Chapter 3
Basic Statistical Properties of Data

This chapter starts with some basic exploratory statistical properties from sample data. Concept of moment and expectation, and moment-generating and characteristic functions are considered afterwards. Different methods for parameter estimation build the foundation for many statistical inferences in the field of hydrology and hydroclimatology.

3.1 Descriptive Statistics

The probabilistic characteristics of random variables can be described completely if the form of the distribution function is known and the associated parameters are specified. However, in the absence of knowledge of any parametric distribution, approximate description about the population is assessed through sample statistics. These are also known as descriptive statistics. Some of the most commonly used descriptive statistics are *central tendency*, *dispersion*, *skewness*, and *tailedness*. Respective population parameters are the properties of the underlying probability distribution (Fig. 3.1). Expressions for sample estimates and population parameters are presented simultaneously to facilitate the readers.

3.1.1 Measures of Central Tendency

The measure of central tendency of a random variable can be expressed in terms of three quantities, namely mean, median, and mode. The mean can be further expressed in different forms as discussed in the following sections.

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_3

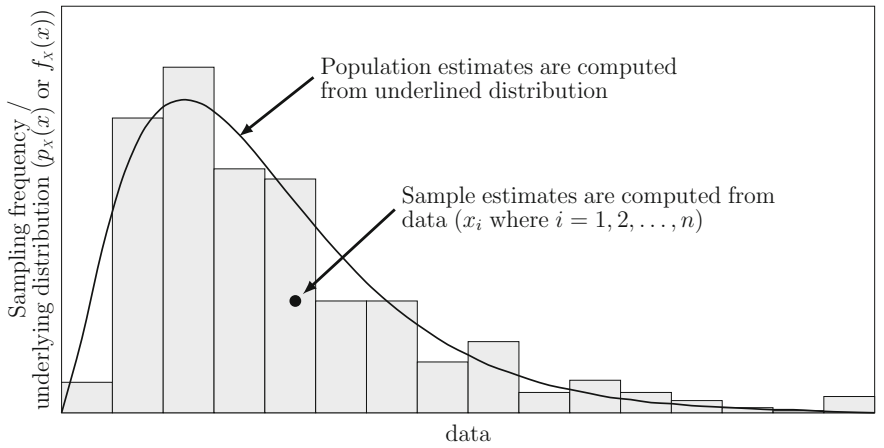

Fig. 3.1 Frequency plot of a data set with the underlying distribution used to evaluate the sample estimates (from the data) and population parameters (from the underlying distribution)

Arithmetic Mean

Arithmetic mean can be defined as the sum of the observations divided by sample size. Let us consider a sample data set with n observations $x_1, x_2, \ldots, x_n$ for a random variable X. The sample estimate of the population mean (μ) is the arithmetic average $\overline{x}$, calculated as

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

In case of grouped data, let us consider k as the number of groups, n as the total number of observations, n_i as the number of observations in the ith group, and x_i as the class mark of the ith group. Class mark is defined as midpoint of the group, i.e., mean of upper and lower bounds of group. For the grouped data, the $\overline{x}$ is given by

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i \tag{3.2}$$

For population, considering $p_X(x_i)$ as the underlying distribution (*pmf*) of a discrete random variable X, the population mean μ is expressed as

$$\mu = \sum_{i=1}^{n} x_i \, p_X(x_i) \tag{3.3}$$

and considering $f_X(x)$ as the underlying distribution (*pdf*) of a continuous random variable X, the population mean μ is expressed as

$$\mu = \int_{-\infty}^{\infty} x f_X(x)\, dx \tag{3.4}$$

Expressions for population mean are further discussed later with respect to the concept of moment.

Geometric Mean

The geometric mean indicates the central tendency of a data set by using the product of their values. The geometric mean can be defined as the nth root of the product of n observations. The sample geometric mean, $\overline{x}_G$, can be evaluated as

$$\overline{x}_G = \left(\prod_{i=1}^{n} x_i \right)^{1/n} \tag{3.5}$$

where the symbol $\prod$ implies multiplication. The geometric mean can also be expressed as the exponential of the arithmetic mean of logarithms. Thereby, the logarithm of $\overline{x}_G$ is equal to the arithmetic mean of the logarithms of the x_i's. Geometric mean of the population is expressed as:

$$\mu_G = \text{antilog}\left[E(\log X) \right] \tag{3.6}$$

where $E(\bullet)$ stands for expectation, which is discussed later in Sect. 3.2.

Weighted Mean

The weighted mean is similar to an arithmetic mean except some data points contribute more than others. The calculation of the arithmetic mean of grouped data as explained before is an example of weighted means where n_i/n is the weighted factor. In general, the weighted mean is

$$\overline{x}_w = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i} \tag{3.7}$$

where w_i is the weight associated with the ith observation or group and k is the number of observations or groups.

Median

The median is the value of the random variable at which the values on both sides of it are equally probable. This can be particularly used if one desires to eliminate

the effect of extreme values as mean is highly influenced by the extreme values. The median of n observations can be defined as the value of $(n+1)/2$ numbered observation (the observations are arranged in ascending order) in case n is odd and average of two observations in position $n/2$ and $n/2+1$ in case n is an even number. Thereby, we can say that sample median $\overline{x}_{md}$ is the observation such that half of the values lie on either side of $\overline{x}_{md}$.

Considering X to be a discrete random variable, the population median $\mu_{md} = x_d$ where d is determined from

$$\sum_{i=1}^{d} p_X(x_i) = 0.5 \tag{3.8}$$

Considering X to be a continuous random variable, the population median μ_{md} would be the value satisfying

$$\int_{-\infty}^{\mu_{md}} f_X(x)\,dx = 0.5 \tag{3.9}$$

Mode

The mode is the most probable or most frequently occurring value of a random variable. It is the value of the random variable with the highest probability density or the most frequently occurring value. A sample or a population may have none, one, or more than one mode. Thus, the population mode, μ_{mo}, would be a value of X maximizing *pmf* or *pdf*.

Considering X to be a discrete random variable with *pmf* $p_X(x)$, the mode is the value of x_i for which $p_X(x_i)$ is maximum, i.e.,

$$\mu_{mo} = \arg\max_{x_i} [p_X(x_i)] \tag{3.10}$$

Considering X to be a continuous random variable with *pdf* $f_X(x)$, the mode is the value of X that satisfies the following equation

$$\frac{df_X(x)}{dx} = 0 \quad \text{and} \quad \frac{d^2 f_X(x)}{dx^2} < 0 \tag{3.11}$$

3.1.2 Measure of Dispersion

The dispersion of a random variable corresponds to how closely the values of a random variable are clustered or how widely it is spread around the central value. Figure 3.2 shows two random variables, X and Y, with same mean but dispersion of Y is more than X.

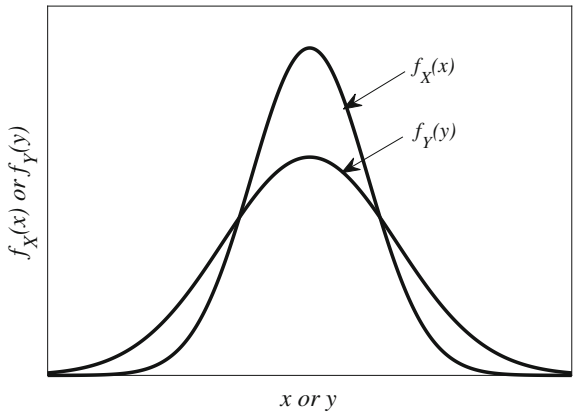

Fig. 3.2 Random variables X and Y with same mean but different dispersion

Range

The range of a sample is the difference between the maximum and the minimum values in the sample. The minimum and the maximum values also convey information about the variability present in data. The range has the disadvantage of not reflecting the frequency or magnitude of values that deviate either positively or negatively from the mean since only the largest and smallest values are used in its determination. Occasionally, the relative range is used which is the range divided by the mean.

Variance

Variance (S^2) is a measure of the dispersion of a random variable taking the mean as the central value. For a sample of size n, the variance is the average squared deviation from the sample mean.

Considering X as a random variable and a sample $x_1, x_2, \ldots, x_n$ with sample mean $\overline{x}$, the differences $x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_n - \overline{x}$ are called the deviations from the mean. The sample estimate of variance can be defined as the average of the squared deviations from the mean. The sample estimate of population variance σ^2 is denoted by S^2 and is given as

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \tag{3.12}$$

The reason for dividing by $n - 1$ instead of n is to make the estimator unbiased. Unbiasedness is one of the four properties that an estimator should possess. These properties are explained later in Sect. 3.6. For the time being, readers may note that one degree of freedom is lost while estimating the sample mean ($\overline{x}$) from the data.

For the grouped data with $x_1, x_2, \ldots, x_k$ as the class mark, the variance can be estimated from the following formula

$$S^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 n_i}{n - 1} \tag{3.13}$$

where k is the number of groups, n is the total number of observations, x_i is the class mark, and n_i is the number of observations in the ith group.

Standard deviation, another measure of dispersion, is the positive square root of variance, and the unit for standard deviation is the same as the unit of the X. The formula for S is as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}} \tag{3.14}$$

A dimensionless measure of dispersion is the coefficient of variation defined as the standard deviation divided by the mean. The coefficient of variation is estimated as

$$C_v = \frac{S}{\overline{x}} \tag{3.15}$$

Higher values indicate more dispersed data, i.e., high variability about mean and *vice versa*. Population estimate of variance (denoted as σ^2) is discussed later in Sect. 3.2.1.

3.1.3 Measure of Symmetry

Distributions of data may not be symmetrical with respect to its mean; i.e., they may tail off to the right or to the left. Such distributions are said to be skewed (Fig. 3.3). Skewness of the data is measured using the coefficient of skewness (γ). For positive skewness (coefficient of skewness, $\gamma > 0$), the data is skewed to the right and similarly for negative skewness ($\gamma < 0$) the data is skewed to the left. The difference between the mean and the mode indicates the skewness of the data. The sample estimate skewness is normally made dimensionless by dividing by S^3 to get the coefficient of skewness. A sample estimate of coefficient of skewness (denoted as C_s) is expressed as

$$C_s = \frac{n \sum_{i=1}^{n} (x_i - \overline{x})^3}{(n - 1)(n - 2) S^3} \tag{3.16}$$

Population estimate for coefficient of skewness (denoted by γ) is discussed later in Sect. 3.2.1.

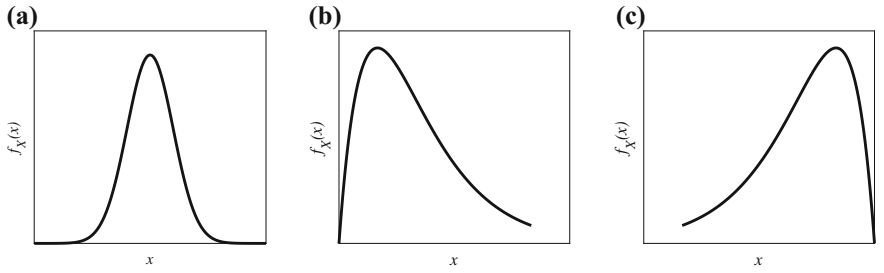

Fig. 3.3 Typical *pdf* plots of a **a** symmetric, **b** positively skewed distribution, and **c** negatively skewed distribution

3.1.4 Measure of Tailedness

The measure of tailedness of a probability distribution function is referred to as kurtosis. Being a measure of tailedness, kurtosis provides important interpretation about the tails, i.e., outlier. For a sample, kurtosis shows the effect of existing outliers. However, for a distribution, kurtosis shows the propensity to produce outliers. The kurtosis is made dimensionless by dividing by S^4 to get the coefficient of kurtosis. Coefficient of kurtosis is a convenient non-dimensional measure of tailedness. The sample estimate of the coefficient of kurtosis is given by

$$k = \frac{n^2 \sum_{i=1}^{n} (x_i - \overline{x})^4}{(n-1)\,(n-2)\,(n-3)\,S^4} \tag{3.17}$$

A particular distribution can be classified on the basis of its tailedness when compared with a standard value. Generally, the standard value taken is the kurtosis of normal distribution that has a value of 3. Thus, sometimes another estimate, $\varepsilon = k - 3$, is also used as a measure of kurtosis. Based on the measure of kurtosis, data or the associated distribution can be divided into three types (Fig. 3.4) as follows:

(i) *Mesokurtic*: If any distribution has same kurtosis as compared to normal distribution, the distribution is called mesokurtic. Thus, for a mesokurtic distribution, $k = 3$ and $\varepsilon = 0$.
(ii) *Leptokurtic*: In case a distribution has a relatively greater concentration of probability near the mean than the normal distribution; the kurtosis will be greater than 3. The value of ε will be positive.
(iii) *Platykurtic*: In case a distribution has a relatively smaller concentration of probability near the mean than the normal distribution; the kurtosis will be less than 3. The value of ε will be negative.

Population estimate for coefficient of kurtosis is discussed later in Sect. 3.2.1.

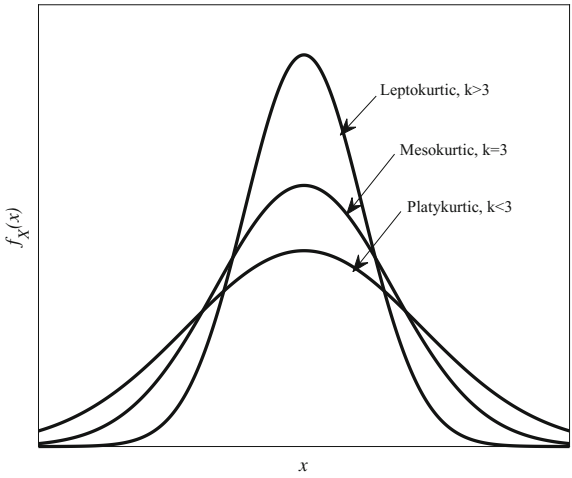

Fig. 3.4 A typical *pdf* plot showing the three zones of kurtosis, namely leptokurtic, mesokurtic, and platykurtic

Example 3.1.1
Consider the following sample data for annual peak discharge (cumec) at a gauging station A. Evaluate the mean, variance, coefficient of skewness, and coefficient of kurtosis for the given sample data. Also, comment regarding the coefficient of skewness and coefficient of kurtosis.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Annual peak discharge (cumec)	4630	2662	1913	3655	3670	4005	4621	1557
Year	2008	2009	2010	2011	2012	2013	2014	2015
Annual peak discharge (cumec)	2405	1625	6216	2602	2157	3120	6403	2934

Solution The mean for the given sample data for peak annual discharge can be evaluated as

$$\begin{aligned}\overline{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{4630 + 2662 + \cdots + 2934}{16} \\ &= 3385.93 \text{ cumec}\end{aligned}$$

The variance of the sample data can be evaluated as

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$
$$= \frac{(46330 - 3385.93)^2 + (2662 - 3385.93)^2 + \cdots + (2934 - 3385.93)^2}{15}$$
$$= 2214860.86 \approx 2.2 \times 10^6 \,\text{cumec}^2$$

The coefficient of skewness of the sample data can be evaluated as

$$C_s = \frac{n\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)(n-2)S^3}$$
$$= \frac{16\left[(46330 - 3385.93)^3 + (2662 - 3385.93)^3 + \cdots + (2934 - 3385.93)^3\right]}{15 \times 14 \times (2214860.86)^{3/2}}$$
$$= 0.745$$

The coefficient of kurtosis of the sample data can be evaluated as

$$k = \frac{n^2\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)(n-2)(n-3)S^4}$$
$$= \frac{16^2\left[(46330 - 3385.93)^4 + (2662 - 3385.93)^4 + \cdots + (2934 - 3385.93)^4\right]}{15 \times 14 \times 13 \times (2214860.86)^2}$$
$$= 2.628$$

As the value of coefficient of skewness is positive so the data is *positively skewed* and coefficient of kurtosis is less than 3, so it is *platykurtic*.

3.2 Concept of Moments and Expectation

In physics, moment is the product of a physical quantity and the distance from a fixed point of reference. While considering mass as the physical quantity, it can be used to locate the center of gravity of any irregularly shaped object. Higher order of moments can also be evaluated. Similar concepts can be utilized to extract some meaningful information from a data set (Fig. 3.5a).

Suppose that the data $x_1, x_2, \ldots, x_n$ is located according to their values on the real line as shown in Fig. 3.5a. Assuming that each data value is equiprobable, the mass of each data can be assumed to be $1/n$, when n is the length of the data. Now, the total moment with respect to origin can be evaluated as $\sum_{\text{all } i} x_i \left(1/n\right)$. We may find out the locations say $\tilde{x}$ of the equivalent total mass, i.e., the mass that will create the

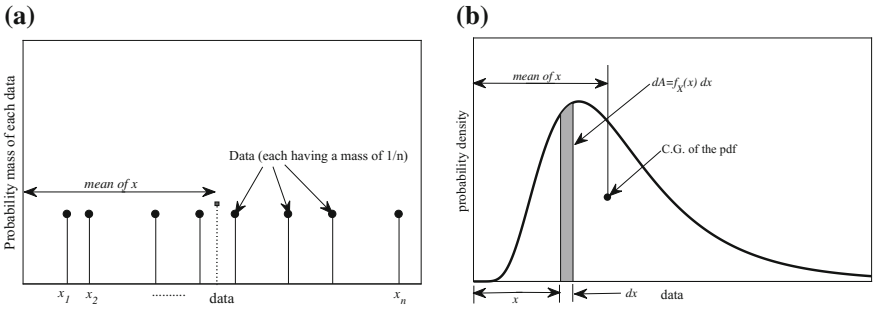

Fig. 3.5 First moment (mean) of the data for **a** discrete data and **b** probability density function of continuous data

same moment (as the total moment) about the origin, expressed as $\left(n \times {}^{1}/_{n}\right)\tilde{x} = \tilde{x}$. Equating these two moments, we get

$$\tilde{x} = \sum_{\text{all } i} x_i (1/n) \tag{3.18}$$

This location ($\tilde{x}$) is equivalent to the mean of the data ($\overline{x}$).

In case of the population which is represented by a *pdf*, mean can be identified following the same concept. Referring to Fig. 3.5b, consider a delta width (dx) located at a distance x from the origin. The total probability mass is equal to the area above dx and below the *pdf* (shaded area $= dA$). Total moment for this area with respect to origin is $x.dA = x.f_X(x)dx$. Integrating for the entire range of the data $(-\infty, \infty)$, the total moment can be written as $\int_{-\infty}^{\infty} x f_X(x)dx$. If it is assumed that the total probability mass is located at a distance x from origin that produces same amount of moment, we may write

$$\mu \times \int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} x f_X(x)dx$$

Since $\int_{-\infty}^{\infty} f_X(x)dx = 1$

$$\mu = \int_{-\infty}^{\infty} x f_X(x)dx \tag{3.19}$$

Following the same concept, higher order moments with respect to origin can also be evaluated using some power of distance from the origin; for example, x^2 and x^3 can be used to evaluate the second- and third-order moments, respectively. The x in Eq. 3.19 can be replaced with x^i to evaluate the ith moment with respect to origin.

However in probability theory, second moment onwards are calculated with respect to the mean. First moment with respect to the mean is zero. The second-order moment with respect to the mean can be evaluated as

$$E\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 \, f_X(x)\, dx \tag{3.20}$$

In general, the ith-order moment with respect to the mean can be evaluated as

$$E\left[(X-\mu)^i\right] = \int_{-\infty}^{\infty} (x-\mu)^i \, f_X(x)\, dx \tag{3.21}$$

3.2.1 Expectation

The expected value of a random quantity intuitively means the averaged value of the outcome of the corresponding random experiment carried out repetitively for infinite times. Mathematically, the expected value of a random variable (X), represented as $E(X)$, can be defined as the first moment about the origin and represented as follows:

$$E(X) = \mu \tag{3.22}$$

Considering X to be a discrete random variable, the expected value of X is given as

$$E(X) = \sum_{all\ j} x_j \, p_X\left(x_j\right) \tag{3.23}$$

and for continuous random variables, the expected value of X is given as

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx \tag{3.24}$$

Any function of X, say $g(X)$, is also a random variable. Thus, the expected value of $g(X)$ is given as

$$E[g(X)] = \sum_{\text{all } j} g\left(x_j\right) p_X\left(x_j\right) \qquad \text{for discrete RV} \tag{3.25}$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx \qquad \text{for continuous RV} \tag{3.26}$$

Relating the concept of moment with expectation, following points can be noted

(i) The *first moment* about *origin* is the mean, i.e.,

$$E(X) = \mu = \begin{cases} \sum_{\text{all } j} x_j p_x(x_j) & \text{for discrete RV} \\ \int_{-\infty}^{\infty} x f_x(x)\, dx & \text{for continuous RV} \end{cases} \tag{3.27}$$

(ii) The *second moment* about the *mean* is the *variance*

$$E\left[(X-\mu)^2\right] = \sigma^2 = \begin{cases} \sum_{\text{all } j} \left(x_j - \mu\right)^2 p_x(x_j) & \text{for discrete RV} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f_x(x)\, dx & \text{for continuous RV} \end{cases} \tag{3.28}$$

It can also be shown that

$$V(x) = E\left(x^2\right) - [E(x)]^2 \tag{3.29}$$

(iii) The *third moment* about the *mean* is the *skewness*

$$E\left[(X-\mu)^3\right] = \begin{cases} \sum_{\text{all } j} \left(x_j - \mu\right)^3 p_x(x_j) & \text{for discrete RV} \\ \int_{-\infty}^{\infty} (x-\mu)^3 f_x(x)\, dx & \text{for continuous RV} \end{cases} \tag{3.30}$$

It can also be shown that

$$E\left[(x-\mu)^3\right] = E\left(x^3\right) - 3E\left(x^2\right) E(x) + 2\{E(x)\}^3 \tag{3.31}$$

The measure of skewness is non-dimensionalized using variance and termed as *coefficient of skewness* (γ). Thus, γ is expressed as

$$\gamma = \frac{E\left[(X-\mu)^3\right]}{\sigma^3} \tag{3.32}$$

(iv) The *fourth moment* about the *mean* is the *kurtosis* (measure of tailedness)

$$E\left[(X-\mu)^4\right] = \begin{cases} \sum_{all\ j} \left(x_j - \mu\right)^4 p_x(x_j) & \text{for discrete RV} \\ \int_{-\infty}^{\infty} (x-\mu)^4 f_x(x)\, dx & \text{for continuous RV} \end{cases} \tag{3.33}$$

It can also be shown that

$$E\left[(x-\mu)^4\right] = E\left(x^4\right) - 4E\left(x^3\right) E(x) + 6E\left(x^2\right)(E(x))^2 - 3\{E(x)\}^4 \tag{3.34}$$

The measure of tailedness (*kurtosis*) is also non-dimensionalized using variance and termed as *coefficient of kurtosis* (κ). Thus, κ is expressed as

$$\kappa = \frac{E\left[(X-\mu)^4\right]}{\sigma^4} \tag{3.35}$$

Table 3.1 Population parameters and sample statistics

Property	Parameter name	Population parameter	Sample statistic
Central tendency	Arithmetic mean	For discrete case, $\mu = \sum x p_x(x)$ For continuous case, $\mu = \int_{-\infty}^{\infty} x f_x(x) dx$	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
	Geometric mean	$\mu_G = \text{antilog}\left[E(\log X)\right]$	$\overline{x}_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$
	Median	X such that $F(x) = 0.5$	50th percentile value of data
	Mode	For discrete case, $\mu_{mo} = \arg\max_{x_i} [p_x(x_i)]$ For continuous case, μ_{mo} is the root of $\frac{df_x(x)}{dx} = 0$ and $\frac{d^2 f_x(x)}{dx^2} < 0$	Most frequently occurring data
Variability	Variance	$\sigma^2 = E\left[(X-\mu)^2\right]$	$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
	Standard deviation	$\sigma = E\left[(X-\mu)^2\right]^{1/2}$	$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$
	Coefficient of variation	$c_v = \frac{\sigma}{\mu}$	$CV = \frac{S}{\overline{x}}$
Symmetry	Coefficient of skewness	$\gamma = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}$	$C_s = \frac{n\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)(n-2)S^3}$
Tailedness	Coefficient of kurtosis	$\kappa = \frac{E\left[(X-\mu)^4\right]}{\sigma^4}$	$k = \frac{n^2\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)(n-2)(n-3)S^4}$

Population parameters and corresponding sample estimates of different descriptive statistics are shown in Table 3.1.
Some useful information on the expected values is:

(i) Expectation of a constant is same as that constant, i.e., $E(C) = C$.
(ii) Expectation of a modified random variable obtained by multiplying with a constant is equal to the product of the constant and the expectation of the original random variable, i.e., $E(CX) = C\,E(X)$.
(iii) Expectation of a random variable obtained by addition/subtraction of two random variables is equal to the sum/difference of their individual expectations, i.e., $E(X \pm Y) = E(X) \pm E(Y)$.

Some useful information on the variance values is:

(i) Variance of a constant is zero, i.e., $V(C) = 0$.
(ii) Variance of a modified random variable obtained by multiplying with a constant is equal to the product of the square of constant and the variance of the original random variable, i.e., $V(CX) = C^2\,V(X)$.

(iii) Variance of a modified random variable obtained by multiplying with a constant (a) and addition to another constant (b) is equal to the product of the square of constant (a) and the variance of the original random variable, i.e., $V\ (aX + b) = a^2\ V\ (X)$.

Example 3.2.1
The number of thunderstorms per year (X) and its *pmf* obtained from the historical data are shown in the following table:

X	0	1	2	3
pmf $(p_X(x))$	0.3	0.4	0.2	0.1

What is the mean and variance of the number of thunderstorms in a year?

Solution The mean of number of thunderstorms per year can be evaluated as

$$\begin{aligned} E(x) &= \sum x p_x\,(x) \\ &= 0 \times 0.3 + 1 \times 0.4 + 2 \times 0.2 + 3 \times 0.1 \\ &= 1.1 \end{aligned}$$

The variance of storms can be evaluated as

$$\begin{aligned} V(x) &= E(x^2) - \{E(x)\}^2 \\ &= [1^2 \times 0.4 + 2^2 \times 0.2 + 3^2 \times 0.1] - 1.1^2 \\ &= 0.89 \end{aligned}$$

Example 3.2.2
The time (T) between two successive floods follows following *pdf*

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

Find the mean, mode, median, and the coefficient of variation of T.

Solution The mean time between successive floods is given by

$$E(T) = \int_0^\infty t\lambda e^{-\lambda t} dt = -\int_0^\infty t\ d\left(e^{-\lambda t}\right)$$

Integrating by parts (i.e., $\int udv = uv - \int vdu$), we get

$$E(T) = -te^{-\lambda t}\Big|_0^\infty + \int_0^\infty e^{-\lambda t}dt = \left[-\frac{e^{-\lambda t}}{\lambda}\right]_0^\infty = \frac{1}{\lambda}$$

Hence, the mean time between successive floods is $\overline{t} = 1/\lambda$.

The mode is the value of t with the maximum value of *pdf*. From the *pdf*, it can be observed that the probability density is highest at $t = 0$. Thus, the mode is $\mu_{mo} = 0$ The median can be evaluated as

$$\int_0^{\mu_{md}} \lambda e^{-\lambda t}dt = 0.5$$

$$\text{or, } 1 - e^{-\lambda\mu_{md}} = 0.5$$

$$\text{or, } \mu_{md} = \frac{-\ln(0.5)}{\lambda} = \frac{0.693}{\lambda}$$

Therefore, median is $\mu_{md} = \frac{0.693}{\lambda}$.

The variance can be evaluated as

$$\sigma_T^2 = \int_0^\infty \left(t - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda t}dt = \int_0^\infty \left(\lambda t^2 - 2t + \frac{1}{\lambda}\right) e^{-\lambda t}dt$$

Integrating by parts (as done for $E(T)$), we get

$$\text{First term, } \int_0^\infty t^2\lambda e^{-\lambda t}dt = -\int_0^\infty t^2 d\left(e^{-\lambda t}\right) = -\frac{2}{\lambda}$$

$$\text{Second term, } -2\int_0^\infty te^{-\lambda t}dt = 2\int_0^\infty \frac{t}{\lambda}d\left(e^{-\lambda t}\right) = \frac{2}{\lambda}$$

$$\text{Third term, } \int_0^\infty \frac{1}{\lambda}e^{-\lambda t}dt = \frac{1}{\lambda^2}$$

$$\text{Hence, } \sigma_T^2 = \frac{2}{\lambda} - \frac{2}{\lambda} + \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

The standard deviation is given by $\sigma_T = \frac{1}{\lambda}$.

The coefficient of variation of the given distribution is $c_v = \frac{\sigma_T}{\mu_T} = \frac{1/\lambda}{1/\lambda} = 1$.

Example 3.2.3

The rainfall depth (in cm) received during thunderstorms at a place (X) is a random variable with the following density function

$$f_x(x) = \begin{cases} \frac{3}{2500}(x-10)(x-20) & 0 \le x \le 10 \\ 0 & \text{elsewhere} \end{cases}$$

Determine the following

(a) Mean value of X;
(b) Median of X;
(c) Mode of X;
(d) Standard deviation of X;
(e) Coefficient of variation of X; and
(f) Skewness coefficient.

Solution The density function is $f_X(x) = \frac{3}{2500}(x-10)(x-20)$ for $0 \le x \le 10$.

(a) Mean (μ)

$$\mu = \int_0^{10} x f_X(x) dx = \frac{3}{2500}\int_0^{10} x(x-10)(x-20)\,dx$$
$$= \frac{3}{2500}\left[\frac{x^4}{4} - 10x^3 + 100x^2\right]_0^{10} = 3$$

(b) Median (μ_{md})

$$\int_0^{\mu_{md}} f_X(x) dx = 0.5$$
$$\text{or, } \frac{3}{2500}\int_0^{\mu_{md}} (x-10)(x-20)\,dx = 0.5$$
$$\text{or, } \frac{3}{2500}\left[\frac{x^3}{3} - 15x^2 + 200x\right]_0^{\mu_{md}} = 0.5$$
$$\text{or, } \mu_{md} = 2.5398$$

(c) Standard deviation σ

$$\sigma^2 = \int_0^{10} (x-\overline{x})^2 f_X(x) dx$$
$$\text{or, } \sigma^2 = \frac{3}{2500}\int_0^{10} (x-3)^2(x-10)(x-20) dx$$
$$\text{or, } \sigma^2 = \frac{1}{12500}\left[3x^5 - 135x^4 + 1945x^3 - 11025x^2 + 27000x\right]_0^{10}$$
$$\text{Hence, } \sigma = \sqrt{5} = 2.2361$$

(d) Coefficient of variation is calculated as

$$c_v = \frac{\sigma}{\mu} = \frac{2.236}{3} = 0.7454 \approx 74.5\%.$$

(e) Coefficient of skewness is obtained as

$$\begin{aligned}\gamma &= \left(\int_0^{10} (x-\mu)^3 f_x(x)dx\right) \Big/ \sigma^3 \\ &= \frac{3}{2500 \times 5\sqrt{5}} \int_0^{10} (x-3)^3 (x-10)(x-20)\, dx \\ &= \frac{1}{250000\sqrt{5}} \left[x(10x^5 - 468x^4 + 7455x^3 - 52740x^2 + 186300x - 324000)\right]_0^{10} \\ &= 0.7155\end{aligned}$$

3.3 Moment-Generating Functions

Moment-generating function of a random variable is generally treated as an alternative to its probability distribution. Though all the random variables may not have moment-generating functions, however, if available, these are sometimes easier to compute moments of the random variables of any desired order.

Expectation of e^{tX}, which is a function of the random variable X, is known as moment-generating function of the random variable X. It can be represented as

$$M_X(t) = E\left(e^{tX}\right) \tag{3.36}$$

In case of discrete random variable, the moment-generating function can be evaluated as

$$M_X(t) = \sum_{all\ j} e^{tx_j} p_x\left(x_j\right) \tag{3.37}$$

In case of continuous random variable, the moment-generating function can be evaluated as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_x(x)\, dx \tag{3.38}$$

We can show that the Taylor series expansion of $M_X(t)$ is

$$M_X(t) = 1 + \mu t + \mu_2^{'} \frac{t^2}{2} + \cdots + \mu_k^{'} \frac{t^k}{k} + \cdots \tag{3.39}$$

The kth moment about origin is then found to be the kth derivative of $M_X(t)$ with respect to t and evaluated at $t = 0$.

$$\mu_k^t = \left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} \tag{3.40}$$

Usefulness of the moment generation function can be explored by evaluating the derivatives of the function. First derivative of $M_X(t)$, evaluated at $t = 0$, results in the expected value, which is first moment of the random variable with respect to origin. Mathematically,

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = \int_{-\infty}^{\infty} x f_X(x)\, dx \tag{3.41}$$

Similarly, second derivative of $M_X(t)$, evaluated at $t = 0$, results in second moment of the random variable with respect to origin. Thus,

$$\left.\frac{d^2M_X(t)}{dt^2}\right|_{t=0} = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx \tag{3.42}$$

In general, nth derivative of $M_X(t)$, evaluated at $t = 0$, results in nth moment of the random variable with respect to origin.

$$\left.\frac{d^nM_X(t)}{dt^n}\right|_{t=0} = \int_{-\infty}^{\infty} x^n f_X(x)\, dx \tag{3.43}$$

Example 3.3.1
Consider a data set to follow the given distribution where λ is a constant. Evaluate the first moment with respect to origin, the second moment with respect to the mean and the moment-generating function.

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots$$

Solution Calculation for the first moment with respect to origin otherwise known as mean

$$E(x) = \sum_0^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = \lambda\sum_1^{\infty}\frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda}\sum_1^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = \lambda$$

Calculation for the second moment with respect to mean otherwise known as variance

$V(x) = E\left(x^2\right) - [E(x)]^2$

In above expression, $E\left(x^2\right)$ can also be expressed as

$E\left(x^2\right) = E[x(x-1)] + E(x)$

$$E[x(x-1)] = \sum_0^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} = \lambda^2\sum_1^{\infty}\frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!} = \lambda^2 e^{-\lambda}\sum_1^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = \lambda^2$$

$$V(x) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Calculation for the moment-generating function

$$E\left(e^{tx}\right) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda\left(e^t - 1\right)}$$

3.4 Characteristic Functions

Similar to moment-generating function, characteristic function of a random variable may also serve as another alternative to its probability distribution. It is the Fourier transform of the probability density function of the random variable.

The expectation of e^{itX} (where $i = \sqrt{-1}$), which is a complex function of the random variable X, is known as characteristic function of that random variable X. It can be defined as

$$\phi_X(t) = E\left(e^{itX}\right) = M_X(it) \tag{3.44}$$

The characteristic function for X can be expressed as

$$\phi_X(t) = \sum_{\text{all } j} e^{itx_j} p_x\left(x_j\right) \qquad \text{for discrete RV} \tag{3.45}$$

$$\phi_X(t) = \int_{-\infty}^{\infty} f_x(x)\, e^{itx} dx \qquad \text{for continuous RV} \tag{3.46}$$

Using the characteristic function, the nth moment of X can be expressed as

$$E\left(X^n\right) = \frac{1}{i^n} \frac{d^n \phi_X(t)}{dt^n}\bigg|_{t=0} \tag{3.47}$$

Example 3.4.1
Consider a random variable X that follows the given distribution. Evaluate the mean, variance, skewness, and kurtosis.

X	0	25	60	75	100
$p_x(x)$	0.5	0.24	0.12	0.08	0.06

Solution Mean can be evaluated as

$$E(x) = 0 \times 0.5 + 25 \times 0.24 + 50 \times 0.12 + 75 \times 0.08 + 100 \times 0.06 = 24$$

Variance can be evaluated as

$$V(x) = E\left[(x-\mu)^2\right] = E\left(x^2\right) - \{E(x)\}^2$$
$$E\left(x^2\right) = 0^2 \times 0.5 + 25^2 \times 0.24 + 50^2 \times 0.12 + 75^2 \times 0.08 + 100^2 \times 0.06 = 1500$$
$$V(x) = 1500 - 24^2 = 924$$

Skewness can be evaluated as

$$E\left[(x-\mu)^3\right] = E\left(x^3\right) - 3E\left(x^2\right)E(x) + 2\{E(x)\}^3$$
$$E\left(x^3\right) = 0^3 \times 0.5 + 25^3 \times 0.24 + 50^3 \times 0.12 + 75^3 \times 0.08 + 100^3 \times 0.06 = 112500$$
$$E\left[(x-\mu)^3\right] = 112500 - 3 \times 1500 \times 24 + 2 \times 24^3 = 32148$$

Kurtosis can be evaluated as

$$E\left[(x-\mu)^4\right] = E\left(x^4\right) - 4E\left(x^3\right)E(x) + 6E\left(x^2\right)(E(x))^2 - 3\{E(x)\}^4$$
$$E\left(x^4\right) = 0^4 \times 0.5 + 25^4 \times 0.24 + 50^4 \times 0.12 + 75^4 \times 0.08 + 100^4 \times 0.06 = 9375000$$
$$E\left[(x-\mu)^4\right] = 9375000 - 4 \times 112500 \times 24 + 6 \times 1500 \times 24^2 - 3 \times 24^4 = 2763672$$

Example 3.4.2
Consider a continuous random variable X having the following marginal distribution. Evaluate the mean, variance, and median.

$$f_X(x) = \begin{cases} \frac{2}{x^3} & \text{for } x > 1 \\ 0 & \text{elsewhere} \end{cases}$$

Solution Mean can be evaluated as

$$\begin{aligned} E(x) &= \int x\, f_X(x) dx \\ &= \int_1^\infty x \times \frac{2}{x^3} dx \\ &= \left[\frac{-2}{x}\right]_1^\infty \\ &= 2 \end{aligned}$$

Variance can be evaluated as

$$V(x) = E\left(x^2\right) - \{E(x)\}^2$$
$$= \int x^2 f_x(x)\,dx - \{E(x)\}^2$$

$$\text{where } \int x^2 f_x(x)\,dx = \int_1^\infty x^2 \times \frac{2}{x^3} dx$$

The integral does not exist; thereby, the variance does not exist.

For the calculation of median, we first need to calculate the *CDF* of X.

$$F_x(x) = \int_1^x f_x(x)dx = 1 - \frac{1}{x^2}$$

$$\text{Hence, } F_x(x) = \begin{cases} 1 - \frac{1}{x^2} & x \geq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Median can be evaluated as

$$F_x(\mu_{md}) = 0.5$$
$$\text{or, } 1 - \frac{1}{\mu_{md}^2} = \frac{1}{2}$$
$$\text{or, } \mu_{md} = \sqrt{2}$$

Example 3.4.3

Evaluate the coefficient of variation, coefficient of skewness, and coefficient of kurtosis for the data supplied in Example 3.4.1. Also, provide an insight into the skewness and tailedness of the distribution.

Solution Calculation for coefficient of variation:

$$c_v = \frac{\sigma}{\mu} = \frac{\sqrt{924}}{24} = 1.266$$

Calculation for coefficient of skewness:

$$\gamma = \frac{E\left[(x-\mu)^3\right]}{\sigma^3} = \frac{32148}{924^{3/2}} = 1.144$$

Calculation for coefficient of kurtosis:

$$\kappa = \frac{E\left[(x-\mu)^4\right]}{\sigma^4} = \frac{2763672}{924^2} = 3.237$$

As γ is positive, so the distribution is positively skewed, and κ-3 is positive so the distribution is leptokurtic.

Example 3.4.4
The 30 years of monthly rainfall data (mm) at rain gauge stations A and B are found to follow the given distribution.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

If the probability of rainfall exceeding 50 mm is 0.135 for station A and 0.188 for station B, which station receives higher mean rainfall?

Solution In order to determine the station receiving higher mean rainfall, we first have to evaluate the mean for the above-mentioned distribution, i.e., $E(X)$. Hence, from Example 3.2.2,

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

Further, the probability that rainfall exceeds 50 mm is given by

$$P(x > 50) = \int_{50}^\infty \lambda e^{-\lambda x} dx = e^{-50\lambda}$$

For station A:

$$e^{-50\lambda} = 0.135$$
$$\lambda = 0.04$$
$$\text{Thus, } f_X(x) = 0.04e^{-0.04x}$$
$$\mu = \int_0^\infty x f_X(x) dx = 25$$

Similarly, for station B:

$$e^{-50\lambda} = 0.188$$
$$\lambda = 0.033$$
$$\text{Thus, } f_X(x) = 0.033e^{-0.033x}$$
$$\mu = \int_0^\infty x f_X(x) dx = 30.30$$

Therefore, station B receives higher mean rainfall.

3.5 Statistical Properties of Jointly Distributed Random Variables

3.5.1 Expectation

If X and Y are considered to be jointly distributed continuous random variable and U is some function of X and Y, $U = g(X, Y)$, then expectation of U, $E(U)$ can be written as

$$E(U) = E[g(X, Y)] = \int u \, f_U(u) \, du \tag{3.48}$$

In case of continuous random variables,

$$E[g(X, Y)] = \int \int g(x, y) \, f_{X,Y}(x, y) \, dx \, dy \tag{3.49}$$

In case of discrete random variables,

$$E[g(X, Y)] = \sum_i \sum_j g\left(x_i, y_j\right) p_{X,Y}\left(x_i, y_j\right) \tag{3.50}$$

In all the cases, the result is the average value of the function $g(X, Y)$ weighted by the probability that $X = x$ and $Y = y$ or the mean of the random variable U.

3.5.2 Moment about the Origin

A general expression for the (r, s)th moment of the jointly distributed continuous random variable X and Y is

$$\mu^1_{r,s} = \int \int x^r y^s f_{X,Y}(x, y) \, dx \, dy \qquad \text{for continuous RV} \tag{3.51}$$

$$\mu^1_{r,s} = \sum_i \sum_j x_j^r \, y_i^s \, p_{X,Y}\left(x_i, y_j\right) \qquad \text{for discrete RV} \tag{3.52}$$

3.5.3 Moment about the Mean (Central Moment)

The central moment for jointly distributed continuous random variables X and Y is given by

$$\mu_{r,s} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)^r (y-\mu_Y)^s f_{X,Y}(x,y)\, dx\, dy \quad \text{for continuous RV} \tag{3.53}$$

$$\mu_{r,s} = \sum_i \sum_j (x_i-\mu_X)^r \left(y_j-\mu_Y\right)^s p_{X,Y}(x,y) \quad \text{for discrete RV} \tag{3.54}$$

3.5.4 Moment-Generating Function

Similar to moment-generating function of a single random variable defined in previous section, the moment-generating function for two random variables is defined for discrete and continuous cases. The moment-generating function for two continuous random variables can be obtained as

$$M_{X,Y}(t,u) = E\left(e^{tX+uY}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{tx+uy} f_{X,Y}(x,y) \tag{3.55}$$

The moment-generating function for two discrete random variables can be obtained as

$$M_{X,Y}(t,u) = E\left(e^{tX+uY}\right) = \sum_{\text{all } x}\sum_{\text{all } y} e^{tx+uy} p_{X,Y}(x,y) \tag{3.56}$$

Example 3.5.1
A reservoir has two inflow points A and B. The streamflow gauging records at station A and B show that inflow at station A (designated by X) and the same at station B (designated by Y) follow the given distributions.

$$f_X(x) = \begin{cases} \frac{1}{50}(10-x) & 0 \le x \le 10 \\ 0 & \text{elsewhere} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{1}{300}(25-y) & 0 \le y \le 20 \\ 0 & \text{elsewhere} \end{cases}$$

Considering the inflow at station A and B to be independent, evaluate the mean of total inflow to the reservoir and the moment-generating function for the same.

Solution As given, X designates the inflow at station A and Y designates the inflow at station B. The total inflow to the reservoir can be designated by another random variable, say Z. Thus, Z is a function of random variables X and Y such that $Z = g(X,Y) = X + Y$.

As the inflows at station A and B are independent, their joint *pdf* can be evaluated as the product of their individual, i.e., $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. The mean of the total inflow to the reservoir can be evaluated as

$$\begin{aligned} E(Z) &= \int_0^{20} \int_0^{10} (x+y) \left(\frac{10-x}{50}\right) \left(\frac{25-y}{300}\right) dx\, dy \\ &= \int_0^{20} -\frac{(3y+10)(y-25)}{900} dy \\ &= \frac{100}{9} \qquad\qquad = 11.11 \end{aligned}$$

The moment-generating function can be written as

$$\begin{aligned} M_Z(t,u) &= E(e^{tX+uY}) \\ &= \int_0^{20} \int_0^{10} e^{tx+uy} \left(\frac{10-x}{50}\right) \left(\frac{25-y}{300}\right) dx\, dy \\ &= \frac{(10t - e^{10t} + 1)(25u - e^{20u} - 5ue^{20u} + 1)}{15000t^2u^2} \end{aligned}$$

3.5.5 *Covariance*

The covariance of jointly distributed random variables X and Y can be written as the expected value of the product of their deviations from their respective mean values as follows:

$$\text{Cov}(X,\ Y) = \sigma_{X,Y} = E\left[(X - \mu_X)(Y - \mu_Y)\right] \tag{3.57}$$

By using the linearity property of expectations, *r.h.s.* of Eq. 3.57 can be transformed to a simpler form, which describes as the expected value of their product minus the product of their expected values, as shown in Eq. 3.58.

$$E\left[(X - \mu_X)(Y - \mu_Y)\right] = E(XY) - E(X)\, E(Y) \tag{3.58}$$

For continuous random variables, covariance can be expressed as

$$\sigma_{X,Y} = \iint (x - \mu_X)(y - \mu_Y)\, f_{X,Y}(x, y)\, dx\, dy \tag{3.59}$$

For discrete random variables, covariance can be expressed as

$$\sigma_{X,Y} = \sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_X)(y - \mu_Y)\, p_{X,Y}(x, y) \tag{3.60}$$

If X and Y are independent, then $f_{X,Y}(x, y) = f_X(x)\ f_Y(y)$ for continuous random variable and $p_{X,Y}(x, y) = p_X(x)\ p_Y(y)$ for discrete random variable.

Thus, covariance for independent continuous random variables can be expressed as

$$\begin{aligned} \sigma_{X,Y} &= \iint (x - \mu_X)(y - \mu_Y)\, f_{X,Y}(x, y)\ dx\, dy \\ &= \int (x - \mu_X)\, f_X(x)\ dx \int (y - \mu_Y)\, f_Y(y)\ dy = 0 \end{aligned} \tag{3.61}$$

Thus, covariance for independent discrete random variables can be expressed as

$$\begin{aligned} \sigma_{X,Y} &= \sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_X)(y - \mu_Y)\, p_{X,Y}(x, y) \\ &= \sum_{all\ x} (x - \mu_X)\, p_X(x) \sum_{all\ y} (y - \mu_Y)\, p_Y(y) = 0 \end{aligned} \tag{3.62}$$

since first central moment with respect to mean is 0. This implies covariance of two independent variables is always 0. However, the reverse is not true, i.e., zero covariance does not necessarily indicate that the variables are independent.

The sample estimate for the covariance $\sigma_{X,Y}$ is $S_{X,Y}$ computed as

$$S_{X,Y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{(n - 1)} \tag{3.63}$$

Example 3.5.2
The joint distribution of two random variables X_1 and X_2 is given as follows. Find out the covariance of X_1 and X_2.

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} 6x_1 & 0 < x_1 < x_2 < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Solution The marginal distributions of X_1 and X_2 are as follows:

$$f_X(x_1) = \int_{x_1}^{1} 6x_1 dx_2 = [6x_1x_2]_{x_1}^{1} = 6x_1(1 - x_1) \qquad 0 < x_1 < 1$$

$$f_X(x_2) = \int_{0}^{x_2} 6x_1 dx_1 = \left[6\frac{x_1^2}{2}\right]_{0}^{x_2} = 3x_2^2 \qquad 0 < x_2 < 1$$

The covariance of X_1 and X_2 can be calculated as follows:

$$\text{Cov}\,(X_1, X_2) = E\,(X_1, X_2) - E\,(X_1)\,E\,(X_2)$$

Expectation for X_1 and X_2 can be calculated as follows:

$$E\,(X_1) = \int_0^1 x_1 6\,(x_1)\,(1 - x_1)\,dx_1 = \frac{1}{2}$$
$$E\,(X_2) = \int_0^1 x_2\left(3x_2^2\right)dx_2 = \frac{3}{4}$$

Expectation of joint distribution of X_1 and X_2 can be evaluated as

$$E\,(x_1 x_2) = \int_0^1 \int_0^{x_2} x_1 x_2 6 x_1 dx_1 dx_2 = \frac{2}{5}$$

Thereby, the covariance can be evaluated as

$$\text{Cov}\,(X_1, X_2) = \frac{2}{5} - \frac{1}{2} \times \frac{3}{4} = \frac{1}{40}$$

3.5.6 Correlation Coefficient

Correlation coefficient is a normalized form of covariance which is obtained by dividing the covariance by the product of standard deviation of X and Y.

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{3.64}$$

The range of $\rho_{X,Y}$ is $-1 \leq \rho_{X,Y} \leq 1$. Actually, $\rho_{X,Y}$ is the measure of linear dependence between X and Y. Thereby, if $\rho_{X,Y} = 0$, and X and Y are linearly independent, however, they might be related by some nonlinear functional form. In this case, X and Y are said to be uncorrelated. A value of $\rho_{X,Y}$ equal to ± 1 implies that X and Y are perfectly related by $Y = a + bX$. In this case, X and Y are said to be correlated. The sample estimate of the population correlation coefficient $\rho_{X,Y}$ is $r_{X,Y}$ computed from

$$r_{X,Y} = \frac{S_{X,Y}}{S_X S_Y} \tag{3.65}$$

where S_X and S_Y are the sample estimates of σ_X and σ_Y, respectively, and $S_{X,Y}$ is the sample covariance.

Example 3.5.3
Let X units denote the rainfall intensity in a particular catchment and Y units denote the runoff from the catchment. The joint *pdf* of X and Y is given as follows. Evaluate the covariance and the correlation coefficient.

$$f_{X,Y}(x, y) = \begin{cases} x^2 + \frac{xy}{3} & 0 \le x \le 1; 0 \le y \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

Solution Evaluation of the marginal *pdf* of X and Y is carried out in Example 3.5.2. In order to evaluate the correlation coefficient, we have to evaluate the variance of X, variance of Y, and covariance of X and Y.

$$\text{Cov}(XY) = E(XY) - E(X)E(Y)$$

$$E(X) = \int_0^1 x\left(\frac{2}{3}x + 2x^2\right) dx = \left[\frac{2}{9}x^3 + \frac{1}{2}x^4\right]_0^1 = \frac{13}{18}$$
$$E\,(Y) = \int_0^2 y\left(\frac{1}{3} + \frac{y}{6}\right) dy = \left[\frac{1}{6}y^2 + \frac{1}{18}y^3\right]_0^2 = \frac{10}{9}$$

$$E(X, Y) = \int_0^1 \int_0^2 xy\left(x^2 + \frac{xy}{3}\right) dy\, dx$$
$$= \int_0^1 \left[\frac{1}{2}x^3y^2 + \frac{1}{9}x^2y^3\right]_0^2 dx = \int_0^1 2x^3 + \frac{8}{9}x^2\, dx = \left[\frac{1}{2}x^4 + \frac{8}{27}x^3\right]_0^1 = \frac{43}{54}$$

$$\text{Cov}(XY) = \frac{43}{54} - \left(\frac{13}{18}\right)\left(\frac{10}{9}\right) = -\frac{1}{162}$$

As $\text{Cov}(X, Y) \neq 0$, thereby, X and Y are correlated.

Calculation of variance of X and Y,

$$\text{Var}(X) = E(X^2) - [E\,(X)]^2$$

$$\text{Var}(Y) = E(Y^2) - [E\,(Y)]^2$$

$$E(X^2) = \int_0^1 x^2\left(\frac{2}{3}x + 2x^2\right) dx = \left[\frac{1}{6}x^4 + \frac{2}{5}x^5\right]_0^1 = \frac{17}{30}$$
$$E\left(Y^2\right) = \int_0^2 y^2\left(\frac{1}{3} + \frac{y}{6}\right) dy = \left[\frac{1}{9}y^3 + \frac{1}{24}y^4\right]_0^2 = \frac{14}{9}$$

$$\text{Var}(X) = \frac{17}{30} - \left(\frac{13}{18}\right)^2 = 0.045$$

$$\text{Var}(Y) = \frac{14}{9} - \left(\frac{10}{9}\right)^2 = 0.321$$

Calculation of correlation coefficient,

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{-1/162}{\sqrt{0.045}\sqrt{0.321}} = -0.051$$

The correlation coefficient is -0.051.

3.5.7 *Further Properties of Moments*

If Z is a linear function of two random variables X and Y such that $Z = aX + bY$, then

$$E(Z) = E(aX + bY) = aE(X) + bE(Y) \tag{3.66}$$

$$\text{Var}(Z) = a^2 Var(X) + b^2 Var(Y) + 2ab\, Cov(X, Y) \tag{3.67}$$

We can generalize the above equations considering Y as a linear function of n random variables such that $Y = \sum_{i=1}^{n} a_i X_i$, then,

$$E(Y) = E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i) \tag{3.68}$$

$$\text{Var}(Y) = \sum_{i=1}^{n} a_i^2\, Var(x_i) + 2\sum_{i<j} a_i a_j\, Cov\left(X_i, X_j\right) \tag{3.69}$$

Now for a special case considering $a_i = 1/n$ in Y, we get $Y = \overline{X}$. Since x_i form a random sample, the $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$ and $\text{Var}(X_i) = \text{Var}(X)$. Thereby,

$$\text{Var}(Y) = \text{Var}\left(\overline{X}\right) = \sum_{i=1}^{n} \frac{1}{n^2}\text{Var}(X) = \frac{n}{n^2} Var(X)$$

or,

$$\text{Var}\left(\overline{X}\right) = \frac{\text{Var}(X)}{n} \tag{3.70}$$

If we consider X and Y to be independent random variables, then the variance of their product XY is given by:

$$\text{Var}\,(XY) = E\,(XY)^2 - E^2\,(XY) \tag{3.71}$$

Now, $E\,(XY)^2 = E\left(X^2\right) E\left(Y^2\right) = \left(\mu_X^2 + \sigma_X^2\right)\left(\mu_Y^2 + \sigma_Y^2\right)$.
And $E^2\,(XY) = E^2\,(X)\;E^2\,(Y) = \mu_X^2\,\mu_Y^2$.
Thus, variance of the product X and Y can also be expressed as

$$\text{Var}\,(XY) = \mu_X^2\sigma_Y^2 + \mu_Y^2\sigma_X^2 + \sigma_X^2\sigma_Y^2 \tag{3.72}$$

3.6 Properties of the Estimator

In general, the probability distribution functions are the functions of a set of parameters and the random variable. To use the probability distribution for the estimation of probability, it is important to calculate the values of the parameters. The general procedure for estimating a parameter is to obtain a random sample from the population and use it to estimate the parameters. Now if we consider $\hat{\theta}_i$ as the estimate for the parameter θ_i, then $\hat{\theta}_i$ is a function of the random variables since $\hat{\theta}_i$ is itself a random variable possessing mean, variance and probability distribution. An ideal estimator should possess the following four characteristics, namely unbiasedness, consistency, efficiency, and sufficiency.

3.6.1 Unbiasedness

An estimator $(\hat{\theta})$ of a parameter (θ) is said to be unbiased if the expected value of the estimate is equal to the parameter $\left(E(\hat{\theta}) = \theta\right)$. As unbiased, estimator implies that an average of many independent estimators for the parameter will be equal to the parameter itself. In case the estimate is biased, the bias can be evaluated as $E(\hat{\theta}) - \theta$.

3.6.2 Consistency

An estimator $(\hat{\theta})$ of a parameter (θ) is said to be consistent if the probability that the estimator differs from the parameter $(\hat{\theta} - \theta)$ by more than a constant (ε) approaches to 0 as the sample size approaches infinity.

3.6.3 Efficiency

An estimator $(\hat{\theta})$ is said to be more efficient estimator for a parameter (θ) if the estimator is unbiased and its variance is at least as small as that of another unbiased estimator $\hat{\theta}_1$. The relative efficiency (RE) of $\hat{\theta}$ with respect to another estimator $\hat{\theta}_1$ can be evaluated as follows:

$$RE = \frac{V\left(\hat{\theta}\right)}{V\left(\hat{\theta}_1\right)} \tag{3.73}$$

If the value of the relative frequency is less than 1, then $\hat{\theta}$ is a more efficient estimator of θ than $\hat{\theta}_1$.

3.6.4 Sufficiency

An estimator $(\hat{\theta})$ is said to be a sufficient estimator for a parameter (θ) if the estimator utilizes all of the information contained in the sample and is relevant to the parameter.

Example 3.6.1
Consider a random variable X such that $X \sim N\left(\mu, \sigma^2\right)$. Check if the estimators of mean $\overline{X} = \frac{1}{n}\sum_i X_i$ and variance $S^2 = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2$ are biased or unbiased.

Solution Estimator of mean (μ) is given as follows:

$$\overline{X} = \frac{1}{n}\sum_i X_i$$

Expectation of the estimator $E\left(\overline{X}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu_i = \mu$, which is equal to population mean. Therefore, $\overline{X}$ is an unbiased estimator of μ.

Estimator of variance $\left(\sigma^2\right)$ is given as follows:

$$S^2 = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2$$

Expectation of the estimator can be evaluated as

$$
\begin{aligned}
E\left(S^2\right) &= \frac{1}{n-1}\sum_i \left(X_i - \mu + \mu - \overline{X}\right)^2 \\
&= \frac{1}{n-1}\sum_i (X_i - \mu)^2 + \left(\mu - \overline{X}\right)^2 + 2\left(X_i - \mu\right)\left(\mu - \overline{X}\right) \\
&= \frac{1}{n-1}\sum_i (X_i - \mu)^2 + \left(\mu - \overline{X}\right)^2 + 2n\left(\overline{X} - \mu\right)\left(\mu - \overline{X}\right) \\
&= \frac{1}{n-1}\sum_i (X_i - \mu)^2 - n\left(\mu - \overline{X}\right)^2 \\
&= \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) \\
&= \sigma^2
\end{aligned}
$$

Therefore, S^2 is an unbiased estimator of σ^2.

3.7 Parameter Estimation

3.7.1 Method of Moments

The method of moments is a popular method of estimation of population parameters. It considers that a good estimate of a probability distribution parameter is that for which central moments of population equal with corresponding central moment of the sample data. Finally, an equation is derived that relates the population moments to the parameters of interest. For this purpose, a sample is drawn and the population moments are estimated from the sample. Then, the equations are solved for the parameters of interest, after replacing (unknown) population moments by sample moments. In case of a distribution with m parameters, the first m moments of the distribution are equated to the sample moments to obtain m equations which can be solved for the m unknown parameters. In other words, let us consider a random variable X that follows a distribution function $f_x\left(x; \theta_1, \ldots, \theta_k\right)$, with parameters $\theta_1, \ldots, \theta_k$ and a random sample $x_1, \ldots, x_n$, and then as per the assumptions of the method of moment, the rth population moment can be equated to the rth sample moment. Thus, we finally get the estimates of that parameter (see Example 3.7.1).

Example 3.7.1
Consider an exponential distribution whose *pdf* is given by $f_x(x) = \lambda e^{-\lambda x}$ for $x > 0$. Determine the estimate of the parameter λ.

Solution Equating the first-order central moment of population to that of sample, we get

$$\mu = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx$$

Using integration by parts (Example 3.2.2)

$$\mu = \frac{1}{\lambda}$$

That yields, $\lambda = 1/\mu$, and thus the corresponding sample estimate is $\lambda = 1/\overline{x}$.

3.7.2 Maximum Likelihood

Maximum-likelihood (ML) method assumes that the best estimator of a parameter of a distribution should maximize the *likelihood* or the joint probability of occurrence of a sample. Let us consider, $x = (x_1, \ldots, x_n)$ is a set of n independent and identically distributed observed sample and $f\ (x,\ \theta)$ is the probability distribution function with parameter θ. The likelihood function can be written as follows:

$$L = \prod_{i=1}^{n} f_X(x_i) \tag{3.74}$$

where the symbol $\prod$ indicates multiplication. Sometimes, it becomes convenient to work with logarithmic of likelihood function, i.e,

$$\ln L = \sum_{i=1}^{n} \ln [f_X(x_i)] \tag{3.75}$$

In this case, $\hat{\theta}$ is said to be the maximum-likelihood estimator (MLE) of θ if $\hat{\theta}$ maximizes the function L or $\ln(L)$.

Example 3.7.2 Consider $x_1, \ldots, x_n$ to follow the following distribution

$$f_X\ (x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty$$

Evaluate MLE for μ and σ^2.

Solution The likelihood function is to be evaluated as follows:

$$\begin{aligned} L &= L\left(\mu, \sigma^2 | x_1, \ldots, x_n\right) \\ &= \prod_{i=1}^{n} f_{X_i}(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^n \left(\sigma^2\right)^{n/2}} e^{-1/2\sigma^2 \sum_{i=1}^{n}(x_i-\mu)^2} \end{aligned}$$

Thereby, the log-likelihood function can be evaluated as follows:

$$\log L = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

The estimator of μ can be evaluated by maximizing the log-likelihood function

$$\begin{aligned} &\frac{\partial \log L}{\partial \mu} = 0 \\ &\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \qquad\qquad \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \hat{\mu} \end{aligned}$$

Therefore, the estimator of the mean is $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$.

$$\begin{aligned} &\frac{\partial \log L}{\delta \sigma^2} = 0 \\ &\left(-\frac{n}{2\sigma^2}\right) + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \\ &\frac{1}{2\sigma^2} \left(-n + \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(x_i - \hat{\mu}\right)^2\right) = 0 \\ &\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \hat{\mu}\right)^2}{n} = \hat{\sigma}^2 \end{aligned}$$

Therefore, the estimator of the variance is $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$.

3.8 Chebyshev Inequality

Certain general statements about random variables can be made without fitting a specific distribution to the random variable. One such statement can be provided by the Chebyshev inequality. It ensures that not more than a certain fraction of values can be away from the mean by certain distance. The Chebyshev inequality states that the probability of getting a value which is away from μ by atleast $k\sigma$ is at most $1/k^2$, where μ is the population mean and σ is the population standard deviation. Thus,

$$P\left(|X-\mu| \geq k\sigma\right) \leq \frac{1}{k^2} \tag{3.76}$$

The Chebyshev inequality provides an upper limit for the probability of a deviation of a specific value from the mean.

3.9 Law of Large Number

Chebyshev's inequality can be written in terms of sample mean (sample size n) as follows:

$$P\left(|\overline{x}-\mu| \geq \frac{k\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2} \tag{3.77}$$

where $\overline{x}$ is the sample mean for a sample of size n. For the above inequality, if $1/k^2 = \delta$ and the sample size is selected such that $n \geq \sigma^2/\varepsilon^2$ where $\varepsilon < 0$ and $0 < \delta < 1$, then we get the law of large number. It can be stated as

$$P\left(\left|\overline{X}-\mu\right| \geq \varepsilon\right) \leq \delta \tag{3.78}$$

The law of large number ensures that by selecting a large enough sample, we can estimate the population mean with the desired accuracy.

3.10 MATLAB Examples

For solving examples in this chapter, symbolic toolbox of MATLAB is required. Some of the important function/commands are listed below.

- `syms`: This command is used for defining new algebraic symbol. For example, `syms x` will define an algebraic symbol `x`.
- `[output1,...,outputN] = eval(expr)`: This function evaluates the expression (`expr` argument). In case of symbolic expressions, this function can be used for simplifying them.

- `[y1,...,yN] = solve(eqns,vars)`: This function is used for solving univariate or multivariate equations (`eqns` argument) for variables `vars`. The variables argument is optional. In case of multiple equations, they are passed as string separated by comma like `[x_value,y_value] = solve('x+y = 7,x-y = 3')` yields `x_value = 5` and `y_value = 2`.
- `output_expr = int(expr,var)`: This function is used for indefinite integration of expression (`expr` argument) with respect to variable (`var` argument). Further, `int(expr,var,a,b)` is used for definite integration of expression (`expr` argument) between the variable (`var` argument) value `a` and `b`.
- `output_expr = diff(expr,var)`: This function is used for symbolic differentiation of expression (`expr` argument) with respect to variable (`var` argument).

Using the functions discussed above, sample MATLAB script for solving Example 3.5.3 is provided in Box 3.1.

Box 3.1 Sample MATLAB script for solving Example 3.5.3

```
clear all;clc

%% Inputs, i.e, definition of all the distribution
    functions.
syms x y
x_fun=(2/3)*x+2*(x^2);
y_fun=(1/3)+(y/6);
joint_fun=(x^2)+(x*y)/3;

%% Evaluation of expectation of x, y and the joint
    distribution %of x and y.
exp_x=int(x*x_fun,x,0,1); % Expectation of x within
    the   %defined support
exp_y=int(y*y_fun,y,0,2); % Expectation of y within
    the %defined support
exp_joint=int(int(x*y*joint_fun,y,0,2),x,0,1);
cov_xy=exp_joint-(exp_y*exp_x); % Covariance of a
    and y

%% Evaluation of the correlation coefficient
exp_x2=int((x^2)*x_fun,x,0,1);
exp_y2=int((y^2)*y_fun,y,0,2);
var_x=exp_x2-(exp_x^2); %Evaluation of variance of x
var_y=exp_y2-(exp_y^2); %Evaluation of variance of y
cc_xy=eval(cov_xy/(sqrt(var_x)*sqrt(var_y))); %
    Evaluation of %the correlation coefficient

%% Display Results
output_file=['output' filesep() 'code_1_result.txt'
    ];
delete(output_file);diary(output_file);diary on;
% Output stating if the variables are correlated
```

```
if cov_xy==0
    disp('The random variables X and Y are not
        correlated.');
else
    disp('The random variables X and Y are
        correlated.');
end
fprintf('The correlation coefficient of X and Y is
    %2.3f.\n',cc_xy)
diary off;
```

The output of the code mentioned in Box 3.1 is provided in Box 3.2. The solution obtained using the MATLAB code is same as the conclusions drawn from the solution of Example 3.5.3.

Box 3.2 Results for Box 3.1

```
The random variables X and Y are correlated.
The correlation coefficient of X and Y is -0.051.
```

Exercise

3.1 Considering the number of storms in an area for the month of June to follow the following distribution

$$p_X(x) = \begin{cases} \frac{2^x e^{-2}}{x!} & x = 1, 2, \ldots, 5 \\ 0.152 & x = 0 \end{cases}$$

Evaluate the mean and median for the number of storms in the given month. (Ans: 0.848; median lies between 1 and 2)

3.2 Soil samples are collected from 15 vegetated locations in a particular area. The moisture content of the samples as obtained from the laboratory tests is shown in the following table. Evaluate the arithmetic mean, geometric mean, range, variance, coefficient of skewness, and coefficient of kurtosis of the soil moisture data. Comment regarding the skewness and kurtosis of the data. (Ans: 0.3207; 0.2926; 0.490; 0.018; 0.4136; 3.496)

Sample no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SMC	0.25	0.40	0.11	0.45	0.36	0.24	0.26	0.31	0.50	0.60	0.39	0.28	0.19	0.14	0.33

3.3 The maximum temperature (in °C) at a city in the month of May follows the distribution as given below

$$f_X(x) = \frac{1}{\beta - \alpha} \qquad 40 \le x \le 45$$

Evaluate the mean, variance, and coefficient of variation of the maximum temperature in the city. (Ans: 42.5 °C; 2.083; 0.034)

3.4 The discharge at a gauging station follows the given distribution

$$f_X(x) = 5e^{-5x} \qquad x \ge 0$$

Determine the nature of the distribution in terms of its coefficient of variation, skewness, and tailedness. (Ans: 1/5; 2; 6)

3.5 A city supplied water from two sources. The joint *pdf* of discharge from two sources is as follows:

$$f_{X,Y}(x, y) = \begin{cases} x^2 + \frac{xy}{3} & 0 \le x \le 1; 0 \le y \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

Evaluate the marginal probability density of each source and the mean discharge from the two sources. Also, evaluate the covariance and the forms of conditional distribution of X given $Y = y$. (Ans: 13/18 units; 30/27 units; $-1/162$)

3.6 Consider a random variable X to follow a two-parameter distribution. The population mean (μ) and standard deviation (σ) are the parameters of the distribution. Evaluate an unbiased estimation of μ and unbiased and biased estimation of σ.

3.7 Let $x_1, x_2, \ldots, x_n$ be a random sample for a distribution with *pdf*

$$f_X(x) = \frac{e^{-x/\beta} \times x^{\alpha-1}}{\beta^{\alpha}\,\Gamma(\alpha)} \qquad \alpha > 1; x, \beta > 0$$

Find estimators for α and β using method of moments.

3.8 Let $x_1, x_2, \ldots, x_n \sim U(0, \theta)$. Find the maximum-likelihood estimate of θ?

3.9 If $x_1, x_2, \ldots, x_n \sim \frac{e^{-\lambda}\lambda^x}{x!}$. Find the maximum-likelihood estimate of λ?

3.10 Considering the peak annual discharge at a location to have a mean of 1100 cumec and standard deviation of 260 cumec. Without making any distributional assumptions regarding the data, what is the probability that the peak discharge in any year will deviate more than 800 cumec from the mean? (Ans: 0.106)

3.11 The random variable X can assume the values 1 and -1 with probability 0.5 each. Find (a) the moment-generating function and (b) the first four moments about the origin. (Ans: (a) $E(e^{tX}) = \frac{1}{2}(e^t + e^{-t})$, (b) 0, 1, 0, 1)

3.12 A random variable X has density function given by

$$f_X(x) = \begin{cases} 2e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Find (a) the moment-generating function and (b) the first four moments about the origin. (Ans: (b) 1/2, 1/2, 3/4, 3/2)

3.13 Find the first four moments (a) about the origin and (b) about the mean, for a random variable X having density function

$$f_X(x) = \begin{cases} 4x(9 - x^2)/81 & 0 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

(Ans: (a) 8/5, 3, 216/35, 27/2 (b) 0, 11/25, 32/875, 3693/8750)

3.14 Find (a) $E(X)$, (b) $E(Y)$, (c) $E(X, Y)$, (d) $E(X^2)$, (e) $E(Y^2)$, (f) $\text{Var}(X)$, (g) $\text{Var}(Y)$, (h) $\text{Cov}(X, Y)$ if the joint *pdf* of random variables X and Y is given as

$$f_{X,Y}(x, y) = \begin{cases} c(2x + y) & 2 < x < 5; 0 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

Use $c = 1/210$. (Ans: (a) 268/63, (b) 170/63, (c) 80/7, (d) 1220/63, (e) 1175/126, (f) 5036/3969, (g) 16225/7938, (h) $-200/3969$)

3.15 Joint distribution between two random variables X and Y is given as follows:

$$f_{X,Y}(x, y) = \begin{cases} 8xy & 2 \leq x \leq 1; 0 \leq y \leq x \\ 0 & \text{otherwise} \end{cases}$$

Find the conditional expectation of (a) Y given X and (b) X given Y. (Ans: (a) $\frac{2x}{3}$ (b) $\frac{2(1+y+y^2)}{3(1+y)}$)

3.16 The density function of a continuous random variable X is

$$f_X(x) = \begin{cases} 4x(9 - x^2)/81 & 0 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Find the (a) mean, (b) median, and (c) mode. (Ans: (a) 1.6 (b) 1.62 (c) 1.73)

3.17 Find the coefficient of (a) skewness and (b) kurtosis of the standard normal distribution which is defined by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad -\infty < x < \infty.$$

(Ans: (a) 0, (b) 3).