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.

The Gap *Test* measures the number of digits between successive occurrences of the same digit. The following example illustrates the length of gaps associated with the digit 3.

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

There are a total of eighteen 3's in the list. Thus only 17 gaps can occur and the first gap is of length 10 (count digit after first 3 digit to second 3 digit), the second gap is of length 7, and so on.

The probability of a particular gap length can be determined by a Bernoulli trail.

$$P(\text{gap of } 10) = P(x \neq 3)\, P(x \neq 3) \ldots P(x \neq 3)\, P(x = 3) = (0.9)^{10}\,(0.1)$$

Since the probability that any digit is not a 3 is 0.9, and that any digit is a 3 is 0.1. In general. If we are only concerned with digits between 0 and 9, then

$$P(\text{gap of } n) = 0.9^{n}\,0.1 \quad \text{where } n = 0.1, 2\ldots$$

The theoretical frequency distribution for randomly ordered digit is given by

$$F(x) = 0.1 \sum_{n=0}^{x} (0.9)^{n} = 1 - 0.9^{(x+1)} \qquad \ldots(5)$$

Note: Observed frequencies for all digits are compared to the theoretical frequency using the Kolmogorov-Smirnov test

Note: Probability of occurrence for certain digit is 0.1.

Note: Observed frequencies for all digits are compared to the theoretical frequency using the Kolmogorov-Smirnov test

Note: Probability of occurrence for certain digit is 0.1.

There are following step for gap test:

1. Specify the cdf for the theoretical frequency distribution given by Equation (5) based on the selected class interval width (See Table 9.6 for an example).
2. Arrange the observed sample of gaps in a cumulative distribution with these same classes.
3. Find D, the maximum deviation between $F(x)$ and S_N (x), as in Equation $D = \max [F(x) - S_N(x)]$.
4. Determine the critical value, D_{α}, from standard table for the specified value of a and the sample size N.
5. If the calculated value of D is greater than the tabulated value of D_{α}, the null hypothesis of independence is rejected.

Example: Based on the frequency with which gaps occur, analyze the 110 digits

1. Specify the cdf for the theoretical frequency distribution given by Equation (5) based on the selected class interval width (See Table 9.6 for an example).
2. Arrange the observed sample of gaps in a cumulative distribution with these same classes.
3. Find D, the maximum deviation between $F(x)$ and S_N (x), as in Equation $D = \max [F(x) - S_N(x)]$.
4. Determine the critical value, D_α, from standard table for the specified value of a and the sample size N.
5. If the calculated value of D is greater than the tabulated value of D_α, the null hypothesis of independence is rejected.

Example: Based on the frequency with which gaps occur, analyze the 110 digits above to test whether they are independent. Use $\alpha = 0.05$. The number of gaps is given by the number of digits minus 10, or 100. The number of gaps associated with the various digits are as follows:

Digit	0	1	2	3	4	5	6	7	8	9
No. of gaps	7	8	8	17	10	13	7	8	9	13

Gap Test Example

Table 9.2

Gap length	Frequency	Relative frequency	Cum. Relative frequency	$F(x)$	$\mid F(x) - SN(x) \mid$
0–3	35	0.35	0.35	0.3439	0.0061
4–7	22	0.22	0.57	0.5695	0.0005
8–11	17	0.17	0.74	0.7176	0.0224
12–15	9	0.09	0.83	0.8147	0.0153
16–19	5	0.05	0.88	0.8784	0.0016
20–23	6	0.06	0.94	0.9202	0.0198
24–27	3	0.03	0.97	0.9497	0.0223
28–31	0	0.00	0.97	0.9657	0.0043
32–35	0	0.00	0.97	0.9775	0.0075
36–39	2	0.02	0.99	0.9852	0.0043
40–43	0	0.00	0.99	0.9903	0.0003
44–47	1	0.01	1.00	0.9936	0.0064

The gap test is presented in table 9.6.

The critical value of D is given by

$$D_{0\ .05} = 1.36/\sqrt{100} \ = 0.136$$

Since $D = \max \mid F(x) - S_N(x) \mid = 0.0224$ is less than $D_{0.05}$, do not reject the hypothesis of independence on the basis of this test.

Runs Tests

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.

1. **Runs up and down:** The runs test examines the arrangement of numbers in a sequence to test the hypothesis of independence. Consider the 40 numbers; both the Kolmogorov-Smirnov and Chi-square would indicate that the numbers are uniformly distributed. But, not quite independent.

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

- A run is defined as a succession of similar events proceeded and followed by a different event. For example, in a sequence of tosses of a coin, we may have

  H T T H H T T T H T

  Six runs: length 1, 2, 2, 3, 1, 1
- Two characteristics: number of runs and the length of run
  - ❑ An up run is a sequence of numbers each of which is succeeded by a larger number;
  - ❑ A down run is a squence of numbers each of which is succeeded by a smaller number
- If a sequence of numbers have too few runs, it is unlikely a real random sequence, e.g., 0.08, 0.18, 0.23, 0.36, 0.42, 0.55, 0.63, 0.72, 0.89, 0.91, the sequence has one run, an up run. It is not likely a random sequence.

- If a sequence of numbers have too many runs, it is unlikely a real random sequence, e.g., 0.08, 0.93, 0.15, 0.96, 0.26, 0.84, 0.28, 0.79, 0.36, 0.57. It has nine runs, five up and four down. It is not likely a random sequence.

For sequence of random numbers, we can define up runs and down runs (successive numbers are increasing or decreasing)

–0.87, + 0.15, + 0.23, + 0.45, –0.69, –0.32, –0.30, + 0.19, 0.24.

+ + + – – – +

–0 87, + 0.15, + 0.23, + 0.05, –0.69, + 0.32, –0.40, + 0.19, 0.24.

+ + + – + – +

Note: 1. If first no. has (–) sign it means second no. is less than first and if first no. has (+) sign it means second no. is greater than the first no. and so on.

2. If N is the number of 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 is given by

$$\mu_a = \frac{2N-1}{3} \qquad ...(6)$$

and

$$\sigma^2_a = \frac{16N-29}{90} \qquad ...(7)$$

Example: In this sequence – + + + – + – + what is the value of a?

Answer: – (1)

+ + + (2)

– (3)

+ (4)

– (5)

+ (6)

Hence value of $a = 6$.

- For $N > 20$, the distribution of a is reasonably approximated by a normal distribution, $N(\mu_a, \sigma_a^2)$. Converting it to a standardized normal distribution by

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

Substituting equation (6) for μ_a and the square root of equation (7) for σ_a yields:

$$Z_0 = \frac{a - [(2N - 1)/3]}{\sqrt{(16N - 29)/90}} \quad \text{where } Z_0 \sim N(0, 1).$$

- Failure to reject the hypothesis of independence occurs when $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$, where the α is the leve of significance.

Acceptance region for hypothesis of independence $- z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$

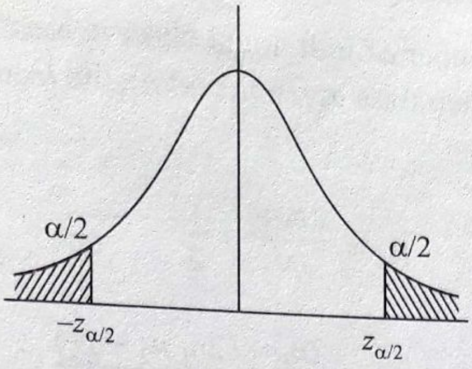

Fig. 9.1

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 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 0.76 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 $a = 26$,

$$\mu_a = \{2(40) - 1\} / 3 = 26.33 \text{ and}$$

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

Then,

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

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.

9.6.2 The Rejection Method

The rejection method is use to obtain samples from a given non-uniform distribution basically works by generating uniform random numbers repeatedly and accepting only those that meet a certain requirement. This requirement of acceptance is so designed that the accepted numbers appear to be drawn from the given distribution. For the rejection method to be applicable, the probability density function $f(t)$ of the distribution must be non-zero only over a finite interval, say (X, Y). Let $f(t)$ be bounded by some upper limit Z as in Fig. 9.3.

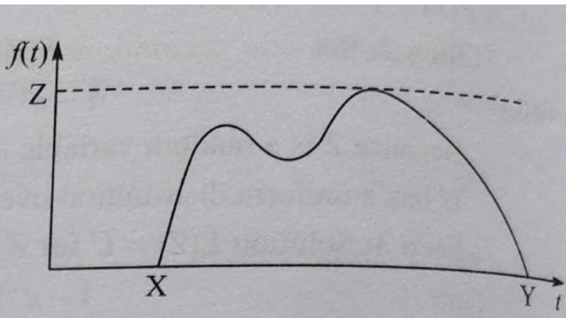

Fig. 9.3 Probability density function.

The procedure of rejection method consists following steps:

Step 1: Generate a pair of uniform random number U_1, U_2 in the interval (0, 1).

Step 2: Using U_1 locate a point m on the horizontal axis as

$$m = X + (Y - X) \cdot U_1$$

Step 3: Using U_2, locate a point n on the vertical axis as

$$n = p \cdot U_2$$

(**Note:** Z is upper limit)

Step 4: If $n > f(m)$ reject the pair and go to the Step 1, otherwise, accept m as the value of a sample from the desired distribution. This procedure is repeated until the required number of samples have been generated.

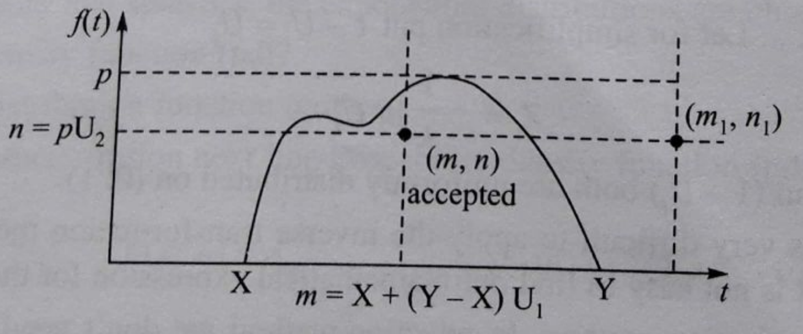

Fig. 9.4 Rejection method.

The above figure shows how the reject method can be used to generate samples from probability density function.

Some Conclusions:

1. All the points above the curve *f*(*t*) in the interval (*X*, *Y*) are rejected, for example point (m_1, n_1).
2. The points that are accepted, are within the boundary of the curve and therefore are distributed according density function *f*(*t*) *e.g.* point (*m*, *n*).
3. The restriction is that, this method will work only for a finite interval.