbers to a uniform distribution.
2. Runs Test: Test the runs up and down or the runs above and below the mean by comparin the actual values to expected values. The Statistics for comparison is the chi-square test.
3. Autocorrelation Test: Test the correlation between numbers and compares the sample correlation to the expected correlation of zero.
4. Gap Test: Counts the number of digits that appear between repetitions of a particular digit and then uses the Kolmogorov Smirnov test to compare with the expected sized of the gaps.
5. Poker Test: Treat numbers grouped together as a poker hand. Then the hands obtained are compared to what is expected using the chi-square test.

First Entry test for Uniformity. Second through Fifth Entry tests for Independence.

7.4.1. Frequency Tests

1. The Kolmogorov-Smirnov Test

Test compares a continuous cdf $F(x)$ to an emperical cdf $S_N(x)$, of the sample of N observations.

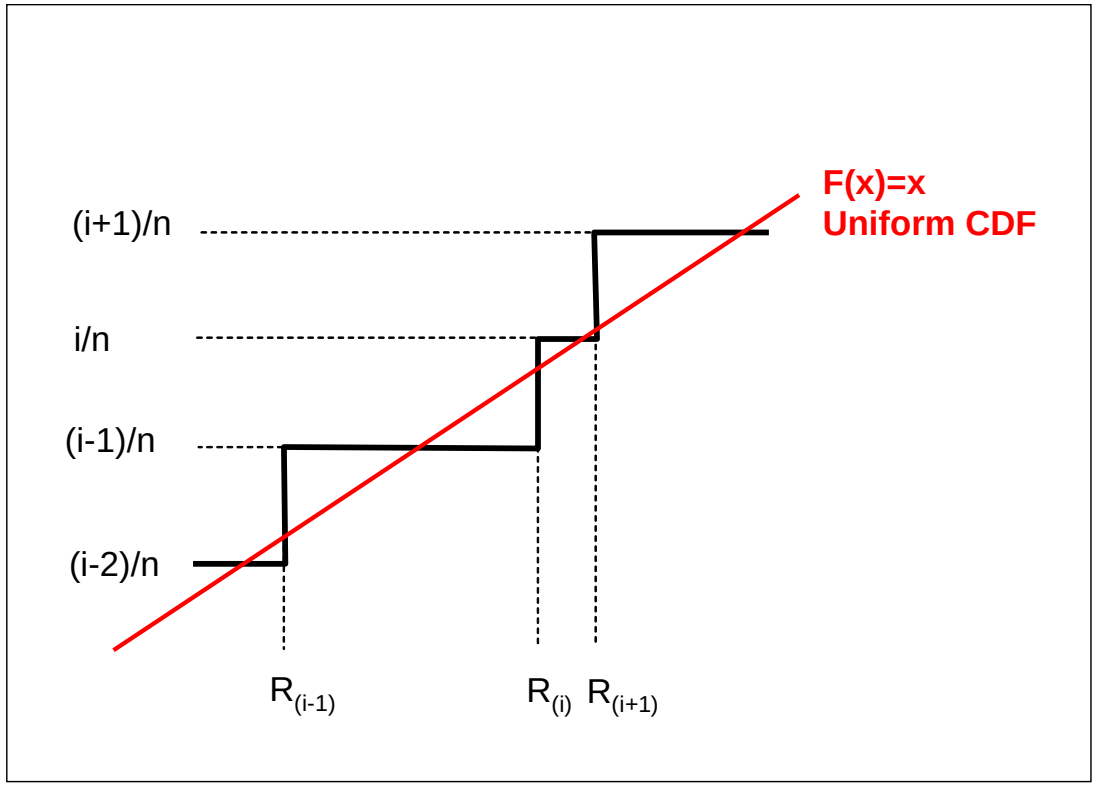
Step 1: Determine Order Statistics

$$R_{(1)} \leq R_{(2)} \leq \ldots \leq R_{(N)} \tag{7.3}$$

Define:

$$S_N(x) = \begin{cases} 0 & x \leq R_{(1)} \\ \frac{i}{N} & R_{(i)} \leq x < R_{(i+1)},\ i = 1, \ldots, N-1 \\ 1 & x \geq R_{(N)} \end{cases}$$

Step 2: Compute $D = \underset{x \in [-\infty, \infty]}{Max} |F(x) - S_N(x)|$ via



$$D^+ = \underset{1 \le i \le N}{Max} \left\{ \frac{i}{N} - R_{(i)} \right\}, D^- = \underset{1 \le i \le N}{Max} \left\{ R_{(i)} - \frac{i-1}{N} \right\}$$

$$D = Max\left\{ D^-, D^+ \right\}$$

Step 3: Compute for significant level $\alpha = 0.10, 0.05$ or 0.01 (for $N \geq 35$)

$$D_{0.10} = \frac{1.22}{\sqrt{N}},\ D_{0.05} = \frac{1.36}{\sqrt{N}},\ D_{0.01} = \frac{1.63}{\sqrt{N}}$$

Step 4: Test Hypothesis

$$\begin{cases} D \leq D_\alpha & \text{Accept: No Difference between } S_N(x) \text{ and } F(x) \\ D > D_\alpha & \text{Reject: No Difference between } S_N(x) \text{ and } F(x) \end{cases}$$

2. The Chi-square Test

The chi-square test uses the sample statistic

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed number in the i-th class, E_i is the expected number in the i-th class, and n is the number of classes.

Step 1: Determine Order Statistics

$$R_{(1)} \leq R_{(2)} \leq \ldots \leq R_{(N)} \tag{7.3}$$

Step 2: Divided Range $R_{(N)} - R_{(1)}$ in n equidistant intervals $[a_i, b_i]$, such that each interval has at least 5 observations.

Step 3: Calculate for $i = 1, \ldots, N$

$$O_i = N \cdot \{S_N(b_i) - S_N(a_i)\},\ E_i = N \cdot \{F(b_i) - F(a_i)\}$$

Step 4: Calculate

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Step 5: Determine for significant level α, $\chi^2_{\alpha,n-1}$

$$\begin{cases} \chi_0^2 \leq \chi^2_{\alpha,n-1} & \text{Accept: No Difference between } S_N(x) \text{ and } F(x) \\ \chi_0^2 > \chi^2_{\alpha,n-1} & \text{Reject: Difference between } S_N(x) \text{ and } F(x) \end{cases}$$

7.4.2. Runs Test

0.08	0.09	0.23
0.11	0.16	0.18
0.02	0.09	0.30
0.12	0.13	0.29

Both KS and Chi-Square test would accept the above sample as from a uniform, but if we look at the sample above column wise we see that each column is strictly larger than the other. Hence, independence may be an issue in that case.

1. Runs up and Runs Down

The runs test examines the arrangement of numbers in a sequence to test the hypothesis of independence. A run is defined as a succesion of similar events preceded and follow by a different event. Length of the run is the number of events that occur in the run

EXAMPLE: Series of coin Tosses

$$HTTHHTTTHT$$

Run 1 has length 1. Run 2 has lenght 2, Run 3 has length 2,
Run 4 has length 3, Run 5 has length 1, Run 6 has length 1.

TWO CONCERNS in a Runs Test:

1. The number of Runs 2. The length of the Runs.

An up run is a sequence of numbers which is succeeded by a larger number. A down run is a sequence of numbers each of which is succeeded by a smaller number.

Example 1	
0.87	-
0.15	+
0.23	+
0.45	+
0.69	-
0.32	-
0.3	-
0.19	+
0.24	-
0.18	+
0.65	+
0.82	+
0.93	-
0.22	+
0.81	

Runs	8
Runs up	4
Runs Down	4

Example 2	
0.08	+
0.18	+
0.23	+
0.36	+
0.42	+
0.55	+
0.63	+
0.72	+
0.89	+
0.91	

Runs	1
Runs up	1
Runs Dowr	0

Example 3	
0.08	+
0.93	-
0.15	+
0.96	-
0.26	+
0.84	-
0.28	+
0.79	-
0.36	+
0.57	

Runs	9
Runs up	5
Runs Down	4

Example 2 and Example 3 are two extremes. If N is the number of observation, the maximum number of runs is $N-1$ and the minimum number of runs is 1. If a is the number of runs in an independent sequence of uniform numbers, the mean and the variance of a is given by

$$\mu_a = \frac{2N-1}{3},\ \sigma_a^2 = \frac{16N-29}{90}$$

For $N > 20$, the distribution of a is well approximated by a Normal distribution, $N(\mu_a, \sigma_a^2)$.

STEP 1: Calculate number of runs a

STEP 2: Calculate

$$Z = \frac{a-\mu_a}{\sigma_a} = \left[a - \frac{2N-1}{3}\right] \Big/ \sqrt{\frac{16N-29}{90}}$$

STEP 3: Determine acceptence region $[z_{-\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}]$ from a $N(0,1)$ distribution

STEP 4:

$$\begin{cases} Z \in \left[z_{-\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right] & \text{Accept: Independence} \\ Z \notin \left[z_{-\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right] & \text{Reject: Independence} \end{cases}$$

2. Runs above and below the mean

Acceptance of the indepence following the previous test is necessary but not sufficient condition. An independent sequence should not contain long runs with observations above the mean and below the mean.

Let b the total number of runs (above or below the mean) and n_1 be the number of observations above the mean and n_2 be the number of observations below the mean. Note that, the maximum number of runs is $N = n_1 + n_2$ and the minimum number of runs is 1.

Given n_1 and n_2

$$\mu_b = \frac{2n_1n_2}{N} + \frac{1}{2}; \sigma_b^2 = \frac{2n_1n_2(2n_1n_2 - N)}{N^2(N-1)}$$

and if $n_1 > 20$ or $n_2 > 20$, b is approximately normal distributed.

STEP 1: Calculate number of runs b

STEP 2: Calculate

$$Z = \frac{b - \mu_b}{\sigma_b} = \frac{a - \frac{2n_1n_2}{N} - \frac{1}{2}}{\sqrt{\frac{2n_1n_2(2n_1n_2 - N)}{N^2(N-1)}}}$$

STEP 3: Determine acceptence region $[z_{-\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}]$ from a $N(0,1)$ distribution

STEP 4:

$$\begin{cases} Z \in [z_{-\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}] & \text{Accept: Independence} \\ Z \notin [z_{-\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}] & \text{Reject: Independence} \end{cases}$$

Homework 1:
Implement the *The Kolmogorov-Smirnov Test* in these notes for a random sample of 100 numbers generated by the RAND() function in EXCEL and test the hypothesis whether we accept or reject uniformity of the random numbers generated by the RAND() function.

Homework 2:
Implement the *Chi-square Test* in these notes for a random sample of 100 numbers generated by the RAND() function in EXCEL and test the hypothesis whether we accept or reject uniformity of the random numbers generated by the RAND() function. Use 10 bins (or classes).