Unit V: Probability Concept and Random Number Generation

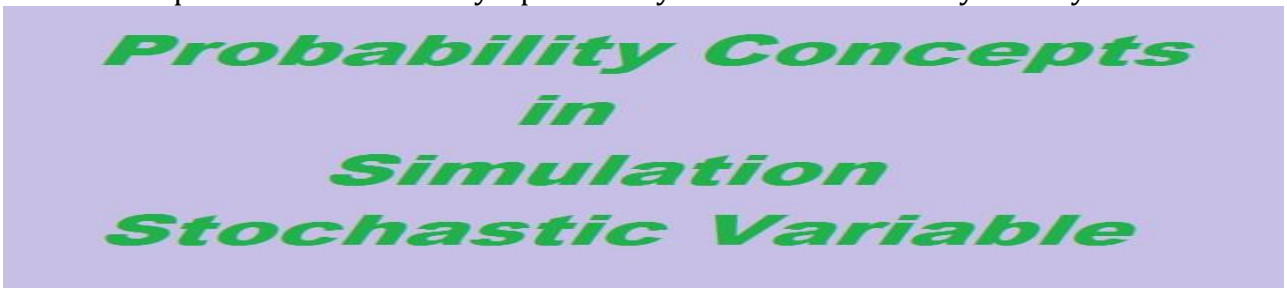
Probability Concepts in Simulation- Stochastic Variable:

The description of activities can be of two types deterministic and stochastic. The process on which, the outcome of an activity can be described completely in terms of its input is deterministic and the activity is said to be deterministic activity. On the other hand, when the outcome of an activity is random, i.e. there can be various possible outcomes, the activity is said to be stochastic activity. In case of an automatic machine, the output per hour is deterministic, while in a repair shop the number of machines repaired will vary from hour to hour, in a random fashion. The terms random and stochastic are interchangeable.

A random variable x is called discrete if the number of possible values of x (i.e. range space) is finite or countably infinite, i.e. possible values of x maybe x1, x2, …, xn.

A random variable x is called continuous if its range space is an interval or a collective of intervals. A continuous variable can assume value over a continuous range.

A stochastic process is described by a probability law called Probability Density Function.



Discrete Probability Function:

If a random variable x can take x_i (i = 1… n) countable infinite number of values with the probability of value x_i being $P(x_i)$ is said to be Probability Distribution or Probability Mass Function of a random variable x.

The number of P (x_i) must satisfy the following two conditions:

i. $P(x_i) >= 0$ for all i
ii. $\sum_{i=1}^{n} P(x_i) = 1$

Cumulative Distribution Function:

It is a function which, gives the probability of a random variable being less or equal to a random variable being less or equal to a given value. In a discrete test, the cumulative distribution function is denoted by $P(x_i)$. This function implies that x takes values less than or equal to x_i.

Continuous Probability Function:

If the random variable is continuous and not limited to discrete values, it will have an infinite number of values in an interval. Such a variable is defined by a function f(x) called a Probability Density Function (pdf). The probability that a variable x, falls between x and x+dx is expressed as f(x)dx and the probability that x falls in the range x1 to x2 is given as:

$$P(x) = \int_{x1}^{x2} f(x)dx$$

Random Variables:

A random variable is a rule that assigns a number to each outcome of an experiment. These numbers are called values of a random variable. Random variables are usually denoted by X.

Example:

1. If a die is rolled out, the outcome has a value from 1 through 6.
2. If a coin is tossed, the possible outcome is head 'H' or tail 'T'.

There are two types of random variables:

1. Discrete Random Variable:

A discrete random variable takes only specific, isolated numerical values. The variables which take finite numeric values are called as Finite discrete random variables and which

takes unlimited values are called as Infinite discrete random variables. The examples are shown in the table below:

Random Variables	Values	Types
Flip a coin three times; X = the total number of heads	{0, 1, 2, 3}	Discrete Finite There are only four possible values for X.
Select a mutual fund; X = the number of companies in the fund portfolio.	{2, 3, 4, ...}	Discrete Infinite There is no stated upper limit to the size of the portfolio.

Let

$X \rightarrow$ discrete random variable

$R_X \rightarrow$ possible values of X, given by range space of X.

$X_i \rightarrow$ the individual outcome in R_X.

A number $P(x_i) = P(X = x_i)$ gives the probability that the random variable equals the value of x_i. The number $P(x_i)$, I = 1, 2, 3 ... must satisfy two conditions:

i. $P(x_i) >= 0$ for all i
ii. $\sum_{i=1}^{\infty} P(x_i) = 1$

The collection of pairs $(x_i, P(x_i))$ i.e. a list of probabilities associated with each of its possible values is called probability distribution of X. $P(x_i)$ is called probability mass function (pmf) of X.

Example:

Consider the experiment of tossing a single die, defining X as the number of spots on up the face of die after a toss.

Solution:

N=total number of observations = 21

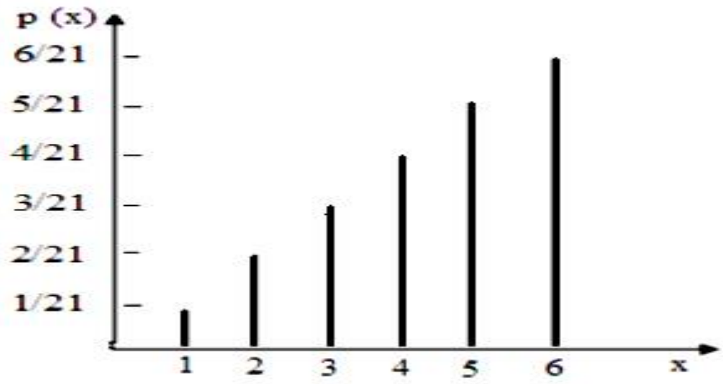
The discrete probability distribution is given by

x_i	1	2	3	4	5	6
$P(x_i)$	1/21	2/21	3/21	4/21	5/21	6/21

The conditions also are satisfied i.e.

1. $P(x_i) >= 0$, for i = 1, 2, ... 6.
2. $\sum_{i=1}^{\infty} p(x_i) = \frac{1}{21} + \frac{2}{21} + \cdots + \frac{6}{21} = 1$

The distribution is shown graphically in the figure below.



2. Continuous Random Variable:

Continuous Random Variable takes any values within a continuous range or an interval. The example is tabulated in the table below.

Random Variable	Values	Type
Measure the length of an object; X = its length in centimetres.	Any positive real number	Continuous. The set of possible measurements can take on any positive value.

For a continuous random variable X, the probability that X lies in the interval [a, b] is given by,

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx - - - - - i$$

The function f(x) is called Probability Density Function (pdf) of random variable X.
The pdf must satisfy the following conditions:

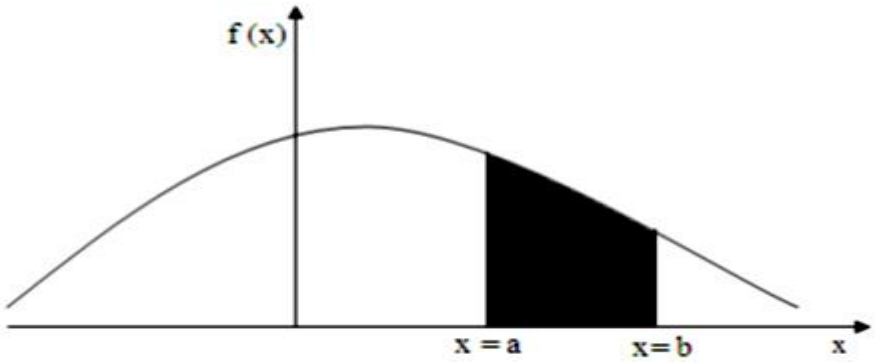
1. f(x) >= 0, for all x in Rx
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (*total area under graph is* 1)
3. $f(x) = 0, if\ x\ is\ not\ in\ Rx$

For any specified value X_0, $P(X = x_0) = 0$ since

$$\int_{x_0}^{x} f(x)\,dx = 0$$

Since $P(X = x_0) = 0$, the following equation also hold:
P(a <= X <= b) = P(a < X <= b) = P(a <= X < b) = P(a < X < b)

The graphical interpretation of equation *i* is shown in the figure below.



Random Numbers:

A random number is a number generated by a process, whose outcome is unpredictable, and which cannot be subsequentially reliably reproduced. Random numbers are the basic building blocks for all simulation algorithms.

Properties of Random Numbers:

The two important statistical properties are:

1. Uniformity
2. Independence

Each random number Ri is an independent sample drawn from a continuous uniform distribution between 0 and 1. The probability density function (pdf) is given by

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

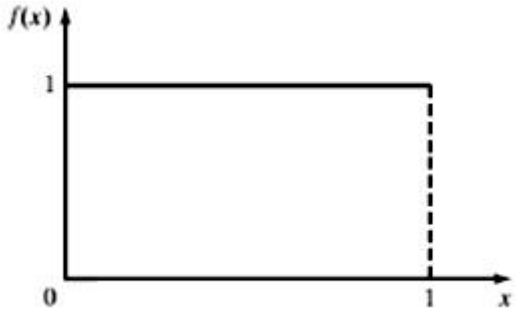


Fig: The Pdf for Random Numbers

The expected value of each Ri is given by

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

The variance is given by

$$\begin{aligned} V(R) &= \int_0^1 x^2\,dx - [E(R)]^2 \\ &= [x^3/3]_0^1 - (1/2)^2 = 1/3 - 1/4 \\ &= 1/12 \end{aligned}$$

The consequences of uniformity and independence properties are:

1. If the interval (0, 1) is divided into n classes or subintervals of equal length, then 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 previous values drawn.

Pseudo-Random Numbers:

Pseudo means false but here pseudo implies that the random numbers are generated by using some known arithmetic operations. Since the arithmetic operation is known and the sequence of random numbers can be repeatedly obtained, the numbers cannot be called truly random. However, the pseudo-random numbers generated by many computer routines very closely fulfil the requirements of the desired randomness.

If the method of random number generation, i.e. the random number generator is defective, the generated pseudo-random number may have the following departures from ideal randomness:

1. The generated random numbers may not be uniformly distributed.
2. The generated random numbers may not be continuous.
3. The mean of the generated numbers may be too high or too low.
4. The variance may be too high or too low.

Generation of Random Number:

In computer simulation where a very large number of random numbers is generally required, can be obtained by the following method.

1. Random numbers maybe drawn from the random number tables stored in a memory of the computer. The process is neither practical nor economical. It is a very slow process and the number occupied considerable space of computer memory. Above all, in the real system many time more random number than available in the table.

2. An electronic device may be constructed as a part of a digital computer to generate truly random numbers. This, however, is considered very expensive.

3. Pseudo-random numbers may be generated using some arithmetic operation. These methods must commonly specify a procedure starting with an initial number, the second number is generated and from that a third number and so on. A number of the recursive procedure are used for generating random numbers.

Qualities of an Efficient Random Number Generator:

1. It should have a sufficiently long cycle i.e. it should generate a sufficiently long sequence of random numbers before beginning to repeat the sequence.

2. The random numbers generated should be replicable i.e. by specifying the starting condition, it should numbers as and when desired. Many times common random numbers are required for the comparison of two systems.

3. The generated random numbers should fulfil the requirement of uniformity and independence.

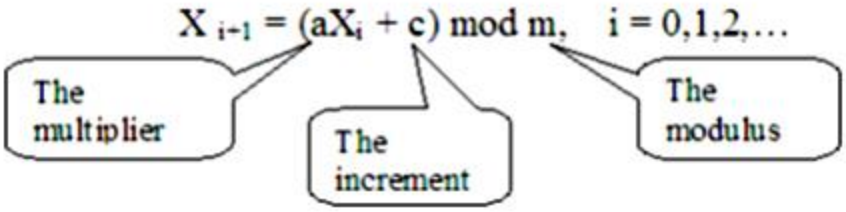
4. The random number generator should be fast and cost-effective.
5. It should be portable to different computers and ideally to a different programming language.

Techniques for Generating Random Numbers:

The most widely used techniques for generating random numbers are:

1. Linear Congruential Method (LCM):

The most widely used technique for generating random numbers, initially proposed by Lehmer [1951]. This method produces a sequence of integers, X1, X2 … between 0 and m-1 by following a recursive relationship:



The initial value X0 is called seed. The selection of the values for a, c, m, and X_0 drastically affects the statistical properties and the cycle length.

a. If $c \neq 0$ in the above equation, then it is called as Mixed Congruential method.

b. If c = 0 the form is known as the Multiplicative Congruential method.

The random numbers (Ri) between 0 and 1 can be generated by

$$R_i = \frac{X_i}{m}, \quad i = 1,2,\ldots$$

Example:

Use linear congruential method to generate sequence of random numbers with X0 = 27, a = 17, c = 43, and m = 100.

Solution:

Random numbers (Ri)

The random integers (Xi) generated will be between the range 0 - 99

Equations → Xi+1 = (a Xi + c) mod m, Ri = Xi / m , i=1,2,…..

X1 = (17 * 27 + 43) mod 100 = 2, R1 = 2 / 100 = 0.02

X2 = (17 * 2 + 43) mod 100 = 77, R2 = 77 / 100 = 0.77

X3 = (17 * 77 + 43) mod 100 = 52, R3 = 52/ 100 = 0.52

Hence the numbers are generated.

The secondary properties to generate random numbers include maximum density and maximum period.

a. Maximum Density:

Maximum Density means values assumed by Ri, i = 1, 2… leave no large gaps on the interval [0, 1].

Problem: The values generated from Ri = Xi / m, is discrete on integers instead of continuous.

Solution: A very large integer for modulus m.

b. Maximum Period:

To achieve Maximum density and avoid cycling, the generator should have the largest possible period. Most digital computers use a binary representation of numbers. Speed and efficiency is aided by a modulus m, to be (or close to) a power of 2. The maximal period is achieved by proper choice of a, c, m and X0.

The Different Cases Are:

1. For *m* a power of 2, say $m = 2^b$ and $c \neq 0$, the longest possible period is $P = m = 2^b$, provided that c is relatively prime to m and a = 1 + 4k, where k is an integer.

2. For *m* a power of 2, say $m = 2^b$ and c = 0, the longest possible period is $P = m / 4 = 2^{b-2}$, which is achieved provided that the seed X_0 is odd and the multiplier a, is given by a = 3 + 8k or a = 5 + 8k, for some k = 0, 1...

3. For *m* a prime number and c = 0, the longest the possible period is P = m - 1, which is achieved provided that 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:

Using the multiplicative congruential method, find the period of the generator for a = 13, m = 2^6 and X_0 = 1, 2, 3, and 4.

Solution:

c=0 (multiplicative congruential method), m = 2^6= 64 and a=13 → (a=5+8*1=13) so 'a' is in the form 5+8k with k=1.

Therefore the maximal period p= m / 4= 64 / 4=16 for odd seeds i.e. for X_0=1 and 3

Equation → $X_{i+1} = (aX_i + c) \bmod m$

When X_0 = 1, i = 1, X_2 = (13 * 1 + 0) mod 64 = 13 mod 64 = 13

When X_0 = 1, i = 2, X_3 = (13 * 13 + 0) mod 64 = 169 mod 64 = 41

When X_0 = 1, i = 3, X_4 = (13 * 41 + 0) mod 64 = 533 mod 64 = 21

When X_0 = 1, i = 16, X_{17} = (13 * 5 + 0) mod 64 = 65 mod 64 = 1

……

……

When X_0 = 2, i = 1, X_2 = (13 * 2 + 0) mod 64 = 26 mod 64 = 26

When X_0 = 2, i = 2, X_3 = (13 * 26 + 0) mod 64 = 338 mod 64 = 18

……

……

When X_0 = 2, i = 8, X_9 = (13 * 10 + 0) mod 64 = 130 mod 64 = 2

Similarly for X_0 =3 and 4 are calculated. The values are tabulated below in the table below

Therefore

For X_0=1, 3, maximal period is 16

For X_0=2, maximal period is 8

For X_0=4, maximal period is 4

i	X_i $X_0 = 1$	X_i $X_0 = 2$	X_i $X_0 = 3$	X_i $X_0 = 4$
0	1	2	3	4
1	13	26	39	52
2	41	18	59	36
3	21	42	63	20
4	17	34	51	4
5	29	58	23	
6	57	50	43	
7	37	10	47	
8	33	2	35	
9	45		7	
10	9		27	
11	53		31	
12	49		19	
13	61		55	
14	25		11	
15	5		15	
16	1		3	

Seed

2. Combined Linear Congruential Generators (CLCG):

As computing power increases, the complexity of the system to simulate also increases. So a longer period generator with good statistical properties is needed. One successful approach is to combine two or more multiplicative congruential generators.

Theorem:

If $W_{i,1}$, $W_{i,2}$,...,$W_{i,k}$ are any independent, discrete-valued random variables and $W_{i,1}$ is uniformly distributed on integers 0 to m_1 -2, then

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

is uniformly distributed on the integers 0 to m_1 -2.

To see how this the result can be used to form combined generators,

Let $X_{i,1}$, $X_{i,2}$ ... $X_{i,k}$ be i^{th} output from k different multiplicative congruential generators, where the j^{th} generator has prime modulus m_j and multiplier a_j is chosen so that the period is m_j -1. Then the j^{th} generator is producing $X_{i,j}$ that are approximately uniformly distributed on 1 to m_j -1 and $W_{i,j} = X_{i,j}$ -1 is approximately uniformly distributed on 0 to m_j -2.

Therefore the combined generator of the form,

$$X_i = \left(\sum_{j=1}^{k} (-1)^{j-1} X_{ij} \right) \bmod m_1 - 1 \qquad \text{Hence, } R_i = \begin{cases} \frac{X_i}{m_1}, & X_i > 0 \\ \frac{m_1 - 1}{m_1}, & X_i = 0 \end{cases}$$

The maximum possible period for a generator is

$$P = \frac{(m_1 - 1)(m_2 - 1)\ldots(m_k - 1)}{2^{k-1}}$$

Note: $(-1)^{j-1}$ coefficient implicitly performs the subtraction $X_{i,1} - 1$

Example:

For 32-bit computers, L'Ecuyer [1988] suggests combining k = 2 generators with m_1 = 2,147,483,563, a_1 = 40,014, m_2 = 2,147,483,399 and a_2 = 40,692. This leads to the following algorithm:

Step 1: Select Seeds

$X_{0,1}$ in the range [1 - 2,147,483,562] for the 1st generator

$X_{0,2}$ in the range [1 - 2,147,483,398] for the 2nd generator

Set i=0

Step 2: For each individual generator, evaluate

$X_{i+1,1} = 40{,}014\ X_{i,1} \bmod 2{,}147{,}483{,}563$

$X_{i+1,2} = 40{,}692\ X_{i,2} \bmod 2{,}147{,}483{,}399$

Step 3:

$X_{i+1} = (X_{i+1,1} - X_{i+1,2}) \bmod 2{,}147{,}483{,}562$

Step 4: Return

$$R_{i+1} = \begin{cases} \frac{X_{i+1}}{2{,}147{,}483{,}563}, & X_{i+1} > 0 \\ \frac{2{,}147{,}483{,}562}{2{,}147{,}483{,}563}, & X_{i+1} = 0 \end{cases}$$

Step 5:

Set i = i+1, go back to step 2.

The combined generator has period: $(m_1-1)(m_2-1)/2 \approx 2 \times 10^{18}$

Tests for Random Numbers:

The two main properties of random numbers are uniformity and independence.

1. Testing for Uniformity:

The hypotheses are as follows
H0 : Ri ~ U [0, 1]
H1: Ri ≁ U [0, 1]

The null hypothesis H0, reads that the numbers are distributed uniformly on the interval [0, 1]. Rejecting the null hypothesis means that the numbers are not uniformly distributed.

2. Testing for Independence:

The hypotheses are as follows
H0: Ri ~ independently
H1: Ri ≁ independently

This null hypothesis, H0, reads that the numbers are independent. Rejecting the null hypothesis means that the numbers are not independent. This does not imply that further testing of the generator for independence is unnecessary.

For each test, a level of significance α must be stated.

$$\text{Level of significance } \alpha = \frac{\text{probability of rejecting the test}}{\text{probability of accepting the test}}$$

= P(reject H_0 | H_0 true)
Frequently, α is set to 0.01 or 0.05.

There are five types of tests. The first is concerned for testing the uniformity whereas second through five with testing for independence.

1. Frequency Test: Compares the distribution of a set of numbers generated to a uniform distribution by using the Kolmogorov-Smirnov or the chi-square test.

2. Runs Test: Tests the runs up and down or the runs above and below the mean by comparing the actual values to expected values. The statistic for comparison is the chi-square test.

3. Autocorrelation Test: The correlation between numbers is tested 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 size of gaps.

5. Poker Test: Treats the numbers grouped together as a poker hand. Then the hands obtained are compared to what is expected using the chi-square test.

Frequency Tests:

The fundamental test performed to validate a new generator is the test for uniformity. The two different methods of testing are:

1. Kolmogorov-Smirnov Test:

It compares the continuous cumulative distribution function (cdf) of the uniform distribution with the empirical cdf, of the N sample observations. The cdf of an empirical distribution is a step function with jumps at each observed value.

Notations Used:

F(x) →Continuous cdf

SN(x) → Empirical cdf

N →Total number of observations

R1, R2 ...RN → Samples from Random generator

D → Sample statistic

Dα→ Critical value

By definition,

$$F(x) = x, \quad 0 \le x \le 1$$

$$S_N(x) = \frac{\text{number of } R_1, R_2 \ldots R_n \text{ which are } \le x}{N}$$

As N becomes larger, SN(x) ≈ F(x).

Maximum deviation over the range of a random variable is given by

D = max | F(x) – SN(x) |

The sampling distribution of D is known and is tabulated as a function of N in the table below.

Procedure For Testing Uniformity Using the Kolmogorov-Smirnov Test:

Step 1: Rank the data from smallest to largest. Let R(i) denote the ith smallest observation, so that

R (1) ≤ R (2) ≤ ….. ≤ R (N)

Step 2: Compute

D+ = max {(i / N) - R (i)}

1 ≤ i ≤ N

D- = max {R (i) – [(i – 1)/ N]}

1 ≤ i ≤ N

Step 3: Compute

D= max (D+, D-)

Step 4: Determine the critical value Dα, from the table A.8 for the specified significance level α and the given sample size N.

Step 5:

a. If D > Dα, the null hypothesis that the data are a sample from a uniform distribution is rejected.

b. If D ≤ Dα then there is no difference detected between the true distribution of {R1, R2 …RN} and the uniform distribution. So it is accepted.

Example:

Suppose 5 generated numbers are 0.44, 0.81, 0.14, 0.05, and 0.93. It is desired to perform a test for uniformity using the Kolmogorov-Smirnov test with a level of significance α = 0.05.

Solution

N=5 , i = 1, 2, 3, 4, 5

Arrange R_i from smallest to largest

Step 1 -	R_i	0.05	0.14	0.44	0.81	0.93
	i/N	0.20	0.40	0.60	0.80	1.00
Step 2 -	$i/N - R_i$	0.15	0.26	0.16	-	0.07
	$R_i - [(i-1)/N]$	0.05	-	0.04	0.21	0.13

$D^+ = \max \{i/N - R_i\}$

$D^- = \max \{R_i - [(i-1)/N]\}$

Step 3- $D = \max (D^+, D^-) = 0.26$

Step 4- For $\alpha = 0.05$, $N = 5$

$D_\alpha = D_{0.05} = 0.565$ (from table A.8)

$D < D_\alpha \rightarrow 0.26 < 0.565$

Therefore H_0 is not rejected, i.e. no difference between the distribution of generated numbers and the uniform distribution.

The calculations in the above table are depicted in the figure below, where empirical cdf SN(x) is compared to uniform cdf F(x). It is seen that D+ is the largest deviation of SN(x) above F(x) and D- is the largest deviation of SN(x) below F(x).

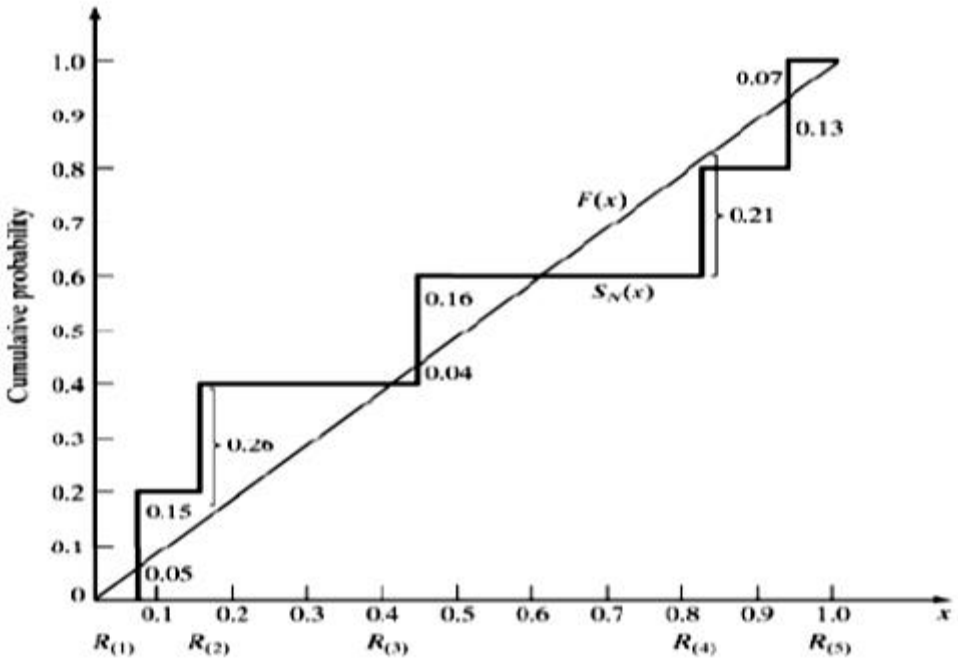


Fig: Comparison of F(x) and $S_N(x)$

2. Chi-Square Test:

It uses the sample statistic.

$$X_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where Oi → observed number in ith class

Ei → expected a number in ith class

n → number of classes

For uniform distribution, Ei is given by

$$E_i = \frac{N}{n}.$$

It can be shown that the sampling distribution X_0^2 is approximately the chi-square distribution with n-1 degrees of freedom (i.e., $X_0^2 \leq X_\alpha^2, n-1$).

Example:

Use a chi-square test with α=0.05 to test whether the data shown below are uniformly distributed.

0.34	0.90	0.25	0.89	0.87	0.44	0.12	0.21	0.46	0.67
0.83	0.76	0.79	0.64	0.70	0.81	0.94	0.74	0.22	0.74
0.96	0.99	0.77	0.67	0.56	0.41	0.52	0.73	0.99	0.02
0.47	0.30	0.17	0.82	0.56	0.05	0.45	0.31	0.78	0.05
0.79	0.71	0.23	0.19	0.82	0.93	0.65	0.37	0.39	0.42
0.99	0.17	0.99	0.46	0.05	0.66	0.10	0.42	0.18	0.49
0.37	0.51	0.54	0.01	0.81	0.28	0.69	0.34	0.75	0.49
0.72	0.43	0.56	0.97	0.30	0.94	0.96	0.58	0.73	0.05
0.06	0.39	0.84	0.24	0.40	0.64	0.40	0.19	0.79	0.62
0.18	0.26	0.97	0.88	0.64	0.47	0.60	0.11	0.29	0.78

Solution:

Let n=10, the interval [0-1] divided in equal lengths, (0.01-0.10), (0.11-0.20), ---, (0.91-1.0)

N = 100

E_i=N/n=100/10=10

The calculations are tabulated below in table below

$X^2_{0.05,9}$ =16.9 (check the table A.6 –using α, n-1)

$$X_0^2 < X^2_{0.05,9} = 3.4 < 16.9$$

Therefore the null hypothesis of the uniform distribution is not rejected.

Interval	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$X_0^2 = (O_i-E_i)^2 / E_i$
0.01 - 0.10	8	10	-2	4	0.4
0.11 - 0.20	8	10	-2	4	0.4
0.21 - 0.30	10	10	0	0	0.0
0.31 - 0.40	9	10	-1	1	0.1
0.41 - 0.50	12	10	2	4	0.4
0.51 - 0.60	8	10	-2	4	0.4
0.61 - 0.70	10	10	0	0	0.0
0.71 - 0.80	14	10	4	16	1.6
0.81 - 0.90	10	10	0	0	0.0
0.91 - 1.00	11	10	1	1	0.1
	100	100	0		$X_0^2 = 3.4$

Note:

a. In general, for any value chooses 'n' such that $E_i \geq 5$.

b. Kolmogorov-Smirnov test is more powerful than the chi-square test because it can be applied to small sample sizes, whereas chi-square requires large sample, say $N \geq 50$.

Runs Tests:

Run - The succession of similar events preceded and followed by a different event is called as run.
Run-length - Number of events that occur in the run.

Example: Tossing coin
Consider the sequence of tossing a coin 10 times: H T T H H T T T H T

No.	Run Length	Run
1	1	H
2	2	T T
3	2	H H
4	3	T T T
5	1	H
6	1	T

There are two possible concerns in run tests. They are
1. Number of runs- Run-up and down & Runs above and below mean
2. Length of runs

1. Runs Up And Down:

a. **Up run**-Sequence of numbers each of which is succeeded by a larger number is called as up run.

b. **Down run**-Sequence of numbers each of which is succeeded by smaller number is called as down run.

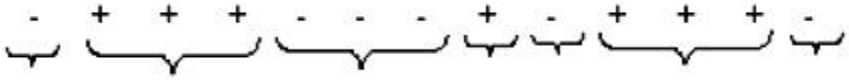
c. If a number is followed by a larger number then it denoted by '+'. If followed by a smaller number then by '-'.

To illustrate the above, consider the sequence of numbers
0.87 0.15 0.23 0.45 0.69 0.32 0.30 0.19 0.24 0.18 0.65 0.82 0.93 0.22

The up run and down run are marked as

-0.87 +0.15 +0.23 +0.45 -0.69 -0.32 -0.30 +0.19 -0.24 +0.18 +0.65 +0.82 -0.93 +0.22

The sequence of '+' and '-'are



It has 7 runs, first run of length one, second run of length three, third run of length 3, and fourth run with one, fifth run with one, sixth run with three and seventh run with one. There are three up runs and four down runs. If N is several numbers in sequence, then maximum numbers of runs are N-1 and a minimum number of runs is one. If 'a' is the total number of runs in a random sequence, Mean is given by

$$\mu_a = \frac{(2N-1)}{3}$$

Variance,

$$\sigma_a^2 = \frac{16N - 29}{90}$$

For N > 20, the distribution of 'a' is reasonably approximated by a normal distribution, $N(\mu a, \sigma a^2)$. This approximation is used to test the independence of numbers from a generator. The test statistic is obtained by subtracting the mean from the observed number of runs 'a' and dividing by standard deviation, i.e. Test statistic is given by,

$$Z_0 = \frac{a - \mu_a}{\sigma_a}$$

Substituting μ_a and σ_a in above equation, we get

$$Z_0 = \frac{a - [(2N-1)/3]}{\sqrt{[(16N-29)/90]}}$$

Where $Z_0 \sim N(0, 1)$

The null hypothesis is accepted when $-Z\alpha/2 \leq Z_0 \leq Z\alpha/2$, where α is the level of significance. The critical values and rejection region is shown in the figure below.

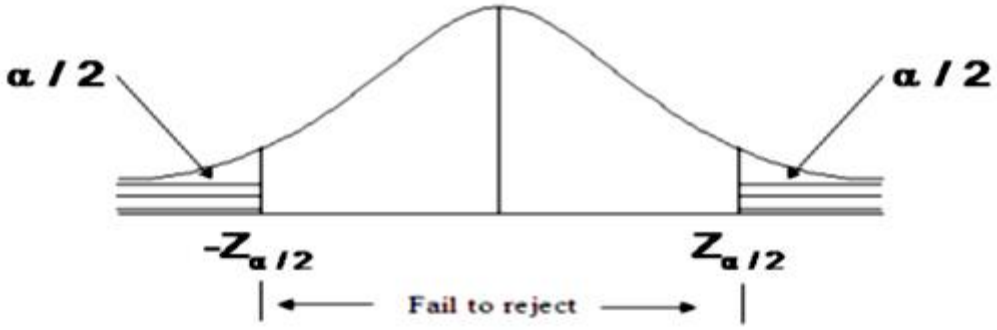


Fig: Accept the Null Hypothesis

Example:

Based on runs up and runs down, determine whether the following sequence of 40 numbers is such that the hypothesis of independence can be rejected or accepted where $\alpha = 0.05$.

0.41	0.68	0.89	0.94	0.74	0.91	0.55	0.62	0.36	0.27
0.19	0.72	0.75	0.08	0.54	0.02	0.01	0.36	0.16	0.28
0.18	0.01	0.95	0.69	0.18	0.47	0.23	0.32	0.82	0.53
0.31	0.42	0.73	0.04	0.83	0.45	0.13	0.57	0.63	0.29

Solution:

The sequence of runs up and down is as follows:

+ + + - + - + - - - + + - +

- - + - + - - + - - + - + +

- - + + - + - - + + -

No. of runs → a = 26

N = 40

$\mu a = \{2(40) - 1\} / 3 = 26.33$

$\sigma a^2 = \{16(40) - 29\} / 90 = 6.79$

$Z_0 = (26 - 26.33) / \sqrt{(6.79)} = -0.13$

Critical value → $Z\alpha/2 \rightarrow Z_{0.025} = 1.96$ (from z - table)

$Z\alpha/2 \leq Z_0 \leq Z\alpha/2 \rightarrow -1.96 \leq -0.13 \leq 1.96$

Therefore independence of the numbers cannot be rejected, we accept the null hypothesis.

Disadvantage Of Runs Up And Down

a. Insufficient to review the independence of a group of numbers

2. Runs Above And Below The Mean

Runs are described with above/below the mean value. A '+' sign is used to indicate above mean and '-' sign for below the mean.

To illustrate the above, consider the sequence of 2-digit random numbers

0.40	0.84	0.75	0.18	0.13	0.92	0.57	0.77	0.30	0.71
0.42	0.05	0.78	0.74	0.68	0.03	0.18	0.51	0.10	0.37

Mean = (0.99+0.00)/2 = 0.495

The runs above and below mean are marked as

-0.40 +0.84 +0.75 -0.18 -0.13 +0.92 +0.57 +0.77 -0.30 +0.71

-0.42 -0.05 +0.78 +0.74 +0.68 -0.03 -0.18 +0.51 -0.10 -0.37

The sequence of '+' and '-'are

- + + - - + + + - + - - +

+ + - - + - -

There are 11 runs, of which 5 are above mean and 6 runs below mean.

Let n_1 → No. of individual observations above mean

n_2 → No. of individual observations below mean

b → Total number of runs

N→ Maximum number of runs, where $N = n_1 + n_2$

The mean is given by

$$\mu_b = \frac{2n_1n_2}{N} + \frac{1}{2}$$

Variance

$$\sigma_b^2 = \frac{2n_1n_2\,(2n_1n_2 - N)}{N^2\,(N-1)}$$

For either n_1 or n_2 greater than 20, b is approximately normally distributed. The test statistic is obtained by subtracting the mean from several runs 'b' and dividing by the standard deviation i.e.

$$Z_0 = \frac{b - (2n_1n_2/N) - 1/2}{\left(\frac{2n_1n_2\,(2n_1n_2 - N)}{N^2\,(N-1)}\right)^{1/2}}$$

The null hypothesis is accepted when $-Z\alpha/2 \leq Z0 \leq Z\alpha/2$, where α is the level of significance.

Example:

Based on runs above and below mean, determine whether the following sequence of 40 numbers is such that the hypothesis of independence can be rejected or accepted where α = 0.05.

0.41	0.68	0.89	0.94	0.74	0.91	0.55	0.62	0.36	0.27
0.19	0.72	0.75	0.08	0.54	0.02	0.01	0.36	0.16	0.28
0.18	0.01	0.95	0.69	0.18	0.47	0.23	0.32	0.82	0.53
0.31	0.42	0.73	0.04	0.83	0.45	0.13	0.57	0.63	0.29

Solution:

Mean= 0.495

The sequence of runs above and below mean is as follows:

-	+	+	+	+	+	+	+	-	-
-	+	+	-	+	-	-	-	-	-
-	-	+	+	-	-	-	-	+	+
-	-	+	-	+	-	-	+	+	-

$n_1 = 18$

$n_2 = 22$

$N = n_1 + n_2 = 40$

$b = 17$

$\mu b = [\{2(18)\ (22)\} / 40] + (1 / 2) = 20.3$

$\sigma b2 = [2\ (18)\ (22)\ \{(2)\ (18)\ (22) - 40\}] / [(40)2\ (40 - 1)] = 9.54$

Since $n_2 > 20$, normal approximation is accepted.

$Z_0 = (17 - 20.3) / \sqrt{(9.54)} = -1.07$

Critical value $\rightarrow Z\alpha/2 \rightarrow Z_{0.025} = 1.96$ (from z - table)

$-Z\alpha/2 \leq Z_0 \leq Z\alpha/2 \rightarrow -1.96 \leq -1.07 \leq 1.96$

Therefore hypothesis of independence cannot be rejected based on this test.

Disadvantage Of Runs Above And Below Mean

a. If two numbers are below mean, two numbers are above mean and so on. Then the numbers are dependent.

3. Runs Test: Length Of Runs

Let Y_i be the number of runs of length i, in a sequence of N numbers. For an independent sequence,

The expected value of Yi for runs up and down is given by

$$E(Y_i) = \frac{2}{(i+3)!}[N(i^2+3i+1)-(i^3+3i^2-i-4)], \quad i \leq N-2$$

$$E(Y_i) = \frac{2}{N!}, \quad i = N-1$$

For runs above and below mean, the expected value of Y_i is given by

$$E(Y_i) = \frac{Nw_i}{E(I)}, \qquad N > 20$$

Where w_i, the approximate probability that a run has length i, is given by

$$w_i = \left(\frac{n_1}{N}\right)^i \left(\frac{n_2}{N}\right) + \left(\frac{n_1}{N}\right)\left(\frac{n_2}{N}\right)^i, \qquad N > 20$$

And E (I), the approximate expected length of a run, is given by

$$E(I) = \frac{n_1}{n_2} + \frac{n_2}{n_1}, \qquad N > 20$$

The approximate expected total number of runs (of all lengths) E (A), is given by

$$E(A) = \frac{N}{E(I)}, \quad N > 20$$

The appropriate test is chi-square test with O_i, the observed number of runs of length i.
The test statistic is given by

$$X_0^2 = \sum_{i=1}^{L} \frac{[O_i - E(Y_i)]^2}{E(Y_i)}$$

Where L = N − 1 for runs up and down
L = N for runs above and below mean.
If null hypothesis of independence is true then X_0^2 is approximately chi-squared distributed with L-1 degrees of freedom.

Example:

Given the sequence of numbers, can the hypothesis that the numbers are independent be rejected on the basis of length of runs up and down at $\alpha = 0.05$?

0.30	0.48	0.36	0.01	0.54	0.34	0.96	0.06	0.61	0.85
0.48	0.86	0.14	0.86	0.89	0.37	0.49	0.60	0.04	0.83
0.42	0.83	0.37	0.21	0.90	0.89	0.91	0.79	0.57	0.99
0.95	0.27	0.41	0.81	0.96	0.31	0.09	0.06	0.23	0.77
0.73	0.47	0.13	0.55	0.11	0.75	0.36	0.25	0.23	0.72
0.60	0.84	0.70	0.30	0.26	0.38	0.05	0.19	0.73	0.44

Solution:

N = 60

The sequence of + and – are as follows

+	-	-	+	-	+	-	+	+	-	+	-
+	+	-	+	+	-	+	-	+	-	-	+
-	+	-	-	+	-	-	+	+	+	-	-
-	+	+	-	-	-	+	-	+	-	-	-
+	-	+	-	-	-	+	-	+	+	-	

The length of runs in the sequence is as follows

1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 2, 1, 1,

1, 2, 1, 2, 3, 3, 2, 3, 1, 1, 1, 3, 1, 1, 1, 3, 1, 1, 2, 1

Calculate O_i

Run Length, i	**1**	**2**	**3**	**4**
Observed Runs, O_i	26	9	5	0

The expected value of Yi,

For run length one,

$$E(Y_1) = \frac{2}{(1+3)!}[60(1^2 + 3(1) + 1) - (1^3 + 3(1)^2 - 1 - 4)] = 25.08$$

Run length two,

$$E(Y_2) = \frac{2}{(2+3)!}[60(2^2 + 3(2) + 1) - (2^3 + 3(2)^2 - 2 - 4)] = 10.77$$

Run length three

$$E(Y_3) = \frac{2}{(3+3)!}[60(3^2 + 3(3) + 1) - (3^3 + 3(3)^2 - 3 - 4)] = 3.04$$

$$\therefore \quad E(Y_1) + E(Y_2) + E(Y_3) = 38.89$$

We find mean (runs up and down)

$$\mu_a = \frac{2N - 1}{3} = \frac{2(60) - 1}{3} = 39.67$$

Expected value, when $i \geq 4$

$$\mu_a - \sum_{i=1}^{3} E(Y_i) = 39.67 - 38.89 = 0.78$$

To find X_0^2, the calculations and procedures are shown in table below:

Run length (i)	Observed number of runs (O_i)	Expected number of runs $E(Y_i)$	$\frac{[O_i - E(Y_i)]^2}{E(Y_i)}$
1	26	25.08	0.03
2	9	10.77	
3	5 } 14	3.82 } 14.59	} 0.02
4	0	0.78	
-	40	39.67	$X_0^2 = 0.05$

$$X^2_{0.05,\,1} = 3.84$$

$$X_0^2 < X^2_{0.05,\,1} = 0.05 < 3.84$$

Therefore hypothesis of independence is accepted.

Example:

Given the sequence of numbers can the hypothesis that the numbers are independent be rejected on the basis of length of runs above and below mean at $\alpha = 0.05$?

0.30	0.48	0.36	0.01	0.54	0.34	0.96	0.06	0.61	0.85
0.48	0.86	0.14	0.86	0.89	0.37	0.49	0.60	0.04	0.83
0.42	0.83	0.37	0.21	0.90	0.89	0.91	0.79	0.57	0.99
0.95	0.27	0.41	0.81	0.96	0.31	0.09	0.06	0.23	0.77
0.73	0.47	0.13	0.55	0.11	0.75	0.36	0.25	0.23	0.72
0.60	0.84	0.70	0.30	0.26	0.38	0.05	0.19	0.73	0.44

Solution

N = 60

Mean = (0.99+0.00)/2 = 0.495

The sequence of + and – are as follows

-	-	-	-	+	-	+	-	+	+	-	+	-
+	+	-	-	+	-	+	-	+	-	-	+	+
+	+	+	+	+	-	-	+	+	-	-	-	-
+	+	-	-	+	-	+	-	-	-	+	+	+
+	-	-	-	-	-	+	-					

$n_1 = 28$

$n_2 = 32$

$N = n_1 + n_2 = 60$

The length of runs in the sequence is as follows

4, 1, 1, 1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 2, 7, 2, 2, 4, 2, 2, 1, 1, 1, 3, 4, 5, 1, 1

Calculate Oi

Run Length, i	**1**	**2**	**3**	**≥ 4**
Observed Runs, O_i	17	8	1	5

The probabilities of runs of various lengths w_i are as follows

$$w_1 = \left(\frac{28}{60}\right)^1 \frac{32}{60} + \frac{28}{60}\left(\frac{32}{60}\right)^1 = 0.498$$

$$w_2 = \left(\frac{28}{60}\right)^2 \frac{32}{60} + \frac{28}{60}\left(\frac{32}{60}\right)^2 = 0.249$$

$$w_3 = \left(\frac{28}{60}\right)^3 \frac{32}{60} + \frac{28}{60}\left(\frac{32}{60}\right)^3 = 0.125$$

The expected number of runs of various lengths is

$$E(Y_1) = \frac{N w_1}{E(I)} = \frac{60(0.498)}{2.02} = 14.79$$

$$E(Y_2) = \frac{N w_2}{E(I)} = \frac{60(0.249)}{2.02} = 7.40$$

$$E(Y_3) = \frac{N w_3}{E(I)} = \frac{60(0.125)}{2.02} = 3.71$$

The expected total number of runs is

$$E(A) = \frac{N}{E(I)} = \frac{60}{2.02} = 29.7$$

For $i \geq 4$,

$$E(A) - \sum_{i=1}^{3} E(Y_i) = 29.7 - 25.9 = 3.8$$

To find X_0^2 the calculations and procedures are shown in table below:

Run length (i)	Observed number of runs (O_i)	Expected number of runs $E(Y_i)$	$\frac{[O_i - E(Y_i)]^2}{E(Y_i)}$
1	17	14.79	0.33
2	8	7.40	0.05
3	1 } 6	3.71 } 7.51	} 0.30
≥4	5 }	3.80 }	
-	31	29.70	$X_0^2 = 0.68$

$$X^2_{0.05,2} = 5.99$$

$$X_0^2 < X^2_{0.05,2} = 0.68 < 5.99$$

Therefore the hypothesis of independence is accepted.

4. Test For Autocorrelation:

The uniformity test of random numbers is only a necessary test for randomness, not a sufficient one. A sequence of numbers may be perfectly uniform and still not random. For example the sequence 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.1, 0.2, 0.3, …, … would give a perfectly uniform distribution with chi-square value perfectly as zero. But the sequence can be no means be regarded as random. The numbers are not independent as the occurrence of one number say 0.3 decides the next, which is to be 0.4, etc. This defect is called serial autocorrelation of an adjacent pair of numbers.

The chi-square test for serial autocorrelation makes use of a 10 * 10 matrix. The 10 class describe in the uniformity test are represented both along the rows and columns. If the classes are to be represented on a bar chart, 100 bars, one for each cell of a matrix will be required. To reduce the number of groups instead of 10 random numbers are divided into a smaller number of a class as 3 or 4. Three class will be as:

a. Less than or equal to 0.33
b. Less than or equal to 0.67
c. Less than or equal to 1.0

With three classes in a row and three classes in a column, there will be 9 cells.