10.4 PROBABILITY DISTRIBUTIONS

In statistics, a probability distribution gives either the probability of each value of a random variable (when the variable is discrete), or the probability of the value falling within a particular interval (when the variable is continuous). The probability distribution describes the range of possible values that a random variable can attain and the probability that the value of the random variable is within any (measurable) subset of that range.

A probability distribution gives important information about the data, how the values are changing, whether they are bunched together or spread out, and whether they are symmetrically disposed on the X-axis or not. Distribution also tells the relative frequency or proportion of various *X* values in the population in the same way that a histogram gives information about a sample. We now describe the distributions that are commonly used in addressing water resources problems.

Commonly used distributions in hydrology are the Normal, Log Normal, Extreme Value type-1 (Gumbel or EV1), Gamma, Pearson Type - III, and Log Pearson Type - III distributions. A brief description of these distributions is given below.

10.4.1 Normal Distribution

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

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \quad -\infty < x < \infty \tag{10.24}$$

the cumulative density function (CDF) of the normal distribution is:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] du \tag{10.25}$$

The origin of the normal organization lies in the central limit theorem which states that if a sequence of random variables x_i, i = 1, 2, …, n are independently and identically distributed with mean μ and standard deviation σ then the distribution of *n* such random variables Y= $\sum_{i=1}^{n} x_i$

tends to the normal distribution with mean nμ and standard deviation $\sqrt{n}\sigma$, as *n* becomes large. This theorem holds good irrespective of the probability distribution of x.

The reduced variate of the normal distribution is defined as Z = (x - μ)/σ. The properties of the reduced variate are mean = 0, standard deviation σ_z = 1, and coefficient of skewness = 0. Fig. 10.8 shows the normal distribution and the area under the curve for three values of the reduced variate. As shown, the area under the curve within μ ± σ is 68.27%, within μ ± 2σ is 95.45 and it is 99.73 within μ ± 3σ.

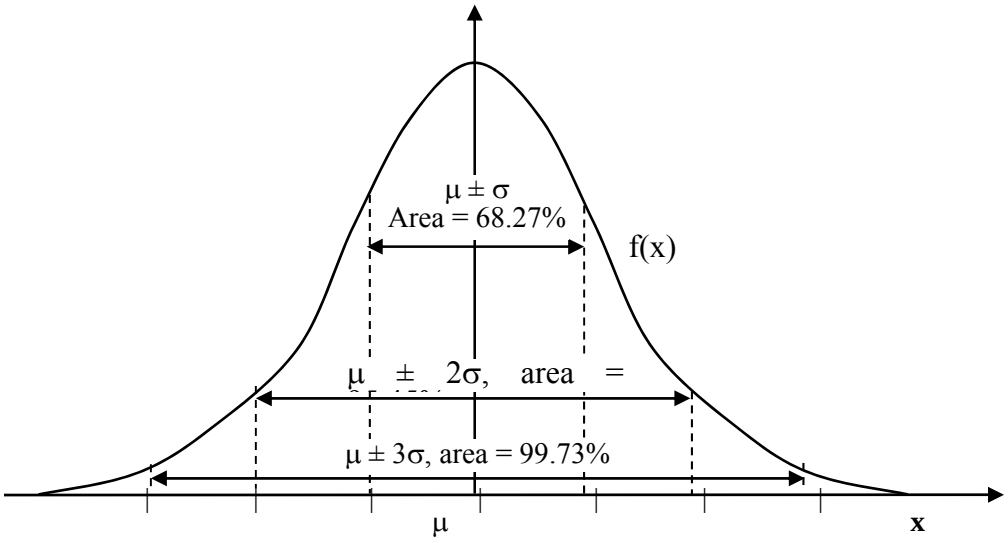

Fig. 10.8 The normal distribution and the area for three values of the standard variate.

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

Example 10.4: Assuming that the data of Sabarmati River follows the normal distribution, find the parameters of the distribution and plot it.

Solution: For the data of Sabarmati river, the mean and SD are:

Mean of the data $\bar{x}$ = 665.37 million cubic m.
Standard deviation σ = 346.9 million cubic m.
Coefficient of variation C_V = 346.9/665.37 = 0.521.
Coefficient of skewness C_s = 0.76 (positively skewed).
Kurtosis C_k = 3.65.

Fig. 10.9 shows the plot of the probability distribution of Sabarmati River data.

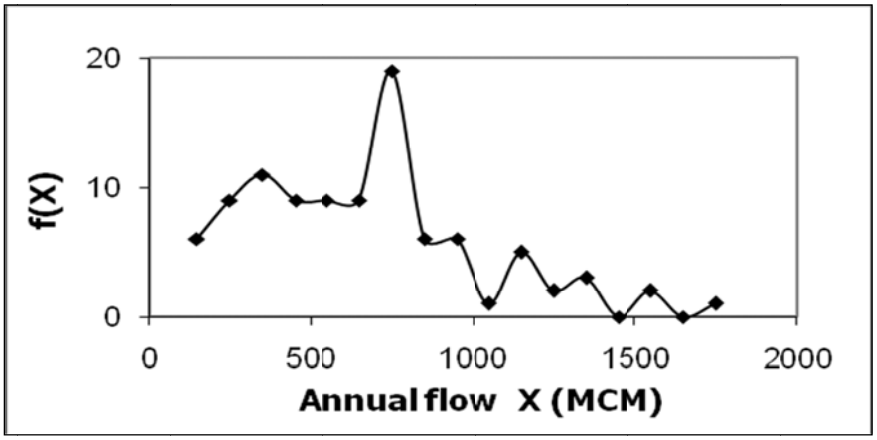

Fig. 10.9 Probability distribution of Sabarmati River data.

For the Sabarmati data, the coefficient of skewness C_s is 0.76 or the data is positively skewed. This is easily verified by Fig. 10.9. Further, kurtosis C_k for the data is 3.65 (kurtosis for the normal distribution is 3). Again, this can also be verified from Fig. 10.9.

10.4.2 Log-Normal Distribution

The log-normal distribution is the probability distribution of a random variable whose logarithm is normally distributed. Let X be a random variable with a normal distribution, then Y = exp(X) has a log-normal distribution. In other words, if Y is log-normally distributed, then X = log(Y) is normally distributed. When a random variable represents a process that is the resultant of multiplicative product of many small effects each of which is positive, then it can be expressed the sum of logarithms of these small effects. The logarithm of such a random variable can be expected to follow a normal distribution. Hence, if the variable is transformed to the log domain, it is likely to follow the normal distribution. An advantage of the log-normal distribution is that it

is often useful to represent quantities that cannot have negative values. It has proven useful to model rainfall amounts, size distributions of aerosol particles, etc.

The PDF of the log-normal distribution is

$$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu_y)^2}{2\sigma_y^2}\right] \quad x > 0 \qquad (10.26)$$

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

$$y_i = \ln x_i \qquad (10.27)$$

10.4.3 Extreme Value Type 1 (EV1) Distribution

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

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

$$F(x) = \exp\{-\exp[-\alpha(x - \beta)]\} \qquad (10.28)$$

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

$$\hat{\alpha} = \frac{1.283}{s}; \qquad \hat{\beta} = \bar{x} - 0.45s \qquad (10.29)$$

It has been shown that the EVI distribution is a special case of a distribution known as the Generalized Extreme Value (GEV) distribution. The CDF of the GEV distribution is given by

$$F(x) = \exp\left[-\left(1 - k\frac{x-u}{\alpha}\right)^{1/k}\right] \quad (10.30)$$

Where k, u, and α are the parameters of the distribution. When k = 0, we get the EV1 distribution. For k<0, the distribution known as EV2 and it is known as EV3 distribution when k > 0.

According to Gumbel, the probability that an event with magnitude larger than x_0 occurs is (Subramanya 2008):

$$\text{Prob}\ (X \geq x_0) = 1 - \exp[-\exp(-y)] \quad (10.31)$$

where y is the reduced variate, given as

$$y = \alpha(x - \beta),$$
$$\beta = \bar{x} - 0.45\ \sigma_x$$
$$\alpha = 1.2825/\ \sigma_x \quad (10.32)$$

Substituting the values of a and α

$$y = 1.285\ (x - \bar{x})/\ \sigma_x + 0.577 \quad (10.33)$$

The expression for the reduced variate y for return period T is

$$y = -\ [0.834 + 2.303\ \log\ (\log\ \frac{T}{T-1})] \quad (10.34)$$

Now we can compute variate x with return period T by

$$X_T = \bar{x} + k\sigma_x \quad (10.35)$$

where $k = (y_T - 0.577)/1.2825$ (10.36)

Equations (10.35) and (10.36) assume that a large data series are available to compute the various statistics. However in practice, the record length is finite. In such cases, the following equation may be used:

$$x_T = \bar{x} + k_n \sigma_{n-1} \quad (10.37)$$

where σ_{n-1} is the standard deviation of the sample of size *n*. Frequency factor for use with sample of size *n* is given as

$$k_n = (y_T - \bar{y}_n)/s_n \qquad (10.38)$$

where $\bar{y}_n$ and s_n are reduced mean and reduced standard deviations which are functions of *n*. Values of these can be obtained from standard tables that are widely available (see, for example, Subramanya 2008). Note that as $n \to \infty$, $\bar{y}_n \to 0.577$ and $s_n \to 1.2825$.

Example 10.5: From the flow data of a river, the mean and standard deviation were computed and these turned out to be 660 million cubic m and 330 million cubic m, respectively. Find the value of parameters of EV1 distribution.

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

$\alpha = 1.2825/330 = 0.0039.$

and $\beta = 660 - 0.45*330 = 511.5.$

Example 10.6: Annual maximum flood discharge data of a river was available for 30 years. Mean and standard deviation were 5250 m^3/s and 1650 m^3/s. Compute the flood discharge with a return period of 100 years by using the Gumbel Extreme Value 1 distribution.

Solution: From standard tables, for n = 30 years

$y_n = 0.5362$, $s_n = 1.1124$.

Hence $y_T = -[0.834 + 2.303 \log (\log \frac{100}{99})] = 4.601$

$k_{100} = (4.601 - 0.5362)/1.1124 = 3.654$

$x_{100} = 5250+3.654*1650 = 11279\ m^3/s$

10.4.4 Log Pearson Type - III (LP3) Distribution

Log Pearson Type III distribution was found to give good results in numerous studies dealing with flood peak data. This distribution is the standard distribution for flood frequency analysis in the USA since its use for flood frequency analysis was recommended by the US Water

Resources Council.

LP3 is a three-parameter distribution and is widely used in hydrology. Its parameters are related to mean, standard deviation, and skewness.

$$f(x) = \frac{1}{a\Gamma(b)}\left(\frac{x-c}{a}\right)^{b-1} \exp\left(-\frac{x-c}{a}\right) \qquad (10.39)$$

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

$$a = \sigma/\sqrt{b} \qquad (10.39a)$$

$$b = (2/C_s)^2 \qquad (10.39b)$$

$$c = \mu - \sigma\sqrt{b} \qquad (10.39c)$$

To determine flood for a return period T by using the LP3 distribution, the procedure described below is followed.

First of all, the frequency factor, K_T is computed by (Chow et al. 1988):

$$K_T = z + (z^2 - 1)k + (z^3 - 6z)k^2/3 + (z^2 - 1)k^3 + zk^4 + k^5/3 \qquad (10.40)$$

Where $k = C_s/6$. To complete z for a given return period T, exceedance probability p is obtained as p = 1/T. Now, complete a variable *w* as

$$w = \sqrt{\ln(1/p^2)} \qquad 0 < p \le 0.5$$

Now z is calculated by (Abramowitz and Stegun, 1965)

$$z = w - \frac{2.515517 + 0.802853w + 0.010328w^2}{1 + 1.432788w + 0.189269w^2 + 0.001308w^3} \qquad (10.41)$$

when p > 0.5, p in eq. (10.41) is replaced by (1-p) and the negative sign is put before z computed by eq. (10.42). Now, by following the frequency factor method, the flood for the return period T years is computed by:

$$y_T = \bar{y} + K_T s_y \qquad (10.42)$$

Example 10.7: For the data of Example 10.1, find the parameters of the Pearson Type III distribution.

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

b = $(2/C_s)^2 = (2/0.76)^2$ = 6.93 million cubic m.
a = 346.9/√6.93 = 131.78 million cubic m.
c = 665.37 – 346.9*√6.93 = -247.84.

Example 10.8: Logarithms of the annual flood peak data of a river were taken and the mean was 4.146, SD was 0.403 and C_s = -0.07. Find 50 year return period flood by using the LP3 distribution.

Solution First we find the value of k_{50} by the following equation:

$$K_{50} = z + (z^2 - 1)k + (z^3 - 6z)k^2/3 + (z^2 - 1)k^3 + zk^4 + k^5/3$$

Here $k = C_s/6 = -0.07/6 = -0.0117$. For T = 50, p = 1/50 = 0.02.

$$w = \sqrt{\ln(1/0.02^2)} = 2.797$$

From eq. (10.41)

$$z = 2.797 - \frac{2.515517 + 0.802853*2.797 + 0.010328*2.797^2}{1 + 1.432788*2.797 + 0.189269*2.797^2 + 0.001308*2.797^3} = 2.054$$

Now K_{50} is calculated as

$K_{50} = 2.054 + (2.054^2 - 1)*(-0.0117) + (2.054^3 – 6*2.054)* (-0.0117)^2/3$
$+ (2.054^2 - 1)* (-0.0117)^3 + 2.054*(-0.0117)^4 + (-0.0117)^5/3$
$= 2.016$

Hence, $y_{50} = 4.146 + 2.016*0.403 = 4.959$
So, the 50-year flood $x_{50} = (10)^{4.959} = 90942$.

10.4.5 Discrete Probability Distributions

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

10.4.6 Binomial Distribution

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

$$P_{r,n} = n\ C_r\ P^r\ q^{n-r} = \frac{n!}{r!(n-r)!}\ p^r\ q^{n-r} \qquad (10.43)$$

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

$$P_{o,n} = q^n = (1 - p)^n \qquad (10.44)$$

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

$$P_1 = 1 - q^n = 1 - (1 - p)^n \qquad (10.45)$$

Example 10.9: An analysis of data on the maximum one-day rainfall depth at a station indicated that a depth of 280 mm had a return period of 50 years. Determine the probability of a one-day rainfall depth equal to or greater than 280 mm occurring (a) once in 20 successive years, and (b) two times in 15 successive years.

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

a) In the first case, *n* = 20, *r* = 1. Therefore, from eq. (39)

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

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

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

Example 10.10: What is the probability that a 5-year flood will not occur at all in a 10-year period?

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

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

10.4.7 Poisson Distribution

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

$$P(X = x) = \frac{\lambda^{x} \exp(-\lambda)}{x!}, \quad \lambda > 0, \quad \text{x} = 0, 1, 2, ... \tag{10.46}$$

The CDF of the Poisson distribution is

$$P(X \le x) = \sum_{i=0}^{x} \frac{\lambda^{i} \exp(-\lambda)}{i!} \tag{10.47}$$

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

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