LECTURE-6

TESTS OF INDEPENDENCE AND GOODNESS OF FIT

OBJECTIVES :

The objectives of this lecture are to explain :

(i) the tests used for checking the randomness of data series in time and space demain, and

(ii) the tests used for checking the goodness of fit of known probability distributions with the empirical frequency distribution.

6.1 INTRODUCTION

Statistical analysis of hydrologic data often assumes certain conditions of the data which must be tested before proceeding with the analysis under consideration. For instance, flood frequency analysis is usually carried out assuming that the annual flood data constitutes a random sample. Strictly speaking one must verify that the flood data is in fact random based on statistical tests of independence, before doing the frequency analysis of the flood data. Likewise, certain decisions for filling in missing data or extending short records are based on the degree of cross-correlation (correlation in space) between two sets of hydrologic data. Thus, it is often necessary to test whether such cross-correlation between the two samples is statistically significant.

Tests of goodness of fit of known probability models to empirical frequency distributions are also presented in this lecture. They may be applicable for any probability model under consideration. In this lecture four tests of independence in time, one test of independence in space, three tests of goodness of fit and test of skewness for normality are presented.

6.2 TEST OF INDEPENDENCE IN TIME

When a sequence of observations is uncorrelated the population autocorrelation function for all lags other than zero is theoretically equal to zero. However, when sampling from an uncorrelated series, the estimated autocorrelation function r_k (correlogram) is not exactly equal to zero, but it has a sampling distribution which depends on the sample size N. This sampling distribution may be used to test the hypothesis that r_k is not significantly different from zero. If the hypothesis is accepted the series is uncorrelated, otherwise it is correlated. When the underlying variable is normal the property of independence implies that the series is uncorrelated and vice versa. Thus, the test of independence is based on the test of the correlogram r_k. In most hydrologic applications the above concept and test is used even if the underlying variable is not normal. In addition of the correlogram test, three other tests of independence are presented subsequently based on other properties of the data under consideration.

6.2.1 Anderson's Correlogram Test

It may be shown (Anderson, 1942) that when the sample size N is large the distribution of r_k is normal with mean zero and variance 1/N. Therefore, the null hypothesis $r_k = 0$, k = 1, 2,....., is tested based on the two-sided tolerance limits given by :

$$\left[\frac{-u_{1-\alpha/2}}{\sqrt{N}}, \frac{u_{1-\alpha/2}}{\sqrt{N}}\right] \tag{6.1}$$

Where, $u_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution, and N is the sample size.

Anderson (1942) also gives the expected value and variance of r_1 as :

$$E(r_1) = -1/(N-1) \tag{6.2}$$

$$\text{Var}(r_1) = (N-2)/(N-1)^2 \tag{6.3}$$

which, under the normal approximation, may be used to test the hypothesis $\rho_1 = 0$. Yevjevich (1972 b) suggests to modify eqs. 6.2 and 6.3 by substituting N by N − k + 1 so that they can be used to test the hypothesis $\rho_k = 0$.

so

$$E(r_k) = -1/(N-k) \tag{6.4}$$

$$\text{Var}(r_k) = (N-k-1)/(N-k)^2 \tag{6.5}$$

Therefore, the $\gamma = (1-\alpha)$ tolerance limits to test the hypothesis of zero autocorrelation are :

$$\frac{-1-u_{1-\alpha/2}\sqrt{N-k-1}}{N-k}, \frac{-1+u_{1-\alpha/2}\sqrt{N-k-1}}{N-k} \tag{6.6}$$

The null hypothesis $\rho_k = 0$, k = 1, . . . , M, where M is the total number of lags, should be rejected if more than α M sample correlation coefficients r_k fall outside of the tolerance limits. If the null hypothesis is accepted, the hypothesis of independence is accepted.

6.2.2 Run Test

Consider the sequence of observations y_i, i = 1, . . . , N with N = sample size and the sample mean = $\bar{y}$. A sequence of ones and zeros may be defined as follows :

$$w_i = 1 \quad \text{if } y_i > \bar{y} \tag{6.7}$$
$$w_i = 0 \quad \text{if } y_i \leqslant \bar{y}$$

for i = 1, . , , , N. For example, for N = 14 a particular one and zero sequence may be formed as 1 1 0 1 1 1 1 0 0 1 0 0 0 1. A run is defined by a consecutive series of zeros or a consecutive series of ones. In the above example, there are 3 runs of zeros and 4 runs of ones or a total of 7 runs.

The run test (Keeping, 1966) is based on the assumption that if a series is independent, the number of total runs U (runs of zeros and runs of ones) is approximately normal with mean and variance given by :

$$E(U) = \frac{2N_1 N_2}{N_1 + N_2} + 1 \tag{6.8}$$

$$\text{Var}(U) = \frac{2N_1 N_2 (2N_1 N_2 - N_1 - N_2)}{(N_1 + N_2)^2 (N_1 + N_2 - 1)} \tag{6.9}$$

Where N_1 is the number of ones in the series w_i , and N_2 is the number of zeros. The test statistic T is computed by :

$$T = \frac{U - E(U)}{(\text{Var}(U))^{1/2}} \tag{6.10}$$

Then, the hypothesis of independence is accepted at the $\gamma = (1-\alpha)$ probability level if :

$$|T| \leqslant u_{1-\alpha/2}$$

where $u_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

6.2.3 Spearman's Rank Correlation Coefficient Test

Consider a sample series y_i , $i=1, 2, \ldots, N$, where N is the sample size and let w_i be the rank of y_i when the series of observations is arranged in ascending order. The Spearman's test (Keeping, 1966) is based on the rank correlatian coefficient R between the pairs (i, w_i) for $i=1, 2, \ldots, N$. This coefficient is computed by :

$$R = 1 - \frac{6 \sum_{i=1}^{N} (1 - w_i)^2}{N (N^2 - 1)} \tag{6.11}$$

If the sample series is independent, the Spearman's rank correlation coefficient R is normally distributed, and $1 - R^2$ has a chi square distribution with $(N - 2)$ degrees of freedom. Then the ratio

$$T = \frac{R\sqrt{N-2}}{\sqrt{1-R^2}} \tag{6.12}$$

follows the student t-distribution with $N - 2$ degrees of freedom. The hypothesis of independence is accepted at the $\gamma = 1 - \alpha$ probability level if :

$$|T| \leqslant t_{1-\alpha/2,\,(N-2)}$$

where, $t_{(1-\alpha/2,\,(N-2)}$ is the $1 - \alpha/2$ quantile of the Student's t-distribution with $N - 2$ degrees of freedom. Percentile values $(t_{\alpha,\nu})$ for the t distribution with ν degrees of freedom is given in Appendix V.

6.2.4 Turning Point Test

For a given series of observations y_i, i=1, ..., N, a peak is defined as the occurrence of a value y_i such that,

$$y_{i-1} < y_i > y_{i+1}$$

and a trough by

$$y_{i-1} > y_i < y_{i+1}$$

If the sample series is independent, the total number of peaks and troughs M is approximately normally distributed with mean and variance given by (Clarke, 1973),

$$E(M) = \tfrac{2}{3}(N-2) \tag{6.13}$$

and

$$\text{Var}(M) = \frac{16N-29}{90} \tag{6.14}$$

respectively, where N is the number of observations.

The test statistic T can be computed by

$$T = \frac{M-E(M)}{(\text{Var}(M))^{1/2}} \tag{6.15}$$

Then, the hypothesis of independence is accepted at the $\gamma = 1 - \alpha$ probability level if

$$|T| \leqslant u_{1-\alpha/2}$$

where $u_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

6.3 TEST OF INDEPENDENCE IN SPACE

The test of independence in space is the test of independence between two sets of hydrologic data. Two of the typical problems in statistical hydrology are is to (i) fill in missing data and (ii) to extend short records. Most procedures for approaching these problems are usually based on the degree of (cross) dependence between two or more sets of variables. Under the normality assumption testing for independence between two variables implies tests for zero correlation between them and vice versa.

This section provides a test for independence of two data sequences.

Consider the sequences of observations $x_1 \ldots, x_N$ and $y_1 \ldots, y_N$ with N = Sample size. The correlation coefficient between the variables x and y can be computed as,

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{s_x \, s_y} \tag{6.16}$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively and s_x and s_y are the corresponding standard deviation which are computed by :

$$s_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{6.17}$$

and

$$s_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{6.18}$$

It may be shown that the statistic

$$T = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{6.19}$$

is student's t-distributed with (N—2) degrees of freedom. A test of independence can be made by comparing T and $t_{1-\alpha/2,\,(N-2)}$ which is the $1-\alpha/2$ quantile of the t-distribution with N—2 degrees of freedom. Then, the hypothesis of independence is accepted at the $\gamma = 1-\alpha$ probability level if

$$|T| \leqslant t_{1-\alpha/2,\,(N-2)}$$

6.4 TESTS OF GOODNESS OF FIT

The validity of a probability distribution function proposed to fit the empirical frequency distribution of a given sample may be tested graphically or by analytical methods. Graphical approaches are usually based on comparing visually the probability density function with the corresponding empirical density function of the sample under consideration. In other words model CDF is compared with empirical CDF. Often these CDF graphs are made on specially designed paper such that the model CDF plots as a straight line. An example of this is the Gumbel paper. If empirial CDF plots as a straiht line on the Gumbel paper it is an indication that the Gumbel distribution may be a valid model for the data at hand. Often, graphical approaches for judging how good a model is, are quite subjective. A number of analytical tests have been proposed for testing the goodness of fit of proposed models. Three of these tests are presented subsequently.

6.4.1 Chi-Square Test

The Chi-square goodness of fit test is one of the most commonly used tests for testing the goodness of fit of probability distribution functions to empirical frequency distributions.

Assume that it is desired to test the goodness of fit of a probability model, with density function $f_y(y, \theta')$ and CDF $F_y(y, \theta')$, to the empirical distribution of a sample $y_1, \ldots, y_N$, where N is the sample size, $\theta' = (\theta'_1, \ldots, \theta'_p)$ is the set of parameters estimated from the sample, and p = number of parameters. The probability space (100% probability) is divided into m

intervals (class intervals) with probabilities $p_1, \ldots, p_m$ in each class interval such that $p_1 + \ldots + p_m = 1$. If such probabilities are the same then $p_j = 1/m$ and the m cumulative probabilities are $\frac{1}{m}, \frac{2}{m}, \frac{3}{m} \ldots 1$. For the first $(m-1)$ cumulative probabilities the corresponding values of y are determined from the model. Let $y'_1, \ldots, y'_{m-1}$ be the set of y's corresponding to probabilities $\frac{1}{m}, \ldots, \frac{1}{m-1}$ which are also the upper class limits for the first m—1 class intervals.

Now let the sample $y_1, \ldots, y_N$ be arranged in increasing order of magnitude and let N_j be the number of sample values that fall in the j—th class interval for $j = 1, \ldots m$. Since the probability is p_j for the j—th class interval, the expected number of sample values that would fall in the j—th interval is equal to $p_j . N$. Then, it may be shown (Benjamin and Cornell, 1970) that the (test) statistic

$$C = \sum_{j=1}^{m} \frac{(N_j - NP_j)^2}{NP_j} \tag{6.20}$$

is approximately Chi-square distributed with m-1-p degrees of freedom. Since $p_j = 1/m$, Eq. 6.20 becomes

$$C = \frac{m}{N} \sum_{j=1}^{m} N_j^2 - N \tag{6.21}$$

The number of classes are selected in such a way that theoretical frequency of each class is not less than 5. The number of classes should not be less than 6 and more than 20 though these rules don't have theoretical basis. The length of class intervals should be selected in such a way that the main characteristic features of the observed distribution are emphasized and chance variations are obscured.

Eq. 6.20 compares the number of sample values in each interval with the number to be expected for the given sample size. So, small values of C would indicate a good fit while large values of C would indicate the contrary.

Eq. 6.20 or 6.21 may be used to test the hypothesis of good fit of a given model to the empirical frequency distribution of a sample by comparing the computed test statistic C with the tabulated Chi-square statistic $x^2_{1-\alpha, (m-1-p)}$ in which α is the significance level and (m-1-p) is the number of degrees of freedom. The, the hypothesis of good fit is accepted at the $\gamma = 1-\alpha$ probability level if

$$C \leqslant x^2_{1-\alpha\ (m-1-p)}$$

The critical values x^2 for different probability levels and degrees of freedom are given in Appendix VI.

6.4.2 Kolmogorov-Smirnov Test

This is a distribution free test widely used in statistical hydrology. It is based on the maximum difference between the cumulative empirical distribution F_e (y) and the cumulative probablity distribution being fitted F_y (y ; θ'). Consider the Statistic

$$D = \max_{i=1}^{N} (F_e(y) - F_y ; (y_i, \theta')) \tag{6.22}$$

where F_e (y_i) and F_y (y_i θ') represent the empirical and model cumulative distribution, respectively, corresponding to the observation value y_i, i=1,....., N which has been arranged in increasing (or decreasing) order of magnitude, N = Sample size and θ' is the parameter set of the model estimated from the given sample. In the Kolmogorov–Smirnov test, the empirical CDF F_e (y_i) is based on the plotting position i/N (Benjamin and Cornell, 1970) although in practice the plotting positions i/(N+1) is often used (Yevjevich, 1972).

The goodness of fit test of the selected probability model to the empirical distribution is accepted if :

$$D \leqslant d_\alpha (N)$$

where d_α (N) is the Kolmogorov-Smirnov statistic corresponding to the sample size N and confidence level $\gamma = 1 - \alpha$. The statistic d_α (N) is listed in Table 6.1.

TABLE—6.1

Kolmogorov-Smirnov Test Statistic d_α (N)

Sample Size	Significance Level			
N	0.20	0.10	0.05	0.01
5	0.45	0.51	0.56	0.67
10	0.32	0.37	0.41	0.49
15	0.27	0.30	0.34	0.40
20	0.23	0.26	0.29	0.35
25	0.21	0.24	0.27	0.32
30	0.19	0.22	0.24	0.29
35	0.18	0.20	0.23	0.27
40	0.17	0.19	0.21	0.25
45	0.16	0.18	0.20	0.24
50	0.15	0.17	0.19	0.23
Large N	$1.07/\sqrt{N}$	$1.22/\sqrt{N}$	$1.36/\sqrt{N}$	$1.63/\sqrt{N}$

6.4.3 General Commenfs about Chi-square and K.S. Test :

Many hydrologists discourage the use of the chi-square and Kolmogorov-Smirnov tests when testing hydrologic frequency distributions. The reason for this is the importance of the tails of hydrologic frequency distribution and the insensitivity of these tests in the tails of the distributions. The sensitivity of the chi-square test can be improved in the tails of the distribution if classes are not combined to get an expected frequency of 3 to 5. The disadvantage of this is that a single observation in a class with a low expectation can result in a chi-square value in excess of the critical value. This single observation can lead to rejecting the hypothesis.

Neither the chi-square test nor the Kolmogorov-Smirnov test is very powerful in the sense that the probability of accepting the hypothesis when it is infact false is very high when these tests are used.

6.4.4 D-Index method

In order to compare the relative fit of different distributions to hydrological data, USWRC (United States Water Resources Council) has suggested the following procedure :

The probability of exceedance of observation is estimated by Weibull plotting position formula.

$$P(X \geqslant x) = m / (N+1) \qquad (6.23)$$

where, P is the probability of exceedance

m is the rank of the flood values arranged in the descending order of magnitude, and

N is the number of observations.

Flood peaks are estimated for a specified series of recurrence intervals viz., 2, 5, 10, 15, 20 and 30 years. For each recurrence interval, the historical value is obtained by interpolation in terms of recurrence interval between the two floods of record of adjacent recurrence intervals. The discharge corrresponding to the same recurrence interval is also calculated on the basis of fitted distribution. The D index for comparison purposes of the fit of different distributions is given as

$$\text{D-Index} = \frac{1}{\bar{x}} - \sum_{i=1}^{6} \text{ABS}\,(X_i, \text{observed} - X_i, \text{computed}) \qquad (6.24)$$

in which, $\bar{x}$ is the mean of the observed series.

Often instead of using flood values corresponding to recurrence intervals of 2, 5, 10, 15, 20 and 30 years, highest six observations/flood values are used, as the aim is to see the fit of the distribution in the upper tail region. From the studies carried out in NIH, it has been found appropriate to use largest six observations and the corresponding values based on the fitted distribution in calculating the D-Index.

The distribution which gives minimum D index is considered as the best fit distribution.

6.5 SKEWNESS TEST OF NORMALITY

The test is based on the fact that normally distributed variables have zero skewness. If the sample comes from a normal distribution, the coefficient of skewness c_s is approximately normally distributed with mean zero and variance equal to 6/N (Snedecor and cochran, 1967). Therefore, the corresponding tolerence limits may be determined by :

$$(-u_{1-\alpha/2}\sqrt{\frac{6}{N}}\ ,\ u_{1-\alpha/2}\sqrt{\frac{6}{N}}) \qquad (6.25)$$

where $u_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. If, for a given sample, the coefficient of skewness C_s fall within the computed tolerance limits, it is assumed that the skewness is not significant, and hypothesis of normality is accepted.

Actually the assumption of normality of statistic C_s is sufficiently accurate for sample sizes greater than 150. For smaller sample sizes more accurate results may be obtained by using tabulated test statistics given in table 6.2. In this case the hypothesis of normality of a given sample is accepted at the $\gamma = 1-\alpha$ confidence level if

$$C_s \leqslant g_\alpha\ (N)$$

Where, g_α (N) is the referred table statistic, a function of the significance level α and the sample size N. Values of this statistic are shown in Table 6.2 for $\alpha = 0.05$ and 0.01 and for various values of N,

TABLE—6.2

Skewness-Coefficient Test Statistic g_α (N)

Sample size	Significance level		Sample size	Significance level		Sample size	Significance level	
N	0.05	0.01	N	0.05	0.01	N	0.05	0.01
25	0.711	1.061	70	0.459	0.673	200	0.280	0.403
30	0.662	0.986	80	0.432	0.631	250	0.251	0.360
35	0.621	0.923	90	0.409	0.596	300	0.230	0.329
40	0.587	0.870	100	0.389	0.567	350	0.213	0.305
45	0.558	0.825	125	0.350	0.508	400	0.200	0.285
50	0.534	0'787	150	0.321	0.464	450	0.188	0.269
55	0.492	0.723	175	0.298	0.430	500	0.179	0.255

References :

1. Anderson, R.L. (1942), 'Distribution of the Serial Correlation Coefficient', Ann. Math. Stati. Vol. 13, No. 1.

2. Benjamin, J.R. and Cornell, C.A. (1970), 'Probability, Statistics and Decision for Civil Engineers', McGraw Hill, New York.

3. Clarke, R.T. (1973), 'Mathematical Models in Hydrology', Irrigation and Drainage Paper 19, FAO, Rome.

4. Haan, C.T. (1977), 'Statistical Methods in Hydrology', Iowa State University Press, Ames, Iowa.

5. Keeping, E.S. (1966), 'Distribution Free Methods in Statistics', in Proceedings of Hydrology Symposium No. 5, McGill University, Canada.

6. Salas, J.D. (1983), 'Statistics in Water Resources Engineering', Lecture Notes, C.S.U., Fort Collins, U.S.A.

7. Snedecor, G.W. and Cochran, W.G. (1967), 'Statistical Methods', The Iowa State University Press, Ames, Iowa.

8. Walsh, E. (1976), 'Handbook of Nonparametric Statistics', Vol. 1, Van Nostrand, New York.

9. Yevjevich, V. (1972 a), 'Probability and Statistics in Hydrology', Water Resources Publications, Fort Collins, Colorado.

10. Yevjevich, V. (1972 b), 'Stochastic Processes in Hydrology', Water Resources Publication, Fort Collins, Colorado.