Chapter 4

Statistical Techniques for Data Analysis

The objectives of this chapter are:

- to explain the basic concepts of statistical analysis of data,
- to describe the frequency distributions that are commonly used in water resources, and describe methods of estimation of their parameters,
- to discuss regression and correlation analysis, and
- to briefly discuss time-series analysis.

The observed behavior of many water resources variables is chance dependent and cannot be adequately explained in terms of known physical laws. This could be because: a) a poor or incorrect understanding of the underlying complex processes, b) non-availability of sufficient data, and c) inherent randomness of the variable. In such cases, techniques for statistical analysis are summoned to make inferences about the behavior of the variables.

Statistics deals with methods to draw inferences about the properties of a *population* based on sample data from that population. Population refers to a collection of objects. It can be finite or infinite, for example, the collection of all flow data of a river at a given site. Often, the measurements of the entire population are not available and what is generally available is a limited number of observations or a finite *sample*. Based on this sample, properties of the population are determined assuming these to be unbiased estimates of the properties of the population.

A variable whose value at any time is not influenced by the value at earlier time(s) is known as a *random variable*. Such a variable can be discrete which can take on only a finite set of values, such as number of rainy days in a year at a place. It can also be continuous and can take on any value, for example, the water level of river at a gauging site or the magnitude of rainfall at a place.

In many problems, the sample data consist of measurements on a single random variable; the techniques of analysis are called univariate analysis and estimation. Univariate analysis is carried out by using the measurements of the random variable, which is called sample information, to identify the statistical properties of the population from which the sample measurements are likely to have come. After the underlying population has been identified, one can make probabilistic statements about the future occurrences of the random variable, this represents univariate estimation. It is important to remember that univariate estimation is based on the assumed population and not the sample, the sample is used only to characterize the population.

The following are the main steps of statistical analysis of data:

i) It is always useful to first plot the sample data.
ii) Select a set of probability distribution functions.
iii) Fit the selected distributions with the sample data. Common methods of parameter estimation are least squares method; methods of moments, linear moments and probability weighted moments; entropy-based methods; and method of maximum likelihood.
iv) Select the best fit distribution using the goodness-of-fit tests.
v) Use the best-fit probability distribution to make inferences about the likelihood of occurrence of the magnitudes of the random variable.

If all the values of a random variable and the corresponding probabilities are known or found, the relation between these values and probabilities is described by a probability distribution. Knowing this distribution, the probability of any value of the random variable can be determined.

In statistical analysis of multivariable data, the functional forms of the relationships are studied. Linear regression analysis is one of the ways to develop a suitable form of the multiple-variable models wherein a dependent variable takes on values caused by variations in one or more independent or predictor variables.

4.1 BASIC CONCEPTS

Let X denote a random variable and x be a possible value of X. The cumulative distribution function (CDF), $F_X(x)$ is the probability that the random variable X is less than or equal to x:

$$F_X(x) = P(X \leq x) \tag{4.1}$$

The probability distribution function (PDF) describes the relative likelihood that a continuous random variable X takes on different values, and is the derivative of the CDF:

$$f_X(x) = d\{F_X(x)\}/dx \tag{4.2}$$

The PDF and CDF of a random variable are shown in Fig. 4.1.

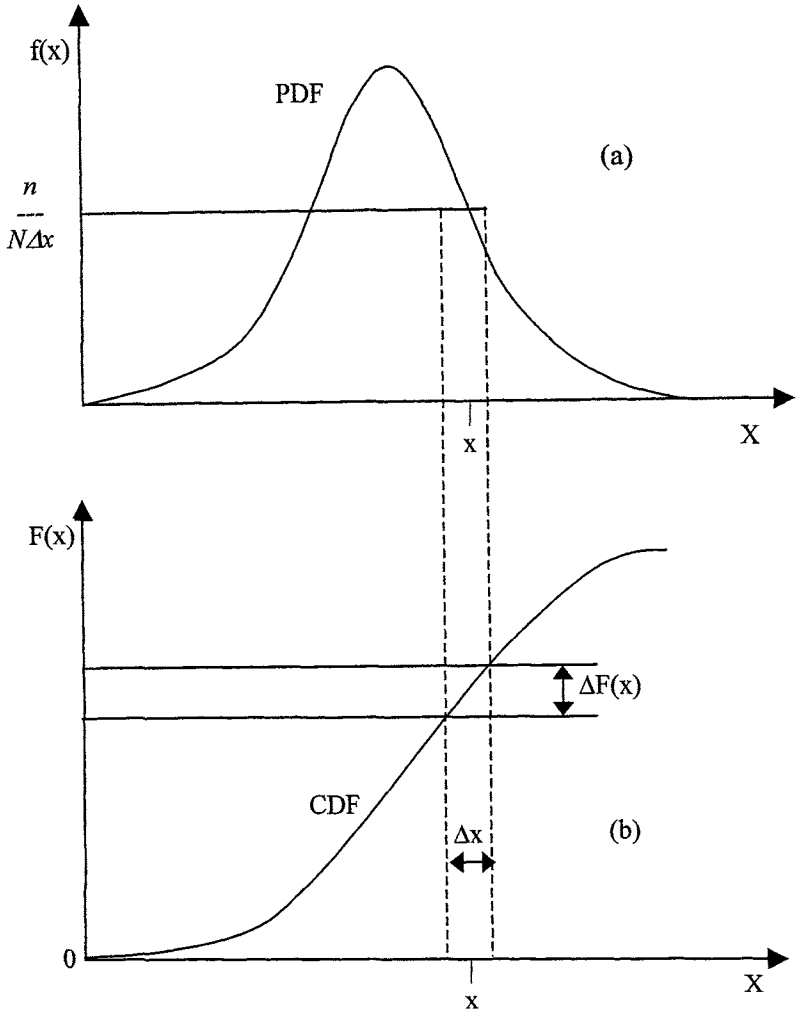

Fig. 4.1 PDF and CDF of a random variable.

At this stage, it is useful to state some of the properties of $f(x)$ and $F(x)$ for continuous random variables.

1. The probability of a random variable cannot be negative

$$f(x) \geq 0, \quad -\infty < x < \infty \tag{4.3}$$

2. The sum of probabilities of all possible outcomes is equal to 1, i.e., the area under the PDF is unity.

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{4.4}$$

3. $$P(X \le x) = F(X \le x) = F(x) = \int_{-\infty}^{x} f(x)dx \tag{4.5}$$

If a and b are any real numbers such that $a < b$, the events $X \le a$ and $a < X \le b$ will be mutually exclusive. Then $P(X \le b) \ge P(X \le a)$ or $F(X \le b) \ge F(X \le a)$, and

$$P(X \le b) = P(X \le a) + P(a < X \le b)$$

$$= \int_{-\infty}^{a} f(x)dx + \int_{-a}^{b} f(x)dx = \int_{-\infty}^{b} f(x)dx = F(X \le b)$$

This yields

$$\begin{aligned} P(a < X \le b) &= P(X \le b) - P(X \le a) \\ &= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx \\ &= \int_{a}^{b} f(x)dx, \quad for\ a < b \end{aligned} \tag{4.6}$$

4. The probability that X (continuous variable) assumes a particular value is zero, that is, $P(X = a) = F(x = a) = 0$,

$$\int_{a}^{a} f(x)dx = F(a) - F(a) = 0 \tag{4.7}$$

5. $$F(+\infty) = \lim_{x \to \infty} F(x) = 1 \tag{4.8}$$

Also $$F(-\infty) = \lim_{x \to -\infty} F(x) = 0 \tag{4.9}$$

This can be verified from the area under the PDF.

For discrete random variables, analogous statements can be made:

1. $$\sum_{i} f(x_i) = 1 \tag{4.10}$$

where $f(x_i)$ represents the probability of $X = x_i$ in the sample space if the observations are finite in the sample. Thus, this can be replaced by $p(x_i)$.

2. $$P(a \le x \le b) = \sum_{x_i \ge a}^{x_i \le b} p(x_i) \tag{4.11}$$

3. $$P(X \le x_k) = \sum_{i=1}^{k} p(x_i) \tag{4.12}$$

4.1.1 Distribution Characteristics

There are four principal moments for characterizing probability distributions:

(i) the central tendency or the value around which all other values are clustered,
(ii) the spread of the sample values around mean,
(iii) the asymmetry or skewness of the frequency distribution, and
(iv) the flatness of the frequency distribution.

These characteristics are expressed in terms of the parameters of distributions, the parameters can themselves be expressed in terms of moments. These parameters are estimated from the observed sample data, and are then used as estimates of the parameters of the population distribution.

Measures of Central Tendency

In statistics various measures of location are described. The important measures are the following.

(i) Arithmetic Mean: If $x_1, x_2 \ldots x_n$ represent a sequence of observations, the mean of this sequence is the ratio the of sum of values and the number of values:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.13}$$

where $\bar{x}$ represents the sample mean; population mean is generally represented by μ.

(ii) Mode: It is the value in the sample (or population) which occurs most frequently. It is the peak value of the PDF. Note that a sample or population may have more than one peak.

(iii) Median: It is the middle value of the ranked values for a sample (or population). The median divides the distribution in two equal parts.

Measure of Dispersion or Variation

Some of the important measures of dispersion or variation include:

(i) Variance: It represents the dispersion of data about the mean and is expressed as:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4.14}$$

(ii) Standard deviation: The unbiased estimate of population standard deviation (s) from the sample is given as the square root of the variance, i.e.,

$$s = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2} \tag{4.15}$$

For $n < 30$, the unbiased estimate of s is found by replacing n by $n-1$ in this equation.

(iii) The coefficient of variation C_V is a dimensionless dispersion parameter and is equal to the ratio of the standard deviation and the mean:

$$C_V = s/\bar{x} \tag{4.16}$$

The variance has the square of the units of the original data. The standard deviation has the same dimensions as of the data. The coefficient of variation is a dimensionless quantity.

Measures of Symmetry

If the data are exactly symmetrically displaced about the mean then the measure of symmetry should be zero. If the data to the right of the mean (larger) are more spread out from the mean than those on the left then, by convention, the asymmetry is positive and vice versa for negative asymmetry.

The third moment of the sample data about the mean is given by:

$$M_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \tag{4.17}$$

This moment is zero if the data are symmetrical. Otherwise it is positive or negative.

Coefficient of Skewness: It is a non-dimensional measure of the asymmetry of the distribution of the data. An unbiased estimate of the coefficient is given by:

$$C_s = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3} \tag{4.18}$$

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 indicates that the frequency distribution of the sample may be approximated by the normal distribution since $C_s = 0$ for this function. The symmetrical and skewed distributions are shown in Fig. 4.2.

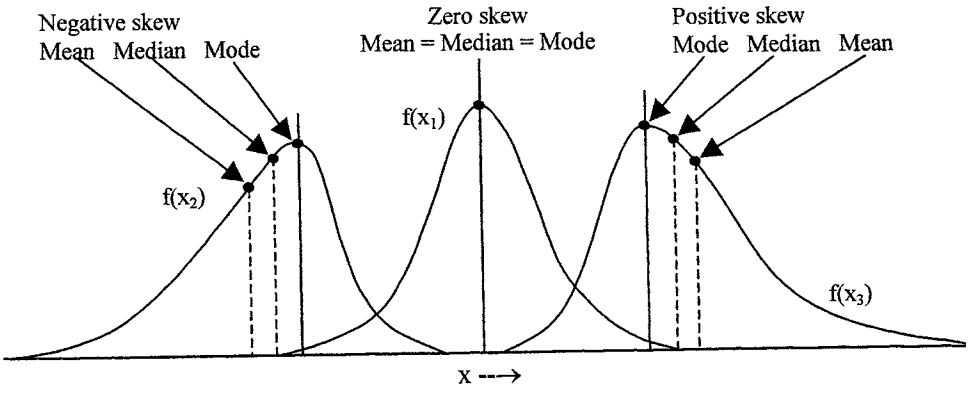

Fig. 4.2 Symmetrical and skewed distributions.

Note that because the third central moment has dimensions equal to the cube of the data, it is not of direct use while comparing different data sets. The coefficient of skewness does not have this disadvantage and is, therefore, preferred.

Measures of Peakedness or Flatness

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

$$C_k = \frac{n^2 \sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} \tag{4.19}$$

The kurtosis for a normal distribution is 3.

Standard Errors of Sample Statistics

Because of the short length of most records, the statistics calculated from the sample are only estimates of the true or population values which would be available if very large samples were available. The reliability of the statistics calculated from the sample can be judged from the standard error of the estimate (SEE). According to the statistical theory, the probability that the true or population value of each statistic is within one standard error of estimate of the value calculated from the available data is about 68%.

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

$$S_e(\bar{x}) = s/\sqrt{n} \tag{4.20}$$

$$S_e(S) = s/\sqrt{2n} \tag{4.21}$$

$$S_e(C_s) = \sqrt{6n(n-1)/[(n+1)(n+2)(n+3)]} \tag{4.22}$$

Clearly, the standard error of estimate for each parameter becomes smaller as the length of record used in the analysis becomes even longer.

Graphical Presentation of Data

For graphical presentation of data in the form of histograms and cumulative histograms of frequency (or relative frequency or probability), a frequency table is prepared. The range of the data is divided into a number of intervals of convenient size and the number of frequencies of values occurring in each interval is entered alongside. This table provides a valuable summary. The selection of class interval can affect the appearance of a frequency histogram. If the class intervals are very large, the table is compact but loses detail. If the intervals are too small, the table may be too bulky and not succinct enough. For the choice of class interval, the following guidelines may be considered:

(a) Brooks and Carruthers' rough guide:

$$\text{No of classes} \leq 5 \log (\text{sample size}) \tag{4.23}$$

(b) Charlier's rule of thumb:

$$w = (\text{maximum value} - \text{minimum value})/20 \tag{4.24}$$

where w is the size of class interval. Number of classes is generally 15 to 25.

(c) According to Sturges (1926), the class interval can be estimated by:

$$m = 1 + 3.3 \text{ Log } (n)$$

where m is the number of classes, and n is the number of observations.

The frequency table can be prepared using the following steps:

(i) Arrange the variable (X_i) in increasing or decreasing order of magnitude.

(ii) Decide the number of class intervals (NC) and thereby the size of the class interval ΔX following the above guidelines.

(iii) Divide the ordered observations X_i into NC intervals (or groups).

(iv) Determine the absolute frequency n_j by counting the observations that fall within the j^{th} class interval for j=1, ... NC.

(v) Determine the corresponding relative frequencies as n_j/n, j=1,... NC.

(vi) Compute the cumulative relative frequencies F_j, j = 1, ... NC. These cumulative frequencies approximate the probabilities as:

$F_j = F(X \leq x)$ if order is increasing, or
$F_j = F(X > x)$ if order is decreasing.

(vii) Prepare the plots for the relative frequencies as well as cumulative relative frequencies on simple graph papers taking the group interval as abscissa and the relative frequencies or cumulative relative frequencies as ordinate.

Example 4.1: The annual flow of Sabarmati River at Dharoi is plotted in Fig. 4.3 for the period 1868-1965. Find the statistical parameters of this data and plot the histogram.

Solution: The histogram of the annual flow of Sabarmati River at Dharoi for the period 1868-1965 is plotted in Fig. 4.4.

Mean of the data $\bar{x}$ = 65206/98 = 665.37 million cubic m.
Variance σ^2 = 11841713/98 = 120833.8 (million cubic m)^2.
Standard deviation $\sigma = (120833.8)^{0.5}$ = 346.9 million cubic m.
Coefficient of variation C_V = 346.9/665.37 = 0.521.
Coefficient of skewness C_s = 0.76 (positively skewed).
Kurtosis C_k = 3.65.

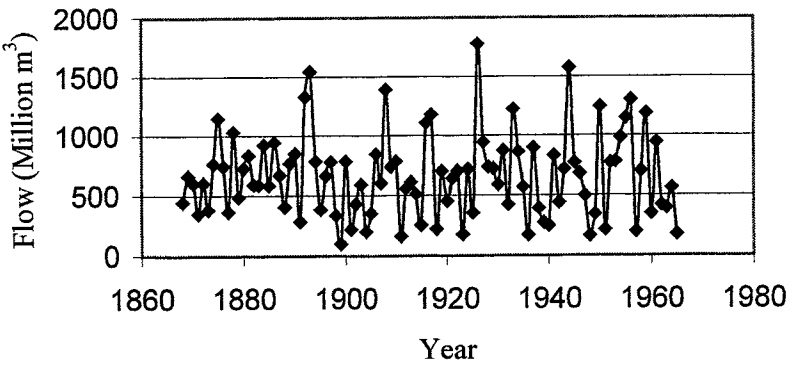

Fig. 4.3 Annual flow of Sabarmati river at Dharoi.

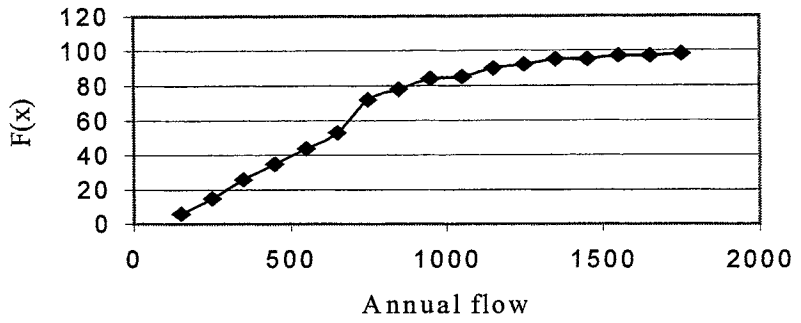

Fig. 4.4 Cumulative Histogram of Annual flows of River Sabarmati.

Standard error of mean = $346.9/(98)^{0.5} = 35.042$.
Standard error of SD = $346.9/(2*98)^{0.5} = 24.85$
Standard error of C_s = $[6*98*97/(99*100*101)]^{0.5} = 0.239$.

4.2 PROBABILITY DISTRIBUTIONS

A distribution is an attribute of a statistical population. It describes the relation between the random variable and the probabilities. A distribution gives important information about the data, whether they are bunched together or spread out, and whether they are symmetrically disposed on the X-axis or not. Distribution also tells the relative frequency or proportion of various X values in the population in the same way that a histogram gives that information about a sample.

The distributions that are commonly used in water resources problems are described in the following. A summary of the distributions is provided in Table 4.1.

Table 4.1 Summary of Distributions Commonly Used in Hydrology

Distribution	Probability density function	Range	Mean	Variance
Binomial	$P(x)=\binom{n}{x}p^x(1-p)^{n-x}$	0≤ x≤ n	np	np(1-p)
Geometric	$P(x)=pq^{x-1}$, q = 1-p	1 ≤ x ≤	1/p	q/p^2
Poisson	$P(x)=\frac{\lambda^x \exp(-\lambda)}{x!}$	0 ≤ x ..	λ	λ
Exponential	$f(x)=\lambda\exp(-\lambda x)$	0≤ x ≤ ∞	1/λ	$1/\lambda^2$
Gamma	$f(x)=\frac{\lambda^n x^{n-1}}{(n-1)!}\exp(-\lambda x)$	0≤ x ≤ ∞	n/λ	n/λ^2
Normal	$f(x)=\frac{1}{\sigma\sqrt{2n}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	-∞< x < ∞	μ	σ^2
Log-Normal (y = ln x)	$f(x)=\frac{1}{\sigma_x\sqrt{2n}}\exp\left[-\frac{(\ln x-\mu_y)^2}{2\sigma_y^2}\right]$	0 < x < ∞	μ_y or $\exp(\mu_y+\sigma_y^2/2)$	σ_y^2 or $\mu_x^2[\exp(\sigma_y^2)-1]$
Gumbel	$f(x)=\alpha\exp\{-\alpha(x-\beta)-\exp[-\alpha(x-\beta)]\}$	-∞< x < ∞	μ+ 0.5772/α	$\pi^2/6\alpha^2$
Pearson Type III	$f(x)=\frac{1}{a\Gamma(b)}\left(\frac{x-c}{a}\right)^{b-1}\exp\left(-\frac{x-c}{a}\right)$	-∞< x < ∞	ab + c	a^2b
Log Pearson Type III (y = ln x)	$f(x)=\frac{1}{ax\Gamma(b)}\left(\frac{\ln x-c}{a}\right)^{b-1}\exp\left(-\frac{\ln x-c}{a}\right)$	0< x < ∞	$\mu_y = c + ab$	$\sigma_y^2=a^2b$

4.2.1 Continuous Probability Distributions

The commonly used continuous distributions are the Normal, Log Normal, Extreme Value type-1 (Gumbel or EV1), Gamma, Pearson type-III, and Log Pearson type-III distributions. The probability density functions (PDF), cumulative density functions (CDF) and other properties of these distributions are given below.

Normal Distribution

Also known as Gaussian distribution, the normal distribution is a symmetrical bell-shaped probability density function. When a hydrologic variable, integrated over a large time period, is used in analysis, the variable is expected to follow a normal distribution. The normal distribution has two parameters, mean μ and standard deviation σ, and its PDF can be expressed as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad -\infty < x < \infty \tag{4.25}$$

Integrating eq. (4.25), the CDF of the normal distribution is:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{(u-\mu)^2}{2\sigma^2}\right] du \tag{4.26}$$

For the normal distribution, the reduced variate is $Z = (x - \mu)/\sigma$. The mean of the reduced variate is 0, standard deviation $\sigma_z = 1$, and its coefficient of skewness is 0. Fig. 4.5 shows the normal distribution and the area for three values of the standard variate.

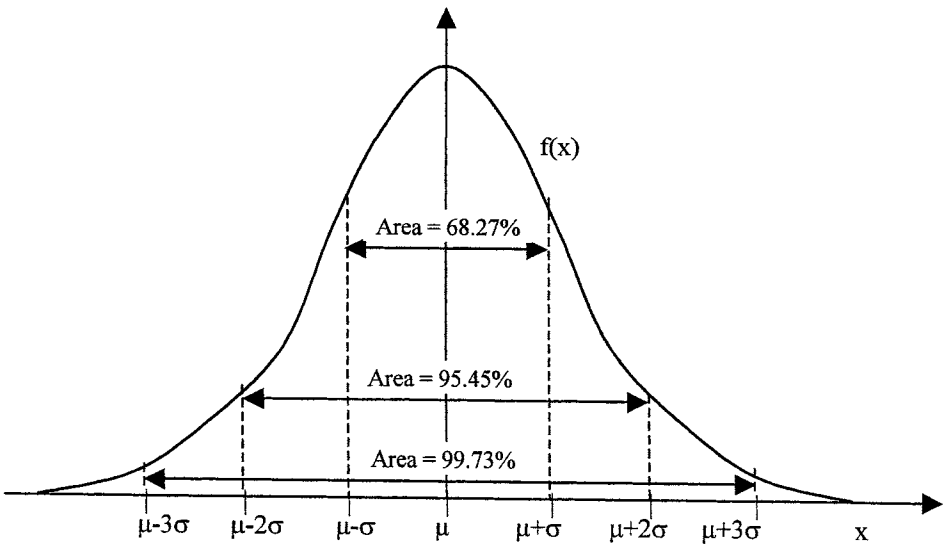

Fig. 4.5 The normal distribution and the area for three values of the standard variate.

The normal distribution is the most widely used distribution and is employed in analysis of variance, estimation of random errors of hydrologic measurements, hypothesis testing, generation of random numbers, etc. A random variable that is made up of the sum of many small independent effects is expected to follow a normal distribution. Many hydrologic variables are not normally distributed, but transformations can, in many cases, make them approximately normally distributed. When the time interval over which a hydrologic variable is measured increases, the variable approximately follows a normal distribution because the number of causative effects increases.

Log-Normal Distribution

When a random variable is the resultant of the product of many small effects, then its logarithm is made up of the sum of logarithms of these small effects. The logarithm of such a random variable can be expected to follow a normal distribution. Hence, if the variable is transformed to the log domain, it is likely to follow the normal distribution. Let $Y = ln X$. If Y is normally distributed, then X is log-normally distributed. The PDF of the log-normal distribution is

$$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu_y)^2}{2\sigma_y^2}\right] \quad x > 0 \tag{4.27}$$

The parameters of log-normal distribution are μ_y and σ_y which can be estimated by transforming all x_i's to y_i's by

$$y_i = \ln x_i \tag{4.28}$$

Extreme Value Type 1 Distribution (EV1)

Let a series of large number of (N) observations of random variable be subdivided into n subsamples of size m each, such that $N = nm$. Each subseries shall have two extreme values: one maximum and one minimum corresponding to, for example, floods and droughts. Gumbel (1958) showed that the n largest values of subsamples asymptotically follow an extreme value type 1 (EV1) distribution. This distribution, also known as the Gumbel distribution or double negative exponential distribution, is widely used for frequency analysis of floods, maximum rainfall, etc. This distribution is essentially a log-normal distribution with constant skewness (approximately 1.14). Its PDF and CDF are as follows:

$$f(x) = \alpha \exp\{-\alpha(x - \beta) - \exp[-\alpha(x - \beta)]\} \quad -\infty < x < \infty;\ -\infty < \beta < \infty;\ \alpha > 0$$

$$F(x) = \exp\{-\exp[-\alpha(x - \beta)]\} \tag{4.29}$$

where α, and β are scale and location parameters. The estimates of parameters using the method of moments are:

$$\hat{\alpha} = \frac{1.283}{s}; \qquad \hat{\beta} = \bar{x} - 0.45s \tag{4.30}$$

Example 4.2: For the Sabarmati River data of Example 4.1, find the value of parameters of EV1 distribution.

Solution: The mean and standard deviation of the data are 665.37 million cubic m and 346.9 million cubic m, respectively. Therefore, the method of moment estimates are:

$\alpha = 1.283/346.9 = 0.00367.$

and $\beta = 665.37 - 0.45*346.9 = 508.$

Gamma Distribution

The probability density function of the gamma distribution, with λ and n as parameters, is given by

$$f(x) = \frac{\lambda^n x^{n-1}}{(n-1)!}\exp(-\lambda x), \quad x>0,\ \lambda>0,\ n=1,2,3... \tag{4.31}$$

This distribution is extensively used in hydrology though it is not widely used in frequency analysis. It can be used to determine the time to nth event which is the time to the first event T_1 plus the time interval between the 1st and 2nd events T_2, plus the time interval between the 2nd and 3rd events T_3, and so on, or $T_1+T_2+ \ldots + T_n$. Because the time interval between events is described by the exponential distribution, the gamma distribution

The mean and the variance of the gamma distribution are

$$E(X) = n/\lambda \tag{4.32}$$
$$Var[X] = n/\lambda^2 \tag{4.33}$$

Example 4.3: The average time interval between floods in some parts of Gujarat is 2 years. Compute the probability that there will be a period less than or equal to 10 years for the occurrence of five floods.

Solution: Here, $\lambda = ½ = 0.5$.

Hence,
$$F(x \le 10) = \int_0^{10} \frac{0.5^5 x^4}{4!}\exp(-0.5x)dx$$
$$= \frac{1}{768}\int_0^{10}\exp(0.5x)dx$$
$$= \left[\frac{\exp(-0.5x)}{768}\left(-2x^4 - 16x^3 - 96x^2 - 384x - 768\right)\right]_0^{10}$$
$$= \frac{1}{768}\{\exp(-5)[2(10)^4 - 16(10)^3 - 96(10)^2 - 384(10) - 768]$$
$$- \exp(0)(0 - 0 - 0 - 768)\}$$
$$= -0.433 + 1 = 0.567.$$

Pearson Type-III Distribution (PT3)

The PT3 is a three-parameter gamma distribution and is widely used in hydrology. Its parameters are related to mean, standard deviation, and skewness.

$$f(x) = \frac{1}{a\Gamma(b)}\left(\frac{x-c}{a}\right)^{b-1} \exp\left(-\frac{x-c}{a}\right) \tag{4.34}$$

where *a, b*, and *c* are scale, shape, and location parameters, respectively, and Γ(b) is a gamma function. If $c = 0$, this distribution becomes a two-parameter gamma distribution. Parameters *a, b*, and *c* are related to mean, standard deviation, and coefficient of skewness as (method of moment estimates)

$$a = \sigma/\sqrt{b} \tag{4.35a}$$
$$b = (2/C_s)^2 \tag{4.35b}$$
$$c = \mu - \sigma\sqrt{b} \tag{4.35c}$$

Example 4.4: For the Sabarmati River data of Example 4.1, find the parameters of the PT3 distribution.

Solution: The estimates of parameters using the method of moments are

$b = (2/0.76)^2 = 6.93$ million cubic m.
$a = 346.9/\sqrt{6.93} = 131.78$ million cubic m.
$c = 665.37 - 346.9*\sqrt{6.93} = -247.84$.

Log Pearson Type-III Distribution (LPT3)

If the random variable $Y = ln\ X$ follows a PT3 distribution, then X follows the LPT 3 distribution. This distribution was recommended by the U. S. Water Resources Council for adoption as the standard distribution to be used in flood frequency analysis. Its PDF is given by

$$f(x) = \frac{1}{ax\Gamma(b)}\left(\frac{\ln x - c}{a}\right)^{b-1} \exp\left(-\frac{(\ln x - c)}{a}\right) \tag{4.36}$$

It is a very versatile distribution and can accommodate a variety of shapes. The mean, standard deviation, and coefficient of skewness of LPT3 distribution are given by

$$\mu_y = c + ab \tag{4.37a}$$
$$\sigma_y = a\sqrt{b} \tag{4.37b}$$
$$\gamma_y = 2/\sqrt{b} \tag{4.37c}$$

Transformation Techniques

In many instances, it is better to transform the data to a particular distribution of known characteristics instead of assuming that a known distribution fits the data. Since the properties of normal distribution are completely defined, the given data are transformed to a

normal distribution. Of the several transformations that are available, the power transformation is most commonly used (Jain and Singh, 1986a, 1986b).

$$y = (x^{\lambda} - 1)/\lambda, \text{ for } \lambda \neq 0 \quad (4.38)$$
$$= \ln x \quad \text{for } \lambda = 0$$

The reciprocal and square-root transformations can be obtained as special cases of eq. (4.38).

4.2.2 Discrete Probability Distributions

The use of discrete probability distributions is restricted generally to those random events in which the outcome can be described as success or failure, i.e., there are only two mutually exclusive events in an experiment. Moreover, the successive trials are independent and the probability of success remains constant from trial to trial. The binomial or Poisson distributions can be used to find the probability of occurrence of an event r times in n successive years.

Binomial Distribution

This distribution arises in Bernoulli processes where in any trial, the event may or may not take place. The probability of occurrence of the event is the same from one trial to another. This distribution usually occurs while dealing with complementary events. A common example is tossing of coins in which the probability of head appearing is the same in each trial. The occurrence of wet and dry days over a given time interval is also a complementary event. The probability of occurrence of the event r times in n successive years is given by:

$$P_{r,n} = n\ C_r\ P^r\ q^{n-r} = \frac{n!}{r!(n-r)!}\ p^r\ q^{n-r} \quad (4.39)$$

where $P_{r,n}$ is the probability of a random event of a given magnitude and exceedance probability P occurring r times in n successive years. The probability of the event not occurring at all in n successive years is:

$$P_{o,n} = q^n = (1 - p)^n \quad (4.40)$$

The probability of an event occurring at least once in n successive years:

$$P_1 = 1 - q^n = 1 - (1 - p)^n \quad (4.41)$$

Example 4.5: An analysis of data on the maximum one-day rainfall depth at a station indicated that a depth of 280 mm had a return period of 50 years. Determine the probability of a one-day rainfall depth equal to or greater than 280 mm occurring (a) once in 20 successive years, and (b) two times in 15 successive years.

Solution: Here, $P = 1/50 = 0.02$.

a) In the first case, $n = 20$, $r = 1$. Therefore, from eq. (4.39)

$$P_{1,20} = \frac{20!}{19!1!} * (0.02) * (0.98)^{19} = 0.272.$$

b) In this case, $n = 15$, $r = 2$. Therefore,

$$P_{2,15} = \frac{15!}{13!2!} * (0.02^2) * (0.980)^{13} = 0.0292 .$$

Example 4.6: What is the probability that a 5-year flood will not occur at all in a 10-year period?

Solution: Here, p = 1/5 = 0.2, n = 10, and r = 0. Hence the probability is

$$P_{0,10} = \frac{10!}{0!10!} * 0.2^0 * (0.8)^{10} = 0.1074$$

Poisson Distribution

The Poisson distribution is a limiting form of the binomial distribution when p is very small and n is very large, and np tends to a constant value λ. This may happen when the interval over which the Bernoulli process is defined gets smaller and smaller and the number of trials becomes greater and greater, keeping np constant. The Poisson distribution has only one parameter λ that denotes the expected mean frequency of occurrence of some event in a given time t. The probability distribution of the number of events in a given time is

$$P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}, \quad \lambda > 0, \quad x = 0, 1, 2, ... \tag{4.42}$$

The CDF of the Poisson distribution is

$$P(X \le x) = \sum_{i=0}^{x} \frac{\lambda^i \exp(-\lambda)}{i!} \tag{4.43}$$

The conditions for application of Poisson distribution are: a) the number of events is discrete, b) two events cannot coincide, c) the mean number of events per unit time is constant, and d) events are independent. Thus, it can be applied to following situations with p relatively small and n relatively large to determine the probability of:

(i) droughts in a given time period,
(ii) number of rainy days at a given location,
(iii) probability of rare flood events, and
(iv) probability of reservoir being empty in any one year out of a long period of record.

4.3 METHODS OF PARAMETER ESTIMATION

A number of methods are available to estimate parameters of hydrologic models. Some of the popular methods used in hydrology include (1) method of moments (Nash, 1959; Dooge, 1973; Harley, 1967; Singh, 1988); (2) method of probability weighted moments (Greenwood, et al., 1979); (3) method of mixed moments (Rao, 1980, 1983; Shrader, et al.,

1981); (4) L-moments (Hosking, 1986, 1990, 1992); (5) maximum likelihood estimation (Douglas, et al., 1976; Sorooshian, et al., 1983; Phien and Jivajirajah, 1984); and (6) least squares method (Jones, 1971; Snyder, 1972; Bree, 1978a, 1978b). A brief review of these methods is given here.

4.3.1 Method of Moments for Continuous Systems

The method of moments is frequently utilized to estimate parameters of linear hydrologic models (Nash, 1959; Dooge, 1973; Singh, 1988). Nash (1959) developed the theorem of moments which relates the moments of input, output and impulse response functions of linear hydrologic models. Moments of functions are amenable to use of standard methods of transform, such as the Laplace and Fourier transforms. The method of moments has been used to estimate parameters of frequency distributions. Wang and Adams (1984) reported on parameter estimation in flood frequency analysis. Ashkar et al. (1988) developed a generalized method of moments and applied it to the generalized gamma distribution. Kroll and Stedinger (1996) estimated moments of a lognormal distribution using censored data.

Let X be a continuous variable (it may or may not be a random variable) and $f(x)$ its function satisfying some necessary conditions. The r^{th} moment of $f(x)$ about an arbitrary point is denoted as M_r^a (f). Here M denotes the moment, the subscript ($r \geq 0$) denotes the order of the moment, the superscript denotes the point about which to take the moment, and the quantity within the parentheses denotes the function, in normalized form, whose moment is to be taken. Thus, the r^{th} moment of the function $f(x)$ can be defined as

$$M_r^a(f) = \int_{-\infty}^{\infty} (x-a)^r f(x)dx \tag{4.44}$$

This is the definition used normally in statistics. If the area enclosed by the function $f(x)$ does not add to unity, the definition of eq. (4.44) becomes

$$M_r^a(f) = \frac{\int_{-\infty}^{\infty} (x-a)^r f(x)dx}{\int_{-\infty}^{\infty} f(x)dx} \tag{4.45}$$

As the denominator in eq. (4.45) defines the area under the curve which is usually unity or made to unity by normalization, the two definitions are numerically the same. In this text the definition of eq. (4.44) is used with $f(x)$ normalized beforehand. It is assumed here that the integral in eq. (4.44) converges. There are some functions which possess moments of lower order; some do not possess any moment except of zero order. However, if a moment of higher order exists, moments of all lower orders must exist. Fig. 4.6 shows the concept of moment of a function about an arbitrary point.

Moments are statistical descriptors of a distribution and reflect on its qualitative properties. For example, if r = 0 then eq. (4.44) yields

$$M_0^a = \int_{-\infty}^{\infty} (x-a)^0 f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1 \tag{4.46}$$

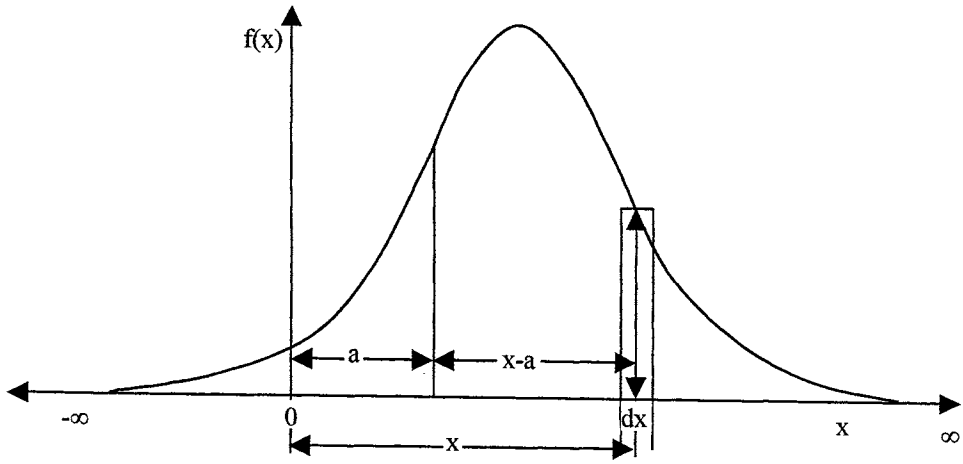

Fig. 4.6 Concept of moment of a function f(x) about an arbitrary point.

Thus, the zero-order moment is the area under the curve defined by $f(x)$ subject to $-\infty < x < \infty$. If $r = 1$, then eq. (4.44) yields

$$M_1^a = \int_{-\infty}^{\infty} (x-a)^1 f(x)dx = \mu - a \tag{4.47}$$

where μ is the centroid of the area or mean. Thus, the first moment is the weighted mean about the point a. If $a = 0$, the first moment gives the mean. When $a = \mu$, the r^{th} moment about the mean is

$$M_r^{\mu} = \int_{-\infty}^{\infty} (x-\mu)^r f(x)dx \tag{4.48}$$

Henceforth, for simplicity of notation, we will drop the superscript if the moment is taken about 0. The descriptive properties of the moments with respect to a specific function can be summarized as follows:

M_0 =Area
M_1 = Mean
M_2^{μ} = Variance, a measure of dispersion of the function about the mean
M_3^{μ} = Measurement of skewness of the function
M_4^{μ} = Kurtosis, a measure of the peakedness of the function

4.3.2 Method of Moments for Discrete Systems

For a discrete function, represented as f_j, $j = -\infty, \ldots, -1, 0, 1, \ldots, \infty$, the r^{th} moment about any other arbitrary point can be defined in a manner analogous to that for continuous functions. When the arbitrary point is the origin, the r^{th} moment is defined as

$$M_r = \sum_{m=-\infty}^{\infty} m^r f_m \tag{4.49}$$

When f_m is normalized:

$$\sum_{m=-\infty}^{\infty} f_m = 1 \tag{4.50}$$

Otherwise,

$$M_r = \sum_{m=-\infty}^{\infty} m^r f_m \Big/ \sum_{m=-\infty}^{\infty} f_m \tag{4.51}$$

It can be noticed that eqs. (4.49) and (4.51) are analogous to eqs. (4.44) and (4.45). Fig. 4.7 explains the concept of moment of a discrete function.

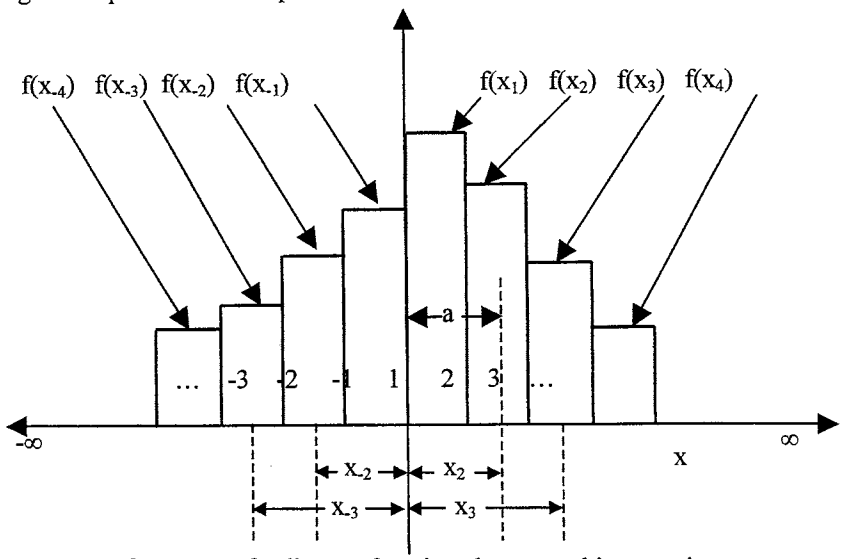

Fig. 4.7 Concept of moment of a discrete function about an arbitrary point.

Example 4.7: The histogram of annual flows of Sabarmati River is given Table 4.2 (A plot is available in Fig. 4.4). Find the mean and variance of the data using the method of moments.

Table 4.2 Histogram of annual flows of Sabarmati River.

Discharge range	Frequency	Discharge range	Frequency
100-200	6	200-300	9
300-400	11	400-500	9
500-600	9	600-700	9
700-800	19	800-900	6
900-1000	6	1000-1100	1
1100-1200	5	1200-1300	2
1300-1400	3	1400-1500	0
1500-1600	2	1600-1700	0
1700-1800	1		

Solution: The first moment of the data = (150*6 + 250*9 + 350*11 + ...+ 1750*1)/98
= 664.2857 cumec.

This is the mean of the data.

The second moment about the mean will give the variance.
Second moment = [$(150-664)^2$*6 + $(250 - 664)^2$*9 + ... + $(1750-664)^2$*1]/98
= 120000 $cumec^2$.

4.3.3 Method of Probability Weighted Moments

Greenwood et al. (1979) introduced the method of probability weighted moments (PWM) and showed its usefulness in deriving explicit expressions for parameters of distributions whose inverse forms X=X (F) can be explicitly defined. They derived relations between parameters and PWMs for Generalized Lambda, Wakeby, Weibull, Gumbel, Logistic and Kappa distributions. Hosking (1986) developed the theory of probability weighted moments and applied it to estimate parameters of several distributions. For flood frequency analysis, Haktanir (1996) modified the conventional method of probability-weighted moments for estimation of parameters of any distribution without the need to use a plotting position formula. Wang (1997) defined partial PWMs and derived them for extreme value type I and III distributions. He applied these moments to lower bound censored samples.

Let a probability distribution function be denoted as F = F(X) = P[X ≤ x]. The PWMs of this function can be defined as

$$M_{i,j,k} = E[x^i F^j (1-F)^k] = \int_0^1 [x(F)]^i F^j (1-F)^k \, dF \tag{4.52}$$

where $M_{i,j,k}$ is the probability weighted moment of order (i, j, k), E is the expectation operator, and i, j and k are real numbers. If $j = k = 0$ and i is a nonnegative integer then $M_{i,0,0}$ represents the conventional moment of order i about origin. If $M_{i,0,0}$ exists and X is a continuous function of F, then $M_{i,j,k}$ exists for all nonnegative real numbers j and k.

For nonnegative integers j, k, we can express

$$M_{i,0,k} = \sum_{j=0}^{k} \binom{k}{j} (-1)^j M_{i,j,0} \tag{4.53a}$$

$$M_{i,j,0} = \sum_{k=0}^{j} \binom{j}{k} (-1)^k M_{i,0,k} \tag{4.53b}$$

If $M_{i,0,k}$ exists and X is a continuous function of F then $M_{i,j,0}$ exists. When the inverse $X = X(F)$ of the distribution $F = F(X)$ cannot be analytically defined, it may, in general, be difficult to derive $M_{i,j,k}$ analytically.

We normally work with the moments $M_{i,j,k}$ into which x enters linearly. In particular, the PWM for hydrologic applications are defined as

$$a_r = M_{1,0,r} = E\,[\,x\,\{1 - F(x)\}^r\,], \qquad r = 0,1,2,\ldots \tag{4.54a}$$

$$b_r = M_{1,r,0} = E\,[\,x\,\{F(x)\}^r\,], \qquad r = 0,1,2,\ldots \tag{4.54b}$$

Here $a_{k-1} = E[x_{1:k}]$ and $b_k = E[x_{k:k}]$ are the expected values of extreme order statistics.

In general, a_r and b_r are functions of each other as

$$a_r = \sum_{k=0}^{r} (-1)^k \binom{r}{k} b_k \tag{4.55a}$$

$$b_r = \sum_{k=0}^{r} (-1)^k \binom{r}{k} a_k \tag{4.55b}$$

Therefore,

$$\begin{aligned} a_0 &= b_0 & b_0 &= a_0 \\ a_1 &= b_0 - b_1 & b_1 &= a_0 - a_1 \\ a_2 &= b_0 - 2b_1 + b_2 & b_2 &= a_0 - 2a_1 + a_2 \\ a_3 &= b_0 - 3b_1 + 3b_2 - b_3 & b_3 &= a_0 - 3a_1 + 3a_2 - a_3 \end{aligned} \tag{4.56}$$

A complete set of these a or b probability-weighted moments characterizes a distribution.

3.3.4 Methods of Mixed Moments

Rao (1980, 1983) proposed the method of mixed moments (MIXM) which is applicable to any log-probability distribution. As the name suggests, the MIXM method is based on mixing the moments of real and logarithmically transformed data. Thus, only the first two moments (mean and variance) of the data are used. For example, if it is desired to fit the log-Pearson type (LP) III distribution to a given set of data then its parameters can be estimated in two ways: (1) The first method uses the mean $(\overline{x})$ and variance S_x^2 of real data and the mean of logarithmically transformed values ($Y = log\ X$). (2) The second method uses the mean of real data $(\overline{x})$ and the mean and variance S_y^2 of logarithmically transformed data (Y = log X). Using Monte Carlo experiments, Rao (1980) showed that the first method possessed superior statistical properties as compared to the second method.

4.3.5 Method of L-Moments

The probability-weighted moments characterize a distribution but are not meaningful by themselves. L-moments were developed by Hosking (1986) as functions of PWMs which provide a descriptive summary of the location, scale, and shape of the probability distribution. These moments are analogous to ordinary moments and are expressed as *linear* combinations of order statistics, hence the name. They can also be expressed by linear combinations of probability-weighted moments. Thus, the ordinary moments, the probability weighted moments, and L-moments are related to each other. L-moments are known to have several important advantages over ordinary moments. L-moments have less bias than ordinary moments because they are linear combinations of ranked observations.

As an example, variance (second moment) and skewness (third moment) involve squaring and cubing of observations, respectively, which compel them to give greater weight to the observations far from the mean. As a result, they result in substantial bias and variance.

If X is a real value ordered random variate of a sample of size n, such that $x_{1:n} \leq x_{2:n} \leq \ldots \leq x_{n:n}$ with the cumulative distribution $F(x)$ and quantile function $x(F)$, then the rth L-moment of X (Hosking 1990) can be defined as a linear function of the expected order statistics as:

$$L_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} E\{X_{r-k:r}\}, \quad r = 1,2,... \tag{4.57}$$

where E{. } is the expectation of an order statistic and is equal to

$$E\{X_{j:r}\} = \frac{r!}{(r-j)!\,j!}\int x\{F(x)\}^{j-1}\{1-F(x)\}^{r-j}\,dF(x) \tag{4.58}$$

As noted by Hosking (1990), the natural estimator of L_r, based on an observed sample of data, is a linear combination of the ordered data values, i.e., an L-statistic. Substituting eq. (4.58) in eq. (4.57), expanding the binomials of $F(x)$ and summing the coefficients of each power of $F(x)$, one can write

$$L_r = E[xP^*_{r-1}\{F(x)\}] = \int_0^1 x(F)P^*_{r-1}(F)dF, \quad r = 1,2,... \tag{4.59}$$

where is $P^*_r(F)$ the r-th shifted Legendre polynomial expressed as

$$P^*_r(F) = \sum_r^k (-1)^{r-k}\binom{r}{k}\binom{r+k}{k}F^k \tag{4.60}$$

Eq. (4.60) can simply be written as

$$P^*_r(F) = \sum_{k=0}^{r} P_{r,k}F^k \tag{4.61}$$

and $$P_{r,k} = (-1)^{r-k}\binom{r}{k}\binom{r+k}{k} \tag{4.62}$$

The shifted Legendre polynomials are related to the ordinary Legendre polynomials $P_r(u)$ as $P^*_r(u) = P_r(2u - 1)$, and are orthogonal on the interval (0,1) with a constant weight function.

The first four L moments are

$$L_1 = E(x) = \int x dF \tag{4.63}$$

$$L_2 = \frac{1}{2}E(x_{2:2} - x_{1:2}) = \int x(2F-1)dF \tag{4.64}$$

$$L_3 = \frac{1}{3}E(x_{3:3} - 2x_{2:3} + x_{1:3}) = \int x(6F^2 - 6F + 1)dF \tag{4.65}$$

$$L_4 = \frac{1}{4}E(x_{4:4} - 3x_{3:4} + 3x_{2:4} - x_{1:4}) = \int x(20F^3 - 30F^2 + 12F - 1)dF \tag{4.66}$$

4.3.6 Method of Maximum Likelihood Estimation (MLE)

The maximum likelihood (ML) estimation method is widely accepted as one of the most powerful parameter estimation methods. Asymptotically, the ML parameter estimates are unbiased, have minimum variance, and are normally distributed, while in some cases these properties hold for small samples. The MLE method has been extensively used for estimating parameters of frequency distributions as well as fitting conceptual models.

Let $f(x; a_1, a_2, \ldots a_m)$ be a PDF of the random variable X with parameters a_i, i=1, 2, …, m, to be estimated. For a random sample of data $x_1, x_2, \ldots x_n$, drawn from this probability density function, the joint PDF is defined as

$$f(x_i, x_2, \ldots x_n; a_1, a_2, \ldots a_m) = \prod_{i=1}^{n} f(x_i; a_1, a_2, \ldots a_m) \tag{4.67}$$

Interpreted conceptually, the probability of obtaining a given value of X, say x_1, is proportional to f(x; a_1, a_2,… a_m). Likewise, the probability of obtaining the random sample $x_1, x_2, \ldots x_n$ from the population of X is proportional to the product of the individual probability densities or the joint PDF. This joint PDF is called the likelihood function, denoted by L.

$$L = \prod_{i=1}^{n} f(x_1; a_1, a_2, \ldots a_m) \tag{4.68}$$

where the parameters a_i, i=1,2,…m, are unknown.

By maximizing the likelihood that the sample under consideration is the one that would be obtained if n random observations were selected from f(x; a_1, a_2, … a_m), the unknown parameters are determined, and hence the name of the method. The values of parameters so obtained are known as MLE estimators. Since the logarithm of L attains its maximum for the same values of a_i, i = 1,2, … m, as does L, the MLE function can also be expressed as

$$\ln L = L^* = \ln \prod_{i=1}^{n} f(x_1; a_1, a_2, \ldots a_m) = \sum_{i=1}^{n} \ln f(x_1; a_1, a_2, \ldots a_m) \tag{4.69}$$

Frequently *ln[L]* is maximized, for it is many times easier to find the maximum of the logarithm of the maximum likelihood function than that of the normal L.

The procedure for estimating parameters or determining the point where the MLE function achieves its maximum involves differentiating L or *ln* L partially with respect to each parameter and equating each differential to zero. This results in as many equations as the number of unknown parameters. For m unknown parameters, we get

$$\begin{aligned} \frac{\partial L(a_1, a_2, \ldots a_m)}{\partial a_1} &= 0 \\ \frac{\partial L(a_1, a_2, \ldots a_m)}{\partial a_2} &= 0 \end{aligned} \tag{4.70}$$

$$\frac{\partial L(a_1, a_2, \ldots a_m)}{\partial a_m} = 0$$

These m equations in m unknowns are then solved for the m unknown parameters.

Applying the method of maximum likelihood, the parameters of EV1 distribution for the Sabarmati data are $\alpha = 0.00354$ and $\beta = 503.6$. Recall that the estimates using method of moments were $\alpha = 0.00367$ and $\beta = 508$.

4.3.7 Method of Least Squares

The method of least squares (MOLS) is one of the most frequently used parameter estimation methods in hydrology. Natale and Todini (1974) presented a constrained MOLS for linear models in hydrology. Williams and Yeh (1983) described MOLS and its variants for use in rainfall-runoff models. Jones (1971) linearized weight factors for least squares (LS) fitting. Shrader et al. (1981) developed a mixed-mode version of MOLS and applied it to estimate parameters of the log-normal distribution. Snyder (1972) reported on fitting of distribution functions by non-linear least squares. Stedinger and Tasker (1985) performed regional hydrologic analysis using ordinary, weighted and generalized least squares.

Let there be a function $Y = f(X; a_1, a_2, \ldots a_m)$, where a_i, $i = 1, 2, \ldots m$, are parameters to be estimated. The method of least squares (MOLS) involves estimating parameters by minimizing the sum of squares of all deviations between observed and computed values of Y. Mathematically, this sum D can be expressed as

$$D = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [y_0(i) - y_c(i)]^2 = \sum_{i=1}^{n} [y_0(i) - f(x; a_1, a_2, \ldots a_m)]^2 \tag{4.71}$$

where $y_0(i)$ is the i^{th} observed value of Y, $y_c(i)$ is the i^{th} computed value of Y, and $n > m$ is the number of observations. The minimum of D in eq. (4.71) can be obtained by differentiating D partially with respect to each parameter and equating each differential to zero, e.g.,

$$\frac{\partial \sum_{i=1}^{n} [y_0(i) - f(x_i; a_1, a_2, \ldots a_m)]^2}{\partial a_1} = 0 \tag{4.72}$$

The resulting m equations, usually called the normal equations, are then solved for estimation of m parameters. This method is frequently used to estimate parameters of linear regression model (see Section 4.7).

4.4 CONCEPT OF ENTROPY

Entropy can be considered as a measure of the degree of uncertainty or disorder associated with a system. Indirectly it also reflects the information content of space-time measurements. Entropy is viewed in three different but related contexts and is hence

typified by three forms: thermodynamical entropy, statistical-entropy, and information-theoretical entropy. In water resources, the most frequently used form is the information-theoretical entropy.

The concept of entropy provides a quantitative measure of uncertainty. To that end, consider a probability density function (PDF) *f(x)* associated with a dimensionless random variable *X*. The dimensionless random variable may be constructed by dividing the observed quantities by its mean value, e.g., annual flood maxima divided by the mean annual flood. As usual, *f(x)* is a possible function for every *x* in some interval *(a, b)* and is normalized to unity such that

$$\int_a^b f(x)dx = 1 \tag{4.73}$$

The most popular measure of entropy was first mathematically given by Shannon (1948) and has since been called the Shannon entropy functional (SEF), denoted as *I[f]* or *I[x]*. It is a numerical measure of uncertainty associated with *f(x)* in describing the random variable *X*, and is defined as

$$I[f] = I[x] = -k\int_a^b f(x)\ln[f(x)/m(x)]dx \tag{4.74}$$

where $k > 0$ is an arbitrary constant or scale factor depending on the choice of measurement units, and *m(x)* is an invariant measure function guaranteeing the invariance of *I[f]* under any allowable change of variable, and provides an origin of measurements of *I[f]*. Scale factor *k* can be absorbed into the base of the logarithm and *m(x)* may be taken as unity so that eq. (4.74) is often written as

$$I[f] = I[x] = -\int_a^b f(x)\ln[f(x)]dx; \qquad \int_a^b f(x)dx = 1 \tag{4.75}$$

We may think of *I[f]* as the mean value of *–ln[f(x)]*. Actually, *-I* measures the strength, *+I* measures the weakness. SEF allows choosing that *f(x)* which minimizes the uncertainty. Note that *f(x)* is conditioned on the constraints used for its derivation. Singh (1988, 1998) has described the theory of entropy and has given expressions of SEF for a number of probability distributions.

4.4.1 Principle of Maximum Entropy

According to the principle of maximum entropy (POME), "the minimally prejudiced assignment of probabilities is that which maximizes entropy subject to the given information." Mathematically, it can be stated as follows: Given *m* linearly independent constraints *C* in the form

$$C_i = \int_a^b y_i(x)f(x)dx, \quad i = 1,2,...,m \tag{4.76}$$

where $y_i(x)$ are some functions whose averages over $f(x)$ are specified. The maximum of I, subject to the conditions in eq. (4.76), is given by the distribution

$$f(x) = \exp[-\lambda_0 - \sum_{i=1}^{m} \lambda_i y_i(x)] \tag{4.77}$$

where λ_i, $i = 0,1,\ldots, m$, are Lagrange multipliers and can be determined from eqs. (4.76) and (4.77) along with the normalization condition in eq. (4.73).

4.4.2 Entropy-Based Parameter Estimation

The general procedure for deriving an entropy-based parameter estimation method for a frequency distribution involves the following steps: (1) Define the given information in terms of constraints. (2) Maximize the entropy subject to the given information. (3) Relate the parameters to the given information. More specifically, let the available information be given by eq. (4.76). Since POME specifies $f(x)$ by eq. (4.77), inserting eq. (4.77) in eq. (4.75) yields

$$I[f] = \lambda_0 + \sum_{i=1}^{m} \lambda_i C_i \tag{4.78}$$

In addition, the potential function or the zeroth Lagrange multiplier λ_0 is obtained by inserting eq. (4.77) in eq. (4.78) as

$$\int_a^b \exp[-\lambda_0 - \sum_{i=1}^{m} \lambda_i y_i]\, dx = 1 \tag{4.79}$$

resulting in

$$\lambda_0 = \ln \int_a^b \exp[-\sum_{i=1}^{m} \lambda_i y_i]\, dx \tag{4.80}$$

The Lagrange multipliers are related to the given information (or constraints) by

$$-\frac{\partial \lambda_0}{\partial \lambda_i} = C_i \tag{4.81}$$

It can also be shown that

$$\frac{\partial^2 \lambda_0}{\partial \lambda_i^2} = \text{var}[y_i(x)]; \quad \frac{\partial^2 \lambda_0}{\partial \lambda_i \partial \lambda_j} = \text{cov}[y_i(x), y_j(x)], i \neq j \tag{4.82}$$

With the Lagrange multipliers estimated from eqs. (4.81) and (4.82), the frequency distribution given by eq. (4.77) is uniquely defined. It is implied that the distribution parameters are uniquely related to the Lagrange multipliers. Clearly, this procedure states that a frequency distribution is uniquely defined by specification of constraints and application of POME.

4.5 PROBLEMS OF PARAMETER ESTIMATION

The parameters of a distribution function are estimated from sample values. There are, of course, myriad ways by which to obtain parameter estimates. The sample data may contain

errors, the hypotheses underlying the method of parameter estimation may not yield accurate estimates, and there may be truncation and round-off errors. These sources of errors may result in errors in parameter estimates. Each estimate of a parameter is a function of sample values which are observations of a random variable. Thus, the parameter estimate itself is a random variable having its own sampling distribution. An estimate obtained from a given set of values can be regarded as an observed value of the random variable. Thus, the goodness of an estimate can be judged from its distribution.

Some important questions arise here. How should we best use the data to form estimates? What do we mean by the best estimates? Also, are these estimates unique? How do we select the best parameter estimator if there is one? A number of statistical properties are available by which to address the above questions. These are discussed below.

Bias

Let the parameter be a and its estimate a_c. The estimate a_c is called an unbiased estimate of a if $E(a_c) = a$. In general, an estimate will have a certain bias $b(a)$ depending on a so that

$$E(a_c = a + b(a) \tag{4.83}$$

Obviously, $b(a) = 0$ for an unbiased estimate. It should, however, be noted that an individual a_c is not equal to or even close to a even if $b(a) = 0$. It simply implies that the average of many independent estimates of a will be equal to a.

The bias in a given quantity is usually measured in dimensionless terms and is often referred to as standardized bias (or BIAS). Thus, BIAS is defined as

$$BIAS = \frac{E(\hat{a}) - a}{a} \tag{4.84}$$

where $\hat{a}$ is an estimate of parameter or quantile of a. In Monte Carlo experimentation, large numbers of samples of different sizes are generated from a given population. For each sample, then, an estimate of a is obtained. If there are, say, 1000 samples of a given size generated then there are 1000 values of parameter a. Thus, E(a) is the average of the 1000 estimates of a for a given sample size and is estimated as

$$E(\hat{a}) = \sum_{i=1}^{n} \hat{a}_i / n \tag{4.85}$$

where n is the number of samples generated or the number of values of the a estimate. The value of a in eq. (4.84) is the true value of a or the value of parameter a of the population.

Consistency

Let there be a sample of size n. The estimate a_c is called a consistent estimate of a if it converges to a with probability one as n tends to infinity. Because many unbiased estimates have variances of the type

$$\text{Var}\,(a_c) \cong C/(n)^{0.5} \tag{4.86}$$

where C is constant. The condition of consistency is satisfied in most cases. In practice, it is desirable to have Var (a_c) as small as possible. This would imply that the probability density function of a_c would be more concentrated about a.

Efficiency

An estimate a_c of a is said to be efficient if it is unbiased and its variance is at least as small as that of any other unbiased estimate of a. If there are two estimates of a, say a_1 and a_2, then the relative efficiency of a_1 with respect to a_2 is defined as

$$e = \frac{E(a_1 - a)^2}{E(a_2 - a)^2} \leq 1 \tag{4.87}$$

if $E(a_2 - a)^2 > E(a_1 - a)^2$, then $e \leq 1$. An efficient estimate has $e = 1$. If an efficient estimate exists, it may be approximately obtained by use of the MLE or entropy method.

Sufficiency

An estimate a_c of a is said to be sufficient if it uses all of the information that is contained in the sample. More precisely, let a_1 and a_2 be two independent estimates of a. Now, a_1 is considered a sufficient estimate if the joint probability distribution of a_1 and a_2 has the property.

$$f(a_1,a_2) = f(a_1)f(a_2 \mid a_1) = f(a_1)K(x_1,x_2,\ldots x_n) \tag{4.88}$$

in which $f(a_1)$ is the distribution of a_1, $f(a_2 \mid a_1)$ is the conditional distribution of a_2 given a_1, and $K(x_1, x_2, \ldots x_n)$ is not a function of a but only of x_i's . If eq. (4.88) holds, then a_2 does not produce any new information about a which is not already contained in a_1. In this case, a_1 is a sufficient estimate.

Standard Error

Another dimensionless performance measure frequently used in hydrology is the standard error (SE), defined as

$$SE = \sigma(\hat{a})/a \tag{4.89}$$

where σ (.) denotes the standard deviation of a and is computed as

$$\sigma(\hat{a}) = \left[\frac{1}{n-1}\sum_{i=1}^{n}\{\hat{a}_i - E(\hat{a}_i)\}^2\right]^{1/2} \tag{4.90}$$

where the summations are over n estimates $\hat{a}$ of a. In Monte Carlo experiments, referred to as above, for each sample size, a value of SE is obtained. Thus, this measure is similar to

the coefficient of variation.

Root Mean Square Error

The root mean square error (RMSE) is one of the most frequently employed performance measures and is defined for parameter a estimate as

$$RMSE = E\left[(\hat{a} - a)^2\right]^{1/2} / a \tag{4.91}$$

where E[.] is the expectation of [.]. It can be shown that RMSE is related to BIAS and SE as

$$RMSE = \left[\frac{n-1}{n} SE^2 + BIAS^2\right]^{1/2} \tag{4.92}$$

Robustness

Kuczera (1982a, b, c) defined a robust estimator as the one that is resistant and efficient over a wide range of population fluctuations. Two criteria for resistant estimator are mini-max and minimum average RMSE. According to the mini-max criteria, the maximum RMSE for all population cases should be minimum. Thus, for a resistant estimator the average RMSE as well as the maximum RMSE should be minimum.

Relative Mean Error

Another measure of error in assessing the goodness of fit of hydrologic models is the relative mean error (RME) defined as

$$RME = \frac{1}{N}\left(\sum_{i=1}^{N}\left[\frac{Q_0 - Q_c}{Q_0}\right]^2\right)^{0.5} \tag{4.93}$$

in which N is the sample size, Q is the observed quantity of a given probability and Q_c is the computed quantity of the same probability. Also, used sometimes is the relative absolute error defined as

$$RAE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Q_0 - Q_c}{Q_c}\right| \tag{4.94}$$

4.6 HYPOTHESIS TESTING

Many times, while analyzing water resources data, questions arise such as: does the flow at a given site follow normal distribution? Is the quality of water in the river violating the relevant standards? Is there significant correlation between two given variables? If the results of the model are very close to the observed one, a conclusion may be reached without using a statistical test but sometimes the difference could be such that well-articulated tests are needed to arrive at a conclusion. Statistical procedures known as hypothesis testing are followed in these situations. The hypothesis tests can be broadly

divided into two categories: parametric tests and non-parameteric tests. The tests that require the specification of the type of distribution of data are termed as parametric test. The examples are the t-test, and the tests related to linear regression. In most cases, it is assumed that the data follow a normal distribution. The non-parametric or distribution free tests do not require specification of the distribution of data. These tests do not require that the data should follow a specified distribution. The examples of these tests are Kendall's Tau test and Kruskal-Wallis test.

The main steps in conducting a statistical tests are: i) formulate of the hypothesis that is to be tested, ii) formulate an alternate hypothesis, iii) formulate the test statistic and significance level, iv) determine the distribution of test statistic, and v) conduct the test.

The first step in parametric tests is to formulate a hypothesis which is termed as null hypothesis and is denoted by H_0. Usually, this is the hypothesis of no change or no difference. For example, the distributions of flow at two stations is identical or there is no correlation between two given variables. The characteristics of the data that is to be examined is quantified by the test statistic. The test checks whether the behavior of the test statistic is similar to what is expected (leaving aside the possibility of this to happen due to chance alone) if H_0 were true. This null hypothesis may be mathematically written as:

$$H_0: \mu_1 = \mu_0 \tag{4.95}$$

This statement says that the statistical measure (in this case, mean) of the parent population from which the sample was drawn is not different from the mean μ_0 of a population.

After choosing the null hypothesis, an alternative hypothesis is formulated such that the null and the alternate hypothesis are mutually exclusive and all inclusive. The alternate hypothesis is a statement of some departure from the null hypothesis that might be expected. The alternate hypothesis can be stated as

$$H_a : \mu_1 \neq \mu_0 \tag{4.96}$$

This states that the mean of the population from which the sample was drawn is not equal to the specified population mean. Some typical examples of null hypothesis are: the distributions of flow at two stations are different or the two variables under examination are related to each other.

Since the rejection of null hypothesis implies that the specific assumption is not true, the alternate hypothesis must be sufficiently general. If the null hypothesis is rejected, the true behavior lies somewhere in the vast set of possibilities stated in the alternate hypothesis. Sometimes the outcome of hypothesis testing is stated as rejection of the null hypothesis versus the failure to reject the null hypothesis.

After the hypothesis has been formulated, statistical analysis is carried out to either accept it or reject it. The hypothesis may be true or false and it might be accepted or rejected. This produces four possible combinations which are indicated in the Table 4.3.

Table 4.3 Hypothesis testing: possible outcomes and their probabilities.

	Hypothesis is correct	Hypothesis is incorrect
Hypothesis is accepted	Correct decision. This outcome has a probability 1-α	Type II error with a probability β
Hypothesis is rejected	Type I error with a probability of α	Correct decision with probability 1-β

If a correct hypothesis is accepted or a wrong hypothesis is rejected, this is a right decision. If, however, a null hypothesis which is true is rejected, this leads to an erroneous conclusion and type I error is said to have been committed. In statistical jargon, the probability of making type I error is termed as level of significance. This probability is to be specified before carrying out the test.

In hydrology, most commonly the significance level of 0.05 (1 in 20) or 0.01 (1 in 100) is adopted. If a level of 0.05 is chosen, it implies that the decision of statistical test may be in error about one time out of 20. In terms of distribution properties, this corresponds to 5% of the area under the curve. This concept is illustrated for a two-sided test in Fig. 4.8 in which the test statistic under null hypothesis is normal and each shaded area near the two tails contains 2.5% of the total area. This shaded area is termed as the area of rejection or the critical region. Since the alternative hypothesis in eq. (4.96) is of inequality type, the null hypothesis is rejected if the test statistic falls in the critical region either because it is too high or too low. The significance level that is chosen in a particular circumstance depends upon the risks associated with a wrong decision.

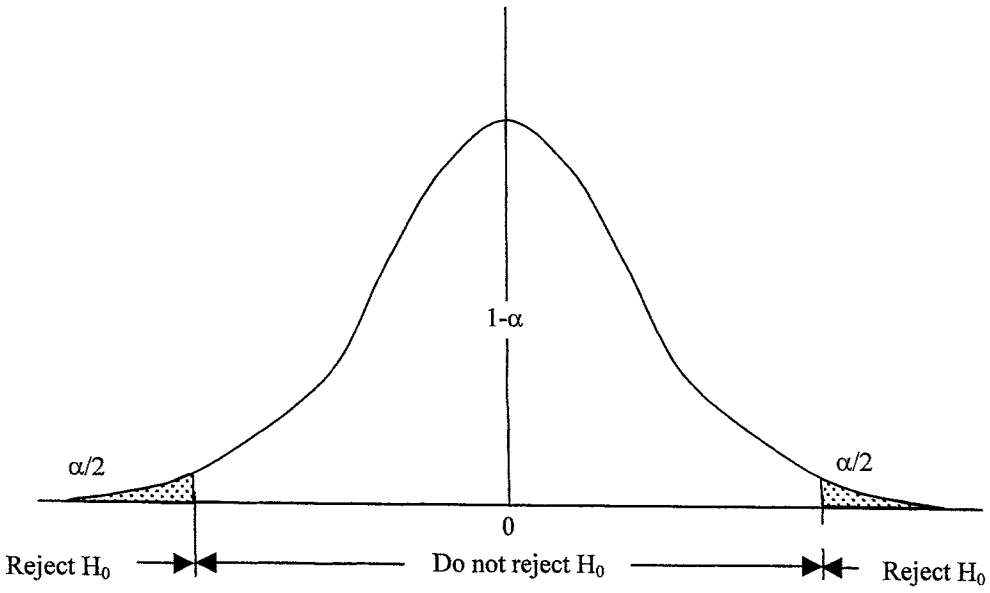

Fig. 4.8 Two-sided test of hypothesis.

If an incorrect hypothesis is accepted, this leads to a type II error, denoted by β. The null hypothesis is formulated with the intention that it will be rejected because this eliminates the possibility of a type II error. The probability of making type II error is not known; this error increases as significance level α increases.

4.6.1 The t-Test

The uncertainty in the estimates of parameters of a probability distribution, such as mean and standard deviation, can be studied using Student's t-distribution. This distribution is similar to the normal distribution but its shape is dependent upon the size of sample; the exact shape depends upon the number of observations in the sample. The t-distribution approaches the normal distribution when there are infinite observations in the sample.

In statistical tests, the same sample is used to estimate the parameters of the distribution and perform the test and thus there is multiple use of observations. The concept of degrees of freedom is used to overcome this limitation. The term *degree of freedom* is defined as the number of observations in a sample less the number of parameters being estimated. The tables of t-distribution list the value of t statistics corresponding to various levels of significance and the degrees of freedom (ν).

Table 4.4 contains values of t-distribution for selected degrees of freedom and significance level. The value of the t-distribution can be read from the row corresponding to ν and the column corresponding to the significance level α. For example, for $\nu = 10$ the t value is 1.81 for 5% significance level. This implies that 95% of the area of the curve lies to the left of value 1.81. Since the t-distribution is symmetric, 5% of the area in the left tail is to the left of t value of -1.81 for $\nu = 10$. For the case where one is interested in 95% of the area but with 2.5% in each tail, the critical t value for $\alpha = 2.5$ is 2.23 for $\nu = 10$.

Table 4.4 Values of t-distribution for selected degrees of freedom and significance level

Degrees of freedom ν	Significance level α (%)			
	10	5	2.5	1
1	3.08	6.31	12.71	31.82
2	1.89	2.92	4.30	6.96
3	1.64	2.35	3.18	4.54
4	1.53	2.13	2.78	3.75
5	1.48	2.02	2.57	3.36
8	1.40	1.86	2.31	2.90
10	1.37	1.81	2.23	2.76
15	1.34	1.75	2.13	2.60
20	1.32	1.72	2.09	2.53
30	1.31	1.70	2.04	2.46
60	1.30	1.67	2.00	2.39
120	1.29	1.66	1.98	2.36
∞	1.282	1.645	1.96	2.32

Example 4.8: Consider a test of hypothesis about sample mean. Assume that the annual runoff data at a site follows normal distribution and 21 observations for this data are available. For these observations, the mean x_m is 15.0 mm and the standard deviation s_d is 5.0 mm. The question now is whether, at 5% significance level, the mean annual runoff can be considered to be drawn from a population whose mean is 17.0 mm.

Solution: To test this, the null hypothesis is $H_0 : \mu = 17.0$ mm against the alternate hypothesis $H_a : \mu \neq 17.0$ mm.

The test statistic is
$t = [n (x_m - \mu)/s_d]^{0.5} = [21*(15.0 - 17.0)/5.0]^{0.5} = -1.83.$

Since this is a two-tailed test, the null hypothesis will be rejected if the test statistic is either too high or too low and hence value of t is needed for $\alpha/2$ and $\nu = n - 1$. From Table 4.4, $t_{\alpha/2,n-1} = t_{2.5,20} = 2.09$. Since $|t| = 1.83$ which is less than 2.09, the statistic does not fall in the region of rejection and the null hypothesis is accepted.

4.6.2 Chi-Square Distribution

Another distribution that is frequently used in hypothesis testing is the Chi square distribution. Let there be a sample of size n and values are taken from a normal population having a mean μ and standard deviation σ. The observations can be standardized using the relation

$$Z = (X - \mu)/\sigma \tag{4.97}$$

If the standardised values are squared and added they follow a new statistic:

$$Y = \sum_{i=1}^{n} Z_i^2 \tag{4.98}$$

The variable Y follows a chi-square (χ^2) distribution with n degrees of freedom. The chi-square distribution is a special case of the gamma distribution. Similar to the t-distribution, this distribution also has a single parameter. However this distribution is not symmetric and is always positive. The chi-square tests are single-tailed and the region of rejection is near the right tail. Table 4.5 lists chi-square values for selected degrees of freedom. For example, for 10 degrees of freedom, 5% of the area in the right tail (region of rejection) from χ^2 values is from 18.31 to ∞.

The goodness of fit test determines whether it is appropriate to use a particular distribution for the given sample data. Visual judgment is one way in which the data are plotted on an appropriate probability paper to check whether the match is acceptable or not. The chi-square test is also widely used for this purpose. The test procedure consists of dividing the sample into a number of segments or classes depending upon the data range. For each segment, the actual number of observations and the expected according to the distribution under test are computed. The test statistic is

$$\chi_c^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i \tag{4.99}$$

where O_i and E_i are the observed and expected number of observations in the i^{th} segment and k is the total number of segments. If p parameters are estimated from data, χ^2_c follows a chi-square distribution with $(k-p-1)$ degrees of freedom. If the difference between the actual and expected observations in the segments is large, it implies that samples were not drawn from the assumed distribution. Therefore, the null hypothesis that the observations follow the assumed distribution is rejected if $\chi^2_c > \chi^2_{1-\alpha,k-p-1}$.

Table 4.5 Values of Chi-square distribution for selected degrees of freedom and significance level

Degrees of freedom ν	Significance level α (%)			
	10	5	2.5	1
1	2.71	3.84	5.02	6.63
2	4.60	5.99	7.38	9.21
3	6.35	7.82	9.35	11.34
4	7.78	9.49	11.14	13.28
5	9.24	11.07	12.83	15.09
8	12.02	14.07	16.01	18.48
10	15.99	18.31	20.48	23.21
15	22.31	25.00	27.49	30.58
20	28.41	31.41	34.17	37.57
30	40.26	43.77	46.98	50.89
40	51.81	55.76	59.34	63.69
50	63.17	67.50	71.42	76.15
100	118.50	124.34	129.56	135.81

Example 4.9: In a goodness of fit test, the data were divided in 10 classes and the value of χ^2_c came out to be 10.44. If two parameters of the distribution were computed, test whether the chosen distribution is appropriate for the data at a significance level of 0.1?

Solution: The degree of freedom is $10 - 2 - 1 = 7$. From the table of chi-square values, $\chi^2_{10,7} = 12.02$. Since this value is greater than 10.44, the null hypothesis cannot be rejected. It is, therefore, concluded that the chosen distribution properly describes the behavior of data.

4.7 LINEAR REGRESSION

It is an approach which is widely used to describe linear cause and effect relations between two variables. The objective is to predict a dependent variable based on an independent variable. The linear regression equation is:

$$y_i = a + bx_i + \epsilon_i \quad i = 1, 2, \ldots n \tag{4.100}$$

where, y_i is the i^{th} value of the dependent or regressed variable, x_i is the i^{th} value of the independent or regressor variable. The regression line crosses the y-axis at a point a (the

intercept), and has a slope b, and $\in_i$ is the random error term for the i^{th} data point. The variables involved in regression should be chosen carefully and there should be a logical reason behind this choice. A scatter plot of y vs. x should be made to ascertain the dependence structure. Sometimes, a transformation of x, such as a power or log transformation, improves the regression relation.

4.7.1 Parameter Estimation

The regression coefficients (a and b) are estimated by minimizing the sum of squares of deviations of y_i from the regression line. For a point x_i, the corresponding $\hat{y}_i$ given by the regression equation will be:

$$\hat{y}_i = a + bx_i \tag{4.101}$$

The residual error at this point is $e_i = y_i - \hat{y}_i$. It provides a measure of how well the least-squares line conforms to the raw data. If the line passes exactly through each sample point, the error e_i would be zero. The sum of square of errors is:

$$S_{se} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.102}$$

Minimizing S_{se} leads to the following values of parameters:

$$b = S_{xy}/S_{xx}$$

and

$$a = \bar{y} - b\,\bar{x} \tag{4.103}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4.104}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

4.7.2 Goodness of Regression

The goodness of regression is measured by the variability of the dependent variable that is explained by the regression relation. Referring to Fig. 4.9, one can write,

$$\begin{aligned} y_i &= \bar{y} + (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \\ y_i &= y_i - \bar{y} - \hat{y}_i + \bar{y} + \hat{y}_i \\ or\ \ y_i - \hat{y}_i &= (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) \end{aligned} \tag{4.105}$$

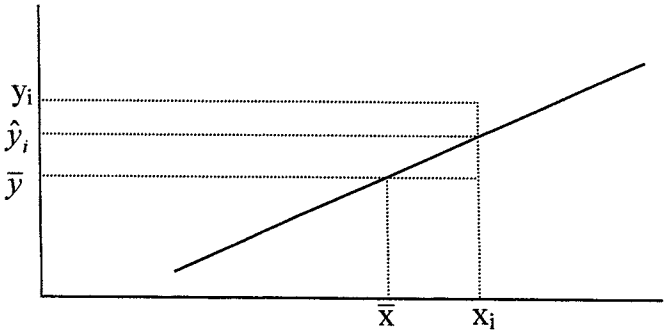

Fig. 4.9 Graphical representation of regression related variables.

Squaring both sides and summing up

$$\sum(y_i - \hat{y}_i)^2 = \sum(y_i - \bar{y})^2 - 2\sum(y_i - \bar{y})(\hat{y}_i - \bar{y}) + \sum(\hat{y}_i - \bar{y})^2$$

Since $\hat{y}_i = a + bx_i \qquad (\hat{y}_i - \bar{y}) = b(x_i - \bar{x})$,

Further, from eq. (4.105), $b\sum(x_i - \bar{x})^2 = \sum(x_i - \bar{x})(y_i - \bar{y})$

Therefore, substituting and simplifying,

$$\sum(y_i - \hat{y}_i)^2 = S_{se} = \sum(y_i - \bar{y})^2 - \sum(\hat{y}_i - \bar{y})^2 \qquad (4.106)$$

However, $\sum(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$

and $S_{sr} = \sum(\hat{y}_i - \bar{y})^2$

Therefore, $S_{se} = \sum y_i^2 - n\bar{y}^2 - S_{sr}$

or

$$\sum y_i^2 = n\bar{y}^2 + S_{se} + S_{sr} \qquad (4.107)$$

Thus, the total sum of squares Σy_i^2 consists of three components. The first is the sum of squares due to mean, the second is the sum of squares of errors, and the third is the sum of squares due to regression S_{sr}. The quantity S_{yy} is also termed as sum of squares about the mean or sum of squares corrected for mean.

Some important indicators of the goodness of regression are:

Mean square error (mse):	$s^2 = S_{se}/(n-2)$	(4.108a)
Standard error of regression	$s = (mse)^{0.5}$	(4.108b)
Correlation coefficient	$r = S_{xy}/(S_{xx}\, S_{yy})^{0.5}$	(4.108c)
Coefficient of determination	$R^2 = 1 - S_{se} / S_{yy}$	(4.108d)

The coefficient of determination represents the fraction of variance that is explained by regression. The closer this ratio is to unity, the 'better' is the regression relation.

4.7.3 Inferences on Regression Coefficients

The variances of coefficients a and b are needed to determine their confidence bands. From eq. (4.103),

$$b = S_{xy}/S_{xx}$$
$$= \sum (x_i - \bar{x})(y_i - \bar{y}) / S_{xx} = \sum y_i (x_i - \bar{x}) / S_{xx}$$

So,

$$\text{var}(b) = \sum (x_i - \bar{x})^2 \, \text{var}(y_i) / S_{xx}^2 = S_{xx} s^2 / S_{xx}^2$$

Hence, $$\sigma_b^2 = s^2 / S_{xx}$$

or $$\sigma_b = s/\sqrt{S_{xx}}$$

Thus, the standard error of b, $S_b = s/\sqrt{S_{xx}}$.

The variance of coefficient a

$$\text{var}(a) = \text{var}(\bar{y} - b\bar{x}) = \text{var}(\bar{y}) - \bar{x}^2 \, \text{var}(b)$$
$$= s^2 / n + s^2 \bar{x} / S_{xx}$$

So, the standard error of a

$$S_a = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \tag{4.109}$$

Test of hypothesis concerning *a* and *b*

Hypothesis H_0: $a = a_0$ versus H_a : $a \neq a_0$ is tested by computing $t = (a - a_0)/S_a$ which has a t distribution with $n-2$ degrees of freedom. H_0 is rejected if $|t| > t_{(1-\alpha/2),\,(n-2)}$.

Hypothesis H_0: $b = b_0$ versus H_b : $b \neq b_0$ is tested by computing $t = (b - b_0)/S_b$. H_0 is rejected if $|t| > t_{(1-\alpha/2),(n-2)}$.

Hypothesis H_0: $b = 0$ is tested by computing $t = (b - 0)/S_b$. H_0 is rejected if $|t| > t_{(1-\alpha/2),\,(n-2)}$ and in this case the regression equation is able to explain a significant amount of variation in y.

4.7.4 Confidence Intervals

The confidence interval at the α% significance level indicates that in repeated applications of the technique, the frequency with which the confidence interval would

contain the true parameter value is (100 - α)%. A typical value of α is 0.05 which corresponds to (1-0.05)*100% = 95% confidence limits. These intervals are defined if the true relationship between the variables is linear and the residuals e_i are independent, normally distributed random variables with constant variance. If the model is correct, then a/S_a and b/S_b should follow t distribution with $(n-2)$ degrees of freedom. Hence, for coefficient a, the lower and upper limits are:

$$(l_a, u_a) = \{a - t_{(1-\alpha/2),(n-2)}\, S_a,\ a + t_{(1-\alpha/2),(n-2)}\, S_a\} \tag{4.110}$$

For coefficient b, the lower and upper limits are:

$$(l_b, u_b) = \{b - t_{(1-\alpha/2),(n-2)}\, S_b\ ,\ b + t_{(1-\alpha/2),(n-2)}\, S_b\} \tag{4.111}$$

where $t_{(1-\alpha/2),\,(n-2)}$ represents Student's t values corresponding to the probability of exceedance $\alpha/2$ and $(n-2)$ degrees of freedom.

Confidence Intervals on Regression Line

These depend on the variance of $\hat{\bar{y}}_k$ which is the predicted mean value of $\hat{y}_k$ for a given x_k:

$$\hat{\bar{y}}_k = a + bx_k \tag{4.112}$$

Then,

$$\begin{aligned} \text{var}(\hat{\bar{y}}_k) &= \text{var}(a) + x_k^2\, \text{var}(b) + 2x_k\, \text{cov}(a,b) \\ &= s^2\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}\right] \end{aligned} \tag{4.113}$$

Hence, the standard error of $\hat{\bar{y}}_k$ would be

$$S_{\hat{\bar{y}}_k} = s\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}\right]^{1/2} \tag{4.114}$$

So, the lower and upper confidence limits on the regression line are:

$$(L,U) = [\hat{\bar{y}}_k - S_{\hat{\bar{y}}_k} t_{(1-\alpha/2),(n-2)},\ \hat{\bar{y}}_k + S_{\hat{\bar{y}}_k} t_{(1-\alpha/2),(n-2)}] \tag{4.115}$$

Confidence intervals on an individual predicted value of y are:

$$S'_{\hat{y}_k} = S\left[1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]^{1/2} \tag{4.116}$$

Example 4.10: The precipitation and runoff for a catchment for the month of July are given below in Table 4.6. (a) Develop the rainfall-runoff relationship in the form: $y = a$

+ *bx*; where *y* represents runoff and *x* represents precipitation. (b) What percent of the variation in runoff is accounted for by the developed regression equation?

Table 4.6 Precipitation runoff data and calculations.

SN	Year	Precipitation (x)	Runoff (y)	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^*$ $(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	1953	42.39	13.26	-0.55	-1.37	0.7535	0.3025	1.8769
2	1954	33.48	3.31	-9.46	-11.32	107.0872	89.4916	128.1424
3	1955	47.67	15.17	4.73	0.54	2.5542	22.3729	0.2916
4	1956	50.24	15.50	7.3	0.87	6.3510	53.2900	0.7569
5	1957	43.28	14.22	0.34	-0.41	-0.1394	0.1156	0.1681
6	1958	52.60	21.20	9.66	6.57	63.4662	93.3156	43.1649
7	1959	31.06	7.70	-11.88	-6.93	82.3284	141.1344	48.0249
8	1960	50.02	17.64	7.08	3.01	21.3108	50.1264	9.0601
9	1961	47.08	22.91	4.14	8.28	34.2792	17.1396	68.5584
10	1962	47.08	18.89	4.14	4.26	17.6364	17.1396	18.1476
11	1963	40.89	12.82	-2.05	-1.81	3.7105	4.2025	3.2761
12	1964	37.31	11.58	-5.63	-3.05	17.1715	31.6969	9.3025
13	1965	37.15	15.17	-5.79	0.54	-3.1266	33.5241	0.2916
14	1966	40.38	10.40	-2.56	-4.23	10.8288	6.5536	17.8929
15	1967	45.39	18.02	2.45	3.39	8.3055	6.0025	11.4921
16	1968	41.03	16.25	-1.91	1.62	-3.0942	3.6481	2.6244
	Total	687.05	234.04	0.01	-0.04	369.4230	570.0559	363.0714

Solution: (a) The various variables required to calculate *a* and *b* are computed in the Table 4.6. Here, $\bar{x}$= 687.05/16 = 42.94, $\bar{y}$= 234.04/16= 14.63. The regression coefficients are:

$b = S_{xy}/S_{xx} = 369.423/570.0559 = 0.648$

and = $a = \bar{y} - b\,\bar{x}$ $14.63 - 0.648*42.94 = -13.195$

Hence, the regression equation is: y = -13.195 + 0.648 x.

(b) The percent of variation in y that is accounted for by the regression is computed as the coefficient of determination (r^2) multiplied by 100. The value of S_{se} has been computed in Table 4.7.

Coefficient of determination $R^2 = 1 - S_{se} / S_{yy} = 1 - 123.668/363.0714 = 0.659$.

Thus, 66 percent of variation in *y* is explained by the regression equation. The remaining 34 percent variation is due to unexplained causes.

The coefficient of correlation (r) = square root of coefficient of determination

$= \sqrt{0.66} = 0.81$.

Table 4.7 Regression computations.

SN	Year	Precipitation (x)	Runoff (y)	$\hat{y}$	$S_{se}=(y-\hat{y})^2$
1	1953	42.39	13.26	14.2737	1.0276
2	1954	33.48	3.31	8.5000	26.9365
3	1955	47.67	15.17	17.6952	6.3764
4	1956	50.24	15.50	19.3605	14.9036
5	1957	43.28	14.22	14.8504	0.3975
6	1958	52.60	21.20	20.8898	0.0962
7	1959	31.06	7.70	6.9319	0.5900
8	1960	50.02	17.64	19.2180	2.4900
9	1961	47.08	22.91	17.3128	31.3282
10	1962	47.08	18.89	17.3128	2.4874
11	1963	40.89	12.82	13.3017	0.2321
12	1964	37.31	11.58	10.9819	0.3577
13	1965	37.15	15.17	10.8782	18.4195
14	1966	40.38	10.40	12.9712	6.6113
15	1967	45.39	18.02	16.2177	3.2482
16	1968	41.03	16.25	13.3924	8.1656
	Total	687.05	234.04	234.0884	123.6680

Example 4.11: Using the data of Example 4.9, (a) Compute the 95% confidence interval on a and b and test the hypothesis that $a = 0$ and the hypothesis that $b = 0.500$ for the above regression; (b) Calculate the 95% confidence limits for the regression line. Calculate the 95% confidence interval for an individual predicted value of y.

Solution: (a) Computation of 95% confidence intervals on a and b.

Mean square error (Table 4.7): $mse = S_{se}/(n-2) = 123.668/14 = 8.83$.
Standard error of regression: $s_r = mse^{0.5} = 8.83^{0.5} = 2.97$. This is a very useful indicator of the quality of regression relationship.

Standard error of b (S_b) = $s_r/\sqrt{(S_{xx})} = 2.97/\sqrt{570.0559} = 0.125$.

Standard error of a (S_a) = $$s_r\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} = 2.97\sqrt{\frac{1}{16}+\frac{42.94^2}{570.0559}} = 5.39$$

From the t-table, for $\alpha = 0.05$, $n-2 = 14$, $t_{(1-0.025),(16-2)} = t_{0.975,14} = 2.14$.

So, 95% confidence intervals on a:

$$(l_a, u_a) = \{a - t_{(1-\alpha/2),(n-2)} \cdot S_a,\ a + t_{(1-\alpha/2),(n-2)} \cdot S_a\}$$
$$= (-13.1951 - 2.14 \times 5.39, -13.1951 + 2.14 \times 5.39)$$
$$= (-24.73, -1.66).$$

Similarly, 95% confidence intervals on b:

$$(l_b, u_b) = \{b - t_{(1-\alpha/2),(n-2)} \cdot S_b,\ b + t_{(1-a/2),(n-2)} \cdot S_b\}$$
$$= (0.648 - 2.14 \times 0.125,\ 0.648 + 2.14 \times 0.125)$$
$$= (0.38, 0.92).$$

(ii) Testing the hypothesis $H_0 : a = 0$ versus $a \neq 0$.

Here, $t = (a - 0.00)/S_a = -13.1951/5.39 = -2.44$. Since $t_{(1-\alpha/2),(n-2)} = t_{0.975,14} = 2.14$, $|t| > t_{0.995,14}$. Hence, the hypothesis $H_0 : a = 0$ is rejected.

Hypothesis $H_0 : b = 0.5$ versus $H_a : b \neq 0.5$.

In this case, $t = (b-0.5)/S_b = (0.648-0.50)/0.125 = 1.184$. Since $|t| < 1.184$, the hypothesis H_0 cannot be rejected.

From the above tests, it is concluded that the intercept is significantly different from zero. However, the slope is not significantly different from 0.5. The significance of the overall regression can be evaluated by testing H_0: $b = 0$. Under this hypothesis,

$$t = \frac{b - 0.00}{S_b} = \frac{0.648}{0.125} = 5.184$$

Since $|t| > t_{0.975,14}$, we reject H_0. The regression equation is able to explain a significant amount of the variation in Y.

4.7.5 Extrapolation

An extrapolation of a regression equation beyond the range of X used in estimating a and b is discouraged for two reasons. First, the confidence intervals on the regression line become very wide as the distance from $\bar{X}$ is increased. Second, the relation between Y and X may be non-linear over the entire range of X and only approximately linear for the range of X investigated. An example of this is shown in Fig. 4.10.

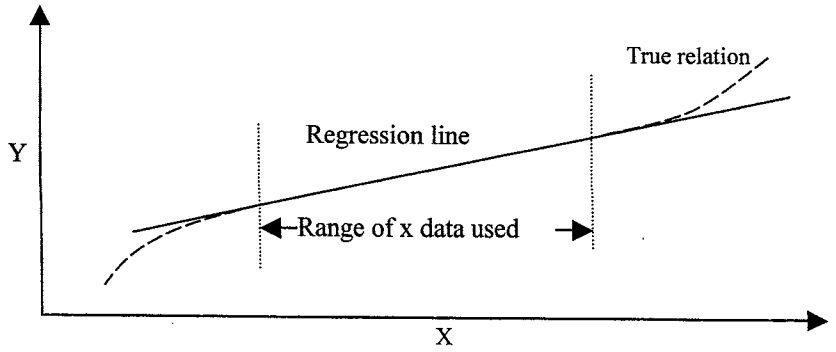

Fig. 4.10 Regression line and extrapolation.

4.8 MULTIPLE LINEAR REGRESSION

The association of three or more variables can be investigated by multiple linear regression and correlation analysis. If all the variables (dependent and independent) are in linear form, the regression is referred to as the multiple linear regression. Often a nonlinear association between the variables is handled by transforming the variables to linear form and applying multiple regression as it is easier to treat linear equations. The general form of the multiple linear regression equation is:

$$y_i = b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p} \tag{4.117}$$

When set of n observations is available on the dependent and each of the p independent variables, there shall be n equations for p unknowns.

$$\begin{aligned} y_1 &= b_1 x_{1,1} + b_2 x_{1,2} + \ldots + b_p x_{1,p} \\ y_2 &= b_1 x_{2,1} + b_2 x_{2,2} + \ldots + b_p x_{2,p} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= b_1 x_{n,1} + b_2 x_{n,2} + \ldots + b_p x_{n,p} \end{aligned} \tag{4.118}$$

where y_i is the i^{th} observation of the dependent variable, and $x_{i,1}$, $x_{i,2}, \ldots x_{i,p}$ are independent variables.

If a regression equation with p parameters is fitted to a set of n data points of variables, the number of degrees of freedom will be n-p. If the number of parameters is equal to the sample size the regression equation will pass through all the points as there is no degree of freedom. It cannot be used for prediction as the errors of parameters are inversely proportional to the number of degrees of freedom.

In designing the multiple linear and nonlinear regression relations, the selection of dependent and independent variables is very important. The dependent variable is defined by the problem itself. The independent variables are selected due to the following two reasons:

(i) They have been observed in the past concurrently with the dependent variable so that the regression equation may be established. In future, they may be used to predict the dependent variable.

(ii) An analysis of physical phenomenon indicates a cause-and-effect relation between dependent and independent variables.

These criteria are necessarily subjective. The variables that are known to have no effect on the dependent variable are neglected. In matrix notation, eq. (4.118) can be written as:

$$\underset{n*1}{Y} = \underset{n*p}{X}\,\underset{p*1}{B} \tag{4.119}$$

4.8.1 Estimation of Regression Coefficients

The difference between the observed and predicted (using regression) value of y or the error = $y_i - \hat{y}_i$. The regression coefficients are obtained by minimizing the sum of squares of errors using the equation:

$$\underset{(px1)}{b} = \underset{(pxn)(nxp)}{(X'X)^{-1}} \underset{(pxn)}{X'} \underset{(nx1)}{Y} \tag{4.120}$$

where X' is transpose of matrix X.

Coefficient of Determination (R^2)

$$\text{Let } Z_{i,j} = (X_{i,j} - \bar{X}_j)/S_j \tag{4.121}$$

where $\bar{X}_j$ and S_j are the mean and standard deviation of the j^{th} independent variable. The correlation matrix is:

$$R = Z'\,Z/(n-1) = [R_{i,j}] \tag{4.122}$$

where $R_{i,j}$ is the correlation between the i^{th} and j^{th} independent variables. R is a symmetric matrix since $R_{i,j} = R_{j,i}$. The coefficient of determination is defined as

R^2 = Sum of squares due to regression / Sum of squares about mean

or $$R^2 = (b'X'Y - n\bar{Y}^2)/(Y'Y - n\bar{Y}^2) \tag{4.123}$$

Here b' is the transpose of vector b of size $(1xp)$, and Y' is the transpose of vector Y of size $(1xn)$. Let the residual error be $\varepsilon = Y - Xb$. The variance of the error, var(ε), is

$$Var(\varepsilon) = S^2 = \Sigma(y_i - \hat{y}_i)^2/(n-p) \tag{4.124}$$

4.8.2 Inferences on Regression Coefficients

(i) Standard errors of b_i $(S_{\hat{b}_i})$

Let $C = X'X$ and C_{ii}^{-1} is the i^{th} diagonal element of $(X'X)^{-1}$. Then

$$Var(\hat{b}_i) = S_{\hat{b}_i}^2 = C_{ii}^{-1} S^2 \tag{4.125}$$

$$S_{\hat{b}_i} = \left| \sqrt{C_{ii}^{-1} S^2} \right| \tag{4.126}$$

(ii) Confidence intervals on b_i

If the model is correct, the quantity $\hat{b}_i / S_{\hat{b}_i}$ follows a t-distribution with $(n-p)$ degrees of freedom. The confidence intervals on b_i are given as

$$(L_{\hat{b}_i}, U_{\hat{b}_i}) = (\hat{b}_i - t_{(1-\alpha/2),(n-p)} S_{\hat{b}_i}, \hat{b}_i + t_{(1-\alpha/2),(n-p)} S_{\hat{b}_i}) \tag{4.127}$$

(iii) Test of hypothesis concerning b_i

The hypothesis that the i^{th} variable is not contributing significantly to explaining the variation in the dependent variable is equivalent to testing the hypothesis $H_o : b_i = b_o$ versus $H_a : b_i \neq b_o$. The test is conducted by computing:

$$t = (\hat{b}_i - b_o) / S_{\hat{b}_i} \tag{4.128}$$

H_o is rejected if $|t| > t_{(1-\alpha/2),\,(n-p)}$. If this hypothesis is accepted, it is advisable to delete the concerned variable from the model.

Significance of the overall regression

The hypothesis $H_o : b_1 = b_2 = \ldots b_p = 0$ versus H_a : at least one of these b's is not zero is used to test whether the regression equation is able to explain a significant amount of variation of Y or not. The ratio of the mean square error due to regression to the residual mean square has an F distribution with $p-1$ and $n-p$ degrees of freedom. Hence, the hypothesis is tested by computing the test statistic:

$$F = \frac{(b'X'Y - n\bar{Y}^2)/(p-1)}{(Y'Y - \hat{b}'X'Y)/(n-p)} \tag{4.129}$$

H_o is rejected if F exceeds the critical value $F_{(1-\alpha),\,(p-1),\,(n-p)}$.

Confidence Intervals on Regression Line:

To put the confidence limits on $Y_k = X_k b$, it is necessary to estimate the variance of $\hat{y}_k$. This is given by

$$S^2_{\hat{Y}_k} = S^2 X_k (X'X)^{-1} X'_k \tag{4.130}$$

where

$$(L,U) = \{\hat{Y}_k - t_{(1-\alpha/2),(n-p)} S_{\hat{Y}_k}, \hat{Y}_k + t_{(1-\alpha/2),(n-p)} S_{\hat{Y}_k}\}$$

Confidence Intervals on Individual Predicted Value of Y

$$\hat{Y}_K = X_K \hat{b} \tag{4.131}$$

$$(L',U') = \{\hat{Y}_k - t_{(1-\alpha/2),(n-p)} S'_{\hat{Y}_k}, \hat{Y}_k + t_{(1-\alpha/2),(n-p)} S'_{\hat{Y}_k}\} \tag{4.132}$$

$$S'^2_{\hat{Y}_k} = S^2[1 + X_k(X'X)^{-1} X'_k]$$

4.8.3 Stepwise Regression

The most common procedure for selecting the best regression equation is stepwise regression. The regression equation is built one variable at a time, adding at each step

the variable that explains the largest amount of the remaining unexplained variation. After each step, all the variables in the equation are examined for significance and any variable that no longer explains a significant amount of variation is discarded. The steps of stepwise regression are:

(i) The variable which has the highest simple correlation with the dependent variable is picked up as the first independent variable.
(ii) The variable which explains the largest of the residual variation in the dependent variable after the first step is added as the next variable.
(iii) Test the significance of the new variable and retain or discard depending on the results of this test.
(iv) Repeat steps (ii) & (iii) until all of the variables not in the equation are found to be insignificant and all the variables in the equation are significant.

Care must be exercised to see that the resulting equation is rational. Of course, the real test of the regression model is its ability to predict those observations of the dependent variable that were not used in estimating the regression coefficients. Transformation of independent variables may significantly improve the regression relationship.

4.9 CORRELATION ANALYSIS

Correlation is a mathematical measure of the strength of relationship between two variables or within the same series. The measure of the relationship is a dimensionless coefficient, called correlation coefficient. However, note that correlation is not an evidence of a causal relationship between two variables. If one variable drives the other, they may be correlated, as rainfall and runoff. The variables may also be correlated if they share the same cause. Examples include dependent variables, such as river discharge, concentration or transport rates of sediment, and concentration or transport rates of substances that are transported in association with suspended sediment (Hirsch et al. 1993).

4.9.1 Cross-Correlation

The correlation between two time-series or cross-correlation $r_{x,y}$ is given by

$$r_{x,y} = s_{x,y}/s_x\, s_y \tag{4.133}$$

where $s_{x,y}$ is the sample covariance between X and Y and s_x and s_y are the sample standard deviations of X and Y, respectively. As with autocorrelation, $r_{x,y}$ can also range from –1 to 1; $r_{xy} = +1$ or -1 implies a perfect linear relationship between X and Y and $r_{x,y} = 0$ implies linear independence, although there may be other types (say, non-linear) of dependence. If observations in a time-series are correlated, this must be kept in mind while drawing any inferences about the data or when modeling the process that has produced the time-series. For example, the correlation for the precipitation-runoff data of Table 4.6 is 0.812 which indicates a high correlation between these two variables.

A high correlation between two variables need not necessarily be due to a cause-and-effect relation between them. If there is a correlation between some variables that do not have a cause and effect relation, this is termed as spurious correlation. The monthly flows of two adjacent streams may be highly correlated but this could be because the influencing external causes are the same. A high correlation in this case does not mean that a change in flow of one stream will force the other stream's flow to change. It is to be noted that independent variables are uncorrelated but uncorrelated variables are not necessarily independent. The dependence in correlated variables is a stochastic dependence and not always physical or cause-and-effect dependence. Any apparent correlation between variables that are in fact uncorrelated is termed as spurious correlation.

4.9.2 Serial or Auto-Correlation

The autocorrelation or serial correlation of a series is defined as linear correlation between a time series and the same series at a later interval of time. Assume that in a time-series, observations are equally spaced in time and that the statistical properties of the process do not change with time. The autocorrelation of a time series (having n observations) at lag k (r_k) is given by

$$r_k = \frac{[\sum_{i=1}^{n-k} x_i x_{i+k} - \sum_{i=1}^{n-k} x_i \sum_{i=1}^{n-k} x_{i+k}]/(n-k)}{[\sum_{i=1}^{n-k} x_i^2 - (\sum_{i=1}^{n-k} x_i)^2/(n-k)]^{0.5}[\sum_{i=1}^{n-k} x_{i+k}^2 - (\sum_{i=1}^{n-k} x_{i+k})^2/(n-k)]^{0.5}} \quad (4.134)$$

Here the lag is the amount of offset when comparing the values of the series. The autocorrelation of lag 1 is determined by computing the correlation between elements 1, 2, ..., *(n-1)* of a series and the elements 2, 3, 4, ... n of the same series. From eq. (4.133), it is clear that r_0 is unity. Note also that as k increases, the number of pairs of observations used in estimating r_k decreases since all of the summations contain $n - k$ terms. Therefore, serial correlation should only be estimated for k sufficiently smaller than n; usually correlation at lags exceeding 20 are not much useful.

A purely random process will have $r_k = 0$ for all k, indicating that all of the observations in the sample are independent of each other. The elements of a sample of data possessing serial correlation are not random elements. The plot of autocorrelations at various lags is known as a correlogram. A typical correlogram begins at a value of +1.0 at 0 lag, and then decays at higher lags. At lags of near coincidence of the elements, the correlogram shows a rise; it falls otherwise. Correlograms help reveal the characteristics of a time-series and disclose intervals of time or distance at which the time series has a repetitive nature.

The annual flows of Sabarmati River at Dharoi have been plotted in Fig. 4.3. The correlogram of this series up to a lag of 20 is plotted in Fig. 4.11. It can be seen from the correlogram that there is very poor auto-correlation in the series. The auto-correlation at lag 1 is 0.0064 and at lag 2, it is –0.0295.

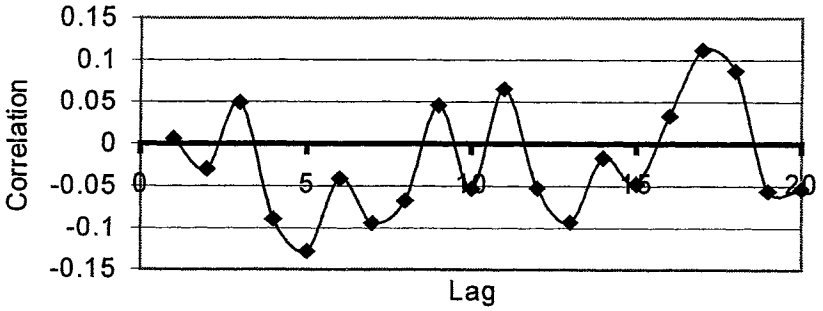

Fig. 4.11 Correlogram of annual flows of Sabarmati River.

4.9.3 Inferences on Correlation Coefficients

For uncorrleated variables, $r_{x,y} = 0$ and for correlated ones, $r_{x,y} \neq 0$. Even in uncorrelated populations, the sample correlation coefficient will rarely be zero. This deviation from zero is likely to be due to chance. Thus, statistical tests are needed to determine if the deviation of the sample correlation coefficient from zero may be ascribed to chance or not. This test can be conducted by making a hypothesis H_o: $r_{x,y} = 0$ or H_o: $r_{x,y} = r^*$ where r^* is known.

Assume that X and Y are random variables from a bivariate normal distribution. The population correlation coefficient is denoted by ρ. If $\rho = 0$, then the quantity

$$t = r[(n - 2)/(1 - r^2)]^{0.5} \tag{4.135}$$

has a t-distribution with $(n-2)$ degrees of freedom. The hypothesis H_o: $\rho = 0$ is rejected if $|t| > t_{1-\alpha/2,n-2}$.

Kendall's Rank Correlation Test

Also known as Kendall's τ test, this is an effective and general measure of correlation between two variables. Since it is a rank-based procedure, it overcomes the problems due to the effect of extreme values and to deviations from a linear relationship. Thus, it is well-suited to use with dependent variables for which the variation around the general relationship exhibits a high degree of skewness or kurtosis. Examples include dependent variables such as river discharge, and concentration or transport rates of sediment.

The steps to conduct Kendall's test for correlation (the null hypothesis H_o is that the distribution of dependent variable y does not change as a function of independent variable x) are as follows:

1. The n data pairs (x_1, y_1), (x_2, y_2), ... (x_n, y_n) are indexed according to the magnitude of the x value, such that $x_1 \leq x_2 \leq \ldots \leq x_n$.

2. Compute the statistic S

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sign(y_j - y_k) \tag{4.136}$$

where

sign (θ) = 1 if θ > 0
= 0 if θ = 0
= -1 if θ < 0.

3. For $n > 10$, the test is conducted using a normal approximation (Hirsch et al., 1993). The standardized test statistics Z is computed as:

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{Var(S)}} & S > 0 \\ 0 & S = 0 \\ \dfrac{S+1}{\sqrt{Var(S)}} & S < 0 \end{cases} \tag{4.137}$$

where Var (S) = n(n – 1) (2n + 5)/18.

4. The null hypothesis is rejected at a significance level α if $|Z| > Z_{(1-\alpha/2)}$, where $Z_{(1-\alpha/2)}$ is the value of the standard normal distribution with a probability of exceedance of α/2. If some of the x and/or y values are tied, the formula for Var (S) is modified. If the sample size is less than 10, then it is necessary to use tables for the S statistic.

Example 4.12: The correlation between the precipitation and runoff data of Table 4.6 is 0.812. Test the null hypothesis H_o: the distribution of runoff does not change as a function of precipitation.

Solution: For this data, 16 pairs of values are available. These are arranged in ascending order. Using eq. (4.136), the S statistic is found to be 72.

Hence, Var(S) = 16*(16-1)*(2*16+5)/18 = 493.33.
Now, applying eq. (4.137)

$$Z = (72 - 1)/\sqrt{493.33} = 3.197.$$

Since |z| > 1.96, the null hypothesis is rejected at 5% significance level.

4.10 FREQUENCY ANALYSIS

Frequency analysis is performed to determine the frequency of the likely occurrence of hydrologic events. This information is required to solve a variety of water-resource problems, for example, design of reservoirs, floodways, bridges, culverts, levees, urban drainage systems, irrigation systems, stream-control works, water-supply systems, and hydroelectric power plants, floodplain zoning, setting of flood-insurance premiums, etc.

Although the frequency analysis of virtually every component of the hydrologic cycle is required, the emphasis here will be on frequency analyses of streamflow extremes and rainfall only.

The hydrologic data to be analyzed for frequency analysis must be treated in light of the objectives of the analysis, length and completeness of record, randomness of data, and homogeneity. The length of record should be more than 25 years for the derived distribution to be acceptable. The hydrologic data must have been controlled by a uniform set of hydrologic and operational factors. For example, the factors causing a winter rainflood are quite different from those during a spring snowmelt flood or a local cloudburst flood. These two types of floods should not be combined into a single record. Sometimes a hydrologic record may have gaps. Missing data may sometimes be estimated using regional analysis or by correlation with other hydrologic data in the region.

Hydrologic data are generally presented in chronological order constituting the complete duration series (CDS). For frequency analysis, CDS is seldom used because the hydrologic design of a project is normally dictated by only a few critical events. Therefore, hydrologic data can be selected in two ways: (1) partial duration series (PDS) and (2) annual duration series (ADS). PDS is comprised of the data exceeding a specified base level. In ADS, one value (usually the highest) is selected from each year. The two series are comparable if the record is longer than 10 years and either can be used.

4.10.1 Point Frequency Analysis

The frequency distributions presented earlier can be fitted to the data. Two commonly used methods of fitting are: (1) the graphical method and (2) frequency factors.

Graphical method

This method involves fitting of an assumed probability distribution to observed data. The sample data are arranged in either ascending or descending order of magnitude and each data point is assigned a rank. If these are arranged in descending order of magnitude, then the highest value will be assigned the rank of 1, the second highest the rank of 2, and so on; the lowest value will have the rank of N, where N is the number of data points in the sample. This arrangement gives an estimate of the exceedance probability, that is, the probability of a value being equal to or greater than the ranked value. If the values are arranged in ascending order, then an estimate of the non-exceedance probability, that is, the probability of a value being less than or equal to the ranked value, is obtained. These data points are plotted on probability paper, with their positions determined from a plotting-position formula.

Many plotting-position formulas are available; some commonly used ones are given in Table 4.8. Adamowski (1981) has shown that all of these formulas are special cases of

$$P_m = \frac{m-a}{N+b} \tag{4.138}$$

where a and b are constants, P_m is the exceedance probability of the m^{th} observation, and m is the m^{th} value of N ordered observations, such as $P_1 < P_2 < ... < P_N$. A commonly used plotting-position formula in hydrology is

$$P_m = \frac{m}{N+1} \tag{4.139}$$

Clearly, the return period of the m^{th} data point, T_m, is

$$T_m = (N + 1)/m \tag{4.140}$$

Table 4.8 Some commonly used plotting-position formulas

Method	Formula $P_m(X > X_m)$	Values for m = 1 and N = 10 P_m	T_m
Hazen (1914)	(m – 0.5)/ N	0.05	20.0
California (1923)	(m/N)	0.10	10.0
Weibull (1939)	(m)/(N + 1)	0.091	11.0
Beard (1943)	(m – 0.31)/(N + 0.38)	0.066	15.0
Chegodayev (1955)	(m – 0.3)/(N + 0.4)	0.067	14.9
Blom (1958)	(m – 0.375)/(N + 0.25)	0.0609	16.4
Gringorten (1963)	(m – 0.44)/(N + 0.12)	0.055	18.1
Cunnane (1978)	(m – 0.4)/(N + 0.2)	0.58	17.2
Adamowski (1981)	(m – 0.25)/(N + 0.5)	0.071	14.1

The observed values and their exceedance probabilities are plotted on the probability paper corresponding to the assumed probability distribution. On the ordinate of the graph paper are observed values and on the abscissa the probabilities or return periods. The objective of using the probability paper is to linearize the distribution so that plotted data can be represented by a straight line. A best-fit straight line is then drawn through the plotted points. The line is assumed to give the probabilities of all values beyond the observed range.

4.10.2 Frequency-Factor Method

Chow (1951) proposed the use of a frequency factor in hydrologic frequency analysis. If a hydrologic variable X is plotted chronologically in time, then a particular value x is found to be composed of two parts: namely, the mean, $\bar{x}$, and the departure from the mean Δx:

$$x = \bar{x} + \Delta x \tag{4.141}$$

The variable Δx can be expressed as the product of the standard deviation S and the frequency factor K. Therefore,

$$x = \bar{x} + S\,K \tag{4.142}$$

where K depends on the return period T and the PDF of X; K literally means the number of standard deviations above and below the mean to achieve the desired quantile. For a distribution, a relation between K and T can be derived. For two-parameter distributions, K varies with T. For skewed distributions, it varies with the coefficient of skewness (C_s) and is very sensitive to the length of record. The frequency factor for some commonly used distributions is given here.

Normal distribution: Recall the definition of the standard normal variate, $Z = (X - \mu)/\sigma$, where μ the population mean, and σ is population standard deviation of the variable X. Its observed values are expressed as

$$z = (x - \bar{x})/S \tag{4.143}$$

or

$$x = \bar{x} + S\,z \tag{4.144}$$

Thus, for the normal distribution, K is the standard normal variate, which can be obtained from the tables of standard normal distribution.

Log-normal distribution: If a variable follows log-normal distribution, its logarithms will follow normal distribution and then the formula for normal distribution can be applied.

Gumbel distribution: If the reduced variate is Y, the frequency factor for this distribution is

$$K = (y - 0.577)/1.283 \tag{4.145}$$

where

$$y = 1.283(x - \bar{x})/S + 0.577 \tag{4.146}$$

Log-Pearson Type 3 Distribution: For the Log-Pearson Type 3 Distribution, K is a function of both the return period and C_s. Values of K for log-transformed data are given by the Water Resources Council (1967). To fit this distribution, transform the data, x_i (annual floods), to their logarithmic values, y_i. Compute the mean, standard deviation S_y, and C_s for the log values. Get the value of K for the desired T from the tabulated values. If C_s falls between –1 and +1, an approximate value of K is obtained from

$$K = \frac{2}{C_s}\left[\left\{\left(z - \frac{C_s}{6}\right)\frac{C_s}{6} + 1\right\}^3 - 1\right] \tag{4.147}$$

Compute y from $y = \bar{y} + KS_y$. Then compute $x = \exp(y)$ for the desired T value.

Example 4.13: The mean annual flood of a river is 32,100 cumec and the standard deviation of flood peaks is 6000 cumec. What is the probability of a flood of magnitude of 40,000 cumec occurring in the river within the next 5 years?

Solution: Given $\bar{x}$ = 32,100 cumec, S = 6000 cumec, and x = 40,000 cumec. Assume that the peaks follow the Gumbel distribution. Hence,

$$y = 1.282(40000 - 32100 + 0.45*6000) = 3.28.$$
$$F(y) = \exp[-\exp(-3.28)] = 0.97$$
$$P = 1 - F(y) = 1 - 0.97 = 0.03$$

The return period of the flood event is 1/0.03 = 33.33. The probability of a 40,000 cumec flood occurring in the next 5 years = $1 - 0.97^5 = 1 - 0.859 = 0.141$.

4.10.3 Confidence Limits

A value of the variate estimated from a probability distribution for a given return period is usually in error due to the limited sample size. Therefore, a statement indicating the limits about the estimated value within which the true value is contained with a specific probability is needed. This statement is made by constructing confidence limits, which are also called the confidence intervals, confidence bands, error limits, or control curves. The confidence interval indicates the limits about the estimated value and the probability with which the true value will lie between those limits. This statement accounts for the sampling errors only.

Let the confidence probability be α. The confidence interval of the variate x corresponding to a return period T is bounded by values x_1 and x_2 (Nemec, 1973) as

$$x_{1,2} = x \pm G(\alpha)\, S_e \tag{4.148}$$

where $G(\alpha)$ is a function of the confidence probability α and can be determined by using the table of normal variates. As an example,

α (%)	50	68	80	90	95	99
$G(\alpha)$	0.674	1.00	1.282	1.645	1.96	2.58

S_e is the probable error expressed as

$$S_e = (1 + 1.3K + 1.1K^2)^{0.5} \frac{S_{N-1}}{\sqrt{N}} \tag{4.149}$$

in which K is the frequency factor of the distribution under consideration, S_N is the standard deviation of the sample, and N is the sample size. By using this method, confidence limits can be placed above and below the fitted distribution curve. If the Gumbel distribution is considered, then for a given sample and T, 80% confidence

limits are about twice as big as 50% ones, and 95% confidence limits are about thrice as big as 50% limits.

4.10.4 Regional Frequency Analysis

For many watersheds, streamflow data are either insufficient or non-existent at the sites of interest. The methods of frequency analysis using data from a single site will have then limited predictive value because of large sampling errors. To overcome the data deficiency, a regional frequency analysis is performed. By defining a region that is hydrologically similar in terms of the variable to be studied, data from several gaging sites within this homogeneous region are pooled together into a single regional frequency analysis. Examples of regional frequency analysis are estimation of design flood from rainfall-runoff relationship, prediction of flood peaks from the relation between observed values and drainage-basin characteristics, and estimation of rainfall depths and frequencies in ungaged areas from characteristics at well-gaged sites in the same area.

The first step in a regional analysis is to define the region itself. The definition of a region depends on the quantities to be estimated. Many methods are available to define a region that is homogeneous. For mean annual precipitation, large physiographic regions can be used, whereas for peak flow, the regions may be confined to drainage basins of certain sizes. Regional boundaries can be defined in terms of similarity of flood-frequency curves or flow curves. Homogeneity tests are used to check if flood-frequency curves in a region can be considered homogeneous.

4.10.5 Index Flood Method

The index-flood (IF) method, developed by the U.S. Geological Survey (Dalrymple, 1960; Benson, 1962), is widely used to perform regional flood-frequency analysis. The basic premise of this method is that a combination of streamflow records maintained at a number of gaging stations will produce a more reliable, not a longer, record, and thus will increase the reliability of frequency analysis within a region. There are two major parts of the IF method. The first is the development of basic dimensionless frequency curves representing the ratio of the flood of any frequency to an index flood (the mean annual flood). The second is the development of relations between geomorphologic characteristics of drainage areas and the mean annual flood by which to predict the mean annual flood at any point within the region. By combining the mean annual flood with the basic frequency curve, a regional frequency curve is produced.

Basic Frequency Curve

In large regions that are homogeneous with respect to flood-producing characteristics, individual streams whose drainage areas vastly differ in size have frequency curves of approximately equal slope if the discharge is expressed as a ratio of the mean. The flood peaks at each gaging station are divided by an index flood (which is often taken as the mean annual flood at the station) and are thus reduced to dimensionless ratios.

The individual curves plotted using the flood ratios can be superimposed and will nearly coincide.

These curves will pass through the recurrence interval of 2.33 years at the ratio of 1.0, but may have different slopes. The variation of these slopes may be used to test the homogeneity of the region. Within the homogeneous region, the ratios are compiled for all stations and then the median ratio for each is obtained. By plotting the median ratios against recurrence interval, a regional frequency curve is obtained. This is supposedly the best representation of a flood-frequency relation obtained by combining all dimensionless curves.

Homogeneity Test

The homogeneity test was developed by Langbein (Dalrymple, 1960) and can be used to test if a region is homogeneous to permit combining individual frequency curves with confidence to form a regional curve. The question is: Do the records differ from one another by amounts attributable to chance? The answer to this question lies in computing these differences and setting limits that will be acceptable statistically.

The standard error of estimate S_e of the reduced variate y for the Gumbel distribution (EV1) can be written as

$$S_e = \exp(y) \{ 1 / \{(T - 1) N]\}^{0.5} \tag{4.150}$$

where T is the recurrence interval, and N is the number of years of record. If a normal distribution of the estimates is assumed, then 95% of the estimates of the T-year flood will lie within $2S_e$ of their most probable value of T. The test employs the 10-year flood because this is the longest recurrence interval for which most records will give dependable estimates. For $T = 10$ years,

$$2S_e = 0.666 \exp (y)/N^{0.5} \tag{4.151}$$

and y for the EV1 distribution is 2.25. Therefore, the confidence limits are specified by

$$2.25 \pm 6.33/ N^{0.5}$$

Dalrymple (1960) listed values of y corresponding to T, and lower and upper confidence limits with the corresponding T for various values of N.

Mean Annual Flood (MAF)

The mean annual flood is defined as the value of the graphical frequency curve at the recurrence interval of 2.33 years. Benson (1962) confirmed experimentally that MAF has a magnitude equivalent to the flood of a 2.33-year recurrence interval. It is dependent on many factors that can be classified as either physiographic or meteorologic. The physiographic factors influencing MAF at a given point are: (1)

drainage area and shape, (2) stream slopes, (3) natural storage in lakes, swamps, or channels, (4) land slope, (5) land use, (6) geology, (7) stream density, (8) stream pattern, (9) elevation, (10) aspect, (11) orographic position, (12) basin relief, and (13) soil cover.

The meterologic factors are connected with the magnitude and distribution of precipitation received by a drainage area, and include: (1) storm types, (2) type of region, whether humid or arid, (3) storm pattern, (4) storm direction, (5) precipitation intensities, (6) snowmelt, (7) storm volume, and (8) extent of ice jams. The evaluation, treatment, and use of some of these factors are difficult. The most commonly used factor is the mean annual precipitation.

Sometimes factors, representing the composite effect of the above factors, have been used in flood frequency studies. The mean annual runoff reflects the effect of precipitation and basin characteristics. Another factor is the lag time, which is the time difference between the centers of rainfall and runoff. It represents the composite effect of all topographic factors.

Median Flood Ratios

The ratios of several floods of different recurrence intervals to the MAF are tabulated for each station. Enough recurrence intervals are selected to define the curve appropriately. By tabulating the flood ratios, the median ratio for each recurrence interval is computed. This median of a recurrence interval is the mid value of its flood ratios if their number is odd or the mean of the two central ratios if their number is even.

Regional Frequency Curve

Each median flood ratio is plotted against its recurrence interval on the probability paper. An average frequency curve is then drawn. This is the regional frequency curve, showing flood discharge in a ratio to MAF, is based on all significant discharge records available, and represents the most likely flood-frequency values for all areas in the region.

Some weaknesses of the IF method are:

1. The flood ratios for comparable streams may differ due to large differences in IF. If IF is not typical and is obtained from a short period of record, the remainder of the frequency curve may be faulty.
2. The homogeneity test cannot be applied at a level much higher than that of the 10-year flood because many individual records are too short to adequately define the frequency curve at higher levels. In many cases, individual curves show wide and sometimes systematic differences at higher levels.
3. Within a flood frequency region, frequency curves are combined for all sizes of drainage areas, excluding the largest. Recent studies using ratios of less frequency

floods have shown in all cases that the ratio of any specified flood to MAF varies inversely with the drainage area. In general, the larger the drainage area, the flatter the frequency curve. The effect of drainage area is relatively greater for floods of higher recurrence intervals.

4.10.6 Multiple-Regression Method

The relation of flood peaks of selected recurrence intervals to basin and climatic parameters is determined by multiple-regression methods. The resulting relation is of the form

$$Q_T = b\prod_{i=1}^{N} x_i^{a_i} \tag{4.152}$$

where Q_T denotes the T-year flood; b is the regression constant; x_i, i = 1, 2, ..., N, are independent basin and climatic parameters; a_i, i = 1, 2, ..., N, are regression coefficients. Many studies have used models similar to eq. (4.152) to estimate the flood magnitude of recurrence interval T. The basin and climatic characteristics are normally evaluated from topographic, geologic, and climatic maps. Although the basin and climatic factors to be used in regression analysis vary from one region to another, the most important factors are drainage area, main-channel slope, and mean annual precipitation. Many of these factors are interrelated. For example, in general, as drainage area increases, slope and rainfall intensity decreases.

Some of the advantages of using a multiple-regression analysis are: (1) It provides a mathematical relation between Q of a specified value and T the independent variables. (2) It provides an evaluation of the independent variables that best define the dependent variable. (3) It provides a measure of the accuracy of the equation in terms of the standard error of estimate, and tests the significance of the coefficients of each independent variable. (4) It evaluates the relative significance of each independent variable by indicating those variables that have a coefficient that is significantly different from zero at a particular percent confidence level. (5) It provides an easy evaluation of the coefficients when the dependent and independent variables are transformed to their logarithms and used in a linear regression.

4.11 TIME SERIES ANALYSIS

A ***time series*** is a set of observations generated sequentially in time. If the set is continuous, the series is said to be *continuous*; if it is discrete, the time series is said to be *discrete*. Here only discrete time series where observations are made at some fixed interval h will be considered. The observations in a discrete series made at equidistant time intervals $\tau_0 + h$, $\tau_0 + 2h$, ... $\tau_0 + th$, ...$\tau_0 + Nh$ may be denoted by $z(\tau_1)$, $z(\tau_2)$, ...$z(\tau_t)$, ...$z(\tau_N)$. For many purposes, the value of τ_0 and h are unimportant; these are needed if the observation times are to be defined exactly. Of course, the information content of a time-series is affected by the choice of h, particularly if the series has rapid changes. If τ_0 is adopted as the origin and h as the unit of time, z_t can be regard as the observation at time t.

A discrete time series may arise in two ways:

(1) By sampling a continuous time series; for example, the continuous river flow at a station may be sampled at hourly intervals.

(2) By accumulating a variable over a period of time; for example, rainfall may be accumulated over a period of a day.

A hydrologic time series can be divided in two basic groups: 1) Univariate (single) time series, e.g., monthly streamflow at a point, and 2) multivariate (multiple) series of different kinds at one point. The examples of the second type are series of flow and water quality variables at a station. If a time-series, e.g., daily precipitation, is composed of nonzero and zero values, it is known as intermittent series. A time series whose values have been observed at regular intervals, such as each day or each hour, is termed as regularly spaced time series.

Time series analysis is useful for many applications, such as forecasting, detecting trends in records, filling-in missing data, and generation of synthetic data. The analysis of a time-series is a subject in itself and only a brief introduction of it is given in the following.

Components of a Time series

A time series can be divided into a number of components. The main components of a hydrologic time series are: trends and other deterministic changes, cycles or periodic changes and autocorrelation, the almost periodic changes, such as tidal effects, and components representing stochastic or random variations. Trend and jumps are introduced in a time series due to gradual or sudden changes in the major factors of the process that is responsible for the time series. For example, a runoff time series will have a trend as a result of major land use changes in the upstream catchment; a water quality time series may show trends if a new factory upstream begins to discharge its effluent in the river. The closure of a diversion dam will lead to a jump in the series because the flow will be reduced due to diversion. The turning point test and Kendall's rank correlation tests are commonly employed to test randomness and trend in a series.

The time series of water resource variables which are measured or accumulated at sub-annual time intervals normally have periodic patterns. Such patterns can be seen in time series, for example, monthly rainfall, daily runoff, daily volumes of urban water demands, and these series are said to have seasonal or periodic patterns. Due to seasonality, the statistical properties of time series vary with time (week or month, etc.). In harmonic analysis, the periodic component of a time-series is represented using a series of sine functions. For a series without trend, the harmonic equation is:

$$z_t = \mu + \sum_{i=1}^{L} \lambda_i \sin(\frac{2\pi ti}{T} + \phi_i) + \varepsilon_i \tag{4.153}$$

where μ is the population mean, λ_i and ϕ_i are the amplitudes and phases of the wave, and i/T is the frequency of the wave.

4.11.1 Stationary Stochastic Processes

A special class of stochastic processes, called *stationary process*, is based on the assumption that the process is in a particular state of statistical equilibrium. A stochastic process is said to be strictly stationary if its properties are unaffected by a change of time origin; that is, if the joint probability distribution associated with m observations $z_{t1}, z_{t2} \ldots z_{tm}$, made at any set of times $t_1, t_2, \ldots t_m$, is the same as that associated with m observations $z_{t1+k}, z_{t2+k}, \ldots, z_{tm+k}$, made at time $t_1 + k, t_2 + k, \ldots t_m + k$. Thus, for a discrete process to be strictly stationary, the joint distribution of any set of observations must be unaffected by shifting all the times of observation forward or backward by any integer amount k. The statistical properties of a non-stationary time-series are time dependent.

Assuming that the stationarity assumption holds true, the joint probability distribution $p(z_{t1}, z_{t2})$ is the same for all times t_1, t_2, which are a constant interval apart. Therefore, the nature of the joint distribution can be inferred by plotting a scatter diagram using pairs of values (z_t, z_{t+k}), of the time series, separated by constant interval or lag k. The covariance between z_t and z_{t+k} is called the autocovariance at lag k and is calculated by

$$\gamma_k = \text{cov}[z_t, z_{t+k}] = E[(z_t - \mu)(z_{t+k} - \mu)] \tag{4.154}$$

For a stationary process, the variance at time $(t + k)$ is the same as at time t. The estimate of the k^{th} lag autocovariance γ_k is

$$c_k = \frac{1}{N}\sum_{t=1}^{N-k}(z_t - \bar{z})(z_{t+k} - \bar{z}),\ k = 0, 1, 2 \ldots, K \tag{4.155}$$

The estimate of lag k autocorrelation is obtained by

$$r_k = c_k/c_0 \tag{4.156}$$

which implies that $r_0 = 1$.

A common cause of autocorrelation or dependence in many hydrologic time-series is the storage effect. In case of river flow series, for example, this storage might be at the catchment surface, in unsaturated zone, or in ground water zone.

4.11.2 Time Series Models

A mathematical model representing a time series or stochastic process is called a *time series model.* The model has a certain structure and a set of parameters. The important categories of time-series models are as follows.

Autoregressive (AR) Models: AR models are extremely useful to represent certain practical series. Let the values of a process at equally spaced times t, $t-1$, $t-2$, ... be Y_t, Y_{t-1}, Y_{t-2}, ... and let z_t, z_{t-1}, z_{t-2} ... be the deviations from the mean μ; for example, $z_t = Y_t$

- μ. In an AR model, the current value of the process is expressed as a finite, linear aggregate of previous values of the process and a shock a_t. Thus

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t \tag{4.157}$$

is called an autoregressive process of order p and is denoted by AR(p). Introducing a *backward shift operator* B, defined by $Bz_t = z_{t-1}$, eq. (4.157) can be written as:

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) z_t = a_t \tag{4.158}$$

or $$\phi(B) z_t = a_t \tag{4.159}$$

Here $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ is termed as an autoregressive operator of order p. The AR models have been extensively used in water resources because this form has an intuitive type of time dependence and the AR models are simple to use.

Moving Average (MA) Models: Another kind of model of great practical importance in the representation of observed time series is the finite moving average process. Here z_t is linearly dependent on a finite number q of previous a's. Thus,

$$z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \tag{4.160}$$

is called a moving average (MA) process of order q and is denoted by MA(q). Similar to the autoregressive operator, a moving average operator of order q can be written as

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \tag{4.161}$$

and the MA(q) process can be written as

$$z_t = \theta(B) a_t \tag{4.162}$$

Autoregressive-moving Average Models: Greater flexibility in fitting time series models is achieved by including both autoregressive and moving average terms in the model. This leads to the mixed autoregressive-moving average ARMA(p, q) model:

$$z_t = \phi_1 z_{t-1} + \ldots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \tag{4.163}$$

or $$z_t - \phi_1 z_{t-1} - \ldots - \phi_p z_{t-p} = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}$$

or $$\phi(B) z_t = \theta(B) a_t \tag{4.164}$$

which employs $p+q+2$ unknown parameters μ, ϕ_1, ..., ϕ_p, θ_1, ..., θ_q, σ_a^2, that are estimated from the data. The simplest member of ARMA(p,q) family is the ARMA(1,1) model which can be written as

$$z_t - \phi_1 z_{t-1} = a_t - \theta_1 a_{t-1} \tag{4.165}$$

The combination of AR and MA models makes it possible to simulate many hydrologic processes by using a small number of parameters. For example, the flow in a stream results due to a number of causes such as precipitation and groundwater effluence. This mixed behavior can be conveniently modelled by ARMA models. In practice, an adequate representation of actually occurring stationary time series can be frequently obtained with autoregressive, moving average, or mixed model, in which p and q are not greater than 2 and often less than 2. Note that ARMA(p,0) model is the same as AR(p) and ARMA(0,q) model is same as MA(q). In eq. (4.165), z_t and a_t may represent time dependent discharge (output) and rainfall (input).

The ARMA models are suitable for stationary hydrologic series. In case of nonstationary series, the periodic or seasonal fluctuation can be removed by taking the differences and the ARMA model can be applied to the resultant series. The resultant model is termed as Autoregressive Integrated Moving Average (ARIMA) model. Consider a time series that is homogeneous except in level, i.e., the various segments of the series look identical except, the difference in level about which it changes. Such a series can be adequately represented by a model of the form:

$$\phi(B)\nabla z_t = \theta(B)\, a_t \tag{4.166}$$

where ∇ is the *backward difference operator* defined as

$$\nabla z_t = z_t - z_{t-1} = (1 - B)z_{t-1} \tag{4.167}$$

Thus, ARIMA(p, d, q) is an ARMA model that is fitted to the data after taking the d^{th} difference of the series:

$$\phi(B)\nabla^d z_t = \theta(B)\, a_t \tag{4.168}$$

where ∇^d indicates that the series is differenced d times. The notation $\nabla_n = 1 - B^n$ indicates differencing with lag of n. The first order differencing [eq. (4.167)] is helpful in removing the trend of a series or non-stationarity in the mean. Two consecutive differencing operations are necessary to remove non-stationarity in the mean and slope. However, it may not always be possible to remove non-stationarity by differencing alone, other transformations may also be needed.

4.11.3 Partial Autocorrelation Function

The partial autocorrelation function is another way of representing the time dependence structure of a series. It is useful in identification of the type and order of the model of a given time series. Let ϕ_{kj} denote the j^{th} coefficient in an autoregressive process of order k; ϕ_{kk} being the last coefficient. Now, ϕ_{kk} satisfies the set of equations

$$\rho_j = \phi_{k1}\, \rho_{j-1} + \phi_{k(k-1)}\, \rho_{j-k+1} + \phi_{kk}\, \rho_{j-k}\,, \qquad j=1,2, \ldots k \tag{4.169}$$

leading to the set of equations known as the Yule-Walker equations. These are written as

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & & \rho_{k-2} \\ \vdots & & & & \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix} \tag{4.170}$$

or

$$P_k \phi_k = \rho_k \tag{4.171}$$

Solving these equations for $k = 1, 2, 3, \ldots$, successively, the values of ϕ_{11}, ϕ_{22} ... are obtained as a function of ρ. The quantity ϕ_{kk}, regarded as a function of the lag k, is called the *partial autocorrelation* function. For an AR(p) process, the partial autocorrelation function ϕ_{kk} will be nonzero for k less than or equal to p and zero for k greater than p. Stated in another way, the partial autocorrelation function of an AR(p) process has a cutoff after lag p. Conversely, the autocorrelation function of an MA(q) process has a cutoff after lag q, while its partial autocorrelation tails off. If both autocorrelations and partial autocorrelations tail off, a mixed process is suggested.

Partial autocorrelations may be estimated by successively fitting autoregressive processes of orders 1, 2, 3, ... by least squares and picking out the estimates ϕ_{11}, ϕ_{22} ... of the last coefficient fitted at each stage. Alternatively, approximate Yule-Walker estimates of the successive autoregressive processes may be employed. The estimated partial autocorrelations can then be obtained by substituting estimates r_j for the theoretical autocorrelations in eq. (4.169) to yield

$$r_j = \hat{\phi}_{k1} r_{j-1} + \hat{\phi}_{k2} r_{j-2} + \ldots + \hat{\phi}_{k(k-1)} r_{j-k+1} + \hat{\phi}_{kk} r_{j-k}, j = 1, 2, \ldots k \tag{4.172}$$

4.11.4 Fitting of ARMA Models

Initially, one does not know the order of the ARMA process for fitting to an observed time series as well as the order of differencing that is required, if any. Therefore, the model is built iteratively, i.e., a set of models is identified using the characteristics of the data and its adequacy is tested. Depending on the results, the model may be adopted or another candidate model is identified. While fitting a time series model, at least 50 observations should be used. If sufficient observations are not available, one proceeds by using experience to build a preliminary model. This model may be updated as more data becomes available. While applying ARMA models, the main stages are:

- *Model Identification* which involves the use of the data and any information on how the series was generated to identify a subclass of parsimonious models worthy to be considered.
- *Parameter Estimation* which involves an efficient use of the data to make inferences about parameters conditional on the adequacy of the considered model.

- *Diagnostic Checking* involves checking the fitted model in its relation to the data with the intent to reveal model inadequacies and to achieve model improvement.

In any time series modeling, it is worthwhile to first plot the data. A visual inspection of the data always gives useful information about the behaviour of the process. The autocorrelation function of the series also gives useful information. If the autocorrelation function fails to die out rapidly, it suggests that the series may be non-stationary and may require differencing to obtain a stationary series. Fig. 4.12 shows a plot of a periodic time series (Box and Jenkins, 1976, airlines data series). The rising trend of the series points to the need of differencing of the series. The correlogram of this time series is plotted in Fig. 4.13.

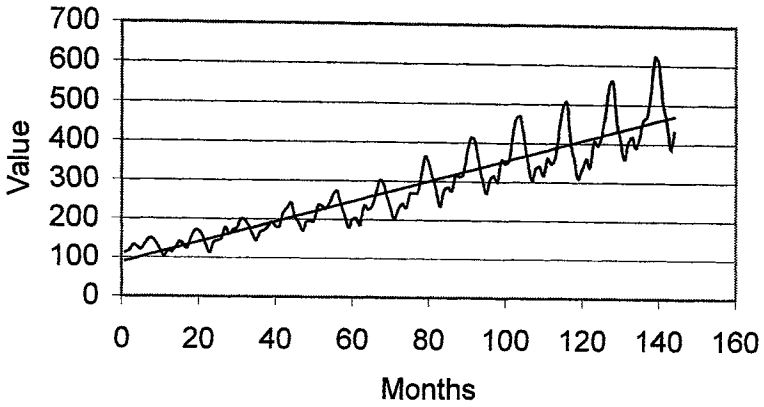

Fig. 4.12 Plot of a periodic time series.

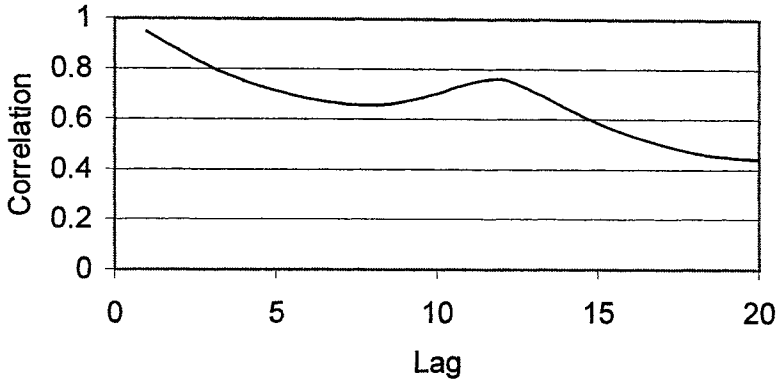

Fig. 4.13 Correlogram of periodic time series of Fig. 4.12.

Having tentatively decided the order of differencing d, the general appearance of the estimated autocorrelation and partial autocorrelation functions of the appropriately differenced series are studied. These provide clues about the choice of the orders p and q for the AR and MA operators. In doing so, the characteristic behaviour

of the theoretical autocorrelation function and of the theoretical partial autocorrelation function for AR, MA, and mixed processes is used. For example, if ϕ_1 of an AR(1) model is +ve, ρ_k decays exponentially to zero while for ϕ_1 -ve, it oscillates in sign. The autocorrelation function of AR(2) model has different forms, depending on the values parameters take on. The properties of autocorrelation and partial autocorrelation functions of a few model types are described in Table 4.8.

Table 4.8 Properties of autocorrelation and PAC functions of some time series models.

Model type	Autocorrelation function	Partial autocorrelation function
AR(1), first-order autoregressive	Decreases exponentially	$\phi_{1,1} \neq 0$ $\phi_{i,i} = 0$ for i = 2,3,4,...
AR(p), pth-order autoregressive	Mixed type of damping from lag 1	$\phi_{i,i} \neq 0$ for $i \leq p$ $\phi_{i,i} = 0$ for $i > p$
MA(1), first-order moving-average	$\rho_1 \neq 0$ $\rho_i = 0$ for i = 2,3,4,...	Decreases exponentially
MA(q), qth-order moving-average	$\rho_i = 0$ for $i > q$ $\rho_i \neq 0$ for $i \leq q$	Mixed type of damping from lag 1
ARMA(1, 1),auto-regressive moving-average	Decreases exponentially after lag 1	Decreases exponentially after lag 1
ARMA (p, q), auto-regressive moving-average	Mixed type of damping after lag q + 1	Mixed type of damping after lag p-q

To illustrate, the series in Fig. 4.12 was differenced at lag 1 and the differenced series is plotted in Fig. 4.14. The correlogram of the differenced series is plotted in Fig. 4.15. Although the differenced series has considerably small correlations, these are still quite high at lags of the multiple of 12 and this shows that further differencing is needed.

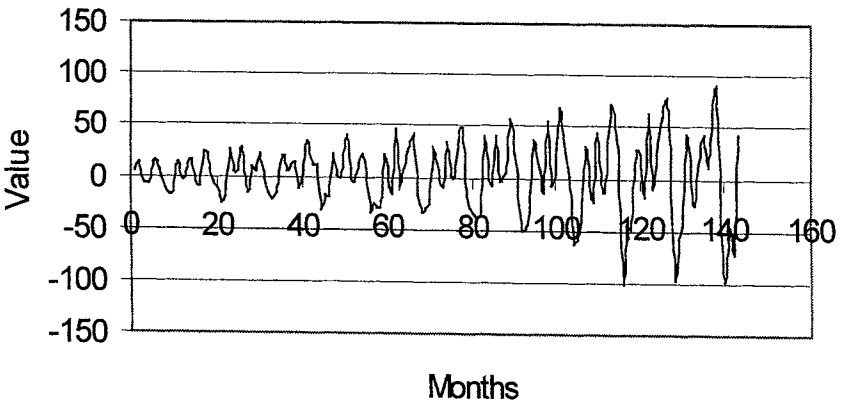

Fig. 4.14 Time series of Fig. 4.12 after differencing at lag 1.

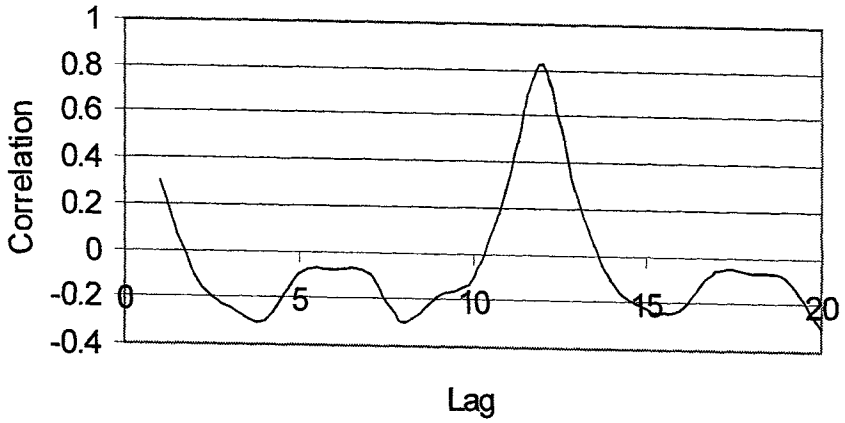

Fig. 4.15 Correlogram of the time series of Fig. 4.14.

For most monthly hydrological series, it is often helpful to first standardise the series by subtracting the mean and dividing by the standard deviation of the corresponding month. A first-order differencing of the resultant series is often adequate to yield a stationary series that can be modelled by the ARMA class of models.

The common techniques to estimate of the parameters of a time series model are the method of moments, the method of least squares, and the method of maximum likelihood. Mostly, the parameters obtained by the method of moments are used as the first approximation and are refined by other methods.

After an ARMA model has been fitted, it is necessary to apply statistical tests to check its adequacy and suitability. The tests which are used for this purpose include *the Porte Manteau Lack of Fit Test, the Akaike Information Criterion (AIC)*, and the test of correlogram. The statistic AIC is calculated by

$$\mathrm{AIC}(p,q) = N \ln\left(\hat{\sigma}_{\epsilon}^2\right) + 2(p+q) \tag{4.173}$$

where N is the sample size, and $\hat{\sigma}_{\epsilon}^2$ is the maximum likelihood estimate of the residual variance. The model which gives the minimum AIC is selected.

Examination of the residuals (difference between observed and computed values of a dependent variable) of a model is always helpful. The residuals of an adequate model should resemble white noise, the lag-one serial correlation should be close to zero, and they should have small variance. A test to check whether the residuals of a model are independent or not is the *Porte Manteau Lack of fit* test. In this, the Q statistic is determined by

$$Q = N\sum_{k=1}^{L} r_k^2 \tag{4.174}$$

where r_k denotes autocorrelation of residuals at lag k and L is the number of lags considered. Q approximately follows a chi-square distribution with (L-p-q) degrees of freedom.

The technique of overfitting, in which a more elaborate model is fitted to the data and then the results are compared, has also been recommended. Box and Jenkins (1976), Salas et al. (1980), and Salas (1993) are good references for time series models.

The water resources literature contains innumerable applications of time series modeling to a wide range of problems. The ARMA models are frequently used in rainfall runoff modelling. A number of well-known hydrologic models are special cases of the ARMA model. For example, the Muskingum model of flood routing is obtained by setting certain parameters of this equation to zero. The ARMA model parameters do have a physical interpretation. As the precipitation event occurs, the MA parameters influence the rising limb of the hydrograph and these parameters will, therefore, depend on the basin physiographic characteristics and its state. Similarly, the AR parameters will more heavily influence the recession limb of the hydrograph. These parameters can vary from storm to storm.

4.12 MARKOV MODELS

Many data sequences consist of a succession of mutually exclusive states, for example, annual streamflow data at a site. The future states of any stochastic process, e.g., the annual flow at a particular location, cannot be predicted with certainty. However, based on past data, it is possible to assess the probability of any particular outcome. Consider a stream in which the annual flow varies from 50 to 90 million m^3 and the interest lies in the nature of transitions from one state to another. To construct transition matrix, the data is classified into several mutually exclusive states. Let the data be classified into four classes of intervals of 10 million m^3 each: 50-60, 60-70, 70-80, and 80-90 million m^3, designated as A, B, C and D. Now it is possible to determine the percent of time the past annual flows were within each of the four intervals. Here it is assumed that the probability distribution of annual streamflows does not significantly change with time and past data provide a reasonable basis to define probability distributions. Denote p_i as the fraction of time the annual flow was in the interval i where i = 1, 2, 3, 4. Then,

$$p_i = [\text{Number of years when flow was in interval i}]/[\text{Total number of years of record}] \quad (4.175)$$

The following is an example of vector p of probabilities that may be obtained:

$$p = \{p_1, p_2, p_3, p_4\} = \{0.186, 0.304, 0.321, 0.189\} \quad (4.176)$$

The probabilities of eq. (4.176) are shown in a histogram in Fig. 4.16. Thus, the probability of an annual streamflow in a year falling in the range from 50 to 60 mcm is 0.186, from 60 to 70 mcm is 0.304, and so on. These streamflow intervals can be thought of as discrete states of the streamflow system. If all possible states are defined by these four intervals, then

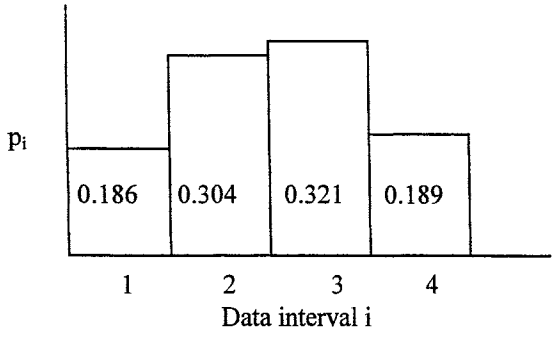

Fig. 4.16 Histogram of probabilities.

$$\sum_{i=1}^{4} p_i = 1 \tag{4.177}$$

The next question is the tendency for one state to succeed another. Since there are four states, a 4*4 matrix can be constructed, each box showing the number of times a given state is succeeded by another. If each element in the i^{th} row is divided by the total of the i^{th} row, each row in this matrix will sum to one. This matrix is called a transition probability matrix. Let P_{ij} be the conditional probability of the current annual streamflow being in interval j given that last year's streamflow was in interval i.

$$P_{ij} = \frac{\text{Number of events in interval j following those in interval i}}{\text{Total number of events in interval i}}$$

Then, $$\sum_{j=1}^{4} p_{ij} = 1 \text{ for all rows } i \tag{4.178}$$

A typical transition probability matrix is shown in Table 4.9.

Table 4.9 Matrix of streamflow transition probabilities.

		Streamflow state j in year $y+1$			
		A (50-60)	B (60-70)	C (70-80)	D (80-90)
Streamflow state i in year y	A	0.40	0.25	0.25	0.10
	B	0.20	0.30	0.40	0.10
	C	0.10	0.40	0.30	0.20
	D	0.10	0.20	0.30	0.40

The matrix tells, for example, that when state A occurs in a year, it is followed by state A 40% of the time, by B 25% of time, and so on. Note that this probability is independent of the total number of times that A occurs. Each row of the matrix is the conditional probability vector corresponding to an initial stream flow state. Each element P_{ij} in the matrix is the probability of a transition from streamflow i in one year to streamflow j in the next year. These conditional probabilities are called transition

probabilities. If the transition from one discrete value to the next can be described by a matrix of transition probabilities as that shown in Table 4.9, this stochastic streamflow process is called a discrete Markov process and the matrix is called a first order Markov chain. These probability estimates improve as more observations are available.

Stochastic processes in which the probability of a future state is dependent only on the present state and not on any of the past states are said to follow a first order (lag one) Markov chains. A useful property of Markov chains in describing ergodic processes, such as stremflows, is that there exists a stationary or steady-state probability distribution that is independent of the initial state. This implies that although next year's flow probabilities are dependent on this year's flow, the distribution of streamflows far in future, say 20 years from now, will be independent of this year's flow.

Coming back to the transition probability matrix of Table 4.9, assume that in year y the streamflow was 64 million m^3. Since this is in the interval between 60 and 70, the state of the streamflow in year y was 2 or B. The initial probability vector $p^{(y)}$ for year y is {0, 1, 0, 0}. Knowing $p^{(y)}$, it is easy to determine the probabilities $p_j^{(y+1)}$ of each of the four possible streamflow states j that can occur in year $(y+1)$. In Table 4.9, the probability vector or row for state corresponding to B is {0.2, 0.3, 0.4, 0.1}. This vector can be calculated by using the fact that in year $(y+1)$, the probability of being in state j is equal to the sum of the probabilities of being in each state i in year y times the probability of a transition from state i in year y to state j in year $(y+1)$. If $p_j^{(y+1)}$ represents the unconditional probability of being in state j in year $y+1$, this can be computed by:

$$p_j^{(y+1)} = p_1^{(y)} p_{1j} + p_2^{(y)} p_{2j} + p_3^{(y)} p_{3j} + p_4^{(y)} p_{4j} = \sum_i p_i^{(y)} p_{ij} \quad \forall j \tag{4.179}$$

where p_{ij} is the probability of a transition from state i to state j; here it is assumed to be independent of year y. Denoting the matrix of p_{ij}'s by P and the probability vector for year y by $p^{(y)}$, eq. (4.175) can be written in a compact way:

$$p^{(y+1)} = p^{(y)} P \tag{4.180}$$

Likewise, the probability of each streamflow state in year $(y+2)$ can be obtained knowing $p^{(y+1)}$:

$$p^{(y+2)} = p^{(y+1)} P \tag{4.181}$$

In a similar manner, it is easy to compute the probabilities of each possible streamflow state for year $(y+1)$, $(y+2)$, $(y+3)$ and so on. The probability vectors for the first 7 years are listed in Table 4.10.

Note that as y increases, the probabilities tend toward a limiting value, that of $(y+7)$ in this example. These are the unconditional steady state probabilities of having any one of the possible streamflow states. These are the same probabilities that are shown in the histogram of Fig. 4.16.

Table 4.10 Successive Streamflow Probability Vectors

Year y	Streamflow State Probabilities			
	p_1^y	p_2^y	p_3^y	p_4^y
Y	0.000	1.000	0.000	0.000
y+1	0.200	0.300	0.400	0.100
y+2	0.190	0.320	0.320	0.170
y+3	0.189	0.306	0.323	0.183
y+4	0.187	0.305	0.321	0.187
y+5	0.187	0.304	0.321	0.188
y+6	0.186	0.304	0.321	0.189
y+7	0.186	0.304	0.321	0.189

The annual streamflow correlation will not be as strong as assumed here. However, the shorter the time duration of streamflow data, e.g., monthly, weekly, and daily, the auto-correlation will increase.

4.13 CLOSURE

Statistical analysis of hydrological variables provides useful information about the nature of distribution that the data tend to follow. This can be used to predict the magnitude and associated frequency. Regression techniques are widely used to develop prediction equations for different hydrological variables. These prediction equations are useful to estimate the dependent hydrological variables, which are difficult to monitor, based on the data of independent variables which can be easily monitored.

4.14 REFERENCES

Adamowski, K. (1981). Plotting formula for flood frequency. Water Resources Bulletin, 17(2), 197-202.

Ashkar, F., Bobee, B., Leroux, D. and Morisette, D. (1988). The generalized method of moments as applied to the generalized gamma distribution. Stochastic Hydrology and Hydraulics, 2, 161-174.

Box, G.E.P., and Jenkins, G.M. (1976). Time Series Analysis-- Forecasting and Control. Holden Day Inc., San Francisco.

Bree, T. (1978a). The stability of parameter estimation in the linear model. Journal of Hydrology, Vol.37, pp.47-66.

Bree, T. (1978b). The general linear model with prior information. Journal of Hydrology, Vol.39. pp.113-127.

Chow, V.T. (1951). A general formula for hydrologic frequency analysis. Transactions, American Geophysical Union, 32:231-237.

Davis, J.C. (1986). Statistics and Data Analysis in Geology, John Wiley & Sons, New York.

Dooge, J.C.I. (1973). Linear theory of hydrologic systems. Technical Bulletin No. 1468. 327 pp., Agricultural Research Service, U.S. Department of Agriculture, Washington, D.C.

Douglas, J.R., Clarke and Newton, S.G. (1976). The use of likelihood functions to fit conceptual models with more than one dependent variable. Journal of Hydrology, Vol. 29, 181-198.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C. and Wallis, J.R. (1979). Probability-weighted moments: definition and relation to parameters to several distributions expressible in inverse form. Water Resources Research, 15, 1049-1054.

Gumbel, E.J. (1958). Statistics of Extremes. Columbia University Press, New York.

Haan, C.T. (2002). Statistical Methods in Hydrology. Iowa State Press, Ames, U.S.A.

Haktanir, T. (1996). Probability-weighted moments without plotting position formula. Journal of Hydrologic Engineering, 1(2), 89-91.

Harley, B.M. (1967). Linear routing in uniform open channels. M. Engg. Sc. Thesis, University College, Cork, Ireland.

Hirsch, R.M., Helsel, D.R., Cohn, T.A., and Gilroy, E.J. (1993). Statistical Analysis of Hydrologic Data, in Handbook of Hydrology, edited by D.R. Maidment. McGraw-Hill Inc., New York.

Hosking, J.R.M. (1986). The theory of probability-weighted moments. Technical Report RC 12210. Mathematics, IBM Thomas J. Watson Research Center, Yorktown Heights, New York.

Hosking, J.R.M. (1992). Moments or L-moments. An example comparing two measures of distributional shape. American Statistician, 46(3), 186-189.

Hosking, J.R.M. (1990). L-Moments: Analysis and estimation of distribution using linear combination of order statistics. J. of Royal Statistical Soc., Series B, 52(1), 105-124.

Jain, D. and Singh, V. P. (1986a). A comparison of transformation methods for flood frequency analysis. Water Resources Bulletin, 22(6), 903-912.

Jain, D. and Singh, V. P. (1986b). Comparing methods of transformation for flood frequency analysis. In Multivariate Analysis of Hydrologic Processes, edited by H. W. Shen, J. T. B. Obeysekara, V. Yevjevich and D. G. DeCoursey, pp. 755-767, Colorado State University, Fort Collins, Colorado.

Jones, L.E. (1971). Linearizing weight factors for least squares fitting. Journal of the Hydraulics Division, ASCE, Vol. 97, No. HY5, 665-675.

Kite, G.W. (1977). Frequency and Risk Analysis in Hydrology. Water Resources Publications, Colorado.

Kroll, C.N. and Stedinger, J.R. (1996). Estimation of moments and quantiles using censored data. Water Resources Research, 32(4), 1005-1012.

Kuczera, G. (1982a). Combining site-specific and regional information: An empirical Bayes approach. Water Resources Research, 18(2), 306-314.

Kuczera, G. (1982b). Robust flood frequency models. Water Resources Research, 18(2), 315-324.

Kuczera, G. (1982c). On the relationship between the reliability of parameter estimates and hydrologic time series data used in calibration. Water Resources Research, 18(1), 146-154.

Landwehr, J.M., Matalas, N.C. and Wallis, J.R. (1979a). Probability-weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. Water Resources Research, 15, 1055-1064.

McCuen, R.H. (1993). Statistical Hydrology. Prentice-Hall, New Jersey.

McGuess, R.H. and Snyder, W.M. (1985). Hydrologic Modelling, Statistical Methods and Applications. Prentice-Hall, Englewood Cliffs, New Jersey.

Nash, J.E. (1959). A systematic determination of unit hydrograph parameters. Journal of Geophysical Research, 64(1), 111-115.

Natale, L. and Todini, E. (1974). A constrained parameter estimation technique for linear models in hydrology. Publication No. 13. Institute of Hydraulics, University of Pavia, Pavia, Italy.

Phien, H.N. and Jivajirajah, T. (1984). Fitting the S_b curve by the method of maximum likelihood. Journal of Hydrology. Vol. 67, 67-75.

Rao, A.R. and Hamed, K.H., (1994. Frequency analysis of upper Cauvery flood data by L-moments. Water Resources Management. Vol.8, 183-201.

Rao, D.V. (1980). Log-Pearson type 3 distribution: method of mixed moments. Journal of Hydraulics Division. ASCE, 106(HY6), pp. 999-1019.

Rao, D.V., (1983). Estimating log Pearson parameters by mixed moments. Journal of Hydraulic Engineering, 109(8), 1118-1132.

Salas, J.D., Delleur, J.R., Yevjevich, Y., and Lane, W.I. (1980). Applied Modeling of Hydrologic Time Series. Water Resources Publications, Colorado.

Salas, J.D. (1993). Analysis and Modeling of Hydrologic Time Series, in Handbook of Hydrology, edited by D.R. Maidment. McGraw-Hill Inc., New York.

Shannon, C.E. (1948). A mathematical theory of communications, I and II. Bell System Technical Journal, 27, 379-443.

Shrader, M.L., Rawls, W.J., Snyder, W.M. and McCuen. R.H., (1981). Flood peak regionalization using mixed-mode estimation of the parameters of the log-normal distribution. Journal of Hydrology. Vol. 52, 229-237.

Singh, V.P. (1992). Elementary Hydrology. Prentice-Hall, Englewood Cliffs, NJ.

Singh, V.P. (1988). Hydrologic Systems, Vol. I: Rainfall-Runoff Modeling. Prentice-Hall, Englewood Cliffs, New Jersey.

Singh, V.P. (1998). Entropy-based Parameter Estimation in Hydrology, Water Science and Technology Library, Volume 30, Kluwer Academic Publishers, Dordrecht.

Snyder, W.M. (1972). Fitting of distribution functions by nonlinear least squares. Water Resources Research, 8(6), 1423-1432.

Sorooshian, S., Gupta, V.K. and Fulton, J.L. (1983). Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility. Water Resources Research, 19(1), 251-259.

Stedinger, J.R. and Tasker, G.D. (1985). Regional hydrologic analysis: 1. Ordinary, weighted and generalized least squares compared. Water Resources Research, 21(9), 1421-1432.

Sturges, H.A. (1926). The choice of a class interval. Journal of the American Statistical Association, 21, 65-66.

Wang, S.X. and Adams, B.J. (1984). Parameter estimation in flood frequency analysis. Publication 84-02, Department of Civil Engineering, University of Toronto, Toronto, Canada.

Wang, Q.J. (1997). LH moments for statistical analysis of extreme events. Water Resources Research. 33(12), 2841-2848.

Williams, B.J. and Yeh, W.W.G. (1983). Parameter estimation in rainfall-runoff models. Journal of Hydrology. Vol. 63, 373-393.

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