LECTURE - 2

STATISTICAL PARAMETERS AND DATA REQUIREMENT OF FREQUENCY ANALYSIS

OBJECTIVES

The objectives of this lecture are :

(i) to define basic terminologies used in frequency analysis

(i i) to discuss the procedures for computing the main sample statistics of a given data set.

(iii) to explain basic assumptions of flood frequency analysis and

(iv) to explain data requirements of frequency analysis.

2.1 DEFINITIONS

a. Peak annual discharge

The peak annual discharge is defined as the largest instantaneous volumetric rate of discharge during a year.

b. Annual flood series

The annual flood series is the sequence of flood formed by abstracting the peak annual discharges for each year of record. Thus for N years of data the annual flood series will consist of N values of discharges.

(c) Design flood

Design flood has been defined as:

(i) The maximum flood that any structure can safely pass,
(i i) The flood adopted to control the design of a structure.

(d) Recurrence interval

The term recurrence interval (also called the return period) is the time on average that elapses between two events that equal or exceed a particular level e.g. T year flood will be equalled or exceeded on an average once every T years.

(e) Population data :

Population data encompass all possible values an event can take.

(f) Sample data

Sample data are available data from the observation of an event.

(g) Random events

Events whose occurrence is not influenced by the occurrence of the same event earlier.

(h) Probability density function

Probability density function (P.D.F.) is the probability of occurrence of an event.

(i) Cumulative density function

Cumulative density function (C.D.F.) is the probability of occurrence of all the events that are equal to or less than an event.

(j) Partial duration series

The partial duration series consists of all recorded peak discharge floods above a selected base value regardless of the number of such floods occurring each year. The various discharges must be independent and this can be achieved by ensuring that consecutive flood peaks are separated by a recession of a suitable length of time.

(k) Probability paper

A probability paper is a special graph paper on which the ordinate usually represents the magnitude of the flood and the abscissa represents the probability P, or the return period T. The ordinate and abscissa scales are so designed that the distribution plots more nearly a straight line permitting better definition of the upper and lower parts of the frequency curve. The probability paper is used to linearize the distribution so that data to be fitted appear close to the straight line. The extreme value and the log normal probability papers are used for linearization of the extreme value and log normal distribution.

(l) Plotting position

Determining the probability to assign a data point is commonly referred to as determining its plotting position.

2.2 SAMPLE STATISTICS

In any analysis of statistical data in general and of hydrologic data in particular, certain calculations are usually made in order to determine some of the basic properties inherent in the data. For instance, the sample mean and variance are two statistics defining the most important characteristics of a given set of statistical data, In general sample statistics provide the basic information about the variability of a given data set, The most useful sample statistics measure the following characteristics.

(i) the central tendency or value around which all other values are clustered.

(ii) the asymetry or skewness of the frequency distribution and

(iii) the flatness of the frequency distribution.

These statistical properties are determined by sample statistics as described below :

2.2.1 Mean

The sample mean measure the central tendency of a given data set. If $X_1, X_2, X_3 \ldots\ldots, X_n$ represent a sequence of observations, the mean of this sequence is determined by

$$\overline{X} = 1/N \sum_{i=1}^{N} X_i \tag{2.1}$$

Here $\overline{X}$ represent the sample mean, Popplation mean is generally represented by

2.2.2 Standard Deviation, Variance and Coefficient of Variation

The standard deviation measures the dispersion of sample values around the mean. The unbiassed estimate of population standard deviation(S) from the sample is given by :

$$S = [1/(N-1) \sum_{i=1}^{N} (X_i - \overline{X})^2]^{1/2} \tag{2.2}$$

Variance is the square of standard deviation. Generally is used for population standard deviation.

The coefficient of variation C_v is a dimensionless dispersion parameter and is equal to the retio of the standard deviation and the mean

$$C_v = S/\overline{X} \tag{2.3}$$

This coefficient is extensively used in hydrology particularly as a regionalisation parameter :

2.2.3 Skewness Coefficient :

The skewness coefficient or Coeffcient of skewness measures the asymetry of the frequency distribntion of the data. An unbiassed estimate of the coefficient is given by :

$$C_s = \frac{N \sum_{i=1}^{N} (X_i - X)^3}{(N-1)\ (N-2)\ S^3} \tag{2.4}$$

The skewness coefficient has an important meaning since it gives indication of the symmetry of the distribution of the data. Symmetrical frequency distributions have very small or negligible sample skewness coefficient C_s while asymmetrical frequency distributions have either positive or negative coefficients. Often a small value of C_s indicate that the frequency distribution of the sample may be approximated by the normal distribution function since $C_s = 0$ for this function,

2.2.4 Kurtosis Coefficient :

The Kurtosis coefficient measures the peakedness or the flatness of the frequency distribution near its centre. An unbiassed estimate of this coefficient is given by :

$$C_K = \frac{N^2 \sum_{i=1}^{N} (X_i - \overline{X})^4}{(N-1)\ (N-2)\ (N-3)\ S^4} \tag{2.5}$$

A related coefficient called the excess coefficient denoted by E is defined by

$$E = C_K - 3 \qquad (2,6)$$

Positive values of E indicate that a frequency distribution is more peaked around its centre than the normal distribution. Frequeocy distribution is known as LEPTOKURTIC. The negative valuas of E indicate that a given frequency distribution is more flat around its centre rhan the normal. Frequency distribution is known as PLATYKURTIC.

Normal distribution is said to be MESOKURTIC. Both kurtosis and excess coefficient are seldom used in statistical hydrology.

2.2.5 Standard Errors of Sample Statistics :

Because of the short period of record the statistics calculated from the sample are only estimates of the true or population values which would be calculated if an infinitely large samples were available. The reliability of the statistics calculated from the sample can be judged from the standard errors of the estimate (SEE). Statistical Theory states there is about 68% probability that the true of population value of each statistic is within one standard error of estimate of the value calculated from the available data.

The standard errorsed of mean, standard deviation and coefficient of skewness are given below :

$$SEE\ (\overline{X}) = S/\sqrt{N} \qquad (2.7)$$

It describes the variability that can be expected in the means of samples by repeated random collection from the same population.

$$SEE\ (S) = S/\sqrt{2N} \qquad (2.8)$$

$$SEE\ (C_s) = \sqrt{6N\ (N-1)\ /\ (N-2)\ (N+1)\ (N+3)} \qquad (2.9)$$

The standard error of estimate for each moment becomes smaller as a longer length of record becomes available for use in the analysis.

2.3 BASIC ASSUMPTIONS OF FREQUENCY ANALYSIS

If the frequency analysis is to provide useful answers, it must start with a data that is relevant, adequate and accurate. The terms, relevance, adequacy and accuracy are explained subsequently.

RELEVANCE implies that the data must deal with the problem. Most flood studies are concerned with peak flows and the data series will consist of selected observed peaks. If the problem is of duration of flooding e.g. for what period of time a highway adjacent to a stream is likely to be flooded, the data series should repreeent the duration of flows in excess of some crrtical value. If the problem is of interior drainage of a leveled area, the data required must consist of those flood volumes occurring when the main river is too high to permit gravity drainage.

ADEQUACY refers primarily to length of record but scarsity of data collecting stations is often a problem. The observed record is merely a sample of the total population of floods that have occurred and may be expected to occur again. If the sample is too small the probabilities

derived cannot be expected to be reliable. Available streamflow rscords are too small to answer tuis question. Table 2.1 gives some estimates derived from synthetic data (Linsley et. al., 1975).

TABLE 2.1

LENGTH OF RECORD IN YEARS REQUIRED TO ESTIMATE FLOODS OF VARIOUS PROBABILITIES WITH 95% CONFIDENCE

Design Probability	Return Period	Acceptable error 10%	25%
0.1	10	90	18
0.02	50	110	39
0.01	100	115	40

Table 2.1 suggests that extrapolation of frequency estimates beyond probability of 0.01 is extremely risky with data series generally available (30–40) years).

Ott (1971) show that with 20 years of record the probability is 80 percent that the design flow will be over estimated and that 45 percent of over estimates will exceed 30 percent so it thus appears that records shorter than 20 years should not be used for frequency analysis, 40–50 years data series have been found to be define event magnitude upto 50 years well.

Accuracy of data series refers primarily to the problem of homogenity. Most of the flow records are not satisfactory in term of their intrinsic accuracy. Obviously if a station is so poor that the reported flows are unreliable, it cannot provide satisfactory data for probability analysis. Even though reported flows are accurate, they may be unsuited for probability analysis if changes in the hydrologic characteristics, i.e. the record is not internaily homogeneous. Dams, levees, diversions, urbanizations and other land use changes may introduce in consistencies, such record should not be used without adjusting to a common set of watershed conditions, usually either natural conditions or current conditioes.

2.4 DATA REQUIREMENT

All frequency techniques are totally data dependent. An assumption must be made of a theoretical frequency distribution suitable for the population events and the statistical parameters of the distribution must be computed from the sample data. Two types of sample date, namely (i) annual peak flood series and (ii) partial duration series may be used for flood frequency analysis Annual peak flood series is arrived at from the recorded flood peaks by picking up only one event from each year of the record. Annual peak flood series ensures complete randomness of the data and thus assumption of randomness is satisfied. But a disadvantage of using this series for analysis is that the second or third highest events in a particular year may be higher than some of the year's annual peak floods and still they are totally disregarded in the analysis. Such a disadvantage is remedied by using the partial duration series in which all the events above a certain threshold are included in the analysis. However care should be taken not to include those peaks which are dependent as the assumption of randomness would be violated. This can be achieved by ensuring that consecutive flood peaks are separated by a recession of a suitable length of time. The procedures for deeling with dependent data are still in research stage.

Criterial for the minimum interval between events and for establishing the base discharge must be decided for individual cases. As a guide, the minimum time separating two consecutive peak discharges may be set two to three times the estimated time of concentration for the catchment and select the base dissharge so that the average number of exceedences is between three and four each year. Some times this depends upon the application also to which the statistical analysis is to be applied. For example, if the job involves the diversion of flood flows for a dam construction project in woich flooding would cause a job delay of one week then two floods occurring one week apart should be regarded as independent events regardless of any hydrologic effects that may persist longer than tnis. Also, if the flooding problem is connected with an urban area which would take one month to return to normal after a flood, then two floods occuring one month apart should be regarded as independent events,

As a preliminary step thc basic data should be screened and adjusted to remove, as far as possible, any non-conformities that may exist. The following are the more important considerations (CWC, 1969).

(i) Effect of man made changes in the regime of flow should be investigated and adjusiment be made as required.

(ii) For small catchment areas a distinction should be made between daily maxima and instantaneous or momentary, flood peaks.

(iii) Changes in the stage discharge relation render stage records non homogeneous and unsuitable for frequency analysis studies. It is therefore preferable to work with discharges and if stage frequencies are required, reter the results to the most recent rating.

(iv) Any useful information contained in data publications and manuscripts should be made use of after proper scrutiny.

References :

1. Central Water and Power Cammission (1969), 'Estimation of Design flood-Recommended Procedures'.

2. Linseley, R.K., Kohler, H.A. and Paulhus, J.L.H. (1975), 'Hydrology for Engineers', McGraw Hill, International Book Company.

3. Ott, R.F. (1971), 'Stream flow frequency using stochastically Generated Hourly Rainfall', Stenford University, Deptt. of Civil Engineering, Technical Report No. 151.

4. Salas, J.D. (1983), 'Statistics in Water Resources Engineering', Lecture notes, CSU, Fort Collins, Co.