CHAPTER 6

Module 3

Random Number Generation

6.1 Introduction

Q. Why is it necessary to test the properties of random numbers ? How would you generate random numbers ?

MU - Dec. 12

Q. Why random numbers used in simulation ? What are techniques used to generate them ?

MU - May 13

- In simulation, random numbers are used in order to introduce randomness in the model. For instance, let us consider for a moment the machine interference simulation model. In this model it was assumed that the operational time of a machine was constant. Also, it was assumed that the repair time of a machine was constant. It is possible that one may identify real-life situations where these two assumptions are valid.
- However, in most of the cases one will observe that the time a machine is operational varies. Also, the repair time may vary from machine to machine. If we are able to observe the operational times of a machine over a reasonably long period, we will find that they are typically characterized by a theoretical or an empirical probability distribution.
- Similarly, the repair times can be also characterized by a theoretical or empirical distribution. Therefore in order to make the simulation model more realistic, one should be able to randomly numbers that follow a given theoretical or empirical distribution.
- In order to generate such random numbers one needs to be able to generate uniformly distributed random numbers, otherwise known as pseudo-random numbers. These pseudo-random numbers can be used either by themselves or they can be used to generate random numbers from different theoretical or empirical distributions, known as random variates or stochastic variates.
- Random numbers are essentials and an important constituent of the mathematical modeling. They are used in simulation of all the discrete systems.

- There are many reasons why we might want to use random numbers. For centuries, people have used randomness, a coin toss, for instance, to make important decisions. Of course, many states run lotteries that use a sequence of random numbers to determine the recipients of huge amounts of money.
- Random numbers are used to form encryption keys when information needs to be passed securely, say, across the Internet. Random sequences of numbers are also important for simulating real-world phenomena.
- For instance, a bank that operates a network of ATM machines may wish to test its software by simulating the actions of customers accessing their accounts at random times through random machines.
- The variability and the uncertainty inherent in the systems studied are modeled using random inputs.
- They are also used in video gambling games, military draft, assigning subjects to treatments in a pharmaceutical experiment, state lotteries and pairing teams in a sports tournaments, etc.

Random Variable as a Measurement

- Think of much more complicated experiments :
  - A chemical reaction.
  - A laser emitting photons.
  - A packet arriving to an IMP.
- We cannot give an exact description of a sample space in these cases, but we can still describe specific measurements on them :
  - The temperature change produced.
  - The number of photons emitted in one millisecond.
  - The time of arrival of the packet.
- Thus a random variable can be thought of as a measurement on an experiment as shown in Fig. 6.1.1.

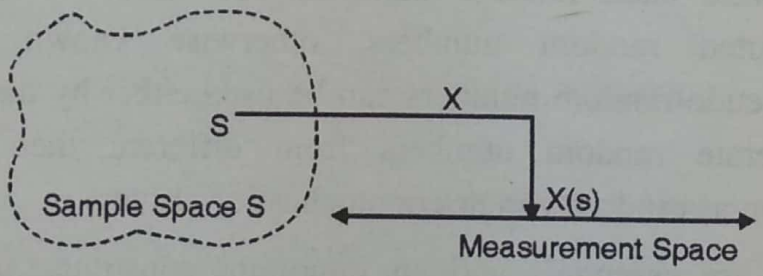

Fig. 6.1.1

Why to use random number ?

There are two reasons for introducing random variables. First, for most real-life system, we do not have exact characterizations of the input parameters. Hence, using probabilistic inputs makes the results of the analysis more robust. Second, even if we do have an exact characterization of the input parameters, it is often computationally too expensively or analytically intractable to take them into account.

6.2 Properties of Random Numbers

Q. State the properties of random numbers.
MU - May 05, Dec. 05, May 06, May 08, Dec. 08, May 09

Q. Why is it necessary to test the properties of random numbers ? How would you generate random numbers ?
MU - Dec. 12

Q. Explain the properties of random numbers.
MU - Dec. 10, May 14, May 15

- An unpredictable sequence of number is called random numbers. Random numbers are of two types : Pseudo random number and True random number.
- The random number which is generated with help of any known mathematical formula or algorithm is called as pseudo random number. Example: Linear Congruential Method.
- The random number which is not generated with the help of known mathematical formula or algorithm is called as true random number. Example : Number from any existing table such as log table.
- Each random number R_i must be an independent sample drawn from a continuous uniform distribution between 0 and 1. Hence the PDF is given by :

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

- The expected value of each R_i is given by :

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

- The variance is given by :

$$V(R) = E(R^2) - [E(R)]^2 = \int_0^1 x^2\,dx - [E(R)]^2 = x^3/3|_0^1 = 1/12$$

- Two important statistical properties that a sequence of random numbers must follow are independence and uniformity.
    - **Independence :** The probability of observing a value in a particular interval is independent of the previous values drawn.
    - **Uniformity :** If the interval [0, 1] is divided into n sub intervals of equal length, the expected number of observation in each interval is N/n, where N is the total number of observation. Each random number R_i is a sample drawn from a uniform distribution.
- Random number generators use iterative deterministic algorithms for producing a sequence of pseudo-random numbers that approximate a truly random sequence. Ideally the sequence should be uniformly distributed, uncorrelated, reproduceable, portable, easily changed by adjusting an initial seed value, easily split into many independent subsequences, have a large period of repetition, pass all empirical tests for randomness, and be generated rapidly using limited computer memory.
- Random number generators are widely used for simulations in computational science and engineering. Randomness is often present in the formulation of the problem, for example random noise or perturbations, and quantum processes. Since "random" numbers are in practice computed using deterministic algorithms, these are more accurately called pseudo-random number generators.

6.3 Generation of Pseudo Random Numbers

Q. Why is it necessary to test the properties of random numbers ? How would you generate random numbers ? **MU - Dec. 12**

Q. How are random numbers generated? **MU - May 15**

A standard pseudo-random number generator aims to produce a sequence of real random numbers that are uncorrelated and uniformly distributed in the interval [0, 1]. Such a generator can also be used to produce random integers and sequences with a probability distribution that is not uniform.

- Pseudo means false, hence this involves the generation of false random numbers.
- Generating numbers using a known method removes the potential for true randomness. That is why they are called "pseudo : random numbers".
- The goal is to produce a sequence of numbers in [0, 1] that simulations, or imitates, the ideal properties of random numbers.

- Some errors/ departure from randomness in generating pseudo – random numbers are :
  - The generated numbers may not be uniformly distributed.
  - The generated numbers may be discrete valued instead of continuous valued.
  - The mean may be too high or too low.
  - The variance may be too high or too low.

Essential Properties of a Random Number Generator

- **Repeatability :** The same sequence should be produced with the same initial values (or seeds). This is vital for debugging.
- **Randomness :** Should produce independent uniformly distributed random variables that pass all statistical tests for randomness.
- **Long period :** A pseudo-random number sequence uses finite precision arithmetic, so the sequence must repeat itself with a finite period. This should be much longer than the amount of random numbers needed for the simulation.
- **Insensitive to seeds :** Period and randomness properties should not depend on the initial seeds.

Further Properties of a Good Random Number Generator

- **Portability :** Should give the same results on different computers.
- **Efficiency :** Should be fast (small number of floating point operations) and not use much memory.
- **Disjoint subsequences :** Different seeds should produce long independent (disjoint) subsequences so that there are no correlations between simulations with different initial seeds.
- **Homogeneity :** Sequences of all bits should be random.

6.4 Techniques for Generating Random Numbers

Q. What are the techniques are used to generate random number ?
MU - May 05, May 07, Dec. 07, May 08, Dec. 08, May 09

Q. Why random numbers used in simulation ? What are techniques used to generate them ?
MU - Dec. 10, May 11, May 13

Q. What are the methods used to generate random numbers?
MU - May 14, Dec. 14

Random number generators

- While some sequences generated by natural phenomena produce truly random numbers, many applications require that we be able to create random numbers efficiently inside a computer. This may sound impossible : computers simply execute a set of instructions whose output is determined by the input. Since we supply the computer with the instructions and the input, the output is determined by our choices. How can such a number be random ?
- The answer is that it's not. Instead, our goal is to use a procedure that hides our footprints so that the numbers create the illusion of randomness. More precisely, we want the numbers to share many properties that we would expect a truly random sequence to enjoy. Such a procedure is often called a *pseudorandom number generator*, since the numbers generated are not truly random, though we will follow convention and use the term random number generator.
- John von Neumann proposed using the following method as one of the first random number generators. Suppose we want to create eight-digit numbers. Begin with an eight-digit number X_0, which we call the *seed*, and create the next integer in the sequence by removing the middle eight digits from X_0^2. For instance, if $X_0 = 35385906$, we find that

$$X_0^2 = 1252\mathbf{16234344}0836$$

- So that our next number is $X_1 = 16234344$. If we repeat this a few times we find :

X_0	16234344
X_1	55392511
X_2	33027488
X_3	81496359
X_4	65653025
X_5	31969165
X_6	02751079
X_7	56843566
X_8	19099559
X_9	79315399

- Since it is difficult, at first glance, to find a pattern in these numbers, we may think that this is an appropriate way to find random numbers. In other words, we have created the

illusion of randomness through a deterministic process. Further study shows, however, that this is not a good random number generator.

- Each term in the sequence depends only on its immediate predecessor and there are only a finite number of possible terms. This means that the sequence will inevitably repeat. The problem is that the sequence can get caught in relatively short cycles. For instance, if the number 31360000 appears in the sequence at some point, we end up with this number again after another 99 iterations and this cycle continues indefinitely.
- The middle square method may give you the idea to ask the computer to perform a sequence of many, unrelated, random operations.

6.4.1 Linear Congruential Method

- This is a very popular method and most of the available computer code for the generation of random numbers uses some variation of this method. The advantage of this congruential method is that it is very simple, fast and it produces pseudo-random numbers that are statistically acceptable for computer simulation.
- The congruential method uses the following recursive relationship to generate random numbers.

$$xi + 1 = axi + c \pmod m$$

where xi, a, c and m are all non-negative numbers. m is the modulus, c is an increment, a is a constant multiplier, x_i is initial value called the seed.

- Given that the previous random number was xi, the next random number xi + 1 can be generated as follows. Multiply xi by a and then add c. Then, compute the modulus m of the result. That is, divide the result by m and set xi + 1 equal to the remainder of this division.
- For example, if x0 = 0, a = c = 7, and m = 10 then we can obtain the following sequence of numbers : 7, 6, 9, 0, 7, 6, 9, 0,...
- The method using the above expression is known as the mixed congruential method.
- A simpler variation of this method is the multiplicative congruential method. If c = 0 then it is multiplicative congruential method, and if $c \neq 0$ then the form is called mixed congruential method. This method utilizes the relation xi + 1= axi(mod m).
- The numbers generated by a congruential method are between 0 and m – 1. Uniformly distributed random numbers between 0 and 1 can be obtained by simply dividing the resulting xi by m.

- The number of successively generated pseudo-random numbers after which the sequence starts repeating itself is called the period. If the period is equal to m, then the generator is said to have a full period.
- Theorems from number theory show that the period depends on m. The larger the value of m, the larger is the period.
- In particular, the following conditions on a, c, and m guarantee a full period :
  - m and c have no common divisor.
  - a = 1 (mod r) if r is a prime factor of m. That is, if r is a prime number (divisible only by itself and 1) that divides m, then it divides a – 1.
  - a = 1 (mod 4) if m is a multiple of 4.
- It is important to note that one should not use any arbitrary values for a, c and m. Systematic testing of various values for these parameters have led to generators which have a full period and which are statistically satisfactory. A set of such values is : a = 314, 159, 269, c = 453, 806, 245, and m = 232 (for a 32 bit machine).
- In order to get a generator started, we further need an initial seed value for x. It will become obvious later on that the seed value does not affect the sequence of the generated random numbers after a small set of numbers has been generated.
- The implementation of a pseudo-random number generator involves a multiplication, an addition and a division. The division can be avoided by setting m equal to the size of the computer word. For, if the total numerical value of the expression axi + c is less than the word size, then it is in itself the result of the operation axi + c (mod m), where m is set equal to the word size. Now, let us assume that the expression axi + c gives a number greater than the word size.
- In this case, when the calculation is performed, an overflow will occur. If the overflow does not cause the execution of the program to be aborted, but it simply causes the significant digits to be lost, then the remaining digits left in the register is the remainder of the division (axi + c)/m. This is because the lost significant digits will represent multiples of the value of m, which is the quotient of the above division.
- In order to demonstrate the above idea, let us consider a fictitious decimal calculator whose register can accommodate a maximum of 2 digits. Obviously, the largest number that can be held in the register is 99. Now, we set m equal to 100.
- For a = 8, x = 2 and c = 10, we have that axi + c = 26, and 26 (mod 100) = 26. However, if x = 20, then we have that axi + c = 170. In this case, the product axi (which is equal to 8×20) will cause an overflow to occur. The first significant digit will be lost and thus the register will contain the number 60.

- If we now add c (which is equal to 10) to the above result we will obtain 70, which is, the remaining of the division 170/100.
- Thus, the choice of the parameters i.e. a, c, m, X_0 drastically affects the statistical properties and cycle length.

6.4.2 Combined Linear Congruential Generators

- These days, longer period generator is needed because of the increasing complexity of simulated systems.
- The problem is overcome by choose multiple generators and combining them.
- The approach used is to combine two or more multiplicative congruential generators.
- Let $X_{i,1}, X_{i,2}, \ldots, X_{i,k}$, be the ith output from k different multiplicative congruential generators.
- The jth generator has prime modulus m_j and multiplier a_j and period is $m_j - 1$.
- Produces integers $X_{i,j}$ is approximately uniform on integers in $[1, m_j - 1]$ $W_{i,j} = X_{i,j} - 1$ is approximately uniform on integers in $[0, m_j - 2]$
- The combined generators take the form:

$$X_i = \left[\sum_{j=0}^{k} (-1)^{j-1} X_{ij} \right] \bmod (m_i - 1)$$

And R_i is :

$$R_i = X_i / m_1, X_i > 0$$
$$= (m_i - 1)/ m_1, \quad X_i = 0$$

- The maximum possible period for this generator is:

$$P = (m_1 - 1)(m_2 - 1).....(m_k - 1)] / (2^{k-1})$$

The algorithm is as follows :

i. Select seed $X_{i,0}$ in the range $[1, m_j - 1]$ for each k generator, such that i = 1 to k. Set j = 0, l = 0.

ii. Evaluate each individual generator. For i = 1 to k,

$$X_{j+1} = (X_{1,j+1} X_{2,j+1}) \bmod m_i$$

iii. Combine the generators,

$$X_{l+1} = \left[\sum_{j=0}^{k} (-1)^{j-1} X_{ij} \right] \bmod (m_i - 1)$$

iv. Return

$$R_{i+1} = X_{i+1}/m_1, \quad X_{i+1} > 0$$
$$= (m_1 - 1)/m_1, \quad X_{i+1} \leq 0$$

v. Set j = j + 1, go back to step(ii)

6.5 Test for Random Numbers

Q. Explain various test for random numbers. MU - Dec. 04

Ideally a pseudo-random number generator would produce a stream of numbers that

- are uniformly distributed,
- are uncorrelated,
- never repeats itself,
- satisfy any statistical test for randomness,
- are reproduceable (for debugging purposes),
- are portable (the same on any computer),
- can be changed by adjusting an initial "seed" value,
- can easily be split into many independent subsequences,
- can be generated rapidly using limited computer memory.

6.5.1 Hypothesis Testing

Q. State hypothesis for testing property of random number. MU - Dec. 07

- Hypothesis testing is used in statistics to decide whether a particular assumption is correct or not. This assumption is called a hypothesis, and it typically is an assertion about a distribution of one or more random variables, or about some measure of a distribution, such as the mean and the variance. The tested statistical hypothesis is called the null hypothesis, and it is denoted as H0. An alternative hypothesis, denoted as Ha, is also formulated. Hypothesis testing is used to test uniformity and independence properties of random numbers.
- The hypothesis for testing uniformity is :

$$H_0: R_i \sim U[0, 1]$$

$$H_1: R_i \nsim U[0, 1]$$

- Failure to reject the null hypothesis, H_0 means that evidence uniformity has been detected.

- The hypothesis for testing independence is:

$$H_0: R_i \sim \text{independently}$$

$$H_1: R_i \nsim \text{independently}$$

- Failure to reject the null hypothesis, H_0 means that evidence of independence has been detected.
- For each test, a level of significance must be stated

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

- For each test, a level of significance must be stated frequently, α is set to 0.01 or 0.05 (1%, 5% failure by chance)
- There are two errors associated with hypothesis testing, namely, type I error and
- Type II error : A type I error occurs when we reject the null assumption, whereas in fact it is correct (which of course we do not know). A type II error occurs when we accept the null hypothesis when in fact it is not correct. Since we do not know the truth, we do not know whether we have committed a type I or a type II error. The type I error is commonly known as a false negative, and the type II error is known as a false positive.
- Table 6.5.1 summarizes the type I and type II errors.

Table 6.5.1 : Type I and type II errors

Real situation	Decision	
	H_0 is not rejected	**H_0 is rejected**
H_0 is true	Valid	Type I error
H_0 is not true	Type II Error	Valid

- Type I Error (α) : It is the error of rejecting H_0 when in fact it is true.
- Type II Error (β) : It is the error of accepting H_0 when in fact it is false.

6.5.2 Testing of Uniformity

Frequency test. Uses the Kolmogorov-Smirnov or the chi-square test to compare the distribution of the set of numbers generated to a uniform distribution.

6.5.2.1 Kolmogorov - Smirnov Test

- It compares the continuous CDF, F(x), of the uniform distribution with the empirical CDF, $S_N(x)$, of the N sample observations.
- The continuous CDF of the uniform distribution $F(x) = x, 0 \leq x \leq 1$.

- If the sample from the R_N generator is $R_1, R_2, \ldots, R_N$, then the empirical CDF,

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

- It can be done in the following two methods :

Graphical method

- This test is based on the largest absolute deviation between f(x) and $S_N(x)$ over the range of the random variable as shown in Fig. 6.5.1.

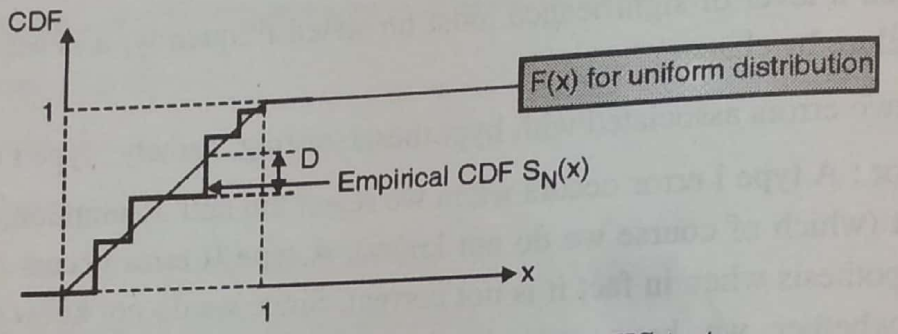

Fig. 6.5.1 : Graphical method for KS test

- Plot f(x) and $S_N(x)$ and get the statistics $D = \max| f(x) - S_N(x) |$ this is the maximum vertical distance.
- Computed value of D is compared with tabulated value of simple distribution, D_α for the significance level and the sample size N given.
- If $D \leq D_\alpha$ accept, otherwise reject null hypothesis.

Algorithm

i. Define the hypothesis for testing the uniformity as :

$$H_0: R_i \sim U[0, 1]$$

$$H_1: R_i \nsim U[0, 1]$$

ii. Rank the data from smallest to largest

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

iii. Compute D^+ and D^-

$$D^+ = \max_{1 \leq i \leq N} \left\{ \frac{i}{N} - R(i) \right\}$$

$$D^- = \max_{1 \leq i \leq N} \left\{ R(i) - \frac{i-1}{N} \right\}$$

iv. Compute $D = \max(D^+, D^-)$

v. Determine the critical value, D_α for the significance level and the sample size N given.

vi. If $D \leq D_\alpha$ accept, otherwise reject H_0.

vii. If $D \leq D\alpha$ no difference is detected between the true distribution and the uniform distribution.

Advantages of the K-S Test

1. The K-S test statistic does not depend on the underlying cumulative distribution function being tested.

2. The K-S test is an exact test (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid).

Limitations of the K-S Test

1 It only applies to continuous distributions.

2. It tends to be more sensitive near the center of the distribution than at the tails.

Ex. 6.5.1 : The sequence of numbers 0.14, 0.05, 0.44, 0.81, 0.93 has been generated. Use the Kolnogorov Smirnov Test with $\alpha = 0.05$ to perform a test of uniformity.

Soln. :

I	1	2	3	4	5
R(i)	0.05	0.14	0.44	0.81	0.93
i/N	0.20	0.40	0.60	0.80	1.00
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

$$D^+ = \max(i/N - R_{(i)})$$

$$= \{0.15, 0.26, 0.16, 0.07\} = 0.26$$

$$D^- = \max(R_{(i)} - (i-1)/N)$$

$$= \{0.05, 0.04, 0.21, 0,13\} = 0.21$$

$$\text{Hence, } D = \max(D^+, D^-)$$

$$= 0.26$$

Critical value of D for $\alpha = 0.05$ and $N = 5$ is 0.565.

Since computed value 0.26 is less than tabulated critical value 0.565, the hypothesis that the distribution is uniform is not rejected.

6.5.2.2 Chi - Square Test

- Chi-square test checks whether a sequence of pseudo-random numbers in [0,1] are uniformity distributed. The chi-square test, in general, can be used to check whether an empirical distribution follows a specific theoretical distribution. Now, we are concerned about testing whether the numbers produced by a generator are uniformly distributed.
- Let us consider a sequence of pseudo-random numbers between 0 and 1. We divide the interval [0, 1] into k subintervals of equal length, where $k > 100$.
- Let *fi* be the number of pseudo-random numbers that fall within the ith subinterval (make sure that enough random numbers are generated so that $fi > 5$). The *fi* values are called the *observed* values.
- Now, if these generated random numbers are truly uniformly distributed, then the mean number of random numbers that fall within each subinterval is n/k, where n is the sample size. This value is called the *theoretical* value.
- The chi-square test measures whether the difference between the observed and the theoretical values is due to random fluctuations or due to the fact that the empirical distribution does not follow the specific theoretical distribution.
- Chi-Square is used to test the uniformity within a data set or to determine the probability of a dependency relationship between two or more distinct data set.
- Chi-Square tests a null hypothesis that the relative frequencies of occurrence of observed events follow a specified frequency distribution.
- The condition should be satisfied while applying χ^2 test is, N the total no of observation (generated random number) must be sufficiently large. Preferably $N \geq 50$.

Algorithm

i. Define the hypothesis for testing the uniformity as :

$$H_0: R_i \sim U[0, 1]$$

$$H_1: R_i \not\sim U[0, 1]$$

ii. Divide the total number of observation (N) into equally numbered classes (n), $(a_0, a_1) \ldots (a_{n-1}, a_n)$, n should be chosen in such a way that $E_i\ (= N/n) \geq 5$.

iii. Compute the sample test statistics :

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

where O_i is the observed number in the ith class

E_i is the expected number in the ith class

Determine the critical value for the specified significance level α with n – 1 degree of freedom.

If $\chi^2_0 > \chi^2_{\alpha,n-1}$, H_0 is rejected else no difference has been detected the sample distributions and the uniform distribution.

Ex. 6.5.2 : Consider the following sequence of numbers given and use Chi-Square test with $\alpha = 0.05$ to test the hypothesis that the number are uniformly distributed [0, 1]

0.01, 0.1, 0.11, 0.62, 0.59, 0.69, 0.42, 0.12, 0.97, 0.13, 0.15, 0.22, 0.23, 0.26, 0.99, 0.16, 0.56, 0.17, 0.45, 0.95, 0.18, 0.47, 0.76, 0.19, 0.9, 0.41, 0.31, 0.41, 0.02, 0.77, 0.48, 0.03, 0.25, 0.52, 0.61, 0.4, 0.61, 0.04, 0.7, 0.86, 0.05, 0.94, 0.47, 0.06, 0.07, 0.54, 0.09, 0.05, 0.93, 0.51, 0.28, 0.92, 0.32, 0.96, 0.33, 0.34, 0.37, 0.36, 0.42, 0.91, 0.73, 0.43, 0.45, 0.46, 0.44, 0.72, 0.93, 0.52, 0.48, 0.95, 0.40, 0.5, 0.88, 0.78, 0.55, 0.98, 0.56, 0.58, 0.51, 0.53, 0.67, 0.63, 0.92, 0.71, 0.64, 0.65, 0.43, 0.08, 0.54, 0.94, 0.66, 0.68, 0.8, 0.87, 0.81, 0.82, 0.83, 0.89. 0.83, 0.84

Soln. :

Interval	Upper Limit	O_i	E_i	$O_i - E_i$	$[(O_i - E_i)^2]$	$[(O_i - E_i)^2]/E_i$
1	0.1	10	10	0	0	0
2	0.2	9	10	– 1	1	0.1
3	0.3	5	10	– 5	25	2.5
4	0.4	6	10	– 4	16	1.6
5	0.5	16	10	6	36	3.6
6	0.6	13	10	3	9	0.9
7	0.7	10	10	0	0	0
8	0.8	7	10	– 3	9	0.9
9	0.9	10	10	0	0	0
10	1.0	14	10	4	16	1.6
S		**100**	**100**	**0**		**11.2**

Since $\alpha = 0.05$ and degree of freedom = 10 – 1 = 9

From table we get,

$$\chi^2_{0.05,9} = 16.9$$

$\chi^2_0 < \chi^2_{\alpha,n-1}$

Ho accepted.

Accepted since, $\chi^2_0 = 11.2 < \chi^2_{0.05,9}$

6.5.2.3 K-S Test vs. Chi-Square Test

Kolmogorov - Smirnov test

- Kolmogorov - Smirnov test is designed for small samples and for continuous distribution.
- Kolmogorov - Smirnov test is based on differences between observed and expected CDF.
- Kolmogorov - Smirnov test uses each sample without grouping.
- Kolmogorov - Smirnov test is exact test, provided parameters of expected distribution known.

Chi-square test

- Chi-square test is designed for large samples and discrete distribution.
- Chi-square test is based on differences between observed and hypothesized PMF or PDF.
- Chi-square test requires samples be grouped into small number of cells.
- Chi-square test is approximate test, sensitive to cell size.

6.5.3 Testing of Independence

Q. Give sample data for random input, how would you test for Independence.

MU - Dec. 06

Following test are used to test independence :

- **Runs test :** Looks for patterns of increasing/decreasing values. Run 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.
- **Autocorrelation test :** Tests the correlation between numbers and compares the sample correlation to the expected correlation of zero.
- **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 number of gaps.
- **Poker test :** Treats numbers grouped together as a poker hand. Then the hands obtained are compared to what is expected using the chi-square test.

6.5.3.1 Runs Test

- Sometimes the generators pass the KS and Chi Square test for uniformity, but the numbers generated are not independent.
- The runs test analyses an orderly grouping of numbers in a sequence to test the hypothesis of independence.
- The runs test can be used to test the assumption that the pseudo-random numbers are independent of each other.
- A run is defined as succession of similar events preceded and followed by different event.
- The number of events that occur in the run gives the length of the run.

6.5.3.2 Runs Up and Runs Down

- An up run is a sequence of numbers each of which is followed by a larger number.
- Similarly, a down run is a sequence of numbers each of which is followed by a smaller number.
- For sequence of random number, we can define up runs and down runs

  0.87, 0.15, 0.25, 0.49, 0.69, 0.35, 0.33, 0.12

  - + + + - - -
- If there is increase put + sign and if a decrease put – sign
- Each succession of + and – forms runs
- In above example there are 3 runs of length 1,3,3
- Let be the total numbers of runs found in the sequence.
- For $N > 20$, a is approximated by a normal distribution

Algorithm

i. The hypothesis for testing independence as :

$$H_0: R_i \sim \text{Independently}$$

$$H_1: R_i \nsim \text{Independently}$$

ii. Write the sequence of runs up and runs down.

iii. Count the total numbers of runs α, present in the sequence.

iv. Compute the mean and variance of α

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

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

v. Compute the standard normal variety.

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

vi. Determine the critical value $z_{\alpha/2}$ and $-z_{\alpha/2}$ for the specified significance level from the table.

vii. If $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$ then H_0 is accepted for the significance level.

Ex. 6.5.3 : Consider the sequence of 40 numbers: 0.52, 0.99, 0.46, 0.58, 0.64, 0.25, 0.88, 0.11, 0.20, 0.18, 0.97, 0.44, 0.43, 0.94, 0.82, 0.60, 0.73, 0.69, 0.21, 0.03, 0.04, 0.81, 0.85, 0.30, 0.47, 0.96, 0.17, 0.72, 0.62, 0.27, 0.10, 0.60, 0.34, 0.65, 0.79, 0.44, 0.02, 0.37, 0.48, 0.50. Determine whether the hypothesis of independence can be rejected based on runs up and down where $\alpha = 0.05$.

Soln. :

The sequence of runs up and down is :

+	–	+	+	–	+	–	+	–	+
–	–	+	–	–	+	–	–	–	+
+	+	–	+	+	–	+	–	–	–
+	–	+	+	–	–	+	+	+	

The total number of runs a = 25.

Mean and variance of a is :

$$\mu_\alpha = (2N - 1)/3 = [2(40) - 1]/3 = 26.33$$

$$\sigma_\alpha^2 = (16N - 29)/90 = [16(40) - 29]/90 = 6.79$$

Compute the standard normal variate.

$$Z_0 = (\alpha - \mu_\alpha)/\sigma_\alpha$$

Determine the critical value $z_{\alpha/2}$ and $-z_{\alpha/2}$ for the specified significance level from the table. Since $\alpha = 0.05$, $\alpha/2 = 0.025$

$$\phi(z_{0.025}) = 1 - 0.025 = 0.975$$

$$z_{0.025} = \phi^{-1}(0.975) = 1.96$$

Since $-z_{0.025} = -1.96 \leq Z_0 = -0.51 \leq z_{0.025} = 1.96$. Hence, H_0 is accepted for the significance level.

6.5.3.3 Runs Above and Below the Mean

- The runs up and down test are inadequate for testing independence, so we use runs above and below the means.
- A + sign is used to denote a number above the mean and a – sign will denote the number below the mean.
- We assume n_1 to be observation above mean and n_2 as observation below mean.
- The maximum number of runs here is given by $N = n_1 + n_2$.
- The minimum number of runs is 1.
- Let b be the total numbers of runs found in the sequence.
- For $n_1 > 20$ or $n_2 > 20$ b is approximated by a normal distribution
- Finally standard normal test statistics Z_0 is developed and is compared with the critical value.

Algorithm

a. The hypothesis for testing independence as :

1. H_0: R_i ~ Independently
2. H_1: R_i ≁ Independently

b. Write the sequence of runs above and runs below mean.

c. Count the total numbers of runs above mean (n_1) and runs below mean (n_2), present in the sequence.

d. Compute the mean and variance of b.

$$\mu_b = \frac{2n_1 n_2}{N} + \frac{1}{2}$$

$$\sigma_b^2 = \frac{2n_1 n_2 (2n_1 n_2 - N)}{N^2 (N-1)}$$

e. Compute the standard normal variate.

$$Z_0 = (b - \mu_b) / \sigma_b$$

f. Determine the critical value $z_{\alpha/2}$ and $-z_{\alpha/2}$ for the specified significance level from the table.

g. If $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$ then H_0 is accepted for the significance level.

Ex. 6.5.4 : Consider the sequence of 40 numbers

0.09	0.42	0.23	0.68	0.89	0.72	0.12	0.45	0.08	0.32
0.53	0.13	0.65	0.97	0.14	0.49	0.55	0.46	0.77	0.28
0.81	0.63	0.40	0.57	0.02	0.16	0.33	0.86	0.99	0.22
0.76	0.48	0.61	0.39	0.43	0.78	0.20	0.35	0.17	0.93

Determine whether there are an excessive number of runs above or below the mean. Use $\alpha = 0.5$ and mean = 0.495

Soln. :

(i) The hypothesis for testing independence is :

$H_0 : R_i \sim$ independently

$H_1 : R_i \not\sim$ independently

(ii) The sequence of runs above and below the means is 0.495 is

$- \; - \; - \; + \; + \; + \; - \; - \; - \; -$

$+ \; - \; + \; + \; - \; - \; + \; - \; + \; -$

$+ \; + \; - \; + \; - \; - \; - \; + \; + \; -$

$+ \; - \; + \; - \; - \; + \; - \; - \; - \; +$

(iii) The number of observations above mean $n_1 = 17$. The number of observation below mean $n_2 = 23$. The total number of runs $b = 24$.

(iv) Mean and variance of b

$$\mu_b = \frac{2 n_1 n_2}{N} + \frac{1}{2} = \frac{2 \times 17 \times 23}{40} + \frac{1}{2}$$

$$= 20.05$$

$$\sigma_b^2 = \frac{2n_1n_2 (2n_1 n_2 - N)}{N^2 (N - 1)}$$

$$= \frac{2 \times 17 \times 23 (2 \times 17 \times 2 - 40)}{40^2 (40 - 1)}$$

$$= 9.3$$

(v) The standard normal statistics

$$Z_0 = \frac{b - \mu_b}{\sigma_b} = \frac{24 - 20.05}{\sqrt{9.3}}$$

(vi) Determine the critical value $z_{\alpha/2}$ and $-z_{\alpha/2}$ for the specified significance level α from. Since $\alpha = 0.05$, $\alpha/2 = 0.025$.

Hence $z_{0.025} = 1.96$

(vii) Since $-z_{0.025} = -1.96 \leq Z_0 = -0.51 \leq z_{0.025} = 1.96$. Hence H_0 is not rejected

6.5.3.4 Length of Runs

- Length of run method can be used to test the hypothesis.
- It can be used in the previous two method.
- Assume Yi be the number of runs length I in a sequence of N numbers.
- The expected value of Yi for runs up and runs down or runs above and runs below is determined.
- To compare the expected value with observed value, the chi-square test is applied.

Algorithm :

i The hypothesis for testing independence as :

H_0: R_i ~ Independently

H_1: R_i ≁ Independently

ii. Write the sequence of runs up and down or runs above and runs below mean.

iii. Count the length of runs in the sequence.

iv Prepare a table for number of observed runs of each length.

Run length	1	2
Observed runs		

Compute the expected value for runs up and down is given by,

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

and

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

The number of runs above and below the mean, also random variables the expected value of Y_i is approximated by,

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

where $E(I)$ the approximate expected length of a run and w_i is the approximate probability of length i.

w_i is given by

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

$E(I)$ is given by,

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

vi) Compute the mean or expected total number of runs in sequence.

a) For runs up and runs down

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

b) The approximate expected total number of runs above and below mean in a sequence of length N is given by

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

vii) Compute the expected number of runs of length greater than or equal to the maximum observed length

a) For runs up and runs down

$$\mu_a = \sum_{i=1}^{m} E(Y_i)$$

Where m is equal to maximum length of observed run.

b) For runs above and below mean

$$E(A) = \sum_{i=1}^{m} E(Y_i)$$

Where m is equal to maximum length of observed run.

viii) Apply the Chi-Square test

Run Length, I	Observed Number of Runs, O_i	Expected Number of Runs, $E(Y_i)$	$\frac{[O_i - E(Y_i)]^2}{E(Y_i)}$
$1 \geq m$			

The test statistics is given by

$$\chi_0^2 = \sum_{i=1}^{n} \frac{[O_i - E(Y_i)]^2}{E(Y_i)}$$, When n is the number of classes

ix) Determine the critical value for the specified significance level.

x) If $-\chi_0^2 < \chi_{\alpha,n-1}^2$ then H_0 is accepted.

Ex. 6.5.5 : Consider the sequence of 40 numbers.

0.42	0.51	0.58	0.10	0.89	0.27	0.65	0.34	0.89	0.45
0.81	0.33	0.58	0.97	0.32	0.95	0.96	0.46	0.77	0.04
0.05	0.63	0.86	0.57	0.65	0.66	0.33	0.16	0.89	0.66
0.19	0.35	0.94	0.77	0.43	0.31	0.78	0.35	0.49	0.93

Determine if the hypothesis that the numbers are independent can be rejected on the basis of the length of runs up and down where $\alpha = 0.05$. **MU - Dec. 10**

Soln. :

(i) Define the hypothesis for testing independence as :

$H_0 : R_i \sim$ independently

$H_1 : R_i \nsim$ independently

ii) The sequence of runs up and down is :

+ + − + − + − + − +

− + + − + + − + − +

+ + − + + − − + − −

+ + − − − + − + +

(iii) The length of run in the sequence is :

2 1 1 1 1 1 1 1 1

1 2 1 2 1 1 1 3

1 2 2 1 2

2 3 1 1 2

(iv) The expected number of observed runs of each length is :

Run Length (i)	1	2	3
Observed Runs (O_i)	17	8	2

(v) The expected number of length one, two and three are :

$$E(Y_i) = \begin{cases} \frac{2}{(i+3)!}[N(i^2+3i+1)-(i^3+3i^2-i-4)], & i \le N-2 \\ \frac{2}{N!}, & i = N-1 \end{cases}$$

$$E(Y_1) = \frac{2}{(1+3)!}[40(1+3+1)-(2+3-1-4)] = 16.75$$

$$E(Y_2) = \frac{2}{(2+3)!}[40(4+6+1)-(8+12-2-4)] = 7.1$$

$$E(Y_3) = \frac{2}{(3+3)!}[40(9+9+1)-(27+27-3-4)] = 1.98$$

(vi) The mean or expected total number of runs of all length in sequence is :

$$\mu_a = \frac{2N-1}{3} = \frac{2(3)-1}{3} = 26.33$$

(vii) The expected number of runs of length greater than or equal to 4 is :

$$\mu_a - \sum_{i=1}^{m} E(Y_i) = 26.33 - (16.75 + 7.1 + 1.98) = 0.5$$

(viii) Apply Chi-Square Test

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	16.75		$3.73 * 10^{-3}$
2	8	7.1		0.0184
3	2	1.98	9.58	0.184
≥4	0	0.5		

It is suggested that the minimum value of expected frequency is 5 in case of Chi-Square test. If it is less than 5, it can be combined with expected frequency of adjacent class interval. The corresponding observed frequency will also be combined accordingly and the value of the number of classes 'n' would be reduced.

In the above table, class 3 and 4 has expected frequency less than 5 so it is combined with class 2. Similarly combine the observed frequency of class 3 and 4 with class2. Reduce the number of classes to 2. Hence n = 2.

The test statistics is :

$$\chi_0^2 = \sum_{i=1}^{n} \frac{[O_i - E(Y_i)]^2}{E(Y_i)} = (3.73 * 10^{-3} + 0.0184) = 0.022133$$

(ix) The critical value for to the specified level of significance level $\alpha = 0.05$ with $n - 1 = 2 - 1 = 1$ degrees of freedom is given by :

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

(x) Since $\chi_0^2 = 0.02213 < \chi^2_{0.05,1} = 3.84$, H_0 is not rejected.

Ex. 6.5.6 : Consider the sequence of 50 numbers

0.89	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

Determine the hypothesis that the number are independent can be rejected the basis of length of runs above and below, mean where $\alpha = 0.05$.

Soln. :

(i) Define the hypothesis for testing independence as follows :

$H_0 : R_i \sim$ independently

$H_1 : R_i \nsim$ independently

(ii) The sequence of runs up and sown is :

+ − + − − + − − − −

− + + − + − + − + −

+ − + + − + + + + −

− − + − − + − − + +

− − + + + − + − − +

(iii) The length of run the sequence is :

1 1 1 2 1 5

2 1 1 1 1 1 1

1 1 2 1 4 3

1 2 1 2 2

2 3 1 1 2 1

(iv) The number of observed runs of each length is :

Run Length (i)	1	2	3	4	5
Observed Runs (O_i)	19	8	2	1	1

(v) The expected number of runs of length one, two, there, four and five are :

The number of observations above mean $n_1 = 24$. The number of observation below mean $n_2 = 26$.

(a) The approximate probability that a run has length 'i' is :

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

$$w_1 = \left(\frac{24}{50}\right)^1 \left(\frac{26}{50}\right) + \left(\frac{24}{50}\right)\left(\frac{26}{50}\right)^1 = 0.4992$$

$$w_2 = \left(\frac{24}{50}\right)^2 \left(\frac{26}{50}\right) + \left(\frac{24}{50}\right)\left(\frac{26}{50}\right)^2 = 0.2496$$

$$w_3 = \left(\frac{24}{50}\right)^3 \left(\frac{26}{50}\right) + \left(\frac{24}{50}\right)\left(\frac{26}{50}\right)^3 = 0.125$$

$$w_4 = \left(\frac{24}{50}\right)^4 \left(\frac{26}{50}\right) + \left(\frac{24}{50}\right)\left(\frac{26}{50}\right)^4 = 0.0627$$

$$w_5 = \left(\frac{24}{50}\right)^5 \left(\frac{26}{50}\right) + \left(\frac{24}{50}\right)\left(\frac{26}{50}\right)^5 = 0.0315$$

(b) The approximate expected length of a run is :

$$E(I) = \frac{n_1}{n_2} + \frac{n_2}{n_1} = \frac{24}{26} + \frac{26}{24} = 2.0064$$

(c) The expected runs of various lengths :

$$E(Y_i) = \frac{N w_i}{E(I)}$$

$$E(Y_1) = \frac{N w_1}{E(I)} = \frac{50\,(0.49952)}{2.0064} = 12.44$$

$$E(Y_2) = \frac{N w_2}{E(I)} = \frac{50\,(0.2496)}{2.0064} = 6.22$$

$$E(Y_3) = \frac{N w_3}{E(I)} = \frac{50\,(0.125)}{2.0064} = 3.115$$

$$E(Y_4) = \frac{N w_4}{E(I)} = \frac{50\,(0.0627)}{2.0064} = 1.5625$$

$$E(Y_5) = \frac{N w_5}{E(I)} = \frac{50\,(0.0315)}{2.0064} = 0.785$$

(vi) The mean or expected total number of runs of all length in a sequence is :

$$E(A) = \frac{N}{E(I)} = \frac{50}{2.0064} = 24.9203$$

The expected number of runs of length grater than or equal to 5 is :

$$E(A) - \sum_{i=1}^{4} E(Y_i) = 24.9203 - 23.3375 = 1.5828$$

(viii) Apply Chi-Square test

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	19	12.44	3.4593
2	8	6.22	0.5094
3	2	3.115	
4	1 4	1.5625	0.8161
≥ 5	1	1.5828	

It is suggested that the minimum value of expected frequency is 5 in case of Chi-Square test. If it is less than 5, it can be combined with expected frequency of adjacent class interval. The corresponding observed frequency will also be combined accordingly and the value of the number of classes 'n' would be reduced.

In the above table, class 3, 4 and 5 has expected frequency less than 5 so it is combined which gives the result as 6.2603. Similarly combined the observed frequency of class 3, 4 and 5. Reduce the number of classes by 2.

Hence n = 5 – 2 = 3.

The test statistics is : $$\chi_0^2 = \sum_{i=1}^{n} \frac{[O_i - E(Y_i)]^2}{E(Y_i)} = 4.7848$$

(ix) The critical value for the specified level of significance level $\alpha = 0.05$ with $n - 1 = 2 - 1 = 1$ degrees of freedom is given by :

$$\chi^2_{0.05,\,1} = 5.99$$

(x) Since $\chi_0^2 = 4.7848 < \chi^2_{0.05,\,1} = 5.99$ Hence H_0 is not rejected.

6.5.3.5 Autocorrelation Test

- Autocorrelation is a statistical test that determines whether a random number generator is producing independent random numbers in a sequence.
- Autocorrelation is concerned with dependence between numbers in a sequence. It tests for correlation between numbers.
- It computes the autocorrelation between every m numbers starting with i.

What does autocorrelation look like?

- Here is a list of random integers generated. See if you can identify a pattern by looking at the values in the sequence.

1	2	3	4	5	6	7	8	9	10
0.63	0.28	0.30	0.42	0.97	0.05	0.71	0.63	0.17	1.0
0.61	0.19	0.94	0.64	0.84	0.54	0.56	0.57	0.09	0.99
0.01	0.10	0.69	0.38	0.93	0.85	0.68	0.14	0.18	0.84
0.19	0.71	0.44	0.72	0.95	0.28	0.96	0.51	0.50	0.89
0.66	0.31	0.50	0.33	0.89	0.54	0.73	0.76	0.62	0.92

- If you look carefully at the following table of random integers generated, you will notice that every number in the 5th, 10th, 15th, and 20th position is a larger value.

1	2	3	4	5	6	7	8	9	10
0.63	0.28	0.30	0.42	0.97	0.05	0.71	0.63	0.17	1.0
0.61	0.19	0.94	0.64	0.84	0.54	0.56	0.57	0.09	0.99
0.01	0.10	0.69	0.38	0.93	0.85	0.68	0.14	0.18	0.84
0.19	0.71	0.44	0.72	0.95	0.28	0.96	0.51	0.50	0.89
0.66	0.31	0.50	0.33	0.89	0.54	0.73	0.76	0.62	0.92

- The autocorrelation between numbers of a sequence R_i, R_{i+m}, R_{i+2m}, $R_{i+(M+1)m}$ is represented by ρ_{im}.
- It then compares the sample correlation to the expected correlation zero.
- Hence, if $\rho_{im} > 0$ then the subsequence has positive autocorrelation whereas if $\rho_{im} < 0$ then the subsequence has negative autocorrelation.
- A non-zero autocorrelation implies lack of independence.

What are the important variables ?

m : is the lag, the space between the numbers being tested.

i : is the index, or the number in the sequence that you start with

N : the number of numbers generated in a sequence

M : is the largest integer such that $i + (M + 1)\, m \leq N$

Algorithm

i. The hypothesis for testing independence as :

H_0: $\rho_{im} = 0$, if numbers are independent

H_1: $\rho_{im} \neq 0$, if numbers are dependent.

ii. Find out the value of 'i' and lag 'm' using the given data.

iii. Using i, m and N estimate the value of M where M is the largest integer such that $i + (M + 1)\, m \leq N$ and N is the total number of values in the sequence.

iv. For large values of M the distribution of estimator ρ_{im}, denoted by $\hat{\rho}_{im}$ is approximately normal if the numbers R_i, R_{i+m}, R_{i+2m}, $R_{i+(M+1)m}$ are uncorrelated where

$$\hat{\rho}_{i,m} = \frac{1}{M+1}\left[\sum_{k=0}^{M} R_{i+km} R_{i+(k+1)m}\right] - 0.25$$

v. Find the standard deviation of the estimator

$$\sigma = \frac{\sqrt{13M+7}}{12\,(M+1)}$$

vi. Compute the test statistics.

$$Z_0 = \frac{\hat{\rho}_{i,m}}{\sigma}$$

vii. Determine the critical value $z_{\alpha/2}$ and $-z_{\alpha/2}$ for the specified significance level from the table.

viii. If $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$ then H_0 is accepted for the significance level.

How do you figure out the value of M ?

To figure out the value of M, you must use some simple algebra. For example, the equation : $i + (M + 1)\, m \leq N$ must be solved using the given values i the index, N the number of elements in the sequence, and m the lag.

For example

$$i = 1$$

$$m = 5$$

$$N = 50;$$

$$M = ?;$$

$$i + (M + 1)\, m \leq N$$

$$1 + (M + 1)\, 5 \leq N$$

$$1 + (5M + 5) \leq 30$$

$$5M + 6 \leq 30$$

$$\_\_\_ - 6 \_\_ - 6 \_\_$$

$$5M \leq 24$$

$$M = 4.8$$

Since the value M must be an integer, the 0.8 is truncated and the 4 is saved as the value of M. Thus, M is equal to 4.

Drawbacks when using autocorrelation

- Autocorrelation is not very sensitive to small values of M, when the values being tested are on the low side. For example, if all the values were equal to zero, then the resulting value would be – 1.95, which is not enough to reject the hypothesis.
- Many sequences can be formed in a set of date (the sequence of all numbers, the sequence from the first, fifth, .. numbers and so forth. If the alpha is equal to 0.05, then there is a possibility of rejecting a true hypothesis. For example, if these are 10 independent events, the possibility of finding no autocorrelation alone is $(0.95)^{10}$ or 0.60. Thus, 40% of the time significant autocorrelation would be detected when it does not exist.

Ex. 6.5.7 : Consider the following sequence of numbers: 0.12, 0.01, 0.23, 0.28, 0.89, 0.31, 0.64, 0.28, 0.83, 0.93, 0.99, 0.15, 0.33, 0.35, 0.91, 0.41, 0.60, 0.27, 0.75, 0.88, 0.68, 0.49, 0.05, 0.43, 0.95, 0.58, 0.19, 0.36, 0.69, 0.87. Test whether 3rd, 8th, 13th numbers in the sequence are auto correlated where $\alpha = 0.05$.

Soln. :

The hypothesis for testing independence as :

H_0: $\rho_{im} = 0$, if numbers are independent

H_1: $\rho_{im} \neq 0$, if numbers are dependent.

Find out the value of 'i' and lag 'm' using the given data.

Given : $N = 30$

Using i, m and N estimate the value of M where M is the largest integer such that

$$i + (M+1)\, m \leq N$$

$$3 + (M+1)\, 5 \leq 30$$

Hence, $M = 4$

$$\hat{\rho}_{im} = 1/(M+1)\left[\sum_{k=0}^{M} R_{i+km}, R_{i+(k+1)m}\right] - 0.25$$

Hence, $\alpha = 0.05$, $i = 3$, $m = 5$, $N = 30$ and $M = 4$

$$\hat{\rho}_{35} = 1/(4+1)\,[(0.23)(0.28) + (0.28)(0.33) + (0.33)(0.27) + (0.27)(0.05) + (0.05)(0.36)] - 0.25$$

$$= -0.1945$$

Find the standard deviation of the estimator

$$\sigma = (\sqrt{(13M+7)}) / (12(M+1)$$

$$= (\sqrt{(13(4)+7)}) / (12((4)+1)$$

$$= 0.128$$

Compute the test statistics.

$$Z_0 = \hat{\rho}_{im} / \sigma$$

$$= (-0.1945) / 0.128$$

$$= -1.516$$

Determine the critical value $z_{\alpha/2}$ and $-z_{\alpha/2}$ for the specified significance level from the table.

If $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$ then H_0 is not rejected for the significance level.

6.5.3.6 Gap Test

- The gap test is concerned with the randomness of the digits in a sequence of numbers. Let U1...UN be such a sequence. We say that any subsequence Uj, Uj + 1,...Uj + r of r + 1 numbers represents a gap of length r if Uj and Uj + r lie between α and β and $(0 \leq \alpha < \beta \leq 1)$ but Uj + i, i = 1...r – 1, does not. For a "true" sequence of random numbers the probability of obtaining a gap of length r is

$$p(r) = (0.9)^r (0.1)$$

- A chi-square test based on the comparison of the expected and actual numbers of gaps of length r maybe used.
- Gap test counts the number of digits that appear between the repetitions of a particular digit.
- It then uses KS test to compare with the expected size of gaps.
- The probability of the gap is determined by :

  P (m followed by exactly x non – m digits) = $(0.9)^x (0.1)$, x = 0, 1, 2 ...
- Every digit 0, 1, 2 ... 9 must be analyzed to test the numbers are independent.

Algorithm :

i. The hypothesis for testing independence as :

$$H_0: R_i \sim \text{Independently}$$

$$H_1: R_i \nsim \text{Independently}$$

ii. Determine the number of gaps and length of each gap associated with each digit.

iii. Select the interval width based on the number of gaps and generate the frequency distribution table for the sample of gaps and apply **Kolmogorov – Smirnov** test.

iv. Compute the test statistics D that is the maximum deviation between F(x) and $S_N(x)$.

$$D = \max | F(x) - S_N(x) |$$

v. Determine the critical value, D_α for the specified value of significance level and the sample size N from the table.

vi. If $D > D_\alpha$, H_0 is rejected.

Ex. 6.5.8 : Consider the following sequence of 120 digits :

1	3	7	4	8	6	2	5	1	6	4	4	3	3	4	2	1	5	8	7
0	7	6	2	6	0	5	7	8	0	1	1	2	6	7	6	3	7	5	9
0	8	8	2	6	7	8	1	3	5	3	8	4	0	9	0	3	0	9	2
2	3	6	5	6	0	0	1	3	4	4	6	9	9	8	5	6	0	1	7
5	6	7	9	4	9	3	1	8	3	3	6	6	7	8	2	3	5	9	6
6	7	0	3	1	0	2	4	2	0	6	4	0	3	9	3	6	8	1	5

Test whether this digits can be assumed to be independent based on the frequency with which gaps occurs. Use $\alpha = 0.05$.

Soln. :

$$\text{Number of digits} = 120$$

$$\text{Total number of gaps} = \text{Number of digits} - \text{number of distinct digits}$$

$$= 120 - 10$$

$$= 110$$

The number of gaps and length of each gap associated with each digit is :

Digit	Length of each gap	Number of gaps
0	4, 3, 10, 12, 1, 1, 7, 0, 10, 24, 2, 3, 2	13
1	7, 7, 13, 0, 15, 19, 10, 8, 16, 13	10
2	8, 7, 8, 10, 15, 0, 44, 10, 1	9
3	10, 0, 22, 11, 1, 5, 4, 6, 17, 2, 0, 5, 6, 9, 1	15
4	6, 0, 2, 37, 16, 0, 13, 22, 3	9
5	9, 8, 11, 10, 13, 11, 4, 16, 21	9
6	3, 12, 1, 8, 1, 8, 17, 1, 6, 4, 4, 9, 0, 6, 0, 9, 5	17
7	16, 1, 5, 6, 2, 7, 33, 2, 10, 7	10
8	13, 9, 12, 0, 3, 4, 22, 13, 5, 22	10
9	14, 3, 13, 0, 9, 1, 12, 15	8

Select the interval width based on the number of gaps and generate the frequency distribution table for the sample of gaps and apply KS test

Gap length	Frequency	Relative frequency	Cumulative relative frequency	CDF $F(x) = 1 - 0.9^{x+1}$	$\lvert F(x) - S_N(x) \rvert$
0 – 3	33	0.3000	0.3000	0.3439	0.0439
4 – 7	23	0.2091	0.5091	0.5695	0.0604

Gap length	Frequency	Relative frequency	Cumulative relative frequency	CDF $F(x) = 1 - 0.9^{x+1}$	$\lvert F(x) - S_N(x) \rvert$
8 – 11	23	0.2091	0.7182	0.7176	0.0006
12 – 15	15	0.1364	0.8546	0.8146	0.0400
16 – 19	7	0.0636	0.9182	0.8784	0.0398
20 – 23	5	0.0455	0.9637	0.9202	0.0435
24 – 27	1	0.0091	0.9728	0.9497	0.0231
28 – 31	0	0	0.9728	0.9657	0.0071
32 – 35	2	0.0182	0.9910	0.9775	0.0135
36 – 39	1	0.0091	1.0	0.9852	0.0148

$$D = \max |F(x) - S_N(x)| = 0.0604$$

Determine the critical value, D_α for the specified value of significance level and the sample size N = 110 from the table.

$$D_\alpha = D_{0.05} = 0.136$$

$$D < D_\alpha, H_0 \text{ is not rejected.}$$

6.5.3.7 Poker Test

Q. How is Pokers test used for testing independence? MU - Dec. 13

- It tests the frequency with which certain digits are repeated in a series of numbers.
- Treats numbers grouped together as a poker hand. The hands obtained are compared to what is expected using Chi-Square test.
- For three digits : three possibilities
  - All different
  - All equal
  - One pair of like digits
- If digits are truly random following result holds

$$P(\text{exactly one pair}) = \underbrace{\binom{3}{2}}_{\text{Number of possibilities}} \underbrace{(0.1)}_{\text{Given a fixed digit, this digit is the same}} \underbrace{(0.9)}_{\text{Given a fixed digit, this digit different}} = 0.27$$

$$P(\text{three different digits}) = P(\text{second different from first})$$
$$P(\text{third different from first and second})$$

Determine the critical value for the specified significance level with n – 1 degree of freedom.

$$\chi^2_{0.05,2} = 5.99$$

Since, $\chi^2_0 > \chi^2_{\alpha,n-1}$, H_0 is rejected

6.6 Solved Problems

Ex. 6.6.1 : Generate random numbers for $X_0 = 27$, $a = 17$, $c = 43$ and $m = 100$.

Soln. :

By using $X_{i+1} = (a\ X_i + c) \bmod m$, $i = 0, 1, 2 \ldots$ the values of random digits can be found.

The X_i and R_i values are

$$X_1 = (17 * 27 + 43) \bmod 100 = 502 \bmod 100 = 2$$

By using $R_i = X_i/m$, $i = 1, 2 \ldots$

$$R_1 = 2 / 100 = 0.02$$

Similarly,

$$X_2 = (17 * 2 + 43) \bmod 100 = 77$$
$$R_2 = 77 / 100 = 0.77$$
$$X_3 = (17 * 77 + 43) \bmod 100 = 52$$
$$R_3 = 52 / 100 = 0.52$$
$$X_4 = 17 * 52 + 43) \bmod 100 = 27$$
$$R_4 = 27 / 100 = 0.27$$

Ex. 6.6.2 : Consider the following sequence of numbers given and use Chi-Square test with $\alpha = 0.05$ to test the hypothesis that the number are uniformly distributed [0, 1]

0.01, 0.1, 0.11, 0.62, 0.59, 0.69, 0.42, 0.12, 0.97, 0.13, 0.15, 0.22, 0.23, 0.26, 0.99, 0.16, 0.56, 0.17, 0.45, 0.95, 0.18, 0.47, 0.76, 0.19, 0.9, 0.41, 0.31, 0.41, 0.02, 0.77, 0.48, 0.03, 0.25, 0.52, 0.61, 0.4, 0.61, 0.04, 0.7, 0.86, 0.05, 0.94, 0.47, 0.06, 0.07, 0.54, 0.09, 0.05, 0.93, 0.51, 0.28, 0.92, 0.32, 0.96, 0.33, 0.34, 0.37, 0.36, 0.42, 0.91, 0.73, 0.43, 0.45, 0.46, 0.44, 0.72, 0.93, 0.52, 0.48, 0.95, 0.40, 0.5, 0.88, 0.78, 0.55, 0.98, 0.56, 0.58, 0.51, 0.53, 0.67, 0.63, 0.92, 0.71, 0.64, 0.65, 0.43, 0.08, 0.54, 0.94, 0.66, 0.68, 0.8, 0.87, 0.81, 0.82, 0.83, 0.89, 0.83, 0.84

Soln. :

Interval	Upper Limit	O_i	E_i	$O_i - E_i$	$[(O_i - E_i)^2]$	$[(O_i - E_i)^2] / E_i$
1	0.1	10	10	0	0	0
2	0.2	9	10	– 1	1	0.1
3	0.3	5	10	– 5	25	2.5
4	0.4	6	10	– 4	16	1.6
5	0.5	16	10	6	36	3.6
6	0.6	13	10	3	9	0.9
7	0.7	10	10	0	0	0
8	0.8	7	10	– 3	9	0.9
9	0.9	10	10	0	0	0
10	1.0	14	10	4	16	1.6
S		**100**	**100**	**0**		**11.2**

Since $\alpha = 0.05$ and degree of freedom $= 10 - 1 = 9$

From table we get,

$$\chi^2_{0.05, 9} = 16.9$$

$$\text{Accepted since, } \chi^2_0 = 11.2 < \chi^2_{0.05, 9}$$

Ex. 6.6.3 : Consider the sequence of 40 numbers : 0.09, 0.42, 0.23, 0.68, 0.89, 0.72, 0.12, 0.45, 0.08, 0.32, 0.53, 0.13, 0.65, 0.97, 0.14, 0.49, 0.55, 0.46, 0.77, 0.28, 0.81, 0.63, 0.40, 0.57, 0.02, 0.16, 0.33, 0.86, 0.99, 0.22, 0.76, 0.48, 0.61, 0.39, 0.43, 0.78, 0.20, 0.35, 0.17, 0.9. Determine whether there are an excessive number of runs above and below the mean. Use $\alpha = 0.05$ and mean = 0.495.

Soln. :

The sequence of runs above and below the mean 0.495 is :

– – – + + + – – – –

+ – + + – – + – + –

+ + – + – – – + + –

+ – + – – + – – – +

The number of runs above mean $n_1 = 17$

The number of runs below mean $n_2 = 23$

The total number of runs b = 24

Mean and variance of b is :

$$\mu_b = (2 n_1 n_2 / N) + (1 / 2)$$
$$= (2 * 17 * 23 / 40) + (1/2)$$
$$= 20.05$$
$$\sigma_b^2 = [2n_1 n_2 (2n_1 n_2 - N)] / N^2 (N - 1)$$
$$= [2 * 17 * 23 (2 * 17 * 23 - 40)] / 40^2 (40 - 1)$$
$$= 9.3$$

Compute the standard normal variate.

$$Z_0 = (b - \mu_b) / \sigma_b$$
$$= (24 - 20.05) / \sqrt{9.3} = 1.295$$

Determine the critical value $z_{\alpha/2}$ and $- z_{\alpha/2}$ for the specified significance level from the table. Since $\alpha = 0.05$, $\alpha/2 = 0.025$

$$\phi (z_{0.025}) = 1 - 0.025 = 0.975$$

$$z_{0.025} = \phi^{-1}(0.975) = 1.96$$

Since $- z_{0.025} = - 1.96 \leq Z_0 = - 0.51 \leq z_{0.025} = 1.96$. Hence, H_0 is not rejected for the significance level.

Ex. 6.6.4 : Consider the following sequence of 120 digits :

1	3	7	4	8	6	2	5	1	6	4	4	3	3	4	2	1	5	8	7
0	7	6	2	6	0	5	7	8	0	1	1	2	6	7	6	3	7	5	9
0	8	8	2	6	7	8	1	3	5	3	8	4	0	9	0	3	0	9	2
2	3	6	5	6	0	0	1	3	4	4	6	9	9	8	5	6	0	1	7
5	6	7	9	4	9	3	1	8	3	3	6	6	7	8	2	3	5	9	6
6	7	0	3	1	0	2	4	2	0	6	4	0	3	9	3	6	8	1	5

Test whether this digits can be assumed to be independent based on the frequency with which gaps occurs. Use $\alpha = 0.05$.

Soln :

Number of digits = 120

Total number of gaps = Number of digits – number of distinct digits

= 120 – 10 = 110

The number of gaps and length of each gap associated with each digit is :

Digit	Length of each gap	Number of gaps
0	4, 3, 10, 12, 1, 1, 7, 0, 10, 24, 2, 3, 2	13
1	7, 7, 13, 0, 15, 19, 10, 8, 16, 13	10
2	8, 7, 8, 10, 15, 0, 44, 10, 1	9
3	10, 0, 22, 11, 1, 5, 4, 6, 17, 2, 0, 5, 6, 9, 1	15
4	6, 0, 2, 37, 16, 0, 13, 22, 3	9
5	9, 8, 11, 10, 13, 11, 4, 16, 21	9
6	3, 12, 1, 8, 1, 8, 17, 1, 6, 4, 4, 9, 0, 6, 0, 9, 5	17
7	16, 1, 5, 6, 2, 7, 33, 2, 10, 7	10
8	13, 9, 12, 0, 3, 4, 22, 13, 5, 22	10
9	14, 3, 13, 0, 9, 1, 12, 15	8

Select the interval width based on the number of gaps and generate the frequency distribution table for the sample of gaps and apply KS test.

Gap length	Frequency	Relative frequency	Cumulative relative frequency	CDF $F(x) = 1 - 0.9^{x+1}$	$\lvert F(x) - S_N(x) \rvert$
0 – 3	33	0.3000	0.3000	0.3439	0.0439
4 – 7	23	0.2091	0.5091	0.5695	0.0604
8 – 11	23	0.2091	0.7182	0.7176	0.0006
12 – 15	15	0.1364	0.8546	0.8146	0.0400
16 – 19	7	0.0636	0.9182	0.8784	0.0398
20 – 23	5	0.0455	0.9637	0.9202	0.0435
24 – 27	1	0.0091	0.9728	0.9497	0.0231
28 – 31	0	0	0.9728	0.9657	0.0071
32 – 35	2	0.0182	0.9910	0.9775	0.0135
36 – 39	1	0.0091	1.0	0.9852	0.0148

$$D = \max | F(x) - S_N(x) | = 0.0604$$

Determine the critical value, D_α for the specified value of significance level and the sample size N = 110 from the table.

$$D_\alpha = D_{0.05} = 0.136$$

$D < D_\alpha$, H_0 is not rejected.

Ex. 6.6.5 : Test the following random numbers for independence by runs up and down test. Take a = 0.05 and critical value $Z_{0.025} = 1.96$

{37, 59, 63, 07, 92, 48,12,86}

Soln. :

(i) Define the hypothesis for testing independence as :

$$H_0 : R_i \sim \text{independently}$$

$$H_1 : R_i \nsim \text{independently}$$

(ii) The sequence of runs up down is :

$$+ \quad + \quad - \quad + \quad - \quad - \quad +$$

(iii) The total number of runs, a = 5.

(iv) Mean and variance of a is

$$\mu_a = \frac{2N-1}{3} = \frac{2(8)-1}{3} = 5$$

$$\sigma_a^2 = \frac{16N-29}{90} = \frac{16(80)-29}{90} = 1.1$$

(v) The standard normal statistics

$$Z_0 = \frac{a-\mu_a}{\sigma_a} = \frac{5-5}{\sqrt{1.1}} = 0$$

(vi) Critical value

$$Z_{0.025} = 1.96 \text{ is obtained}$$

(vii) Since $-z_{0.025} = -1.96 \leq Z_0 = 0 \leq z_{0.025} = 1.96$. Hence H_0 is not rejected.

Review Questions

Q. 1 State the properties of random numbers.

Q. 2 What are the techniques are used to generate random number ?

Q. 3 State hypothesis for testing property of random number.

Q. 4 Give sample data for random input, how would you test for Independence.

Q. 5 How would you generate random numbers to test the reliability of a system? State the hypothesis for testing the property of random numbers. What do you understand by level of significance.

Q. 6 State the properties of random numbers. How are random numbers generated ?

Q. 7 What do you understand by "Goodness of fit test"? Write the procedure for the same.

Q. 8 State the properties of random numbers. What are the methods used to generate random numbers ?

Q. 9 Test the following random numbers for independence by runs up and down test. Take $\alpha = 0.05$ and critical value $Z_{0.025} = 1.96$ {0.12, 0.01, 0.23, 0.28, 0.89, 0.31, 0.64, 0.28, 0.33, 0.93}

Q. 10 Test the following random numbers for independence by poker test :

{0.594, 0.928, 0.515, 0.055, 0.507, 0.351, 0.262, 0.797, 0.788, 0.442, 0.097, 0.798, 0.227, 0.127, 0.474, 0.825, 0.007, 0.182, 0.929, 0.852}

$\alpha = 0.05$, $\chi^2_{0.05, 2} = 5.99$

6.7 University Questions and Answers

May 2010

Q. 1 Which problems are characteristic of pseudo random numbers ? Why ? How are random numbers generated ? **(Sections 6.1, 6.2 and 6.3)** **(10 Marks)**

Dec. 2010

Q. 2 Consider the following sequence of 40 numbers. **(8 Marks)**

0.67	0.31	0.53	0.91	0.80	0.27	0.61	0.49	0.76	0.85
0.62	0.28	0.55	0.77	0.38	0.65	0.29	0.55	0.83	0.92
0.09	0.33	0.24	0.07	0.30	0.54	0.43	0.66	0.71	0.52
0.11	0.36	0.12	0.78	0.95	0.44	0.50	0.19	0.22	0.38

Based on runs up and runs down, determine whether the hypothesis of independence can be rejected, where $\alpha = 0.05$. **(Ex. 6.5.5)**

Q. 3 State and explain the properties of random number and give the method for generating pseudo random numbers. **(Sections 6.2 and 6.4)** **(6 Marks)**

May 2011

Q. 4 Explain linear congruential method also list down tests for Random numbers. **(Sections 6.4)** **(10 Marks)**

Dec. 2012

Q. 5 Why is it necessary to test the properties of random numbers ? How would you generate random numbers ? **(Sections 6.1, 6.2 and 6.3)** **(10 Marks)**

May 2013

Q. 6 Why random numbers used in simulation ? What are techniques used to generate them ? **(Sections 6.1 and 6.4)** **(10 Marks)**

Dec. 2013

Q. 7 How is Pokers test used for testing independence? **(Section 6.5.3.7)** **(5 Marks)**

May 2014

Q. 8 Explain the properties of random numbers. **(Section 6.2)** **(5 Marks)**

Q. 9 Describe the characteristics of queuing systems. Name and explain some of the useful statistical models for queuing system. **(Sections 6.1 and 6.2)** **(10 Marks)**

Q. 10 What are the methods used to generate random numbers? **(Section 6.4)** **(10 Marks)**

Dec. 2014

Q. 11 Explain various methods for random numbers generation. **(Section 6.4)** **(10 Marks)**

May 2015

Q. 12 State the properties of random numbers. How are random numbers generated? **(Sections 6.2 and 6.3)** **(10 Marks)**

❑❑❑