Chapter-5

Probability Concept and Random Number Generation

Generation of random number:

Random number can be generated by following methods:

1. Random numbers may be drawn from the random number tables stored in the computer memory. It is very slow process and the number considerably occupy space of computer memory.
2. An electronic device may be constructed as part of a digital computer to generate truly random number. This is however considered as expensive.
3. Pseudo random numbers (PRN) may be generated by using mathematical formulas and arithmetic operation. This method commonly specifies a procedure, where starting with an initial number, the second number is generated and from that third number and so on. A number of recursive procedures are used for generating random number.

One of the methods for generating Pseudo Random Number is Mid Square method. It starts with fixed initial value, say 4-digit integer called seed. The number is squared and the middle four digit of this square becomes the second number. The middle digit if this second number are then squared again to generate third random number and so on. We may also have to add zero to make the digit's length eight if necessary. Finally, we get realization from the uniform (0,1) distribution after placement of decimal points i.e., after division by 10000.

Example: if we take seed Z_0 = 1234, then we will get the sequence of numbers as 0.1234, 0.5227, 0.3215,0.3362 0.3030, 0.1809 .....

#Generate the random number sequence of number for Z_0 = 2100 using mid square method.

Quality of an efficient random number Generator

- It should have a sufficiently long cycle i.e.; it should be sufficiently long sequence of random numbers before beginning to repeat the sequence.
- The random numbers generated should be replicable i.e., by specifying a starting condition, it should be possible to obtain the same set of random numbers. Many times, common random numbers are required for the comparison of two systems.
- The generated random number should fulfill the requirement of uniformity and independence.
- The random number generator should be fast and cost-effective.
- It should be portable to different computer and ideally to different programming language.

Testing number for randomness

A sequence of random number is considered to be random if

- The number are uniformly distributed i.e., every number has an
- 0. equal chance of occurrence.
- The number are not serially auto-corelated i.e., there is no correlation between adjacent pair or number, or the appearance of one number doesn't influence the appearance of next number.

There are a number of tests, which are used to ensure that random numbers are uniformly distributed and are not serially auto co-related.

Uniformity Test (Frequency Test)

- The test of uniformity or frequency test is basic test that should always be performed to validate a random number generator. The uniformity test counts how often numbers in a given range occur in the sequence to ensure that the number are uniformly distributed. Two frequency tests are available and they are
  - Kolmogorov-Smirnov test i.e., K-S test.
  - Chi-square test

Both of these tests compare the generated random number with the theoretical uniform distribution. 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_0, known in statistics as null hypothesis. The other hypothesis says the random number generator is not uniformly distributed. We call this H_a, known in statistics as alternative hypothesis. We are interested in testing result of H_0, reject it or fail to reject it.

K-S Test

- It is a statistical hypothesis test.
- The test is non-parametric and entirely agnostic.
- K-S test can be used to compare actual data to normal distribution.
  - The cumulative probabilities of value in the data are compared with the cumulative probabilities in a theoretical normal distribution.
- **Null hypothesis:** Sample is taken from a normal distribution.
- The critical value of $D\alpha$ is found from the K-S table values for one sample test(default=0.565), where $\propto$ is level of significance.
- **Acceptance Criteria:** If calculated value is less than critical value accepts null hypothesis ($D < D\alpha$).

- **Rejection Criteria:** If calculated value is greater than the table value rejects null hypothesis.

K-S Test Limitation:

- It only applies to continuous distributions.
- It tends to be more sensitive near the center of distributions than the tails.
- It typically determined by simulation.

The K-S test is defined by:

- **H_0**: The data follow a specified distribution.
- **H_a**: The data don't follow a specified distribution.

Test Statistic:

The K-S test statistic is defined as

D= max $_{1 \le Y \le N}$ $[F(yi) - \frac{i-1}{N}, \frac{i}{N} - F(yi)]$

Example:

Consider the sequence of 5 numbers

0.15, 0.94, 0.05, 0.51 and 0.29

Given $\propto$ = 0.05

Critical value D$\propto$ = 0.565

Null hypothesis: Whether the hypothesis of uniformity can be rejected

Solution:

i	**1**	**2**	**3**	**4**	**5**
F(yi)	0.05	0.15	0.29	0.51	0.94
$\frac{i}{N}$	0.2	0.4	0.6	0.8	1
$\frac{i}{N} - F(Yi)$	0.15	0.25	0.31	0.29	0.06
$\frac{i-1}{N}$	0	0.20	0.40	0.60	0.8
F(Yi)-[$\frac{i-1}{N}$]	0.05	-0.05	-0.11	-0.09	**0.14**

We know,

$$D = \max_{1 \le Y \le N} \left[F(yi) - \frac{i-1}{N}, \frac{i}{N} - F(yi)\right]$$

Now,

$$\max_{1 \le Y \le N} \left[F(yi) - \frac{i-1}{N}\right] = 0.14$$

$$\max_{1 \le Y \le N} \left[\frac{i}{N} - F(yi)\right] = 0.31$$

Therefore $\mathbf{D = \max_{1 \le Y \le N} \left[F(yi) - \frac{i-1}{N}, \frac{i}{N} - F(yi)\right]}$

=max (0.14,0.31)

i.e., D=0.31

Since D< D∝ i.e., 0.31 < 0.565 at $\propto = 0.05$; So the Null hypothesis is Accepted

Chi-Square Test:

- The sampling method used is simple random sampling.
- The variables under study are each categorical that is they belong to each category.
- **Significance level**: Significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- **Test method:** Use the chi-square test for independence to determine whether there is significant relationship between two categorical variables it can be even done for single variables of time now .
- Degree of freedom: The degree of freedom (DF) is DF = (r-1) * (c-1)

  where, r stands for rows and **c** stands for columns

- Test Statistic: The test statistic is a chi-square random variable

$$(X^2) = \Sigma\left[(O_{r,c} - E_{r,c})^2 / E_{r,c}\right]$$

where,
$O_{r,c}$ -> Observation value in rows and column
$E_{r,c}$ -> Estimated value in rows and column

- P-value: The P-value is the probability of observing a sample statistic as extreme as the test statistic.

Example:

256 people were surveyed to find out their zodiac sign. The results were: Aries(29), Taurus(24), Gemini(22), Cancer(19), Leo(21), Virgo(18), Libra(19), Scorpio(20),

Sagittarius(23), Capricorn(18), Aquarius(20), Pisces(23). Test the hypothesis that zodiac signs are evenly distributed across visual artists.

Count	Observed	Expected	Observation(Obs)-Expectation(Exp)	$(Obs - Exp)^2$	$\frac{(Obs - Exp)^2}{Exp}$
1	29	21.33333	7.66667	58.7778289	2.75521116
2	24	21.33333	2.66667	7.11112889	0.333334219
3	22	21.33333	0.66667	0.444448889	0.0208335449
4	19	21.33333	-2.33333	5.44442889	0.255207645
5	21	21.33333	-0.33333	0.11110889	0.00520823003
6	18	21.33333	-3.33333	11.1110889	0.520832374
7	19	21.33333	-2.33333	5.44442889	0.255207644
8	20	21.33333	-1.33333	1.77776889	0.0833329297
9	23	21.33333	1.66667	2.77778889	0.130208875
10	18	21.33333	-3.33333	11.1110889	0.520832374
11	20	21.33333	-1.33333	1.77776889	0.0833329297
12	23	21.33333	1.66667	2.77778889	0.130208875
	256				5.09375

Expected = $\frac{\text{Observed}}{\text{Number of Observation Presented}}$ = $\frac{256}{12}$ = 21.33333

Degree of freedom=(12-1)=11

Here we have only one values so there was countless strengths well minus 1 is 11.

Chi-Square Statistics = 5.094

CHI SQUARE TABLE

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

Now, from the table chi-square value will be 0.900 and 0.950. So, the p-value for the chi-square statistics or 5.094 and the degree of freedom lies between 0.900 and 0.950.

As the P-value is greater than the level of significance ($\alpha = 0.05$). Hence, the null hypothesis cannot be rejected. Therefore, the value is accepted.

Independence Test

- Sometime the random number generator passes the K-S test and Chi-Square tests for Uniformity, but the number generated may not be independent.
- Hence there is need of independence test.
- There are many methods to check the independence of generators. Two among them are:
  - Auto correlation test
  - Run test

Auto Correlation test

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

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

- Three classes will be as:
  a) Less than or equal to 0.33
  b) Less than or equal to 0.67
  c) Less than or equal to 1.0

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

Runs Tests:

Run- The succession of similar events preceded and followed by a different event is called as run.
Run-length- Number of events that occur in the run.
Example- Tossing coin
Consider the sequence of tossing a coin 10 times: H T T H H T T T H T

Number	Run length	Run
1	1	H
2	2	T T
3	2	H H
4	3	T T T
5	1	H
6	1	T

There are two possible concerns in run tests. They are
a) **Number of runs- Runs Up and Down** & **Runs Above and Below the Mean**
b) Length of runs

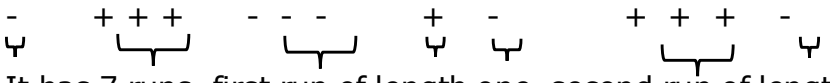
1.Runs Up and Down:

a) **Up-run**- Sequence of numbers each of which is succeeded by a large number is called as up run.
b) **Down-run-** Sequence of numbers each of which is succeeded by a small number is called as down run.
c) If a number is followed by a larger number, then it denoted by '+'. If followed by a smaller number then by '-'.

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

The up run and down run are marked as
-0.87 +0.15 +0.23 +0.45 -0.69 -0.32 -0.30 +0.19 -0.24 +0.18 +0.65 +0.82 -0.93 +0.22

The sequence of '+' and '-' are



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

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

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

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

$$z_0 = \frac{a-\mu_a}{\sigma_a}$$

Substituting μ_a and σ_a in above equation we get,

$$z_0 = \frac{a - \frac{(2N-1)}{3}}{\sqrt{\frac{(16N-29)}{90}}}$$

where z_0 ~ N(0,1)

The null hypothesis is accepted when $-Z_{\frac{\alpha}{2}} \leq Z_0 \leq Z_{\frac{\alpha}{2}}$, *where α is the level of significance.*

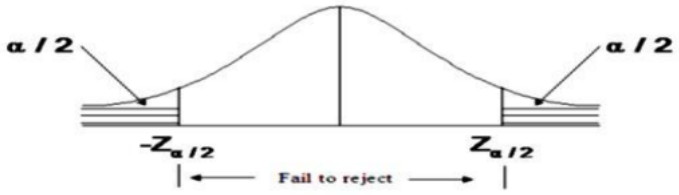


Fig: Accept the Null Hypothesis

Example:

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

0.41	0.68	0.89	0.94	0.74	0.91	0.55	0.62	0.36	0.27
0.19	0.72	0.75	0.08	0.54	0.02	0.01	0.36	0.16	0.28
0.18	0.01	0.95	0.69	0.18	0.47	0.23	0.32	0.82	0.53

0.31	0.42	0.73	0.04	0.83	0.45	0.13	0.57	0.63	0.29

Solution:

The sequence of runs up and down is as follows: -

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

Number of runs -> a =26

N =40

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

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

$z_0 = \frac{a - \frac{(2N-1)}{3}}{\sqrt{\frac{(16N-29)}{90}}} = \frac{26 - 26.33}{\sqrt{6.79}} = -0.13$

Critical value-> $Z_{\frac{a}{2}} = Z_{0.025} = 1.96$ (from z-table)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

1.1 – Negative Z Table

$$Z_{\frac{a}{2}} \leq Z_0 \leq Z_{\frac{a}{2}} = -1.96 \leq -0.13 \leq 1.96$$

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

Disadvantage of Runs Up and Down:

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

2. Runs Above and Below the Mean

- Runs are described with above/below the mean value. A'+' sign is used to indicate above mean and '-' sign for below the mean.
- To illustrate the above consider the sequence of 2-digit random numbers

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

For decimal eg: 0.1 ; we take mean as 0.495 and for whole integer eg: 10; we take mean as 49.5.

Therefore, mean = 0.495

- The runs above and below mean are marked as:

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

- The sequence of '+' and '-' are

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

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

Let n1-> No. of individual observations above mean
n2-> No. of individual observations below mean
b -> Total number of runs.
N -> Maximum number of runs, where N = n1+n2

The mean is given by

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

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

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

$$Z_0 = \frac{b - (\frac{2n_1 n_2}{N}) - \frac{1}{2}}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - N)}{N^2(N-1)}}}$$

  - The null hypothesis is accepted when $-Z_{\frac{\alpha}{2}} \leq Z_0 \leq Z_{\frac{\alpha}{2}}$, where α is the level of significance.

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

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

Solution:
Mean = 0.495
The sequence of runs above and below the means is as follows:

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

$n_1 = 18$
$n_2 = 22$
$N = n_1 + n_2 = 40$
$b = 17$

$\mu_b = \frac{2n_1 n_2}{N} + \frac{1}{2} = \frac{2*18*22}{40} + \frac{1}{2} =$ **20.3**

$\sigma_b^2 = \frac{2n_1 n_2 (2n_1 n_2 - N)}{N^2(N-1)} = \frac{2*18*22(2*18*22-40)}{40^2(40-1)} =$ **9.54**

Since $n_2 \geq 20$, normal approximation is accepted.

$Z_0 = \frac{b - (\frac{2n_1 n_2}{N}) - \frac{1}{2}}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - N)}{N^2(N-1)}}} =. \frac{17 - 20.3}{\sqrt{9.54}} =$ **- 1.07**

Critical value -> $Z_{\frac{a}{2}}$ -> $Z_{0.025}$ = 1.96(from z-table)

$-Z_{\frac{a}{2}} \leq Z_0 \leq Z_{\frac{a}{2}}$ =. -1.96 ≤ **-1.07** ≤ 1.96

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

Disadvantages of Runs Above and Below Mean

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

3 Runs Test length of Runs

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

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

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

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

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

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

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

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

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

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

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

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

The appropriate test is chi-square test with 0_i, the observed number of runs length i. The test statistic is given by

$$\chi_0{}^2 = \sum_{i=1}^{L} \frac{[O_i - E(Y_i)]^2}{E(Y_i)}$$

Where,
L = N-1 for runs up and down
L = N for runs above and below mean
If null hypothesis of independence is true then $\chi_0{}^2$ is approximately chi-squared distributed with L-1 degrees of freedom.

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

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

Solution:
N = 60
The sequence of '+' and '-' are as follows

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

The length of run in the sequence as follows
1,2,1,1,1,1,2,1,1,1,2,1,2,1,1,1,1,2,1,1,
1,2,1,2,3,3,2,3,1,1,1, 3,1,1,1,3,1,1,2,1

Calculate 0_i.

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

The expected value of Y_i,
For run length one,
E(Y_1) = $\frac{2}{(1+3!)}$[60(1^2+ 3(1)+1)-(1^3+3$(1)^2$-1-4)] = 25.08

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

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

Therefore, $E(Y_1) + E(Y_2) + E(Y_3) = 38.89$

We find means (runs up and down)
$\mu_\alpha = \frac{(2N-1)}{3} = \frac{(2*60-1)}{3} = 39.67$

Expected value, when i≥ 4
$\mu_\alpha = \sum_{i=1}^{3} E(Y_i)$ = 39.67-38.89 =0.78

To find χ_0^2, the calculations and procedures are shown in the table below:

Run length(i)	**Observed number of runs(O_i)**	**Expected number of runs E(Yi)**	$[\frac{Oi- E(Yi)]2}{E(Yi)}$]
1	26	25.08	0.03
2	9	10.77	
3	5 **14**	3.04 **14.59**	**0.02**
4	0	0.78	
-	40	39.67	χ_0^2 = 0.05

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

$\chi_0^2 < \chi^2_{0.05,1}$ = 0.05<3.84
Therefore, the hypothesis of independence is accepted.

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

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

Solution
N = 60
Mean = 0.495
The sequence of + and - are as follows

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

n_1 = 28
n_2 = 32
N = n_1+ n_2 = 60
The length of runs in the sequence is as follows

4,1,1,1,1,2,1,1,1,2,2,1,1,1,1,1,2,7,2,2,4,2,2,1,1,1,3,4,5,1,1
Calculate O_i

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

The probabilities of runs of various lengths w_i are as follows

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

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

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

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

$$E(I) = \frac{n_1}{n_2} + \frac{n_2}{n_1} = \frac{28}{32} + \frac{32}{28} = 2.02$$

The expected number of runs of various lengths is

$$E(Y_1) = \frac{Nw_1}{E(I)} = \frac{60*0.498}{2.02} = 14.79$$

$$E(Y_2) = \frac{Nw_2}{E(I)} = \frac{60*0.249}{2.02} = 7.40$$

$$E(Y_3) = \frac{Nw_3}{E(I)} = \frac{60*0.125}{2.02} = 3.71$$

The expected total number of runs is

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

For i≥ 4

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

To find χ_0^2, the calculations and procedures are shown in the table below:

Run length(i)	**Observed number of runs(O_i)**	**Expected number of runs E(Yi)**	$\left[\frac{Oi - E(Yi)]2}{E(Yi)}\right]$
1	17	14.79	0.33
2	8	7.40	0.05
3	1	3.71	**0.30**
≥4	5	3.80	
	6	**7.51**	
-	31	29.70	$\chi_0^2 = 0.68$

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

$\chi_0^2 < \chi^2_{0.05,2} = 0.68 < 5.99$

Therefore, the hypothesis of independence is accepted.