

Chapter 2

Probability Theories

Abstract This chapter constitutes a review of methods employed within probability analysis and presents principles of probability theory as well as statistics concepts for risk and reliability analysis of hydrosystem engineering. Next in this chapter, the concept and application of discrete and continuous random variables are briefly offered. Furthermore, commonly used probability distributions are presented with a number of straightforward examples to better understand the implementation of random variables and probability distributions in water resources engineering.

2.1 Review of Probability Theory

Probability theory is the division of mathematics that deals with random phenomena analysis and can be classified in three different concepts: classical, empirical, and subjective probability. The *classical* concept of probability originates in experiment physics, while it is not compulsory to carry out experiments. A very descriptive example is coin tossing, where there is a 0.5 probability of a balanced coin turning up heads even without having to do a trial. Based on the classical concept, the ratio of the number of successful events to the total number of possible events is the probability of occurrence of an event. For a complex situation where calculating probability is difficult and it is possible to measure an object's properties with adequately large sample, the likelihood that event will occur can be evaluated by running a large number of trials and observing the outcomes. This is *empirical* probability, and it works based on the relative frequencies in the long run as the ratio of frequency of occurrence of the event to the total number of observations. Although larger-number trials help to achieve more accurate results, empirical and classical concepts are not always appropriate

methods in measuring probability of a range of problems and sometimes personal judgment is needed. Personal judgment on an event's occurrence is called *subjective* probability. Subjective probability may be used in making logically constant decisions in the lack of better information and the quality of those decisions still depends on estimator knowledge regarding the nature of the problem (INC 2008).¹ The following sections provides a summary of methods employed within probability analysis, and further presents principles of probability theory as well as statistics concepts for risk and reliability analysis of hydrosystem engineering. Before going on to the next section, the definition of some basic concepts used in probability analysis is presented.

An experiment denotes the observational process in the probability theory which can be infinitely repeated, and the total possible outcomes of any particular experiment express the sample space of that experiment. An event is any subset of sample space that could be an empty set, \emptyset or the whole sample space and its complement. For a random experiment, three main operations that can be applied to construct new events from given events and make relationships among them are union, intersection and complement. Union of A and B ($A \cup B$) means the occurrence of event A or B , while the intersection of A and B ($A \cap B$) is the joint occurrence of events A and B which can also be simply shown as (A, B) . The complement of event A involves all outcomes in the sample space that are not included in the outcomes of event A and it is symbolized as A' . The event A and its complement are mutually exclusive since they cannot occur at the same time. In other words, when there is no outcomes in common, the two events A and B are mutually exclusive and is shown as $(A \cap B) = \emptyset$. If the occurrence of event A depends on the occurrence of event B , then it is a conditional event which is commonly notated as $A|B$.

2.2 Probability Concepts

Consider an experiment includes n possible outcomes $(X_1, X_2, \dots, X_i, \dots, X_n)$, the probability of event X_i is the relative number of X_i occurring in a large number of trials. In other words, this probability can be estimated as $P(X_i) = n_i/n$, where n_i is the total number of successes of event X_i in a sequence of n trials. When there are a finite number of discrete or countable events, the probability of these types of events which is simply termed a discrete probability comes into play. Let's say one defines $P(X_i)$ as the probability of a random event X_i , the discrete probabilities of this random event over the sample space of n possible outcomes would be held by the following conditions:

$$0 \leq P(X_i) \leq 1 \quad (2.1)$$

¹ Internet Center for Management and Business Administration.

$$\sum_{i=1}^n P(X_i) = 1 \quad (2.2)$$

Sometimes it is useful to know the probability of either events A or B, or both will occur. In this case, the probability of the union of two mutually exclusive events is computed as follow:

$$P(X_1 \cup X_2) = P(X_1) + P(X_2) \quad (2.3)$$

If the occurrence of one event (X_1) does not influence the incidence of the other (X_2), they are independent; otherwise they are dependent events. The probability of two independent events is:

$$P(X_1 \cap X_2) = P(X_1) \cdot P(X_2) \quad (2.4)$$

If the events are neither independent nor mutually exclusive, then:

$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2) \quad (2.5)$$

If event X_2 is the necessary condition for event X_1 to take place, the probability of occurring event X_1 can be determined by:

$$P(X_1|X_2) = \frac{P(X_1 \cap X_2)}{P(X_2)} \quad (2.6)$$

and for independent events, we have:

$$P(X_1|X_2) = \frac{P(X_1) \cdot P(X_2)}{P(X_2)} = P(X_1) \quad (2.7)$$

Example 2.1 Consider Y and X as dam overflowing and windy condition events, respectively and $P(Y) = 0.4$, $P(X) = 0.7$ and $P(Y|X) = 0.6$. What is the probability that both events will occur together?

Solution The probability of both events occurring together can be calculated by Eq. (2.6), as:

$$P(X \cap Y) = P(Y|X) \times P(X) = 0.6 \times 0.7 = 0.42$$

If overtopping and windy conditions are considered as independent events, then we have:

$$P(X \cap Y) = P(X) \cdot P(Y) = 0.7 \times 0.4 = 0.28$$

Example 2.2 Consider the following probability values at a culvert site during its service life:

1. Probability of flood occurring: $P_f = 0.2$,
2. Probability of failure: $P_r = 0.35$,

3. Probability of failure when there is a flood: $P(r|f) = 0.5$.

Determine the probabilities of no flooding, no failure, flooding and failure, flooding without failure, no flooding and failure, no flooding and no failure, and flooding or failure.

Solution

$$P_{no\ flood} = P(\bar{f}) = 1 - P(f) = 1.0 - 0.2 = 0.80$$

$$P_{no\ failure} = P(\bar{r}) = 1 - P(r) = 1 - 0.35 = 0.65$$

If flooding and failure are considered as two independent events, the desired probabilities can be computed as follows:

$$P_{flood\ and\ failure} = P(f \cap r) = 0.2 \times 0.35 = 0.07$$

$$P_{flood\ without\ failure} = P(f \cap \bar{r}) = 0.2 \times 0.65 = 0.13$$

$$P_{no\ flood\ and\ failure} = P(\bar{f} \cap r) = 0.8 \times 0.35 = 0.28$$

$$P_{no\ flood\ and\ no\ failure} = P(\bar{f} \cap \bar{r}) = 0.8 \times 0.65 = 0.52$$

The probability of flood or failure occurring is:

$$\begin{aligned} P(f \cup r) &= P(f) + P(r) - P(f \cap r) \\ &= 0.2 + 0.35 - 0.07 = 0.48 \end{aligned}$$

The probability of failure and flood occurring when they are considered as dependent events, is:

$$P(f \cap r) = P(f) \times P(r|f) = 0.2 \times 0.5 = 0.1$$

Finally, the union of probabilities for all dependent events will be computed as:

$$P(f \cup r) = P(f) + P(r) - P(f \cap r) = 0.2 + 0.35 - 0.1 = 0.45$$

2.3 Random Variables

Many significant events can be defined using appropriate random variables in the analysis of statistical characteristics of system performance. A random variable is a variable whose possible value is subject to variations, and outcome of a random phenomenon refers to a real valued function defined on the sample space in the probability theory. There are two broad categories of random variables as discrete and continuous variables. Discrete random variables are counting variable such as 0, 1, 2, ... with no in-between values. Examples of discrete random variables are the number of defective sprinklers in an irrigation system, the number of gates in a dam, or the number of rain gauges in a watershed. On the other hand, continuous random variables are infinite from an uncountable set with possible in-between values.

The examples for this type of variables are the amount of rainfall, discharge of a river, or the amount of water passing through a pipe. It is important to note that if given random variables have both discrete and continuous attributes, they can also be mixed in desired probability analysis. To assign a probability to each of the possible outcomes of a discrete or continuous random variable, a probability distribution is used. In other words, a probability distribution is the relationship between a statistical experiment outcome and its occurrence probability. In this case, the first and foremost concept is the cumulative distribution function (CDF) of random variable X which is obtained as follow:

$$F(x) = P(X \leq x) \quad (2.8)$$

where $F(x)$ is a cumulative function (appears as a staircase), and its upper and lower bounds vary between 0 and 1. The probability mass function (PMF) of a discrete random variable (x) is defined as:

$$p(x) = P(X = x) \quad (2.9)$$

In which $p(x)$ is the probability mass at a discrete point, $X = x$. The probability mass of discrete random variables must satisfy the following conditions:

$$\begin{cases} p(x_i) \geq 0 \\ \sum p(x_i) = 1 \end{cases} \quad \text{for all } x_i$$

The probability density function (PDF) of a continuous random variable can be determined as:

$$f(x) = \frac{dF(x)}{dx} \quad (2.10)$$

in which $f(x)$ and $F(x)$ are the PDF and CDF of random variable X , respectively. The PDF of a continuous random variable should also satisfy the two following conditions:

$$\begin{cases} f(x) \geq 0 \\ \int_{-\infty}^{+\infty} f(x)dx = 1 \end{cases} \quad \text{for all } x_i$$

The CDF of a continuous and discrete random variable (X) with regards to a known PDF is obtained by using the following equations:

$$F(x) = \int_{-\infty}^x f(x)d(x) \quad (2.11)$$

and

$$F(x_n) = \sum_{1 \leq i \leq n} p(x_i) \quad (2.12)$$

2.4 Moments of Distribution

As previously mentioned, the probability distribution of a random variable is a mathematical function that quantifies the likelihood a desired event will occur. However, the problem is practically faced by engineers have been in determining the mathematical function of probability distribution for a particular random variable. In other words, the probability distribution of random variables is often unknown, and usually only a sample of data is known. Hence, descriptive parameters (sometimes called moments), are used to define the distribution at the nearest approximation. The N th moment from the origin of a discrete event is defined as:

$$u'_N = \sum_{i=-\infty}^{\infty} x_i^N \cdot p(x_i) \quad (2.13)$$

and for a continuous event, is:

$$u'_N = \int_{-\infty}^{\infty} x^N f(x) dx \quad (2.14)$$

The first moment about the origin is called the mean or expected value, and it is obtained as:

$$E(x) = \mu_x = \sum_{i=-\infty}^{\infty} x_i P(x_i) \quad (\text{for a discrete PDF}) \quad (2.15)$$

where $E(\cdot)$ is expectation operator. For a continuous case, the expected value can be determined as follow:

$$E(x) = \mu_x = \int_{-\infty}^{\infty} x f(x) dx \quad (\text{for a continuous PDF}) \quad (2.16)$$

The mean (sometimes referred to location parameter) connotes a measure of central tendency and it indicates where the distribution bulk is located along the x -axis. Similar to the presented formula for random variable X , the expected value for a function $g(x)$ of a random variable x is:

$$E[g(x)] = \sum_{i=-\infty}^{\infty} g(x_i) P(x_i) \quad (2.17)$$

in which x is a discrete random variable.

When x is a continuous random variable, $E[g(x)]$ can be estimated from the following equation:

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx \quad (2.18)$$

Two useful and practical expectation operational properties for the expected value are presented in Eqs. (2.19–2.23).

$$E\left(\sum_{i=1}^k a_i X_i\right) = \sum_{i=1}^k a_i E(X_i) \quad (2.19)$$

$$E\left(\prod_{i=1}^k X_i\right) = \prod_{i=1}^k E(X_i) \quad (2.20)$$

As expectation is a linear operator, for constant values of a and b we have (Mays and Tung 1992):

$$E(a) = a \quad (2.21)$$

$$E(bx) = bE(x) \quad (2.22)$$

$$E(a + bx) = a + bE(x) \quad (2.23)$$

In higher order problems, the moments about the central or mean (central moments) are more interesting than the moments about the origin. The n th central moment of the probability mass function (PMF) of random variable x is calculated as:

$$\mu_n = E[(x - E[x])^n] = \sum_{-\infty}^{\infty} (x_i - \mu)^n \cdot P(x_i) \quad (2.24)$$

and for a probability density function (PDF):

$$\mu_n = E[(x - \mu_x)^n] = \int_{-\infty}^{\infty} (x - \mu)^n f(x) dx \quad (2.25)$$

For instance, the zero and first central moments are:

$$\begin{aligned} \mu_0 &= E[(x - E[x])^0] = E[1] = 1 \\ \mu_1 &= E[(x - E[x])^1] = E[(x - E[x])] = E[x] - E[\mu] = 0 \end{aligned}$$

The second central moment which is called variance, is calculated as follow for discrete random variables:

$$Var(x) \equiv \sigma^2 = \mu_2 = E[(x - \mu_x)^2] = \sum_{-\infty}^{\infty} (x_i - \mu)^2 \cdot P(x_i) \quad (2.26)$$

and for a continues case, is:

$$Var(x) \equiv \sigma^2 = \mu_2 = E[(x - \mu_x)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \quad (2.27)$$

Variance (σ_x^2) is the expected value of the squared difference between a random variable value and its mean; and it measures the dispersion of a set of data around their expected value. In other words, it shows how far a set of numbers is spread out. The square root of the variance is called standard deviation (σ) and can simply be calculated as $\sigma_x = \sqrt{\sigma_x^2}$. The expectation operator properties is used to compute the variance of random variable x as:

$$\begin{aligned} Var(x) &= E[(x - \mu_x)^2] = E(x^2 - 2x\mu_x + \mu_x^2) \\ &= E(x^2) - E(2x\mu_x) + E(\mu_x^2) \\ &= E(x^2) - 2\mu_x E(x) + \mu_x^2 = E(x^2) - 2\mu_x^2 + \mu_x^2 \\ &= E(x^2) - \mu_x^2 \\ &= E(x^2) - [E(x)^2] \end{aligned} \quad (2.28)$$

In this case, useful and applicable variance relationships are presented through the Eqs. (2.29–2.32):

$$Var(x = a) = E(a^2) - [E(a)]^2 = a^2 - a^2 = 0 \quad (2.29)$$

and

$$Var(bx) = E[(bx)^2] - [E(bx)]^2 = b^2 E(x^2) - b^2 [E(x)]^2 = b^2 Var(x) \quad (2.30)$$

and

$$Var(a + bx) = Var(a) + Var(bx) = b^2 Var(x) \quad (2.31)$$

finally, for all independent random variables, we have:

$$Var \sum_{i=1}^n c_i X_i = \sum_{i=1}^n c_i^2 \sigma_i^2 \quad (2.32)$$

where a , b and c_i are constant values.

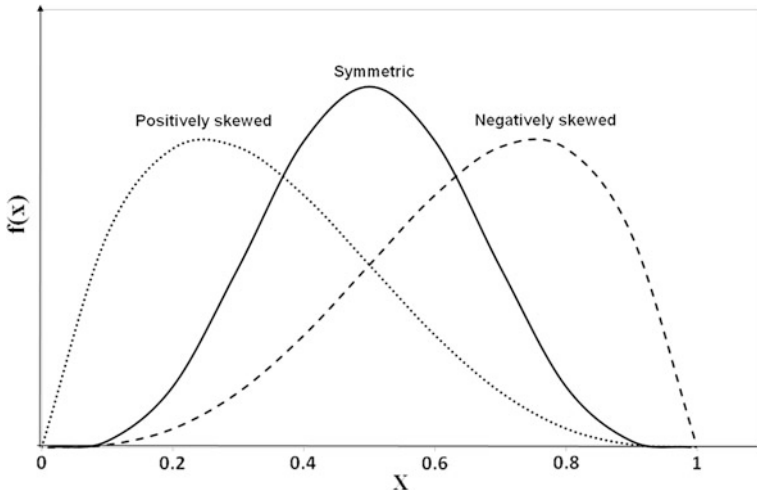


Fig. 2.1 The positive, negative, and symmetric skewed distributions

The third central moment is skewness that measures degree of asymmetry of a probability distribution for desired random variables. The skewness of continuous random variable x is obtained as:

$$\mu_3 = E[(x - \mu_x)^3] = \int_{-\infty}^{\infty} (x - \mu_x)^3 \cdot f(x) dx \quad (2.33)$$

and for a discrete case, is:

$$\mu_3 = \sum_{i=1}^n (x_i - \mu_x)^3 f(x_i) \quad (2.34)$$

The skewness coefficient is a dimensionless value which determines the coefficient of skewness of the probability distribution for random variable x , and it is the ratio of third moment to the cube of standard deviation, as:

$$\gamma_x = \frac{E[(x - \mu_x)^3]}{\sigma_x^3} \quad (2.35)$$

The skewness coefficient is a dimensionless coefficient that its sign shows the degree of PDF symmetry. If $\gamma_x < 0$ distribution is skewed to the left or has a long tail to the left (negatively skewed); while for $\gamma_x > 0$ the distribution is skewed to the right and its tail will be heavier there (positively skewed). If the distribution is neither negatively or positively skewed, the mean of distribution is equal to the median and the skewness will be close to zero Fig. 2.1 (Mays and Tung 1992).

The fourth central moment is called kurtosis and it shows the degree of tallness (or peakedness) and flatness of a probability distribution function. High kurtosis connotes a distribution with sharper peak, while low kurtosis means a distribution with flatter peak. The kurtosis value of a density function is:

$$k_x = \frac{E[(X - \mu_x)^4]}{\sigma_x^4} \quad (2.36)$$

where, k_x is kurtosis and $k_x > 0$.

The dispersion of random data around their mean is measured by coefficient of variation (CV_x) and it is defined as the ratio of standard deviation to the mean of data series:

$$CV_x = \Omega_x = \frac{\sigma_x}{\mu_x} \quad (2.37)$$

This coefficient is commonly used for positive variables to see the degree of variation between desired data series. The other parameter in statistics is median or x_m , which is the middle value of distribution or given values in sample data. If the number of sample data is odd, the median is the middle number, and if it is even, the median is the average of the two middle values. In other words, the median splits the distribution into two equal parts and accordingly the area under the distribution function (probability of occurrence) up to the median point equals to 0.5:

$$F(x_m) = \int_{-\infty}^{x_m} f(x)dx = 0.5 \quad (2.38)$$

2.4.1 Estimate Moments From Data Series

The total set of observation, x_1, x_2, \dots, x_n , with finite or infinite sample data length is referred to the *population*. As previously mentioned, finding appropriate probability distribution for series of observed data or population is a complex issue. For instance, hydrological data are often a combination of different physical processes (e.g. direct runoff is combination of rainfall, snowmelt and infiltration), and each process has its own probability density function. Hence the target function usually is an integrated mixture of those density functions. What is more obvious is the fact that observed data are naturally prone to observational errors and it is too hard to find a perfect fit for them, and therefore, attaining probability distribution is very difficult and sometimes even remains unknown. Regarding this difficulty, various statistical parameters must be obtained directly from observed data and then applied in the desired analysis. The first and foremost parameter is mean or expected value of n independent data which is calculated as:

$$\mu_x = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.39)$$

in which μ_x or \bar{x} is the mean operator and x_i is the value of each observation. In addition to the presented formula, the AVERAGE function in Excel can also be applied to evaluate average or mean of samples as:

$$= \text{AVERAGE} (\text{select the range of data from cells})$$

The variance of observations is obtained by:

$$\text{Var}(X) = \sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} \quad (2.40)$$

in which σ_x^2 is the variance of desired data series. The VAR and STDEV functions in Excel can be used to calculate the variance σ_x^2 and standard deviation σ_x of observations as:

$$\begin{aligned} &= \text{VAR}(\text{select the range of data from cells}) \\ &= \text{STDEV}(\text{select the range of data from cells}) \end{aligned}$$

The variance of the mean of a variable shows how far is the estimated mean from the true mean in desired sample space. In other words, this parameter can be applied to measure the error of expected value of a data set and it is defined as:

$$\text{Var}(\bar{x}) \equiv \sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n} \quad (2.41)$$

To measure asymmetry of a data series, skewness is estimated as:

$$\gamma_x = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum (x_i - \bar{x})^3}{\sigma_x^3} \quad (2.42)$$

in which γ_x is the skewness.

Skewness includes a cubed summation of deviations from the mean of data and is subject to large computation errors. It is important to note that when the mean of the observations is less than the median, skewness is negative, while positive skewness indicates that the mean of the observations is larger than the median. To compute the skewness of desired data sets, the SKEW function in Excel can be applied as:

$$= \text{SKEW} (\text{select the range of data from cells})$$

Example 2.3 Observed flow data of the Kor River in Nourabad, Iran, is available from 1960 to 1989 and presented in the following table. Determine the mean, variance, standard deviation and skewness.

Years	Flow (Q , $cf s$)	Years	Flow (Q , $cf s$)
1960	5759.278	1975	8742.511
1961	1113.213	1976	5984.455
1962	2531.58	1977	8685.196
1963	1716.61	1978	8752.519
1964	4718.362	1979	7366.032
1965	9248.881	1980	868.9257
1966	399.9453	1981	6121.915
1967	2306.414	1982	8277.696
1968	2947.42	1983	4654.374
1969	5472.883	1984	6820.767
1970	6096.213	1985	2575.634
1971	2589.628	1986	7508.591
1972	2111.904	1987	8238.025
1973	3945.78	1988	29.47444
1974	9789.477	1989	7021.186

Solution The mean and variance are:

$$\bar{Q} = \frac{1}{n} \sum_{i=1}^n Q_i = \frac{1}{30} \times 152394.9 = 5079.83$$

and

$$\begin{aligned} \sigma_Q^2 &= \frac{\sum Q_i^2 - n\bar{Q}^2}{n-1} = \frac{(1029079151.172) - (30 \times 5079.83^2)}{30-1} \\ &= 8790998.838 \end{aligned}$$

Thus, standard deviation will be computed as:

$$\sigma_Q = \sqrt{\text{Var}(Q)} = \sqrt{8790998.838} = 2964.96$$

The skewness of observed flows can be calculated by using Eq. (2.43) as:

$$\gamma_Q = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum (Q_i - \bar{Q})^3}{\sigma_Q^3} = -0.11913$$

Furthermore, EXCEL functions can be applied to calculate desired parameters quickly.

Example 2.4 Consider the following density function over the interval $[0, 1]$ as:

$$f(x) = x + 3$$

Calculate $E[x]$, $E[4x]$, and $E[4x + 5]$.

Solution As the desired density function is in continuous form, the expectation of x is computed from Eq. (2.16) as follow:

$$\begin{aligned} E[X] &= \int_0^1 xf(x)dx = \int_0^1 x \times (x+3)dx \\ &= \left[\frac{x^3}{3} + \frac{3x^2}{2} \right]_0^1 = 1.83 \end{aligned}$$

and from Eq. (2.22):

$$E[4X] = 4E[X] = 4 \times 1.83 = 7.33$$

and based on Eq. (2.23):

$$E[4X + 5] = 4E[X] + 5 = 4 \times 1.83 + 5 = 12.32$$

2.5 Two Random Variables

When the nature of a problem includes two or more random variables, the relationships among these random variables are determined by joint probability distributions. All statistical properties of one random variable presented above can also be defined for two or more random variables. If two dependent random variables are considered, then μ_x and μ_y are obtained by applying the following equations:

$$\mu_x = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x,y)dx dy \quad (2.43)$$

$$\mu_y = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yf(x,y)dx dy \quad (2.44)$$

The correlation coefficient is a measure of how good trends are among two variables. When there are two random variables, the degree of linear dependence between X and Y is measured by the correlation coefficient $\rho(X, Y)$ as:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \rightarrow Cov(X, Y) = \rho(X, Y) \cdot \sigma_X \sigma_Y \quad (2.45)$$

where $Cov(X, Y)$ is the covariance between random variables X and Y . Covariance shows the correlation strength between random variables and it is defined as:

$$\begin{aligned} Cov(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X - \mu_X)(Y - \mu_Y)f(X, Y)dxdy \\ &= E[XY] - \mu_X\mu_Y \end{aligned} \quad (2.46)$$

or

$$Cov(X, Y) = \sum_{i=1}^n \sum_{j=1}^m (X_i - \bar{X})(Y_j - \bar{Y})\rho(X_i, Y_j) \quad (2.47)$$

The covariance value for two independent random variables is zero since there is no correlation between them, or, $\rho(X, Y) = 0$. The CORREL and COVAR functions in Excel can be used to calculate the correlation coefficient and covariance between two sets of data, respectively.

= CORREL (select two arrays of data from cells)

= COVAR (select two arrays of data from cells)

Some useful relations for estimating the variance of two dependent random variables are:

$$Var(X + Y) = Var(X) + Var(Y) + 2COV(X, Y) \quad (2.48)$$

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCOV(X, Y) \quad (2.49)$$

If Y is considered as sum of n random variables, $Y = \sum_{i=1}^n X_i$, then the variance of Y is estimated using the following formula:

$$Var(Y) = \sum_{i=1}^n Var(X_i) + 2 \sum_{i < j}^n Cov(X_i, X_j) \quad (2.50)$$

Therefore, Eq. (2.32) will be completed as (Mays and Tung 1992):

$$Var\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \sigma_i^2 + 2 \sum_{i < j}^k c_i c_j Cov(X_i, X_j) \quad (2.51)$$

Example 2.5 The inflow and outflow data with related hydrographs for a particular reservoir are presented in the following table and figure.

$t(\text{day})$	$I(\text{cfs})$	$Q(\text{cfs})$	$t(\text{day})$	$I(\text{cfs})$	$Q(\text{cfs})$
1	0.00	0.00	13	78.30	137.46
2	8.70	0.00	14	52.20	115.71
3	30.45	2.610	15	34.80	93.96
4	65.25	12.18	16	17.40	74.82
5	126.15	31.32	17	8.70	53.94
6	182.70	64.38	18	0.00	38.28
7	195.75	106.14	19	0.00	26.10
8	143.55	140.94	20	0.00	17.40
9	139.20	143.55	21	0.00	10.44
10	136.59	139.20	22	0.00	6.09
11	121.80	142.68	23	0.00	2.61
12	113.10	149.64	24	0.00	7.83

Compute the covariance and correlation coefficient between these two series of data (Fig. 2.2).

Solution The covariance is computed as:

$$\text{Cov}(I, Q) = \frac{1}{n-1} \sum_{i=1}^n (I_i - \bar{I})(Q_i - \bar{Q}) = 2642.444$$

$$\sigma_I = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (I_i - \bar{I})^2} = 66.4$$

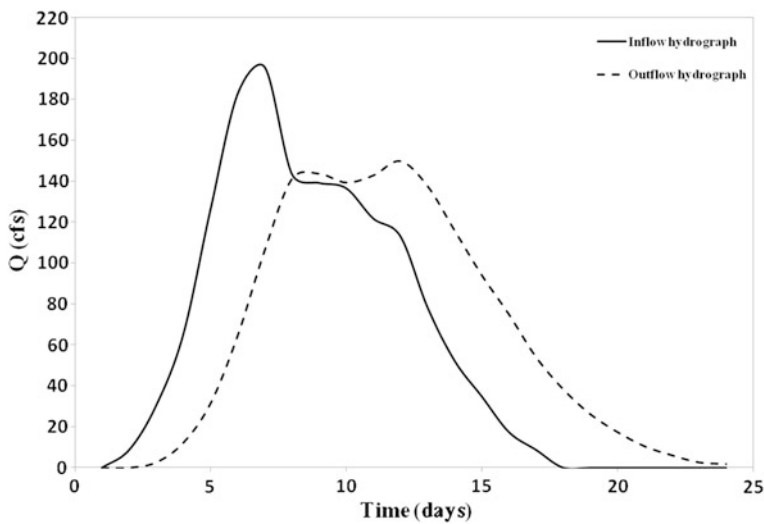


Fig. 2.2 The inflow and outflow hydrographs

$$\sigma_Q = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Q_i - \bar{Q})^2} = 57.295$$

$$\rho(I, Q) = \frac{\text{Cov}(I, Q)}{\sigma_I \sigma_Q} = \frac{2642.444}{66.4 \times 57.295} = 0.694$$

All of the process of computing covariance and correlation coefficient are presented in the following table.

t	$I(\text{cfs})$	$Q(\text{cfs})$	$I - \bar{I}$	$O - \bar{O}$	$I - \bar{I} \times O - \bar{O}$	$(I - \bar{I})^2$	$(O - \bar{O})^2$
1	0.00	0.00	-60.61	-63.22	3831.764	3,673.572	3,996.768
2	8.70	0.00	-51.91	-63.22	3281.75	2,694.648	3,996.768
3	30.45	2.61	-30.16	-60.61	1,827.998	909.6256	3,673.572
4	65.25	12.18	4.64	-51.04	-236.826	21.5296	2,605.082
5	126.15	31.32	65.54	-31.9	-2,090.73	4,295.492	1,017.61
6	182.70	64.38	122.09	1.16	141.6244	14,905.97	1.3456
7	195.75	106.14	135.14	42.92	5,800.209	18,262.82	1,842.126
8	143.55	140.94	82.94	77.72	6,446.097	6,879.044	6,040.398
9	139.20	143.55	78.59	80.33	6,313.135	6,176.388	6,452.909
10	136.59	139.20	75.98	75.98	5,772.96	5,772.96	5,772.96
11	121.80	142.68	61.19	79.46	4,862.157	3,744.216	6,313.892
12	113.10	149.64	52.49	86.42	4,536.186	2,755.2	7,468.416
13	78.30	137.46	17.69	74.24	1,313.306	312.9361	5,511.578
14	52.20	115.71	-8.41	52.49	-441.441	70.7281	2,755.2
15	34.80	93.96	-25.81	30.74	-793.399	666.1561	944.9476
16	17.40	74.82	-43.21	11.6	-501.236	1,867.104	134.56
17	8.70	53.94	-51.91	-9.28	481.7248	2,694.648	86.1184
18	0.00	38.28	-60.61	-24.94	1,511.613	3,673.572	622.0036
19	0.00	26.10	-60.61	-37.12	2,249.843	3,673.572	1,377.894
20	0.00	17.40	-60.61	-45.82	2,777.15	3,673.572	2,099.472
21	0.00	10.44	-60.61	-52.78	3,198.996	3,673.572	2,785.728
22	0.00	6.09	-60.61	-57.13	3,462.649	3,673.572	3,263.837
23	0.00	2.61	-60.61	-60.61	3,673.572	3,673.572	3,673.572
24	0.00	7.83	-60.61	-55.39	3,357.188	3,673.572	3,068.052
<i>Average</i>	60.61	63.22		<i>Sum</i>	60,776.29	101,418	7,5504.81

Example 2.6 The volume of storm runoff in a watershed is dependent on several variables. Based on the SCS rainfall-runoff relationship, the total rainfall (P) over the watershed can be divided into three parts: direct runoff (Q), initial abstraction (I_a), and actual retention (F). Hence, the actual retention is computed as:

$$F = P - I_a - Q$$

Each of these parameters is shown schematically in Fig. 2.3 (McCuen 2005).

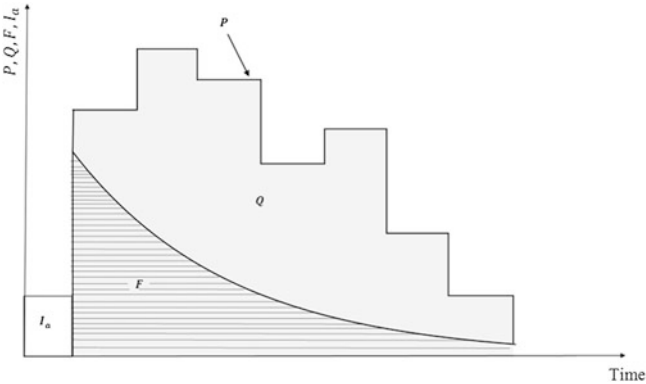


Fig. 2.3 Different parts of total rainfall (McCuen 2005)

Consider the following values for mean and standard deviation of P , I_a , and Q in a small watershed:

Variables	Mean (in)	Standard deviation (in)
P	40	6.0
I_a	5.0	1.2
Q	25	3.4

Determine the mean and standard deviation of the actual retention by considering the total rainfall, direct runoff, and initial abstraction as:

1. Independent random variables,
2. Dependent random variables with $\rho(P, Q) = 0.5, \rho(P, I_a) = 0.15, \rho(Q, I_a) = -0.4$

Solution

1. For independent random variables, the mean and variance can be computed as:

$$\begin{aligned} E(F) &= E(P) + E(-I_a) + E(-Q) \\ &= E(P) - E(I_a) - E(Q) \\ &= 40 - 5 - 25 = 10 \text{ in} \end{aligned}$$

and

$$\begin{aligned} Var(F) &= Var(P) + Var(-I_a) + Var(-Q) \\ &= Var(P) + Var(I_a) + Var(Q) \\ &= (6)^2 + (1.2)^2 + (3.4)^2 = 49 \text{ in}^2 \end{aligned}$$

Therefore, the standard deviation of F is:

$$\sigma_F = \sqrt{49} = 7 \text{ in}$$

2. When there is dependency between random variables, we have:

$$\begin{aligned} \text{Var}(F) &= \text{Var}(P) + \text{Var}(-I_a) + \text{Var}(-Q) + 2[1 \times (-1)]\text{Cov}(P, Q) \\ &\quad + 2[1 \times (-1)]\text{Cov}(P, I_a) + 2[(-1) \times (-1)]\text{Cov}(Q, I_a) \\ &= \text{Var}(P) + \text{Var}(I_a) + \text{Var}(Q) - 2\text{Cov}(P, Q) - 2\text{Cov}(P, I_a) + 2\text{Cov}(Q, I_a) \\ &= \text{Var}(P) + \text{Var}(I_a) + \text{Var}(Q) - 2\rho(P, Q) \cdot \sigma_P \sigma_Q - 2\rho(P, I_a) \cdot \sigma_P \sigma_{I_a} + 2\rho(Q, I_a) \cdot \sigma_Q \sigma_{I_a} \\ &= [(6)^2 + (1.2)^2 + (3.4)^2] - [2 \times 0.5 \times 6 \times 3.4] - [2 \times 0.15 \times 6 \times 1.2] \\ &\quad + [2 \times (-0.4) \times 3.4 \times 1.2] = 49 - 20.4 - 2.16 + (-3.264) = 23.176 \end{aligned}$$

Hence, the standard deviation of actual retention is:

$$\sigma_F = \sqrt{23.176} = 4.814$$

2.6 Commonly Used Probability Distributions in Hydrosystem Engineering

There are several probability distributions that are frequently used in hydrosystem engineering, and they can be classified into discrete and continuous types. Binomial and Poisson distributions are two of the most common discrete probability distributions, while there are a variety of continuous probability distributions including Normal, Log-Normal, Gamma, Beta, Exponential, etc. that are applied for various statistical analyses in hydrosystem engineering. In the following sections, the mathematical formulas and statistical properties of a number of discrete and continuous probability distributions are presented.

2.6.1 Binomial Distribution

Binomial distribution describes the probability distribution of a discrete event and is used to estimate the probability of possible numbers of success (x) through n trials. Binomial distribution is used when the occurrence of an event, not its magnitude, is desired. As this distribution has a binary base, the outcomes is either success or fail. If the probability of success and fail occurrence is denoted by p and q , respectively, the binomial probability mass function (PMF) can be written as:

$$p(x) = f_x(x) = {}_n C_x p^x q^{n-x} = B(n, p), \quad x = 0, 1, 2, \dots, n \quad (2.52)$$

where $q = 1 - p$ and ${}_nC_x$ is a Binomial coefficient:

$${}_nC_x = \binom{n}{x} = \frac{n!}{x!(n-x)!} \quad (2.53)$$

If random variable X follows a Binomial distribution, then mean, variance and skewness of X are estimated as follows:

$$E[X] = np \quad (2.54)$$

$$Var(X) = npq \quad (2.55)$$

$$\gamma_x = \frac{1 - 2p}{\sqrt{np(1-p)}} \quad (2.56)$$

The values p and q produce the shape of the probability mass function (PMF) of a Binomial random variable; in which for $p < q$ the PMF is positively skewed, and when $p > q$ the PMF is negatively skewed, and for $p = q = 0.5$ it is symmetric (Figs 2.4 and 2.5).

Example 2.7 Calculate the probability of having x successes in seven trials ($n = 7$) for the following p values; $p = 0.1$, $p = 0.5$, and $p = 0.8$. Plot the Binomial PMFs and compare the results.

Solution All calculations are shown in the following table.

n	p	x	$p(x)$	n	p	x	$p(x)$	n	p	x	$p(x)$
7	0.1	0	4.783E-01	7	0.5	0	7.813E-03	7	0.8	0	1.280E-05
		1	3.720E-01			1	5.469E-02			1	3.584E-04
		2	1.240E-01			2	1.641E-01			2	4.301E-03
		3	2.296E-02			3	2.734E-01			3	2.867E-02
		4	2.552E-03			4	2.734E-01			4	1.147E-01
		5	1.701E-04			5	1.641E-01			5	2.753E-01
		6	6.300E-06			6	5.469E-02			6	3.670E-01
		7	1.000E-07			7	7.813E-03			7	2.097E-01

If $p = 0.1$, then $q = 1 - 0.1 = 0.9$, hence $p < q$, and the PMF is positively skewed. Otherwise, if $p = 0.8$, then $q = 1 - 0.8 = 0.2$. Thus, $p > q$ and the PMF is negatively skewed. The BINOMDIST function in Excel can be used to calculate the probability of having x successes in n trials as:

$$= \text{BINOMDIST}(\text{select : the range of } x, n, p, \text{ True or False})$$

It should be noted that TRUE returns the cumulative distribution function, and FALSE returns the probability mass function.

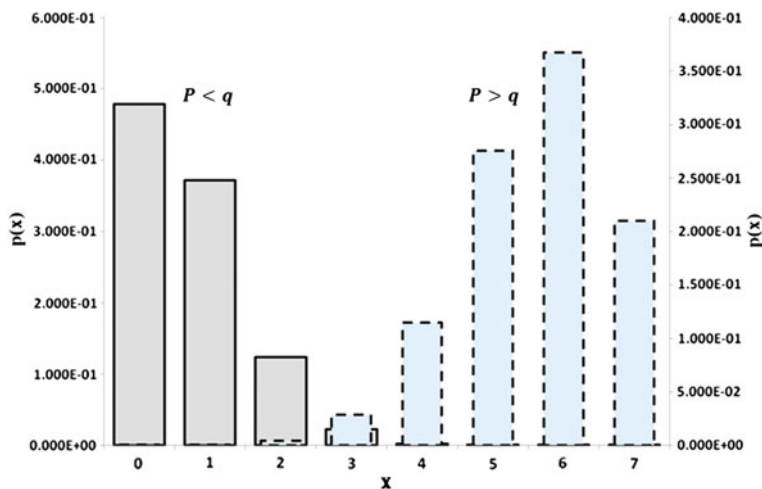


Fig. 2.4 The positive and negative skewed of PMF

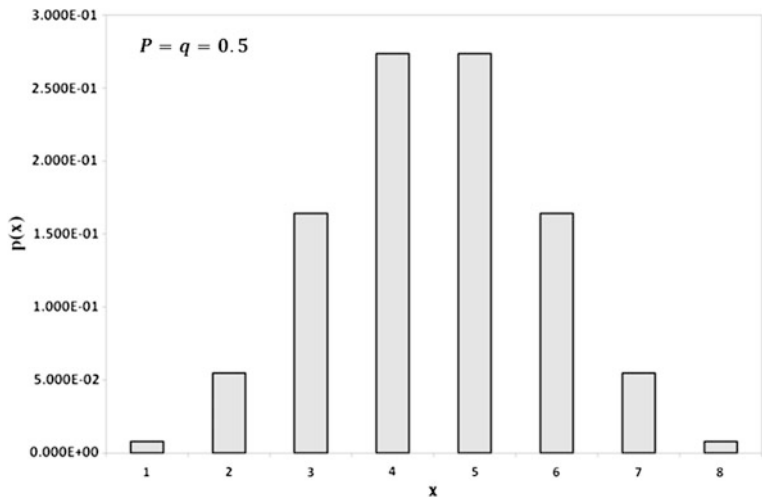


Fig. 2.5 Symmetric distribution

Example 2.8 A small dam is built to withstand a flood of up to 10,000 cfs. If two floods greater than 10,000 cfs occur in 10 years, the downstream facilities are repairable, otherwise, for more than two floods, the damage is considerable and property loss occurs. Assume the annual probability of occurring a flood more than 10,000 cfs is 0.2. What is the probability that no property loss occurs?

Solution An interesting event would be the occurrence of a flood exceeding 10,000 cfs in 10 years, with probability of 0.2 in each trial ($p = 0.2$). The desired

period is 10 years, therefore there are 10 trials ($n = 10$) and, there will be no losses if maximum two floods occur during the 10 year period. The probability of having at most two floods for $n=10$ is calculated by:

$$\begin{aligned}
 p(x \leq 2) &= p(x = 0) + p(x = 1) + p(x = 2) \\
 &= {}_{10}C_0(0.2)^0(1 - 0.2)^{10} + {}_{10}C_1(0.2)^1(1 - 0.2)^9 + {}_{10}C_2(0.2)^2(1 - 0.2)^8 \\
 &= \frac{10!}{0!(10 - 0)!} \times 0.8^{10} + \frac{10!}{1!(10 - 1)!} \times 0.2 \times 0.8^9 + \frac{10!}{2!(10 - 2)!} \times 0.2^2 \times 0.8^8 \\
 &= 0.677
 \end{aligned}$$

Example 2.9 Return period (T) is the expected time between particular event occurrences like floods or rainfalls with a definite size or intensity. In general, several floods with a return period of T -year may occur in a watershed during the course of a year. What is the probability that a 100 year flood will occur at least once in 10 years? Compute the probability of a 100 year flood not occurring in 10 years.

Solution The probability of flooding occurs in known return period T is defined as $p = 1/T$. Therefore, for x as number of T -year flood occurring in n years, x is function of n and p as $x \sim B(n, p)$. The probability of a 100 year flood occurring in 10 years is computed as:

$$p = \frac{1}{T} = \frac{1}{100} = 0.01$$

and

$$\begin{aligned}
 p(x \geq 1) &= 1 - p(0) = 1 - \binom{10}{0} (0.01)^0 (1 - 0.01)^{10} \\
 &= 1 - 0.9044 = 0.0956
 \end{aligned}$$

and the probability that a 100 year flood will not occur in 10 years is:

$$\begin{aligned}
 q &= p(x = 0) = (1 - p)_{1st\ year} \times (1 - p)_{2nd\ year} \times \dots \times (1 - p)_{nth\ year} \\
 &= (1 - p)^n = \left(1 - \frac{1}{T}\right)^n
 \end{aligned}$$

The incidence probability of at least one desired flood in n years is known as risk of flooding. Thus, the risk is defined as the sum of probabilities of one flood, two floods ... n floods occurring during the n year period. Therefore, in the case of this example risk is:

$$\begin{aligned}
 \text{Risk} &= 1 - p(x = 0) = 1 - \left(1 - \frac{1}{T}\right)^n \\
 \text{Risk} &= 1 - \left(1 - \frac{1}{100}\right)^{10} = 0.0956
 \end{aligned}$$

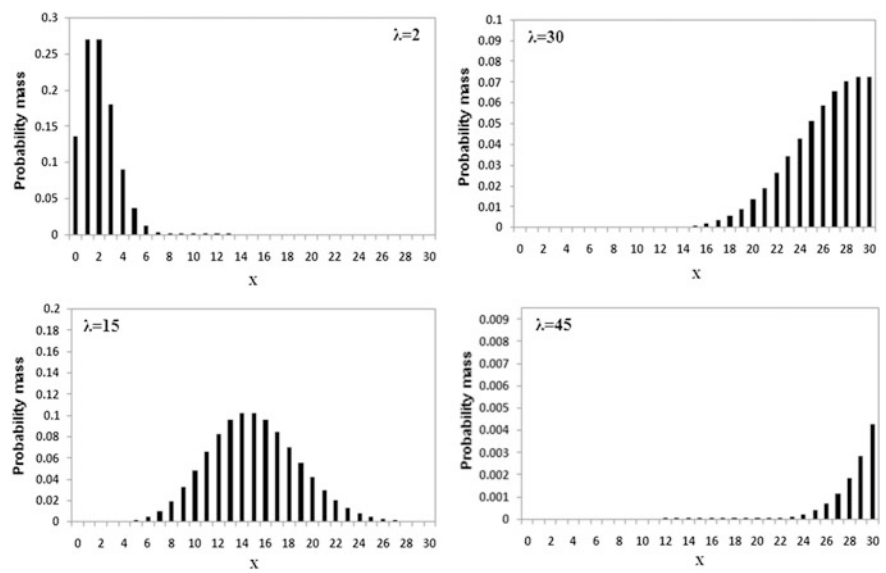


Fig. 2.6 The poisson distribution with different values of λ

2.6.2 Poisson Distribution

The Binomial distribution is not appropriate when there is small probability of occurrence (p) and large number of observations (n). On this basis, it is better to approximate the Binomial distribution with Poisson distribution as they come close to each other for large n and small p . This distribution is characterized only with λ as distribution shape parameter and it shows the number of successes in a particular time interval $[(0, t)]$. The probability mass function (PMF) of the Binomial distribution is:

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots \tag{2.57}$$

in which x is a discrete random variable and λ is a positive value. The probability mass functions of Poisson distribution are plotted for different values of λ in Fig. 2.6.

The mean, variance, coefficient of variation, and skewness of Poisson distribution are presented in Table 2.1.

Table 2.1 The statistical features of Poisson distribution

$E(x)$	$Var(x)$	CV_x	γ_x
λ	λ	$\frac{1}{\sqrt{\lambda}}$	$\frac{1}{\sqrt{\lambda}}$

The POISSON function in Excel can be used to calculate the probability of having x successes in n trials as follow:

$$= \text{POISSON}(\text{select} : \text{the range of, mean value } (\lambda), \text{ True or False})$$

It should be noted that TRUE returns the cumulative Poisson probability, and FALSE returns the Poisson probability mass function.

Example 2.10 Consider the random variable In Example 2.8 follows the Poisson distribution. Compute the probability of no losses.

Solution The parameter λ is computed as:

$$\lambda = np = 10 \times 0.2 = 2.0$$

This value shows the expected (or average) number of occurrences of a flood larger than 10,000 cfs over a 10 year period.

$$P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2)$$

and

$$\begin{aligned} p(x) &= \frac{2^0 \times e^{-2}}{0!} + \frac{2^1 \times e^{-2}}{1!} + \frac{2^2 \times e^{-2}}{2!} \\ &= 0.135 + 0.270 + 0.270 = 0.676 \end{aligned}$$

2.6.3 Normal Distribution

Normal distribution is one of the most commonly used probability distributions by water resources engineers. It is also known as Gaussian distribution in honor of Carl Friedrich Gauss. Normal distribution is identified from its mean (μ_x) which shows the location of the distribution center, and the variance (σ_x^2). Hence, the normal random variable x with mean (μ_x) and variance (σ_x^2) can be shown as $N \sim (\mu_x, \sigma_x^2)$. The variance is always positive or zero because the squares are positive or zero, while the mean could have a negative or positive value. The PDF of a normal distribution is defined as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp \left[-\frac{1}{2} \left(\frac{x - \mu_x}{\sigma_x} \right)^2 \right] \quad \text{for} \quad -\infty < x < \infty \quad (2.58)$$

The PDF of normal distribution is a bell-shaped curve with a peak at the mean, uni-modal, symmetric, and extends to $\pm\infty$. The biggest distribution concentration is located in the center and it decreases along the x -axis. This characteristic demonstrates that normal distribution does not produce uncommonly extreme values in comparison to some other distributions. Figure 2.7 shows the probability

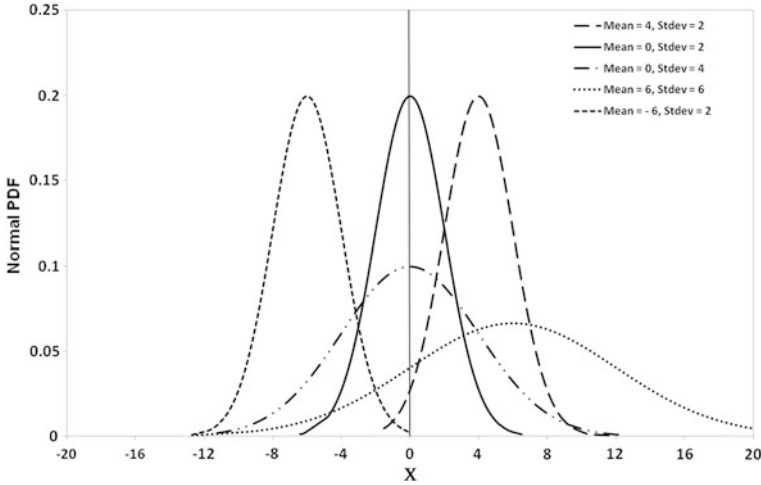


Fig. 2.7 the PDF of normal distribution in different values of μ_x and σ_x

density function (PDF) of normal distribution for different values of mean and standard deviation. Based on the central limit theorem, normal distribution can be used as a simple model to explain complex events when there is sufficiently large number of independent random variables. This theory describes the population of the means of a large number of independent random variables which all of them are drawn from a given parent distribution, have mean always equal to the mean of the parent population, and standard deviation equal to the standard deviation of the parent population divided by the square root of the sample size. This theory demonstrates distribution of means will approximate a normal distribution as the size of samples increases.

Standard normal distribution is a normal distribution with mean and standard deviation of 0 and 1, respectively. The standard type uses a transferring factor (z) in the following form:

$$z = \frac{x - \mu_x}{\sigma_x} \quad (2.59)$$

in which x is the normal random variable with mean μ_x and standard deviation σ_x . On the other hand, for a known z value, the normal random variable x with mean μ_x and standard deviation σ_x is computed as:

$$x = \sigma_x z + \mu_x \quad (2.60)$$

The probability density function (PDF) of standard normal distribution z is:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad \text{for } -\infty < z < \infty \quad (2.61)$$

where $z \sim N(0, 1)$. The process of normal random variable standardization allows to calculate the PDF and CDF of normal distribution using the existing standard normal PDF and CDF tables. The probability of a normal random variable $X \sim N(\mu_x, \sigma_x^2)$ is performed as:

$$\begin{aligned} P(X \leq x) &= P\left[\frac{X - \mu_x}{\sigma_x} \leq \frac{x - \mu_x}{\sigma_x}\right] \\ &= P[Z \leq z] = \Phi(z) \end{aligned} \quad (2.62)$$

where $\Phi(z)$ is the CDF of standard normal random variable z and it is computed as:

$$\Phi(z) = \int_{-\infty}^z \phi(s) ds \quad (2.63)$$

In addition to Eq. (2.63), the standard normal tables are used to compute $\Phi(z)$ for different values of z (see appendix B). Furthermore, the NORMDIST and NORMINV functions in Excel return the normal distribution and the inverse of the normal cumulative distribution for the specified mean and standard deviation, respectively as follow:

$$\begin{aligned} &= \text{NORMDIST}(\text{select : the range of } x, \mu_x, \sigma_x, \text{ True or False}) \\ &= \text{NORMINV}(\text{probability}, \mu_x, \sigma_x) \end{aligned}$$

On the other hand, the NORMSDIST and NORMSINV functions in Excel return the standard normal cumulative distribution function and the inverse of the standard normal cumulative distribution, respectively as:

$$\begin{aligned} &= \text{NORMSDIST}(z) \\ &= \text{NORMSINV}(\text{probability}) \end{aligned}$$

Example 2.11 Assume the normal distribution is the best fit for inflows into a particular reservoir with mean and standard deviation of 15,352 and 4,785 respectively. Determine:

1. The probability of a 200 year flood, and
2. The probability of a flood less than or equal to 20,000 cfs occurring.

Solution The probability of a 200 year flood is:

$$F(Q_{200}) = 1 - \frac{1}{T} = 1 - \frac{1}{200} = 0.995$$

based on Table B-2 (see Appendix B), we have:

$$F(Q_{200}) = 0.995 \rightarrow z = 2.576$$

then,

$$\begin{aligned} Z &= \frac{Q_{200} - \bar{Q}}{\sigma_Q} \Rightarrow Q_{200} = \bar{Q} + Z \cdot \sigma_Q \\ &= 15352 + 2.576 \times 4785 = 27678.16 \text{ cfs} \end{aligned}$$

the probability of occurring a flood less than or equal to 20,000 cfs is:

$$Z = \frac{20000 - 15352}{4785} = 0.971$$

by linear interpolation in Table B-2, we have:

$$F(0.971) = 0.834 = \text{Prob}(Q \leq 20,000)$$

and finally the return period of 20,000 cfs is:

$$T = \frac{1}{1 - 0.834} \approx 6 \text{ yr}$$

2.6.4 Log-Normal Distribution

Log-normal distribution is a statistical continuous distribution of random variables of which logarithm of variables follows normal distribution. In other words, variable x is log-normally distributed if $y = \ln(x)$ is normally distributed. Hence, various properties of log-normal distribution can be derived from the normal probability distribution. This continuous distribution is commonly used when random variables cannot have a negative value. The PDF of the log-normal random variable is:

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma_{\ln x}^2}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu_{\ln x}}{\sigma_{\ln x}}\right)^2\right], \quad \text{for } 0 < x < \infty \quad (2.64)$$

Three useful relationships are presented in the following section to compute the statistical moments of $\ln(x)$ from variable x as:

$$\sigma_{\ln x}^2 = \ln\left(\frac{\mu_x^2 + \sigma_x^2}{\mu_x^2}\right) = \ln(1 + CV_x^2) \quad (2.65)$$

$$\mu_{\ln x} = \ln(\mu_x) - \frac{1}{2}\sigma_{\ln x}^2 = \ln(\mu_x) - \frac{1}{2}\ln(1 + CV_x^2) \quad (2.66)$$

Figure 2.8 shows the shapes of the log-normal density function for different values of Ω_x in which Ω_x is:

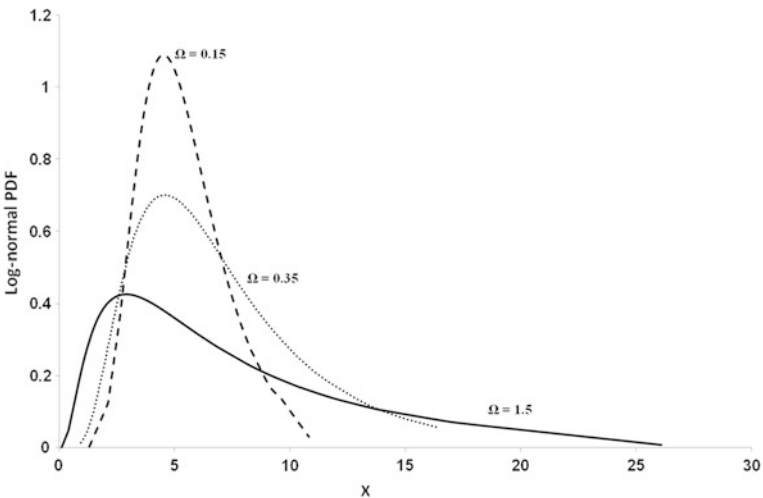


Fig. 2.8 Shape of the log-normal distribution in different values of Ω_X

$$\Omega_x = \sqrt{\exp(\sigma_{\ln x}^2) - 1} \tag{2.67}$$

This figure demonstrates that the log-normal distribution is always positively skewed.

Example 2.12 Assume the log-normal distribution is the best fit to the annual maximum inflows (I) into a reservoir.

- 1. Calculate the probability of inflow exceeding 9,000 cfs each year.
- 2. What is the magnitude of inflow with return period of 200 years?

Years	Inflow(I , cfs)	Years	Inflow (I , cfs)
1960	5,759.278	1975	8,742.511
1961	1,113.213	1976	5,984.455
1962	2,531.58	1977	8,685.196
1963	1,716.61	1978	8,752.519
1964	4,718.362	1979	7,366.032
1965	9,248.881	1980	868.9257
1966	399.9453	1981	6,121.915
1967	2,306.414	1982	8,277.696
1968	2,947.42	1983	4,654.374
1969	5,472.883	1984	6,820.767
1970	6,096.213	1985	2,575.634
1971	2,589.628	1986	7,508.591
1972	2,111.904	1987	8,238.025
1973	3945.78	1988	29.47444
1974	9789.477	1989	7021.186

Solution The mean and variance of presented data are:

$$\mu_I = 5079.82 \text{ cfs}$$

$$\sigma_I = 2964.96 \text{ cfs}$$

1. As observations follow log-normal distribution, the logarithm of observations follows normal distribution with the following mean and standard deviation:

$$\begin{aligned}\mu_{\ln I} &= \ln(5079.82) - \frac{1}{2} \ln \left(\frac{5079.82^2 + 2964.96^2}{5079.82^2} \right) \\ &= 8.386\end{aligned}$$

and

$$\sigma_{\ln I}^2 = \ln \left(\frac{5079.82^2 + 2964.96^2}{5079.82^2} \right) = 0.293$$

The probability that inflow magnitude exceeds 9,000 cfs is:

$$\begin{aligned}P(I \geq 9000) &= P[\ln(I) \geq \ln(9000)] \\ &= 1 - P[\ln I \leq \ln(9000)]\end{aligned}$$

and z will be computed as:

$$z = \frac{\ln(I) - \mu_{\ln I}}{\sigma_{\ln I}}$$

Then, based on Eq. (2.62);

$$\begin{aligned}P(I \geq 9000) &= 1 - P\left(Z \leq \frac{[\ln(I) - \mu_{\ln I}]}{\sigma_{\ln I}}\right) \\ &= 1 - P\left(Z \leq \frac{\ln(9000) - 8.386}{\sqrt{0.293}}\right) \\ &= 1 - \Phi(1.328) = 1 - 0.9079 \cong 0.092\end{aligned}$$

The value of $\Phi(z)$ can be computed from Table B-2 (Appendix B) or NORMSDIST function of Excel.

2. As there is an inverse relation between the return period and the probability of exceeding in any one year, the probability of inflow with a return period of 200 years is:

$$\begin{aligned}P(I \geq q_{200}) &= \frac{1}{T} = \frac{1}{200} = 0.005 \\ P(I \leq q_{200}) &= 1 - P(I \geq q_{200}) = 0.995\end{aligned}$$

Hence,

$$\begin{aligned}
 P(I \leq q_{200}) &= 1 - P(\ln I \geq \ln q_{200}) = 0.995 \\
 0.995 &= P\left[Z \leq \left(\frac{\ln(q_{200}) - \mu_{\ln I}}{\sigma_{\ln I}}\right)\right] \\
 0.995 &= P\left(Z \leq \frac{\ln(q_{200}) - 8.386}{\sqrt{0.293}}\right) \\
 0.99 &= \Phi(z)
 \end{aligned}$$

The value of z can be calculated using the NORMSINV function of Excel. For $\Phi(z) = 0.995$, z equals to 2.575. Then, the magnitude of flow with a 200 year return period is:

$$\begin{aligned}
 \left[\frac{\ln(q_{200}) - 8.386}{\sqrt{0.293}}\right] &= 2.575 \\
 q_{200} &= 17673.718 \text{ cfs}
 \end{aligned}$$

2.6.5 Exponential Distribution

Exponential distribution is a type of continuous probability distribution that describes the time between events in a Poisson process. Exponential distribution is widely employed in different fields of reliability engineering regarding its simplicity. If the random variable t represents the time between events, the probability density function (PDF) of the exponential distribution is defined as:

$$f(t, \lambda) = \begin{cases} f(t) = \lambda \cdot e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (2.68)$$

Parameter λ is the distribution parameter and it varies with the interval $(0, \infty)$. A random variable which follows an exponential distribution can be shown as $x \sim \exp(\lambda)$. The PDF of exponential distribution for different values of λ is presented in Fig. 2.9.

The mean and variance of random variables that follow exponential distribution is determined as:

$$E(t) = \frac{1}{\lambda} \quad (2.69)$$

and

$$\text{Var}(t) = \frac{1}{\lambda^2} \quad (2.70)$$

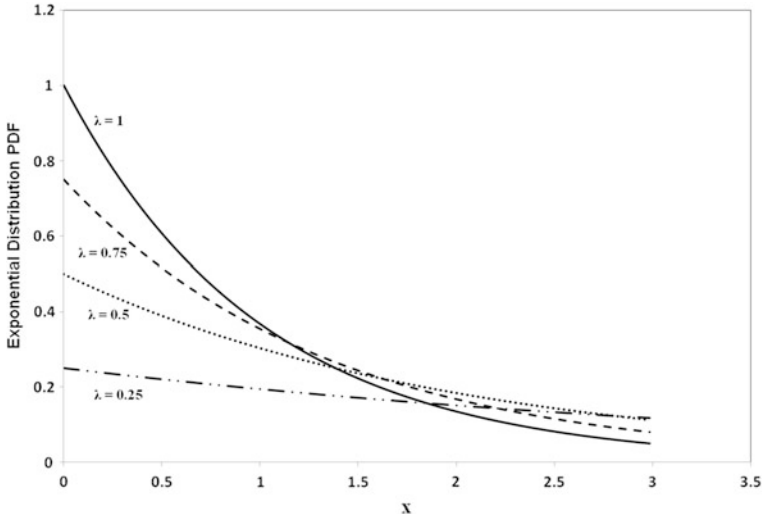


Fig. 2.9 The PDF of exponential distribution for different values of λ

Since the mean and standard deviation of this distribution are same, the coefficient of variation (CV_x) will be equal to one. The cumulative density function (CDF) of exponential distribution can also be evaluated analytically by applying Eq. (2.68) as follow:

$$F(t) = \int_0^t \lambda \cdot e^{-\lambda\tau} d\tau = 1 - e^{-\lambda t} \quad (2.71)$$

Clearly, when t approaches to infinity, $F(t)$ approaches to one, and therefore the whole area under the probability distribution equals to one. Analytically, exponential distribution is easily manipulated and sometimes uses to approximate more complex skewed distributions such as gamma or extreme values. The appropriate function of exponential distribution in Excel is EXPONDIST and it is used as:

$$= \text{EXPONDIST} (\text{select} : \text{ the range of } x, \lambda, \text{ True or False})$$

It should be noted that TRUE returns the cumulative distribution function and FALSE returns the probability density function.

Example 2.13 Assume the number of rainy days follows the exponential distribution with the mean 9 days, and compute the probability of having less than 10 and 15 rainy days.

Solution First, the value of λ is computed as:

$$E(t) = \frac{1}{\lambda} = 9$$

$$\lambda = 0.111$$

Thus,

$$\begin{aligned} F_t(10) &= 1 - e^{-(0.111) \times 10} = 0.670 \\ F_t(15) &= 1 - e^{-(0.111) \times 15} = 0.810 \end{aligned}$$

2.6.6 Gamma Distribution

Gamma distribution is used to model positive and continuous random variables. This distribution involves two positive parameters known as shape factor (α) and scale factor (β). Depending on the value of α , the shape of this distribution can be varied. When $\alpha > 1$, the gamma distribution has a skewed and uni-modal shape, if $\alpha < 1$, it is exponentially shaped and asymptotic at both the vertical and horizontal axes, and for $\alpha = 1$, the gamma distribution is the same as exponential distribution with a scale factor of β (Tung et al. 2005). The scale factor can change the distribution shape by affecting the condensing and stretching of the probability density function (Fig. 2.10).

This distribution is commonly used in hydrology and water resources engineering regarding its shape and well-known mathematical properties. The PDF of Gamma distribution with two parameters α and β is defined as:

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \quad \text{for } x > 0 \quad (2.72)$$

in which $\beta > 0$, $\alpha > 0$ and $\Gamma(\cdot)$ is a gamma function as:

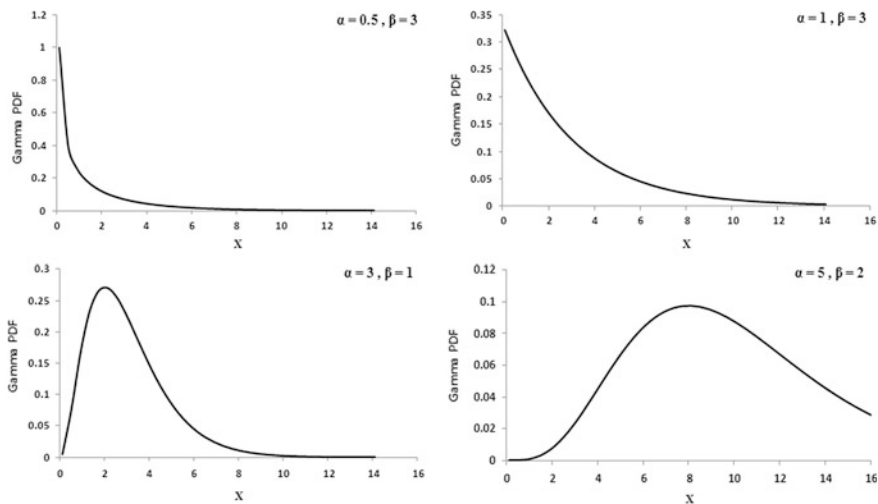


Fig. 2.10 Gamma distribution in different values of α and β

Table 2.2 The statistical features of gamma distribution

$E(x)$	$Var(x)$	CV_x	γ_x
$\alpha\beta$	$\alpha\beta^2$	$\frac{1}{\sqrt{\alpha}}$	$\frac{2}{\sqrt{\alpha}}$

$$\Gamma(\alpha) = \int_0^{\infty} e^{-x} x^{\alpha-1} dx \quad (2.73)$$

For the positive integer value of α , $\Gamma(\alpha)$ equals to:

$$\Gamma(\alpha) = (\alpha - 1)! \quad (2.74)$$

When the PDF of desired random variable is known, the CDF of Gamma distribution will be computed as:

$$F(x, \alpha, \beta) = \int_0^x \frac{u^{\alpha-1} e^{-u/\beta}}{\beta^\alpha \Gamma(\alpha)} du \quad (2.75)$$

The mean, variance, and skewness coefficient of this distribution are presented in the Table 2.2.

If $\alpha = 1$ in Eq. (2.72), the PDF of Gamma distribution will be the same as the exponential distribution. In other words, the exponential distribution is a particular case of Gamma distribution, as:

$$f(x, \beta) = \frac{e^{-x/\beta}}{\beta} = \lambda e^{-\lambda x} \quad t > 0 \quad (2.76)$$

The GAMMADIST and GAMMAINV functions in Excel return the Gamma distribution and the inverse of the Gamma cumulative distribution as:

$$\begin{aligned} &= \text{GAMMADIST}(\text{select : therangeof } x, \alpha, \beta, \text{TrueorFalse}) \\ &= \text{GAMMAINV}(\text{probability}, \alpha, \beta) \end{aligned}$$

To estimate gamma of desired random variable x or $\Gamma(x)$ in Excel, the following command can be applied:

$$= \text{EXP}(\text{GAMMALN}(X))$$

in which GAMMALN(X) returns the natural logarithm of the gamma function or $\gamma(x)$. Hence, EXP[GAMMALN(X)] returns the gamma value. In addition, there are different tools to solve Eq. (2.75) and compute the integral or CDF of Gamma distribution. One of the simplest tools which can be used is the *Wolfram Alpha computational knowledge engine*. It is an online service that answers many mathematical questions and it can be found at: <http://www.wolframalpha.com>.

Example 2.14. Assume the best fit to annual maximum inflows into a small reservoir follow Gamma distribution with $\alpha = 1.5$ and $\beta = 175$.

1. Calculate the mean and variance of annual maximum inflows.
2. Determine the probability of floods less than 100, 200, 500 and 1,000 cfs in any year.
3. What is the return period of each flood?

Solution

1. The mean and standard deviation are:

$$\mu_Q = \alpha\beta = 1.5 \times 175 = 262.5$$

and

$$\sigma_Q = \sqrt{\alpha\beta^2} = \sqrt{1.5 \times 175^2} = 214.33$$

2. The process of computing the probability of occurring floods in given return periods are presented in the following table.

$Q(\text{cfs})$	$F(Q) = P(Q < q)$	T
100	0.2332	1.30
200	0.4847	1.94
500	0.8736	7.91
1000	0.9903	103.09

For example, the probability of flood less than 100 cfs and its return period are computed as:

$$\begin{aligned}
 P(Q \leq 100) = F(100) &= \int_0^{x=100} \frac{u^{\alpha-1} e^{-u/\beta}}{\beta^\alpha \cdot \Gamma(\alpha)} du \\
 &= 0.2332
 \end{aligned}$$

and

$$\begin{aligned}
 T &= \frac{1}{1 - F(Q)} \rightarrow \\
 T_{Q=100} &= \frac{1}{1 - 0.2332} \cong 1.30
 \end{aligned}$$

The PDF of desired Gamma distribution in this example is shown in Fig. 2.11.

$$f(Q, \alpha, \beta) = \frac{Q^{0.5} e^{-Q/175}}{2051.644}$$

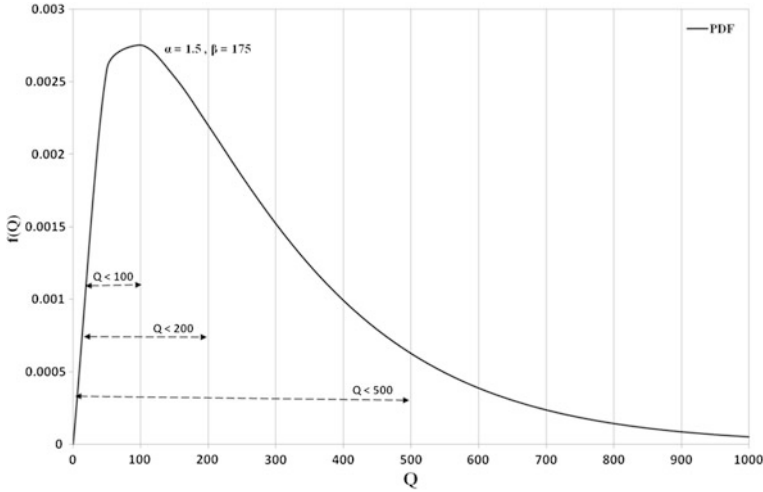


Fig. 2.11 The PDF of presented gamma distribution in Example 2.14

2.6.7 Beta Distribution

Beta distribution is a continuous distribution used for modeling random variables with maximum and minimum values. The probability distribution function (PDF) of Beta is presented in two forms as non-standard and standard Beta. The PDF for non-standard form is defined as:

$$f(x, a, b, \alpha, \beta) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{B(\alpha, \beta)(b-a)^{\alpha+\beta-1}} \quad a \leq x \leq b \quad (2.77)$$

in which a and b are the maximum and minimum values of desired Beta random variables, respectively; α and β are positive shape parameters, and $B(\alpha, \beta)$ is the beta function with parameters α and β . The beta function is:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt \quad (2.78)$$

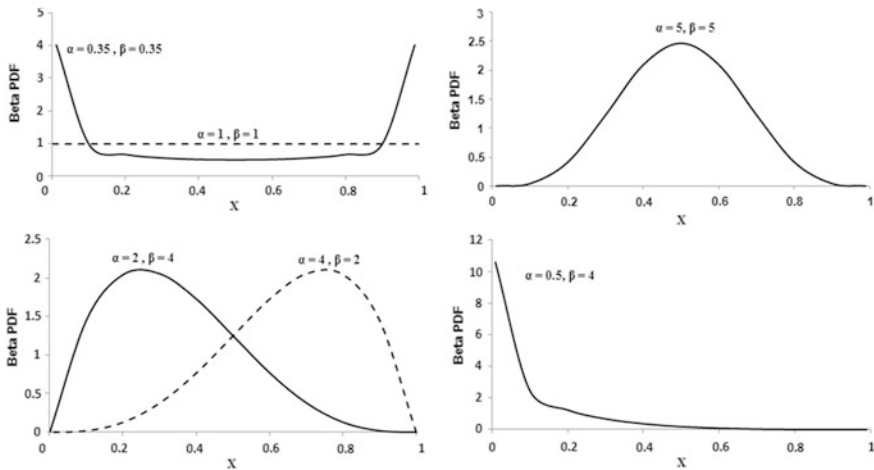
Furthermore, the beta function can be evaluated based on the gamma function as:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad (2.79)$$

The standard PDF of Beta distribution for random variable $x \sim B(\alpha, \beta)$ is defined in the following form:

Table 2.3 The statistical features of standard Beta distribution

$E(x)$	$Var(x)$	CV_x	γ'_x
$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\sqrt{\frac{\beta}{\alpha(\alpha+\beta+1)}}$	$\frac{2(\beta-\alpha)}{(2+\alpha+\beta)} \sqrt{\frac{1+\alpha+\beta}{\alpha\beta}}$

**Fig. 2.12** The shapes of beta distribution for different values of α and β

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad 0 \leq x \leq 1 \quad (2.80)$$

The mean, variance, coefficient of variation, and skewness of standard Beta distribution are presented in Table 2.3.

In addition, Fig. 2.12 shows several shapes of Beta distribution for different values of α and β .

2.6.8 Uniform Distribution

Uniform or rectangular distribution is a distribution with equal probability of occurrence for all random variables. The PDF and CDF of a continuous uniform distribution on the interval (a, b) are:

$$PDF = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \leq x \leq b \\ 0 & x > b \end{cases} \quad (2.81)$$

and

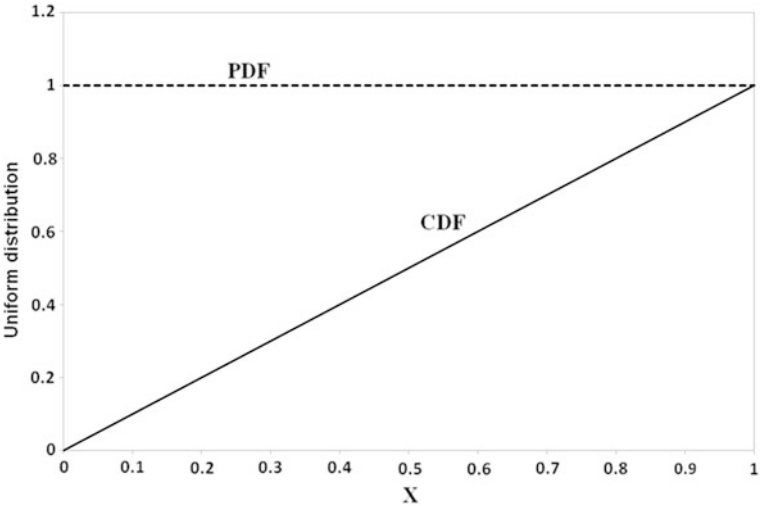


Fig. 2.13 The PDF and CDF of standard uniform distribution

Table 2.4 The statistical features of uniform distribution

$E(x)$	$Var(x)$	γ_x
$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	0

$$CDF = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases} \tag{2.82}$$

in which a and $(b - a)$ are the location and scale parameters, respectively. Uniform distribution is referred to the standard uniform distribution if $a = 0$ and $b = 1$, and its probability density function has following form:

$$f(x) = 1 \quad 0 \leq x \leq 1 \tag{2.83}$$

The PDF and CDF of standard uniform distribution, and its mean, variance, and skewness are presented in Fig. 2.13 and Table 2.4, respectively.

References

Internet Center for Management and Business Administration. (2008). Quick MBA Statistics. Retrieved from <http://www.quickmba.com/stats/probability>.

Mays, L. W., & Tung, Y. K. (1992). *Hydrosystems engineering and management*. New York: Mc Grow-Hill.

McCuen, R. (2005). *Hydrologic analysis and design*. New Jersey: Pearson Prentice Hall.

Tung, Y.K., & Yen, B.C., Melching, C.S. (2005). *Hydrosystems engineering reliability assessment and risk analysis*, McGraw-Hill Professional, New York.

Introduction to Risk and Uncertainty in Hydrosystem
Engineering

Goodarzi, E.; Ziaei, M.; Teang Shui, L.

2013, XIII, 157 p., Hardcover

ISBN: 978-94-007-5850-6