10.2 Key Statistical Measures of Data

Four principal features which characterize a set of observations on a random variable are:

(i) the central tendency or the value around which all other values are bunched,

(ii) the spread of the sample data around mean,

(iii) the asymmetry or skewness of the spread of data, and

(iv) the peakedness of the data.

These characteristics are expressed in terms of statistical properties which are estimated from the sample data.

10.2.1 Measures of Central Tendency

In statistics various measures of central tendency are employed. Three important measures are the following.

(i) Arithmetic Mean: If $x_1, x_2 \ldots x_n$ represent a series of observations, the mean of this series is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (10.15)$$

Where $\bar{x}$ represents the sample mean; the mean of population is generally denoted by μ.

(ii) Mode: It is the value which occurs most frequently. It is the peak value of the PDF. A data set may have more than one peak.

(iii) Median: It is the middle value of the ranked observations for a data set. The median divides the distribution in two equal parts.

10.2.2 Measure of Dispersion or Variation

Three statistical measures of variation of data are commonly used.

(i) Variance: It represents the scatter of the data are about the mean. Variance is computed by:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (10.16)$$

A small value of variance implies that values are bunching close to the mean.

(ii) Standard Deviation (SD): The unbiased estimate of population standard deviation (s) is given computed as the square root of the variance:

$$s = [\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2]^{0.5} \qquad (10.17)$$

when $n < 30$, the unbiased estimate of s is found by replacing n by $n-1$ in the denominator. Greek letter σ is used to denote the standard deviation of population.

(iii) Coefficient of Variation (CV) is a dimensionless parameter and is obtained by dividing the standard deviation by the mean:

$$C_V = s / \bar{x} \qquad (10.18)$$

When the mean of the data is zero, C_v is not defined. This coefficient is useful to compare different populations. Given two samples of data, the one with larger C_v will have more spread of the values around the mean.

Example 10.1: Average annual flows (in cumec) at a river gauging site are given in the table below. Compute the mean, variance, standard deviation, and the coefficient of variation of the flows.

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Flow (cumec)	195.5	145.4	148.1	324.7	205.6	302.9	210.3	194.4	71.2	126.8	216.0
Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Flow (cumec)	136.4	403.9	145.3	161.6	112.5	110.0	90.2	129.8	80.5	136.3	243.3

Solution: We have a total of 22 values. The mean of the flows can be computed as

Mean = (195.5 + 145.4 + 148.1 + …. 80.5 + 136.3 + 243.3)/22 = 176.8 cumec.

The variance can be computed by eq. (10.16)

Variance $s^2 = [(195.5 - 176.8)^2 + (145.4 - -176.8)^2 + (148.1-176.8)^2 + \ldots + (136.3-176.8)^2 + (243.3-176.8)^2)]/22 = 6926.46$ cumec^2

SD $s = (6926.46)0.5 = 83.22$ cumec

CV = 83.22/176.8 = 0.47.

10.2.3 Measures of Symmetry

Usually the hydrologic data are not distributed symmetrically around the mean. If the data to the

right of the mean are more spread out than those on the left then, by convention, the asymmetry is positive and vice versa for negative asymmetry (see figure 10.4).If the data are symmetrically placed around the mean then the measure of symmetry would be zero.

The third moment of the data about the mean is used in indicating symmetry and is given by:

$$M_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \qquad (10.19)$$

It is easy to see that this moment is zero if the data are symmetrical. Otherwise, M_3 will have certain value, a positive or negative.

Note that because the third central moment has dimensions equal to the cube of the data, it is not useful while comparing different data sets. Being non-dimensional the coefficient of skewness does not have this disadvantage and is preferred.

Coefficient of Skewness: A non-dimensional measure of the asymmetry of the distribution of the data is helpful when various data are to be compared and the coefficient skewness is one such measure. The coefficient of skewness (C_s) is given by:

$$C_S = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3} \qquad (10.20)$$

Symmetrical frequency distributions have very small or negligible value for skewness coefficient C_s, while asymmetrical frequency distributions have either positive or negative coefficients. When C_s has a small value, it indicates that the probability distribution may be approximated by the normal distribution since $C_s = 0$ for this distribution. The symmetrical and skewed distributions are shown in Fig. 10.4.

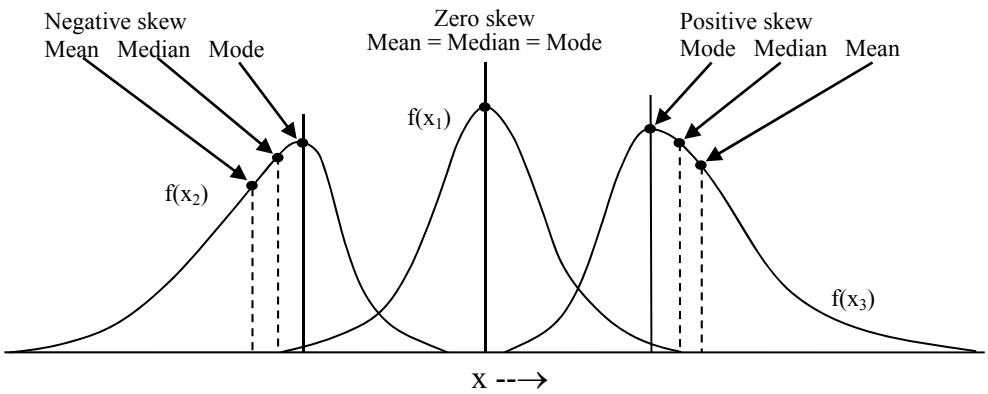

Fig. 10.4 Symmetrical and asymmetrical (+ve and –ve) skewed distributions.

10.2.4 Measures of Peakedness or Flatness

The measure used to denote the peakedness or the flatness of the frequency distribution near its centre is known as the kurtosis coefficient. This coefficient is computed by:

$$C_k = \frac{n^2 \sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} \tag{10.21}$$

Normal distribution has the kurtosis 3. If a data set has a relatively greater concentration near the mean than the normal distribution, the kurtosis will be greater than 3. Conversely, if the data have a relatively smaller concentration near the mean than the normal distribution, the kurtosis will be less than 3.

Example 10.2: Compute the coefficient of skewness and the coefficient kurtosis of the data of example 10.1.

Solution: The coefficient of skewness can be computed by eq. (10.20)

$$C_s = [22/(21*20*83.22^3)]*[(195.5 - 176.8)^3 + (145.4 - -176.8)^3 + \ldots + (136.3-176.8)^3 + (243.3-176.8)^3)]$$
$$= 1.238$$

A positive value of C_S implies that the probability distribution of the data has heavy tail to the right.

Kurtosis can be computed by eq. (26)

$$C_k = [22*22/(21*20*19*83.22^4)]*[(195.5 - 176.8)^4 + (145.4 - -176.8)^4 + \ldots + (136.3-176.8)^4 + (243.3-176.8)^4)]$$
$$= 1.45$$

Since kurtosis is less than 3, it means that the data values are less concentrated around the mean than the normal distribution or the peak of the distribution will be flatter compared to the

normal distribution.

10.3 Graphical Presentation of Data

Graphically presentation helps in a good insight in the behavior and variation of the data. To graphically present the data in the form of histograms, a frequency table is prepared. For this purpose the range of the data is divided into a number of intervals of convenient size and frequencies of values occurring in each interval is entered alongside. The appearance of a frequency histogram depends upon the selection of class interval. If the class intervals are very large, the table is compact but details may be lost. If the intervals are too small, the table may be too bulky. The following guidelines may be considered while choosing the class interval

(a) Brooks and Carruthers' rough guide:

$$\text{Number of classes} \leq 5 \log (\text{sample size}) \qquad (10.22)$$

(b) Charlier's rule of thumb:

$$w = (\text{maximum value} - \text{minimum value})/20 \qquad (10.23)$$

where w is the size of class interval. In general the number of classes varies between and 25.

To prepare the frequency table, steps given below can be followed:

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

(ii) Decide the number of class intervals (NC) and the size of the class interval ΔX.

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

(iv) Determine the absolute frequency n_j as the number of observations that fall in the j^{th} class interval, j=1, ... NC.

(v) Compute the relative frequencies of various classes as n_j/n, j=1,... NC and n is the number of observations.

(vi) Compute the cumulative relative frequencies F_j, j = 1, ... NC.

(vii) Plot the relative frequencies as well as cumulative relative frequencies with group interval as abscissa and the relative frequencies or cumulative relative frequencies as ordinate.

Example 10.3: The annual flow of Sabarmati River at Dharoi is plotted in Fig. 10.5 for the

period 1868-1965. Plot the histogram and the cumulative histogram.

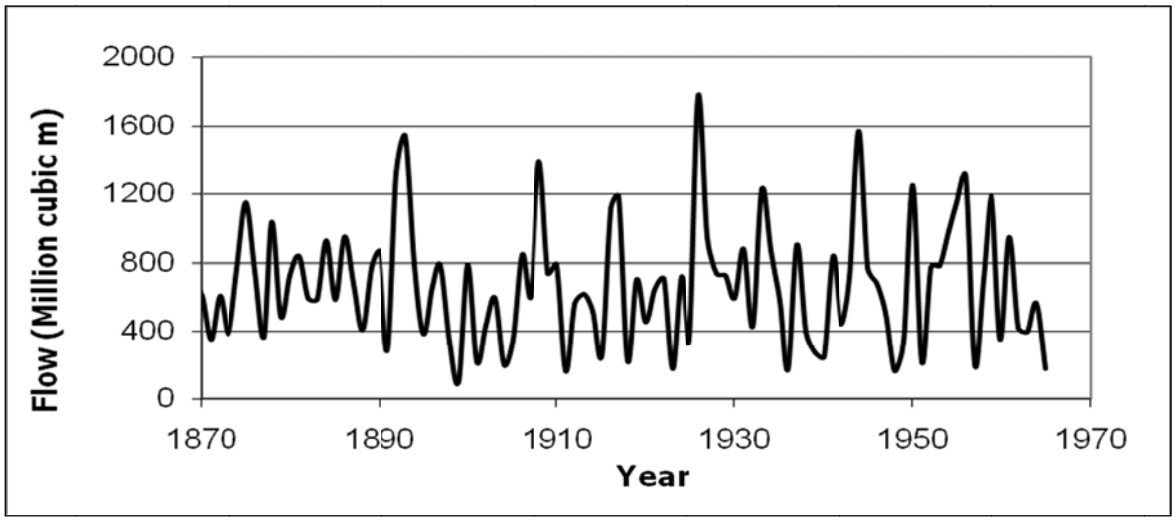

Fig. 10.5 Plot of the annual flow of the river.

Solution: After examining the data having 98 values, the class interval was chosen as 100 MCM. Table 10.1 shows the mid-values of classes in which the data has been divided, the frequency of values in each class and the cumulative frequencies. There are 17 classes.

Table 10.1 Mid-values and frequencies of various classes of example data.

Mid-value of class (MCM)	Frequency	Cumulative frequency
150	6	6
250	9	15
350	11	26
450	9	35
550	9	44
650	9	53
750	19	72
850	6	78
950	6	84
1050	1	85
1150	5	90
1250	2	92

1350	3	95
1450	0	95
1550	2	97
1650	0	97
1750	1	98

The cumulative histogram of the annual flow of Sabarmati River at Dharoi for the period 1868-1965 (98 years) is plotted in Fig. 10.6.

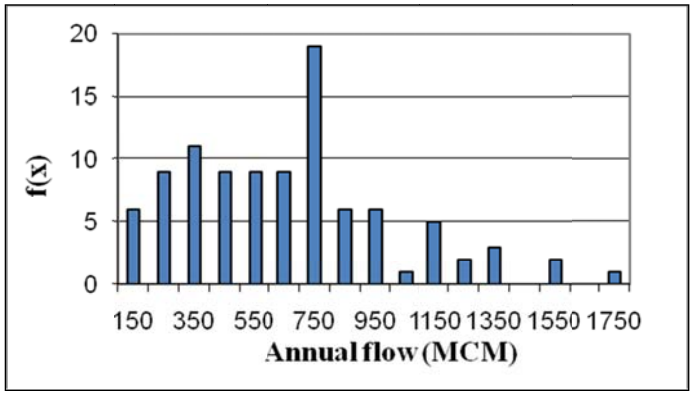

Fig. 10.6 Histogram of the annual river flows.

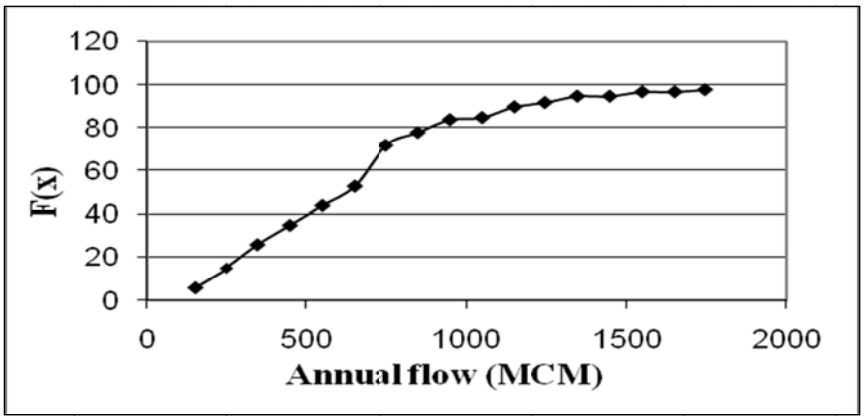

Fig. 10.7 Cumulative histogram of annual flows of the river.