B.N.Bandodkar College of Science, Thane

Random-Number Generation

Mrs M.J.Gholba

Properties of Random Numbers

A sequence of random numbers, R_1, R_2,…., must have two important statistical properties, uniformity and independence. Each random number R_i is an independent sample drawn from a continuous uniform distribution between zero and 1. This is, the pdf is given by

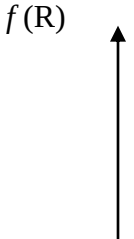
$$f(\mathrm{R}) = \begin{cases} 1, & 0 \le \mathrm{R} \le 1 \\ 0, & \text{otherwise} \end{cases}$$

This density function is shown in Figure 7.1. The expected value of each R_i is given by

$$E(R) = \int_0^1 R\,dR = \frac{R2}{2}\Big|_0^1 = \frac{1}{2}$$

and the variance is given by

$$V(R) = \int_0^1 R^2 dR - [E(R)]^2 = \frac{R3}{3}\Big|_0^1 - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$



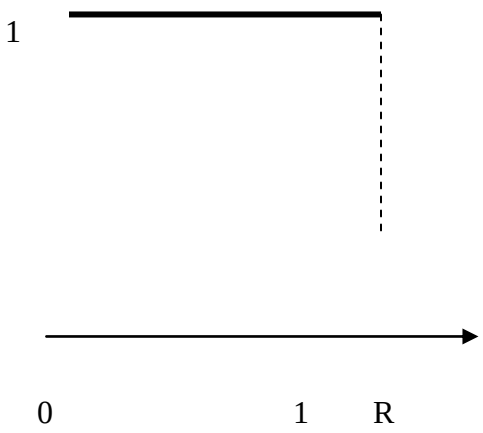


Fig 1 The pdf for random numbers.

Some consequences of the uniformity and independence properties are the following:

1. If the interval (0, 1) is divided into *n* classes, or subintervals of equal length, the expected number of observations in each interval is *N* / *n* where *N* is the total number observations.

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

Generation of Pseudo-Random Numbers

"Pseudo" means false, so false random numbers are being generated. "Pseudo" is used to imply that the very act of generating random numbers by a known, method removes the potential for true randomness. If the method is known, the set of random numbers can be replicated. Then an argument

can be made that the numbers are not truly random. The goal of any generation scheme, however, is to produce a sequence of numbers between zero and 1 which simulates, or imitates, the ideal properties of uniform distribution and independence as closely as possible.

1)Linear Congruential Method

The linear congruential method, initially proposed by Lehmer [1951], produces a sequence of integers, $X_1, X_2, \ldots$ between zero and m-1 according to the following recursive relationship:

$$X_{i+1} = (a X_i + c) \bmod m, \quad i = 0, 1, 2, \ldots \qquad (1)$$

The initial value X_0 is called the seed, a is called the constant multiplier, c is the increment, and m is the modulus. If $c \neq 0$ in Equation (1), the form is called the *mixed congruential method*. When $c = 0$, the form is known as the *multiplicative congruential method*. The selection of the values for a, c, m, and X_0 drastically affects the statistical properties and the cycle length. Variations of Equation (1) are quite common in the computer generation of random numbers. An example will illustrate how this technique operates.

Example 1

Use the linear congruential method to generate a sequence of random numbers with $X_0 = 33$, $a = 17$, $c = 52$, and $m = 100$. Here, the integer values generated will all be between zero and 99 because of the value of the modulus. Also, notice that random integers are being generated rather than random numbers. These random integers should appear to be uniformly distributed on the integers zero to 99. Random numbers between zero and 1 can be generated by

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

The sequence of X_i and subsequent R_i values is computed as follows:

$X_0 = 33$

$X_1 = (17*33 + 52) \bmod 100 = 613 \bmod 100 = 13$

$R_1 = \frac{13}{100} = 0.13$

$X_2 = (17*13 + 52) \bmod 100 = 273 \bmod 100 = 73$

$R_2 = \frac{73}{100} = 0.73$

] $X_3 = (17*73 + 52) \bmod 100 = 1293 \bmod 100 = 93$

$R_3 = \frac{93}{100} = 0.93$

.

.

.

Example 2

Let $m = 10^2 = 100$, $a = 19$, $c = 0$ and $X_0 = 63$, and generate a sequence of random integers using Equation (1).

$X_0 = 63$

$X_1 = (19)\ (63) \text{mod } 100 = 1197 \text{mod } 100 = 97$

$X_2 = (19)\ (97) \text{mod } 100 = 1843 \text{mod } 100 = 43$

$X_3 = (19)\ (43) \text{mod } 100 = 817 \text{mod } 100 = 17$

.

.

.

When m is a power of 10, say $m = 10^b$, the modulo operation is accomplished by saving the b rightmost (decimal) digits. By analogy, the modulo operation is most efficient for binary computers when $m = 2^b$ for some $b > 0$.

2) Combined Linear Congruential Generators: As computing power has increased, the complexity of the systems that we are able to simulate has also increased One approach is to combine two or more multiplicative congruential generators in such a way that the combined generator has good statistical properties and a longer period. The following result from L'Ecuyer [1988] suggests how this can be done

If $W_{i,1}$, $W_{i,2}$,…, $W_{i,k}$ are any independent, discrete-valued random variables (not necessarily identically distributed), but one of them, say $W_{i,1}$, is uniformly distributed on the integers 0 to $m_1 - 2$, then

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

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

To see how this result can be used to form combined generators, let $X_{i,1}$, $X_{i,2}$,….., $X_{i,k}$ be the *i*th output from *k* different multiplicative congruential generators, where the *j*th generator has prime modulus *mj*, and the multiplier *aj* is chosen so that the period is *mj* -1. Then the *j*th generator is producing integers $X_{i,j}$ that are approximately uniformly distributed on 1 to *mj* -1, and $W_{i,j} = X_{i,j}$ - 1is approximately uniformly distributed on 0 to *mj* -2. L'Ecuyer [1988] therefore suggests combined generators of the form

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

With

$$R_i = \begin{cases} \dfrac{X_i}{m_1}, X_i > 0 \\ \dfrac{m_1 - 1}{m_1}, X_i > 0 \end{cases}$$

Notice that the "$(-1)^{j-1}$" coefficient implicitly performs the subtraction $X_{i,1} - 1$; for example, if k = 2, then $(-1)^0 (X_{i,1} -1)-(-1)^1(X_{i,2} -1) = \sum_{j=1}^{2} (-1)^{j-1} X_{i,j.}$

The maximum possible period for such a generator is

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

The algorithms of testing a random number generator are based on some statistics theory, i.e. testing the hypotheses. The basic ideas are the following, using testing of uniformity as an example.

We have two hypotheses; one says the random number generator is indeed uniformly distributed. We call this H_o, known in statistics as *null hypothesis*. The other hypothesis says the random number generator is not uniformly distributed. We call this H_1, known in statistics as *alternative hypothesis*.

We are interested in testing result of H_0, reject it, or fail to reject it.

To see why we don't say *accept H null*, let's ask this question: what does it mean if we had said *accepting H null*? That would have meant the distribution is truly uniform. But this is impossible to state, without exhaustive test of a *real* random generator with infinite number of cases. So we can

only say *failure to reject H null*, which means no evidence of non-uniformity has been detected on the basis of the test. This can be described by the saying ``so far so good".

On the other hand, if we have found evidence that the random number generator is not uniform, we can simply say *reject H null*.

It is always possible that the H_0 is true, but we rejected it because a sample landed in the H_1 region, leading us to reject H_0 . This is known as *Type I* error. Similaly if H_0 is false, but we didn't reject it, this also results in an error, known as *Type II* error.

With these information, how do we state the result of a test? (How to perform the test will be the subject of next a few sections)

- A level of statistical significance α has to be given. The level α is the probability of rejecting the H null while the H null is true (thus, Type I error).

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

- We want the probability as little as possible. Typical values are 0.01 (one percent) or 0.05 (five percent).
- Decreasing the probability of Type I error will increase the probability of Type II error. We should try to strike a balance

Tests for Random Numbers

The desirable properties of random numbers- uniformity and independence – were discussed earlier. To insure that these desirable properties are achieved, a number of tests can be performed

The tests can be placed in two categories according to the properties of interest. The first entry in the list below concerns testing for uniformity. The second through fifth entries concern testing for independence. The five types of tests discussed in this chapter are as follows:

1. ***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.
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.
3. ***Autocorrelation test***. Tests the correlation between numbers and compares the sample correlation to the expected correlation of zero.

In testing for uniformity, the hypotheses are as follows:

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

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

The null hypothesis, H_0, reads that the numbers are distributed uniformly on the interval [0, 1]. Failure to reject the null hypothesis means that no evidence of no uniformity has been detected on the basis of this test. This does not imply that further testing of the generator for uniformity is unnecessary.

In testing for independence, the hypotheses are as follows:

$H_0 : R_i \sim$ independently

$H_1 : R_i \sim$ independently

This null hypothesis, H_0, reads that the numbers are independent. Failure to reject the null hypothesis means that no evidence of dependence has been detected on the basis of this test. 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. The level α is the probability of rejecting the null hypothesis given that the null hypothesis is true.

Or $\alpha = P$ (reject $H_0 \mid H_0$ true)

The decision maker sets the value of α for any test. Frequently, α is set to 0.01 or 0.05.

If several tests are conducted on the same set of numbers, the probability of rejecting the null hypothesis on at least one test, by chance alone [i.e., making a Type I (α) error], increases. Say that $\alpha = 0.05$ and that five different tests are conducted on a sequence of numbers. The probability of rejecting the null hypothesis on at least one test, by chance alone, may be as large as 0.25.

Frequency test.

A basic test that should always be performed to validate a new generator is the test of uniformity. Two different methods of testing are available. They are the Kolmogorov-Simirnov and the chi-square test. Both of these tests measure the degree of agreement between the distribution of a sample of generated random numbers and the theoretical uniform distribution. Both tests are based on the null hypothesis of no significant difference between the sample distribution and the theoretical distribution.

1. ***The Kolmogorov-Smirnov test***. This test compares the continuous cdf, $F(x)$, of the uniform distribution to the empirical cdf, $S_N(x)$, of the sample of N observations. By definition,

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

If the sample from the random-number generator is $R_1, R_2, \ldots, R_N$, then the empirical cdf, $S_N(x)$, is defined by

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

As N becomes larger, $S_N(x)$ should become a better approximation to $F(x)$, provided that the null hypothesis is true.

The cdf of an empirical distribution is a step function with jumps at each observed value.

The Kolmogorov-Simirnov test is based on the largest absolute deviation between $F(x)$ and $S_N(x)$ over the range of the random variable. That is, it is based on the statistic

$$D = \max \left| F(x) - S_N(x) \right| \qquad (3)$$

The sampling distribution of D is known and is tabulated as a function of N in a table . For testing against a uniform cdf, the test procedure follows these steps:

Step 1. Rank the data from smallest to largest. Let $R_{(i)}$ denote the smallest observation, so that $R_{(1)} \le R_{(2)} \le \ldots \le R_{(N)}$

Step 2 Compute,

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

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

Step 3 Compute, $\mathbf{D = max (D^+, D^-)}$

Step 4 Determine the critical value, D_α from the table of Critical values. For the specified significance level α and the given sample size N.

Step 5 If the calculated statistic D is greater than the critical value D_α , the null hypothesis that the data are a sample from a uniform distribution is rejected. If $D <= D_\alpha$, conclude that no difference has been detected between the true distribution of $\{R_1, R_2, \ldots R_N\}$ and the uniform distribution.

Example:

The sequence of numbers 0.54,,0.73, 0.98 ,0.11 and 0.68 has been generated. Use *Kolmogorov-Smirnov test to determine whether the hypothesis that the numbers are uniformly distributed over (0,1) can be rejected. (*$D_{0.05}$=0.565)

Solution

H_0 : The observations are from Uniform distribution(0,1)

H_1: The observations are not from Uniform distribution(0,1)

$R_{(i)}$	0.11	0.54	0.68	0.73	0.98
i/N	0.2	0.4	0.6	0.8	1.0
i/N- $R_{(i)}$	**0.09**	-	-	0.07	0.02
$R_{(i)}$-(i-1)/N	0.11	**0.34**	0.28	0.13	0.18

D^+=Max{ i/N- $R_{(i)}$}=0.09 and D^- = Max{ $R_{(i)}$-(i-1)/N=0.34

D=Max{D^+, D^- }=0.34 < $D_{0.05}$=0.565

Hence H_0 is not rejected.

2. Chi-Square test:- The chi square test uses the test statistic

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

Where

O_i are observed frequencies and E_i are the expected frequencies for i^{th} class.

For uniform distribution E_i= N/n

Sampling distribution of χ^2 is chi square distribution with (n-1) degrees of freedom.

To apply this test the essential conditions are N $\geq$ 50 and each Ei $\geq$5

If Ei < 5 then the frequencies of the consecutive classes should be combined to make Ei $\geq$ 5.

Example: Using chi square test with α=0.05 test whether the data shown below are uniformly distributed. Test is run for 10intervals of equal length.

0.34, 0.90, 0.89, 0.44, 0.46, 0.67, 0.83, 0.76, 0.70, 0.22,

0.96, 0.99, 0.17, 0.26, 0.40, 0.11, 0.78, 0.18, 0.39, 0.24

0.64, 0.72, 0.51, 0.46, 0.05, 0.66, 0.10, 0.02, 0.52, 0.18,

0.43, 0.37, 0.71, 0.19, 0.22, 0.99, 0.02, 0.31, 0.82, 0.67

0.46, 0.55, 0.08, 0.16, 0.28, 0.53, 0.49, 0.81, 0.64,0.75

Solution

H0: The observations are Uniformly distributed.

H1: The observations are not Uniformly distributed

Classes	Tally marks	Frequency Oi	Exp.freq Ei	(Oi-Ei)	$(Oi-Ei)^2/Ei$
0.0 -0.1		4	5	-1	1/5
0.1-0.2		7	5	2	4/5
0.2-0.3		5	5	0	0
0.3-0.4		4	5	-1	1/5
0.4-0.5		7	5	2	4/5
0.5-0.6		4	5	-1	1/5
0.6-0.7		5	5	0	0
0.7-0.8		6	5	1	1/5
0.8-0.9		4	5	-1	1/5
0.9-1.0		4	5	-1	1/5
Total		50	50	0	2.8

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

= 2.8 < Tab $\chi^2_{0.05,9} = 16.9$ Therefore do not reject H0.

1. ***Runs up and runs down***. Consider a generator that provided a set of 40 numbers in the following sequence:

0.08	0.09	0.23	0.29	0.42	0.55	0.58	0.72	0.89	0.91
0.11	0.16	0.18	0.31	0.41	0.53	0.71	0.73	0.74	0.84
0.02	0.09	0.30	0.32	0.45	0.47	0.69	0.74	0.91	0.95
0.12	0.13	0.29	0.36	0.38	0.54	0.68	0.86	0.88	0.91

Both the Kolmogorov-Smirnov test and the chi-square test would indicate that the numbers are uniformly distributed. However, a glance at the ordering shows that the numbers are successively larger in blocks of 10 values. If these numbers are rearranged as follows, there is far less reason to doubt their independence:

0.41	0.68	0.89	0.84	0.74	0.91	0.55	0.71	0.36	0.30
0.09	0.72	0.86	0.08	0.54	0.02	0.11	0.29	0.16	0.18
0.88	0.91	0.95	0.69	0.09	0.38	0.23	0.32	0.91	0.53
0.31	0.42	0.73	0.12	0.74	0.45	0.13	0.47	0.58	0.29

The runs test examines the arrangement of numbers in a sequence to test the hypothesis of independence.

Before defining a run, a look at a sequence of coin tosses will help with some terminology. Consider the following sequence generated by tossing a coin 10 times:

H T T H H T T T H T

There are three mutually exclusive outcomes, or events, with respect to the sequence. Two of the possibilities are rather obvious. That is, the toss can result in a head or a tail. The third possibility is "no event". The first head is preceded by no event and the last tail is succeeded by no event. Every sequence begins and ends with no event.

A run is defined as a succession of similar events preceded and followed by a different event. The length of the run is the number of events that occurs in the run. In the coin-flipping example above there are six runs. The first run is of length one, the second and third of length two, the forth of length three, and the fifth and sixth of length one.

There are two possible concerns in a runs test for a sequence of numbers. The number of runs is the first concern and the length of runs is a second concern. The types of runs counted in the **first case might be runs up and runs down.** An up run is a sequence of numbers each of which is succeeded by a larger number. Similarly, a down run is a sequence of numbers each of which is succeeded by a smaller number. To illustrate the concept, consider the following sequence of 15 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 0.81

The numbers are given a "+" or a "-"depending on whether they are followed by a larger number or a smaller number. Since there are 15 numbers, and they are all different, there will be 14 +'s and –'s. the last number is followed by "no event" and hence will get neither a + nor a -. The sequence of 14 +'s and –'s is as follows:

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

Each succession of +'s and –'s forms a run. There are given eight runs. The first run is of length one, the second and third are of length three, and so on. Further, there are four runs up and four runs down.

There can be too few runs or too many runs. Consider the following sequence of numbers:

0.08 0.18 0.23 0.36 0.42 0.55 0.63 0.72 0.89 0.91

This sequence has **one run, a run up.** It is unlikely that a valid random-number generator would produce such a sequence. Next, consider the following sequence:

0.08 0.93 0.15 0.96 0.26 0.84 0.28 0.79 0.36 0.57

This sequence has **nine runs, five up and four down**. It is unlikely that a sequence of 10 numbers would have this many runs. What is more likely is that the number of runs will be somewhere

between the two extremes. These two extremes can be formalized as follows: if *N* is the numbers in a sequence, the maximum number of runs is $N - 1$ and the minimum number of runs is one.

If *a* is the total number of runs in a truly random sequence, the mean and variance of *a* are given by

$$\mu_a = \frac{2N-1}{3} \tag{4}$$

and

$$\sigma_a^2 = \frac{16N-29}{90} \tag{5}$$

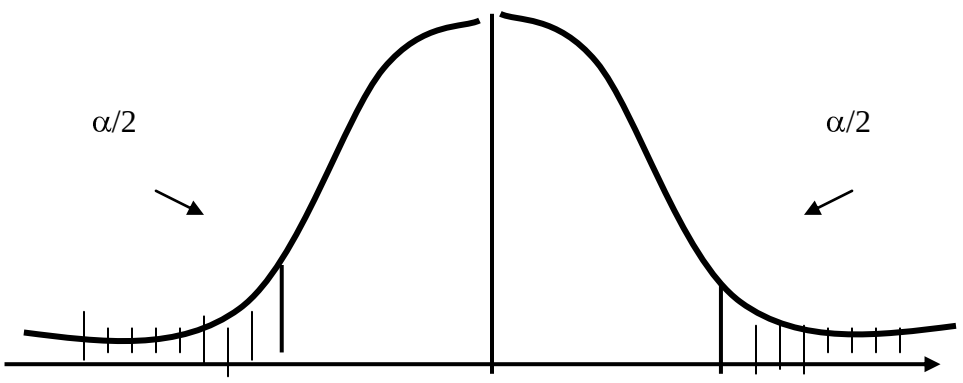
For $N > 20$, the distribution of *a* is reasonably approximately by a normal distribution, $N(\mu_a, \sigma_a^2)$. This approximation can be used to test the independence of numbers from a generator. In that case the standardized normal test statistic is developed by subtracting the mean from the observed number of runs, *a*, and dividing by the standard deviation. That is, the test statistic is

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

Substituting Equation (4) for μ_a and the square root of Equation (5) for σ_a yields

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

Where $Z_0 \sim N(0, 1)$. Failure to reject the hypothesis of independence occurs when $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$, where α is the level of significance. The critical values and rejection region are shown in Figure .



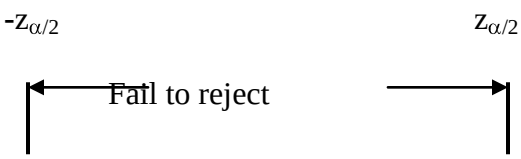


Figure . Failure to reject hypothesis

Example 1

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 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

The sequence of runs up and down is as follows:

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

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

There are 26 runs in this sequence. With $N = 40$ and $a = 26$, Equation (7.4) and (7.5) yield

$$\mu_a = \frac{2(40)-1}{3} = 26.33$$

and

$$\sigma_a^2 = \frac{16(40)-29}{90} = 6.79$$

Then,

$$Z_0 = \frac{26 - 26.33}{\sqrt{6.79}} = -0.13$$

Now, the critical value is $z_{0.025} = 1.96$, so the independence of the numbers cannot be rejected on the basis of this test.

2 Runs above and below mean.

The test for runs up and down is not completely adequate to assess the independence of a group of numbers. Consider the following example for 40 nos

0.63 0.72 0.79 0.81 0.52 0.94 0.83 0.93 0.87 0.67

0.54 0.83 0.89 0.55 0.88 0.77 0.74 0.95 0.82 0.86

0.43 0.32 0.36 0.18 0.08 0.19 0.18 0.27 0.36 0.34

0.31 0.45 0.49 0.43 0.46 0.35 0.25 0.39 0.47 0.41

Mean=0.5565

The sequence of runs up and runs down is as follows

\+ + + - + - + - - - + + - + - - + - +

\- - + - - + - + + - - + + - + - - + + -

Exactly same as example .8

Thus numbers would pass the runs up and runs down test. However the runs can be observed that the first 20 numbers are all above the mean[0.99+00]/2 =0.495 and the last 20 numbers are below the mean .Such an occurrence is highly unlikely .The runs described as being up and down the mean value . A + sign will be used to denote an observation above the mean and a - sign will be denote an observation below the mean

Consider n_1, n_2 be individual observations above and below mean. Let *b* be the total number of runs. Swed and Eisenhart 1943 showed that variance of truly independent sequence is given by

$$\mu_b = \frac{2\,n1\ n2}{N} + \frac{1}{2} \quad \text{----------.6}$$

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

For either n_1 or n_2 greater than 20 b is approximately normally distributed .So the test statistics will be

$$\mathbf{z_b = (b- \mu_b)/ \sigma_b^2}$$

Failure of rejection of hypothesis of independence occurs when

$z_{\alpha/2} \leq \mathbf{z_b} \leq z_{\alpha/2}$ where α is level of significance

Example 7.9

Determine there is an excessive number of a run above and below the means for the sequence of numbers given by Example 7.8. The assignment of+'s and –'s results in the following

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

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

$n_1 = 18$, $n_{2=}$ **22, b= 17 , N= 40**

$$\mu_b = \frac{2(18)(22)}{40} + \frac{1}{2} = 20.3$$

$$\sigma_b^2 = \frac{2(18)(22)(2(18)(22) - 40)}{40^2(40-1)} = 9.54$$

Since, n_2 **>20 normal approximation can be used ,**

Z o = (17- 20.3)/ $\sqrt{9.54}$ **= -1.07**

Since $z_{0.025} = \pm 1.96$, the hypothesis of independence cannot be rejected

3 . ***Runs test: length of runs***

Yet another concern is length of runs. Say two numbers below mean two numbers above the mean. A test of runs above and below the mean would detect no departure from independence. However it is expected that runs other than the length 2 should occur**.**

Here the length of runs are taken into accounts. Let Yi be the number of runs of length in the sequence of N numbers for independence, the expected value of Y_i for runs up and down is given by

$$E(Y_i) = \frac{2}{(i+3)!}\left[N(i^2+3i+1) - (i^3+3i^2-i-4)\right], \quad i \le N-2 \qquad (3.1)$$

$$= \frac{2}{N!}, \quad i = N-1 \qquad (3.2)$$

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

$$E(Y_i) = \frac{Nw_i}{E(I\}}, \quad N>20 \qquad (3.4)$$

$$w_i = \left[\frac{n_1}{N}\right]^i \left[\frac{n_2}{N}\right] + \left[\frac{n_1}{N}\right]\left[\frac{n_2}{N}\right]^i, \quad \mathbf{N>20} \qquad \mathbf{(3.5)}$$

and where E(I), the approximate total number of runs (of all lengths) in sequence of length N , E(A) is given by

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

The approximate test is chi-square test with Oi being observred number of runs of length i . Then the test statistics is

$$\chi^2 = \sum_{i=1}^{L} \left[\frac{\{Oi - E(Yi)\}^2}{E(Yi)} \right]$$

where L=N-1 for runs up and down and L= N for run above and below the mean If the null hypothesis of independence is true , then χ_0^2 is approximately chi-square distributed with L-1 degrees of freedom

Example

Given the following sequence of numbers, can the hypothesis that the numbers are independent be rejected on the basis of the length of runs up and down at α=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

For this sequence the +'s and –'s are as follows:

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

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

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

The length of runs in the sequence is 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

The number of observed runs of each length is as follows:

Run Length , *i*	1	2	3

Observed Runs, O_i	26	9	5

The expected numbers of runs of lengths one, two, and three are computed from Equation () as

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

$$= 25.08$$

$$E(Y_2) = \frac{2}{5!}[60(4+6+1) - (8+12-2-4)$$

$$= 10.77$$

$$E(Y_3) = \frac{2}{6!}[60(9+9+1) - (27+27-3-4)$$

$$= 3.04$$

The mean total number of runs (up and down) is given by Equation (7.4) as

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

Thus far, the $E(Y_i)$ for i = 1, 2, and 3 total 38.89. The expected number of runs of length 4 or more is the difference μ_a - $\sum_{i=1}^{3} E(Y_i)$, or 0.78.

As observed by Hines and Montgomery [1990], there is no general agreement regarding the minimum value of expected frequencies in applying the chi-square test. Values of 3, 4, and 5 are widely used, and a minimum of 5 was suggested earlier in this chapter. Should an expected frequency be too small, it can be combined with the expected frequency in an adjacent class interval. The corresponding observed frequencies would then be combined also, and L would be reduced by one. With the foregoing calculations and procedures in mind, we construct Table 7.4. The critical value $\chi^2_{0.05,2}$ is 3.84. (The degrees of freedom equals the number of class intervals minus

one.)Since $\chi_0^2 = 0.05$ is less than the critical value, the hypothesis of independence cannot be rejected on the basis of this test.

Table. Length of Runs Up and Down: χ^2 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	26	25.08		0.03
2	9	14	10.77	14.59	0.02
≥ 3	5	3.82		
	40	39.67		0.05

Example

Given the same sequence of numbers in above Example, can the hypothesis that the numbers are independent be rejected on the basis of the length of runs above and below the mean at α=0.05? For this sequence, the +'s and –'s are as follows:

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

Mean=0.51

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

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

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

The number of runs of each length is as follows:

Run Length , i	1	2	3	≥ 4
Observed Runs, O_i	17	9	1	5

There are 28 values above the mean (n_1 = 28) and 32 values below the mean (n_2 = 32). The probabilities of runs of various lengths, w_i, are determined from Equation (7.11) as

$$\omega_1 = \left(\frac{28}{60}\right)^1 \frac{32}{60} + \frac{28}{60}\left(\frac{32}{60}\right)^1 = 0.498$$

$$\omega_2 = \left(\frac{28}{60}\right)^2 \frac{32}{60} + \frac{28}{60}\left(\frac{32}{60}\right)^2 = 0.249$$

$$\omega_3 = \left(\frac{28}{60}\right)^3 \frac{32}{60} + \frac{28}{60}\left(\frac{32}{60}\right)^3 = 0.125$$

The expected length of a run, $E(I)$, is determined from Equation () as

$$E(I) = \frac{28}{32} + \frac{32}{28} = 2.02$$

Now, Equation () can be used to determine the expected numbers of runs of various lengths as

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

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

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

The total number of runs expected is given by Equation () as

$E(A) = 60/2.02 = 29.7$.

This indicates that approximately 3.8 runs of length four or more can be expected. Proceeding by combining adjacent cells in which $E(Y_i) < 5$ produces following Table

Table Length of Runs Above and Below the Mean: χ^2 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	14.79	0.33
2	9	7.40	0.35
3	1 6	3.71 7.51	0.30
≥ 4	5	3.80	
	32	29.70	0.98

The critical value $\chi^2_{0.05,2}$ is 5.99. (The degrees of freedom equals the number of class intervals minus one.) Since $\chi^2_0 = 0.98$ is less than the critical value, the hypothesis of independence cannot be rejected on the basis of this test.

Autocorrelation

Example

Test whether the 3rd, 8th, 13th, and so on, numbers in the sequence at the beginning of this section are auto correlated.

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

(Use $\alpha = 0.05$.) Here, $i = 3$ (beginning with the third number), $m = 5$ (every five numbers), $N = 30$ (30 numbers in the sequence), and $M = 4$ (largest integer such that $3 + (M + 1)\,5 \leq 30$). Then,

$$\hat{\rho}_{35} = \frac{1}{4+1}\left[(0.23)(0.28) + (0.28)(0.33) + (0.33)(0.27) + (0.27)(0.05) + (0.05)(0.36)\right] + 0.25$$
$$= -0.1945$$

and

$$\sigma_{\hat{\rho}35} = \frac{\sqrt{13(4)+7}}{12(4+1)} = 0.1280$$

Then, the test statistic assumes the value

$$Z_0 = \frac{-0.1945}{0.1280} = -1.516$$

Now, the critical value is

$$z_{0.025} = 1.96$$

Therefore, the hypothesis of independence cannot be rejected on the basis of this test.

It can be observed that this test is not very sensitive for small values of *M*, particularly when the numbers being tested are not on the low side. Imagine what would happen if each of the entries in the foregoing computation $\hat{\rho}im$ of were equal to zero. Then, $\hat{\rho}im$ would be equal to -0.25 and

the calculated Z would have the value of -1.95, not quite enough to reject the hypothesis of independence.

Many sequences can be formed in a set of data, given a large value of *N*. for example, beginning with first number in the sequence, possibilities include (1) the sequence of all numbers, (2) the sequence formed from the first, third, fifth,…., numbers, (3) the sequence formed from the first, fourth,…, numbers, and so on. If $\alpha = 0.05$, there is a probability of 0.05 of rejecting a true hypothesis. If 10 independent sequences are examined, the probability of finding no significant auto correlation, by chance alone, is $(0.95)^{10}$ or 0.60. Thus, 40% of the time significant auto correlation would be detected when it does not exist. If α is 0.10 and 10 tests are conducted, there is a 65% chance of finding auto correlation by chance alone. In conclusion, when "fishing" for auto correlation, upon performing numerous tests, auto correlation may eventually be detected, perhaps by chance alone, even when no auto correlation is present.

Run test :

Example

Consider the following sequence of 40 numbers.

0.90, 0.89, 0.44, 0.21, 0.67, 0.17, 0.46, 0.83, 0.79, 0.40, 0.94, 0.22, 0.66, 0.42, 0.99, 0.67, 0.41, 0.73, 0.02, 0.72, 0.43, 0.47, 0.17, 0.56, 0.45, 0.78, 0.56, 0.30, 0.71, 0.19, 0.93, 0.37, 0.42, 0.96, 0.73, 0.47, 0.60, 0.29, 0.78, 0.26

Based on the runs ups and downs, determine whether the hypothesis of independence (random) can be rejected.

Solution:

Ho: Sequence is random

H1: Sequence is not random

0.90	0.89	0.44	0.21	0.67	0.17	0.46	0.83	0.79	0.40
	-	-	-	+	-	+	+	-	-

0.94	0.22	0.66	0.42	0.99	0.67	0.41	0.73	0.02	0.72
+	-	+	-	+	-	-	+	-	+
0.43	0.47	0.17	0.56	0.45	0.78	0.56	0.30	0.71	0.19
-	-	-	+	-	+	-	-	+	-
0.93	0.37	0.42	0.96	0.73	0.47	0.60	0.29	0.78	0.26
+	-	+	+	-	-	+	-	+	-

Number of runs= r = 29, sample size =n=40

$\mu = (2n-1)/3 = 26.33$

$\sigma = \sqrt{\frac{16n-29}{90}} = 2.6055496$

$$Z = \frac{r-\mu}{\sigma} = \frac{29-26.33}{2.6055} = 1.024735$$

At 5% los $Z_{\alpha/2}=1.96$

Cal Z< tab Z so do not reject the hypothesis H0

Tests for Auto-correlation

- The tests for auto-correlation are concerned with the dependence between numbers in a sequence.
- The list of the 30 numbers on page 311 appears to have the effect that every 5th number has a very large value. If this is a regular pattern, we can't really say the sequence is random.
- The test computes the auto-correlation between every *m* numbers (*m* is also known as the lag) starting with the *i*th number.

  Thus the autocorrelation ρ_{im} between the following numbers would be of interest.

$$R_i, R_{i+m}, R_{i+2m}, \ldots R_{i+(M+1)m}$$

The value *M* is the largest integer such that i+(M+1)m ≤ N where *N* is the total number of values in the sequence.

E.g. N = 17, i = 3, m = 4, then the above sequence would be 3, 7, 11, 15 (M = 2). The reason we require *M+1* instead of *M* is that we need to have at least two numbers to test (M = 0) the autocorrelation.

- Since a non-zero autocorrelation implies a lack of independence, the following test is appropriate

$$H_0: \ \rho_{im} = 0$$

$$H_1: \ \rho_{im} \neq 0$$

- For large values of *M*, the distribution of the estimator ρ_{im}, denoted as $\hat{\rho_{im}}$, is approximately normal if the values $R_i, R_{i+m}, R_{i+2m}, \ldots R_{i+(M+1)m}$ are uncorrelated.
- Form the test statistic

$$Z_0 = \frac{\hat{\rho_{im}}}{\sigma_{\hat{\rho_{im}}}}$$

which is distributed normally with a mean of zero and a variance of one.

- The actual formula for $\hat{\rho_{im}}$ and the standard deviation is

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

and

$$\sigma_{\hat{\rho_{im}}} = \frac{\sqrt{13M+7}}{12(M+1)}$$

- After computing Z_0, do not reject the null hypothesis of independence if

$$-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$$

where $\alpha$is the level of significance.

Gap Test

- The gap test is used to determine the significance of the interval between recurrence of the same digit.
- A gap of length *x* occurs between the recurrence of some digit.
- See the example on page 313 where the digit 3 is underlined. There are a total of eighteen 3's in the list. Thus only 17 gaps can occur.
- The probability of a particular gap length can be determined by a Bernoulli trail.

$$P(\text{gap of } n) = P(x \neq 3)P(x \neq 3)...P(x \neq 3)P(x = 3)$$

If we are only concerned with digits between 0 and 9, then

$$P(\text{gap of } n) = 0.9^n 0.1$$

The theoretical frequency distribution for randomly ordered digits is given by

$$P(gap \leq x) = F(x) = 0.1 \sum_{n=0}^{x} (0.9)^n = 1 - 0.9^{x+1}$$

- Steps involved in the test.

Step 1. :Specify the cdf for the theoretical frequency distribution given by Equation () based on the selected class interval width ().

Step 2. :Arrange the observed sample of gaps in a cumulative distribution with these same classes.

Step 3. :Find *D*, the maximum deviation between *F(x)* and SN(x)as in Equation 8.3

Step 4. :Determine the critical value, D_α, from Table A.8 for the specified value of $\alpha$and the sample size N.

Step 5. :If the calculated value of *D* is greater than the tabulated value of D_{α}, the null hypothesis of independence is rejected.

Poker Test

- The poker test for independence is based on the frequency in which certain digits are repeated in a series of numbers.
- In a three digit number, there are only three possibilities.
    1. The individual digits can be all different. Case 1.
    2. The individual digits can all be the same. Case 2.
    3. There can be one pair of like digits. Case 3.
- P(case 1) = P(second differ from the first) * P(third differ from the first and second)

  = 0.9 * 0.8 = 0.72

  P(case 2) = P(second the same as the first) * P(third same as the first) = 0.1 * 0.1 = 0.01
  P(case 3) = 1 - 0.72 - 0.01 = 0.27

  Consider the data on three digits

    1. All 4 digits are different P(A1)=0.9*0.8*0.7=0.504
    2. One pair is same. P(A2)=6*0.1*0.9*0.8=0.432
    3. Two pairs are same P(A3)= 6*0.1*1/9=.0666
    4. Three like digits P(A4)=4*.1*.1*.9=0.036
    5. All digits are same P(A5)=0.1*0.1*0.1=.001