# 6

# *FREQUENCY ANALYSIS*

The term *frequency analysis* refers to the techniques whose objective is to analyze the occurrence of hydrologic variables within a statistical framework, i.e., by using measured data and basing predictions on statistical laws. These techniques are applicable to the study of statistical properties of either rainfall or runoff (flow) series. In engineering hydrology, however, frequency analysis is commonly used to calculate flood discharges.

In principle, techniques of frequency analysis are applicable to gaged catchments with long periods of streamflow record. In practice, these techniques are primarily used for large catchments, because these are more likely to be gaged and have longer record periods. Frequency analysis is also applicable to midsize catchments, provided the record length is adequate. For ungaged catchments (either midsize or large), frequency analysis can be used in a regional context to develop flow characteristics applicable to *hydrologically homogeneous* regions. These techniques comprise what is referred to as regional analysis (Chapter 7).

The question to be answered by flow frequency analysis can be stated as follows: Given $n$ years of daily streamflow records for stream S, what is the maximum (or minimum) flow $Q$ that is likely to recur with a frequency of once in $T$ years on the average? Or, what is the maximum flow $Q$ associated with a $T$-year return period? Alternatively, frequency analysis seeks to answer the inverse question: What is the return period $T$ associated with a maximum (or minimum) flow $Q$?

/ S (italics)

In more general terms, the preceding questions can be stated as follows: Given $n$ years of streamflow data for stream $S$ and $L$ years of design life of a certain structure, what is the probability $P$ of a discharge $Q$ being exceeded at least once during the design life $L$? Alternatively, what is the discharge $Q$ which has the probability $P$ of being exceeded during the design life $L$?

This chapter is divided into three sections. Section 6.1 contains a review of statistics and probability concepts useful in engineering hydrology. Section 6.2 describes techniques of flood frequency analysis. Section 6.3 discusses low-flow frequency and droughts.

## 6.1 CONCEPTS OF STATISTICS AND PROBABILITY

Frequency analysis uses random variables and probability distributions. A *random variable* follows a certain probability distribution. A *probability distribution* is a function that expresses in mathematical terms the relative chance of occurrence of each of all possible outcomes of the random variable. In statistical notation, $P(X = x_1)$ is the probability $P$ that the random variable $X$ takes on the outcome $x_1$. A shorter notation is $P(x_1)$.

An example of random variable and probability distribution is shown in Fig. 6-1. This is a discrete probability distribution because the possible outcomes have been arranged into groups (or classes). The random variable is discharge $Q$; the possible outcomes are seven discharge classes, from 0-100 m$^3$/s to 600-700 m$^3$/s. In Fig. 6-1, the probability that $Q$ is in the class 100-200 m$^3$/s is 0.25. The sum of probabilities of all possible outcomes is equal to 1.

A cumulative discrete distribution, corresponding to the discrete probability distribution of Fig. 6-1, is shown in Fig. 6-2. In this figure, the probability that $Q$ is in a class less than or equal to the 100-200 class is 0.40. The maximum value of probability of the cumulative distribution is 1.

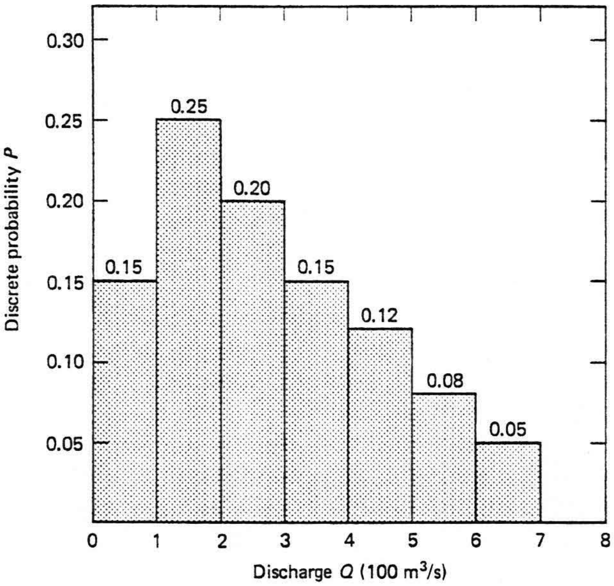


**Figure 6-1** Discrete probability distribution.

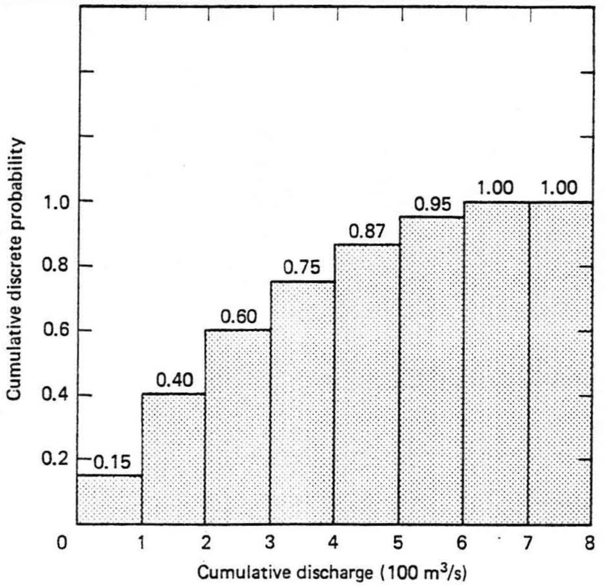


**Figure 6-2** Cumulative discrete probability distribution.

### Properties of Statistical Distributions

The properties of statistical distributions are described by the following measures: (1) central tendency, (2) variability, and (3) skewness. Statistical distributions are described in terms of *moments*. The first moment describes central tendency, the second moment describes variability, and the third moment describes skewness. Higher-order moments are possible but are seldom used in practical applications.

The first moment about the origin is the *arithmetic mean,* or mean. It expresses the distance from the origin to the centroid of the distribution (Fig. 6-3(a)):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{6-1}$$

in which $\bar{x}$ is the mean, $x_i$ is the random variable, and $n$ is the number of values.

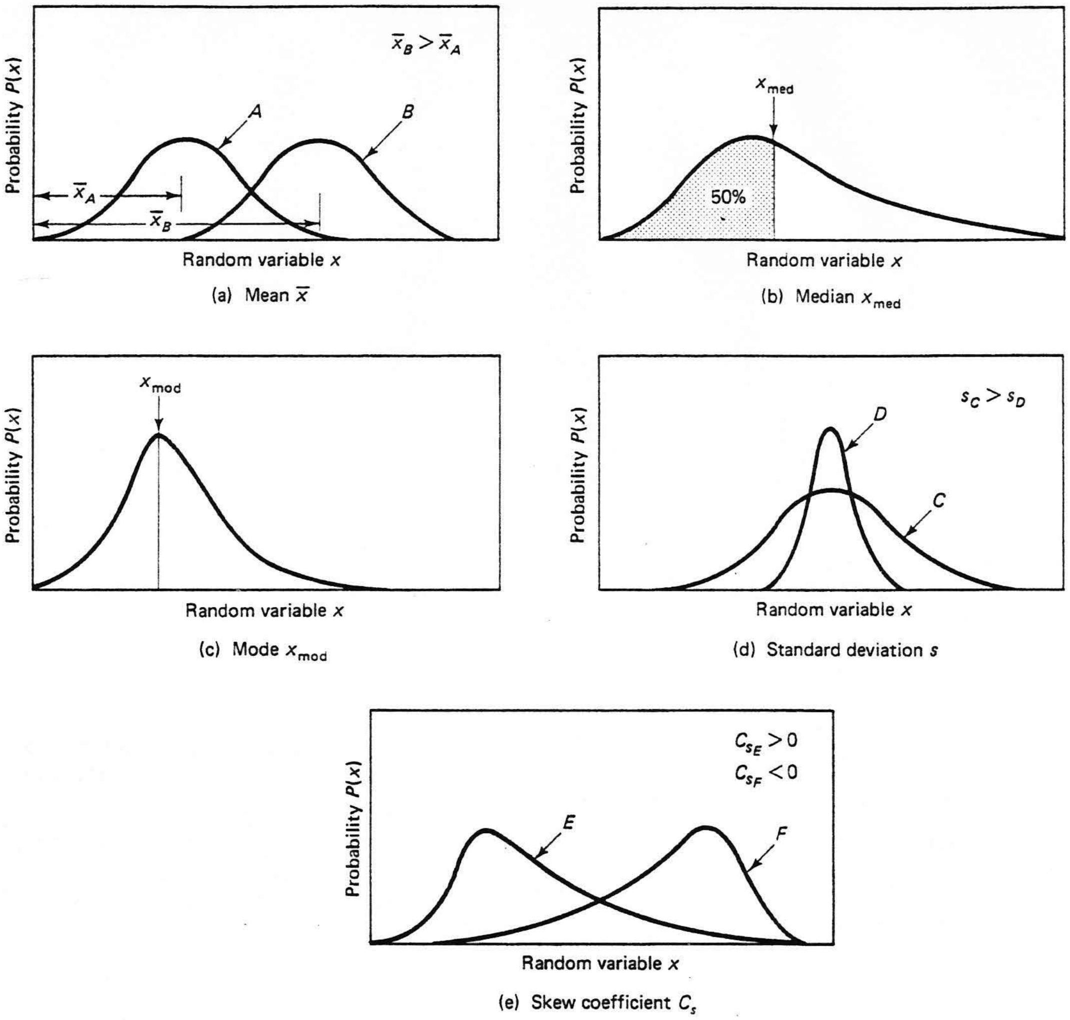
The *geometric mean* is the $n$th root of the product of $n$ terms:

$$\bar{x}_g = (x_1\, x_2\, x_3 \cdots x_n)^{1/n} \tag{6-2}$$

The logarithm of the geometric mean is the mean of the logarithms of the individual values. The geometric mean is to the lognormal probability distribution what the arithmetic mean is to the normal probability distribution.

The *median* is the value of the variable that divides the probability distribution into two equal portions (or areas) (Fig. 6-3(b)). For certain skewed distributions (i.e. one with third moment other than zero), the median is a better indication of central tendency than the mean. Another measure of central tendency is the *mode,* defined as the value of the variable that occurs most frequently (Fig. 6-3(c)).

Statistical moments can be defined about axes other than the origin. The second moment about the mean is the *variance,* defined as



(a) Mean $\bar{x}$

(b) Median $x_{\text{med}}$

(c) Mode $x_{\text{mod}}$

(d) Standard deviation $s$

(e) Skew coefficient $C_s$

**Figure 6-3** Properties of statistical distibutions: (a) mean $\bar{x}$; (b) median $x_{\text{med}}$; (c) mode $x_{\text{mod}}$; (d) standard deviation $s$; (e) skew coefficient $C_s$.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{6-3}$$

in which $s^2$ is the variance. The square root of the variance, $s$, is the *standard deviation.* The *variance coefficient* (or coefficient of variation) is defined as

$$C_v = \frac{s}{\bar{x}} \tag{6-4}$$

The standard deviation and variance coefficient are useful in comparing relative variability among distributions. The larger the standard deviation and variance coefficient, the larger the spread of the distribution (Fig. 6-3(d)).

The third moment about the mean is the *skewness*, defined as follows:

$$a = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i - \bar{x})^3 \tag{6-5}$$

in which $a$ is the skewness. The *skew coefficient* is defined as

$$C_s = \frac{a}{s^3} \tag{6-6}$$

For symmetrical distributions, the skewness is 0 and $C_s = 0$. For right skewness (distributions with the long tail to the right), $C_s > 0$; for left skewness (long tail to the left), $C_s < 0$ (Fig. 6-3(e)).

Another measure of skewness is *Pearson's skewness*, defined as the ratio of the difference between mean and mode to the standard deviation.

**Example 6-1.**

Calculate the mean, standard deviation, and skew coefficient for the following flood series: 4580, 3490, 7260, 9350, 2510, 3720, 4070, 5400, 6220, 4350, and 5930 $m^3/s$.

The calculations are shown in Table 6-1. Column 1 shows the year and Col. 2 shows the annual maximum flows. The mean (Eq. 6-1) is calculated by summing up Col. 2 and dividing the sum by $n = 11$. This results in $\bar{x} = 5171$ $m^3/s$. Column 3 shows the flow deviations from the mean, $x_i - \bar{x}$. Column 4 shows the square of the flow deviations, $(x_i - \bar{x})^2$. The variance (Eq. 6-3) is calculated by summing up Col. 4 and dividing the sum by $(n - 1) = 10$. This results in: $s^2 = 3{,}780{,}449$ $m^6/s^2$. The square root of the variance is the standard deviation: $s = 1944$ $m^3/s$. The variance coefficient (Eq. 6-4) is $C_v = 0.376$. Column 5 shows the cube of the flow deviations, $(x_i - \bar{x})^3$. The skewness (Eq. 6-5) is calculated by summing up Col. 5 and multiplying the sum by $n/[(n-1)(n-2)] = 11/90$. This results in $a = 6{,}717{,}359{,}675$ $m^9/s^3$. The skew coefficient (Eq. 6-6) is equal to the skewness divided by the cube of the standard deviation. This results in $C_s = 0.914$.

**TABLE 6-1** CALCULATION OF MEAN, STANDARD DEVIATION, AND SKEW COEFFICIENT: EXAMPLE 6-1

| (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|
| Year | Peak Flow ($m^3/s$) | $(x_i - \bar{x})$ ($m^3/s$) | $(x_i - \bar{x})^2$ ($m^6/s^2$) | $(x_i - \bar{x})^3$ ($m^9/s^3$) |
| 1 | 4,580 | −591 | 349,281 | −206,425,071 |
| 2 | 3,490 | −1,681 | 2,825,761 | −4,750,104,241 |
| 3 | 7,260 | 2,089 | 4,363,921 | 9,116,230,969 |
| 4 | 9,350 | 4,179 | 17,464,041 | 72,982,227,340 |
| 5 | 2,510 | −2,661 | 7,080,921 | −18,842,330,780 |
| 6 | 3,720 | −1,451 | 2,105,401 | −3,054,936,851 |
| 7 | 4,070 | −1,101 | 1,212,201 | −1,334,633,301 |
| 8 | 5,400 | 229 | 52,441 | 12,008,989 |
| 9 | 6,220 | 1,049 | 1,100,401 | 1,154,320,649 |
| 10 | 4,350 | −821 | 674,041 | −553,387,661 |
| 11 | 5,930 | 759 | 576,081 | 437,245,479 |
| Sum | 56,880 | | 37,804,491 | 54,960,215,521 |

## Continuous Probability Distributions

A continuous probability distribution is referred to as a probability density function (PDF). A PDF is an equation relating probability, random variable, and parameters of the distribution. Selected PDFs useful in engineering hydrology are described in this section.

**Normal Distribution.** The *normal distribution* is a symmetrical, bell-shaped PDF also known as the Gaussian distribution, or the natural law of errors. It has two parameters: the mean, $\mu$, and the standard deviation, $\sigma$, of the population. In practical applications, the mean $\bar{x}$ and the standard deviation $s$ derived from sample data are substituted for $\mu$ and $\sigma$. The PDF of the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{6-7}$$

in which $x$ is the random variable and $f(x)$ is the continuous probability.

By means of the transformation

$$z = \frac{x-\mu}{\sigma} \tag{6-8}$$

the normal distribution can be converted into a one-parameter distribution, as follows:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{6-9}$$

in which $z$ is the standard unit, which is normally distributed with zero mean and unit standard deviation.

From Eq. 6-8,

$$x = \mu + z\sigma \tag{6-10}$$

in which $z$, the standard unit, is the *frequency factor* of the normal distribution. In general, the frequency factor of a statistical distribution is referred to as $K$.

A cumulative density function (CDF) can be derived by integrating the probability density function. From Eq. 6-9, integration leads to

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du \tag{6-11}$$

in which $F(z)$ denotes cumulative probability and $u$ is a dummy variable of integration. The distribution is symmetrical with respect to the origin; therefore, only half of the distribution needs to be evaluated. Table A-5 (Appendix A) shows values of $F(z)$ versus $z$, in which $F(z)$ is integrated from the origin to $z$.

**Example 6-2.**

The annual maximum flows of a certain stream have been found to be normally distributed, with mean 90 m$^3$/s and standard deviation 30 m$^3$/s. Calculate the probability that a flow larger that 150 m$^3$/s will occur.

To enter Table A-5, it is necessary to calculate the standard unit. For a flow of 150 m$^3$/s, the standard unit (Eq. 6-8) is: $z = (150 - 90)/30 = 2$. This means that the flow of 150 m$^3$/s is located two standard deviations to the right of the mean (had $z$ been negative, the flow would have been located to the left of the mean). In Table A-5, for $z = 2$, $F(z) = 0.4772$. This value is the cumulative probability measured from $z = 0$ to $z = 2$, i.e., from the mean (90 m$^3$/s) to the value being considered (150 m$^3$/s). Because the normal distribution is symmetrical with respect to the origin, the cumulative probability measured from $z = -\infty$ to $z = 0$, is 0.5. Therefore, the cumulative probability measured from $z = -\infty$ to $z = 2$, is $F(z) = 0.5 + 0.4772 = 0.9772$. This is the probability that the flow is less than 150 m$^3$/s. To find the probability that the flow is larger than 150 m$^3$/s, the complementary cumulative probability is calculated: $G(z) = 1 - F(z) = 0.0228$. Therefore, there is a $(0.0228 \times 100) = 2.28\%$ chance that the annual maximum flow for the given stream will be larger than 150 m$^3$/s.

**Lognormal Distribution.** For certain natural phenomena, values of random variables do not follow a normal distribution, but their logarithms do. In this case, a suitable PDF can be obtained by substituting $y$ for $x$ in the equation for the normal distribution, Eq. 6-7, in which $y = \ln x$. The parameters of the lognormal distribution are the mean and standard deviation of $y$: $\mu_y$ and $\sigma_y$.

**Gamma Distribution.** The gamma distribution is used in many applications of engineering hydrology. The PDF of the gamma distribution is the following:

$$f(x) = \frac{x^{\gamma-1} e^{-x/\beta}}{\beta^{\gamma}\Gamma(\gamma)} \qquad (6\text{-}12)$$

for $0 < x < \infty$, $\beta > 0$, and $\gamma > 0$. The parameter $\gamma$ is known as the shape parameter, since it most influences the peakedness of the distribution, while the parameter $\beta$ is called the scale parameter, since most of its influence is on the spread of the distribution [4]. /[4]

The mean of the gamma distribution is $\beta\gamma$, the variance is $\beta^2\gamma$, and the skewness is $2/(\gamma)^{1/2}$. The term $\Gamma(\gamma) = (\gamma - 1)!$

, where $\gamma$ is a positive integer, ← add

is an important definite integral referred to as the *gamma function*, defined as follows:

$$\Gamma(\gamma) = \int_0^{\infty} x^{\gamma-1} e^{-x}\, dx \qquad (6\text{-}13)$$

**Pearson Distributions.** Pearson [23] has derived a series of probability functions to fit virtually any distribution. These functions have been widely used in practical statistics to define the shape of many distribution curves. The general PDF of the Pearson distributions is the following [5]:

/24

/6

$$f(x) = e^{\int_{-\infty}^{x} [(a + x)/(b_0 + b_1 x + b_2 x^2)]\, dx} \qquad (6\text{-}14)$$

in which $a$, $b_0$, $b_1$, and $b_2$ are constants. The criterion for determining the type of distribution is $\kappa$, defined as follows:

are the mean and standard deviation of $y$: $\mu_y$ and $\sigma_y$.

**Gamma Distribution.** The gamma distribution is used in many applications of engineering hydrology. The PDF of the gamma distribution is the following:

$$f(x) = \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^{\gamma}\Gamma(\gamma)} \tag{6-12}$$

for $0 < x < \infty$, $\beta > 0$, and $\gamma > 0$. The parameter $\gamma$ is known as the shape parameter, since it most influences the peakedness of the distribution, while the parameter $\beta$ is called the scale parameter, since most of its influence is on the spread of the distribution [4].

/[4]

The mean of the gamma distribution is $\beta\gamma$, the variance is $\beta^2\gamma$, and the skewness is $2/(\gamma)^{1/2}$. The term $\Gamma(\gamma) = (\gamma - 1)!$, where $\gamma$ is a positive integer, is an important definite integral referred to as the *gamma function*, defined as follows:

add

$$\Gamma(\gamma) = \int_0^{\infty} x^{\gamma-1}e^{-x}\,dx \tag{6-13}$$

**Pearson Distributions.** Pearson [23] has derived a series of probability functions to fit virtually any distribution. These functions have been widely used in practical statistics to define the shape of many distribution curves. The general PDF of the Pearson distributions is the following [5]:

/24

/6

$$f(x) = e^{\int_{-\infty}^{x} [(a + x)/(b_0 + b_1x + b_2x^2)]\,dx} \tag{6-14}$$

in which $a$, $b_0$, $b_1$, and $b_2$ are constants. The criterion for determining the type of distribution is $\kappa$, defined as follows:

$$\kappa = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)} \tag{6-15}$$

in which $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$, with $\mu_2$, $\mu_3$, and $\mu_4$ being the second, third, and fourth moments about the mean. With $\mu_3 = 0$ (i.e., zero skewness), $\beta_1 = 0$, $\kappa = 0$, and the Pearson distribution reduces to the normal distribution.

The Pearson Type III distribution has been widely used in flood frequency analysis. In the Pearson Type III distribution, $\kappa = \infty$, which implies that $2\beta_2 = (3\beta_1 + 6)$. This is a three-parameter skewed distribution with the following PDF:

$$f(x) = \frac{(x - x_o)^{\gamma-1} e^{-(x-x_o)/\beta}}{\beta^{\gamma}\Gamma(\gamma)} \tag{6-16}$$

and parameters $\beta$, $\gamma$, and $x_o$. For $x_o = 0$, the Pearson Type III distribution reduces to the gamma distribution (Eq. 6-12). For $\gamma = 1$, the Pearson Type III distribution reduces to the exponential distribution, with the following PDF:

$$f(x) = \left(\frac{1}{\beta}\right) e^{-(x-x_o)/\beta} \tag{6-17}$$

The mean of the Pearson Type III distribution is $x_o + \beta\gamma$, the variance is $\beta^2\gamma$, and the skewness is $2/(\gamma)^{1/2}$.

**Extreme Value Distributions.** The extreme value distributions Types I, II, and III are based on the theory of extreme values. Frechet (on Type II) in 1927 [9] and Fisher and Tippett (on Types I and III) in 1928 [7] independently studied the statistical distribution of extreme values. Extreme value theory implies that if a random variable Q is the maximum in a sample of size $n$ from some population of $x$ values, then, provided $n$ is sufficiently large, the distribution of $Q$ is one of three asymptotic types (I, II, or III), depending on the distribution of $x$.

/10 /8 /Q (italic)

The extreme value distributions can be combined into one and expressed as a general extreme value (GEV) distribution [22]. The cumulative density function of the GEV distribution is:

/23

$$F(x) = e^{-[1-k(x-u)/\alpha]^{1/k}} \tag{6-18}$$

in which $k$, $u$ and $\alpha$ are parameters. The parameter $k$ defines the type of distribution, $u$ is a location parameter, and $\alpha$ is a scale parameter. For $k = 0$, the GEV distribution reduces to the extreme value Type I (EV1), or Gumbel, distribution. For $k < 0$, the GEV distribution is the extreme value Type II (EV2), or Frechet, distribution. For $k > 0$, the GEV distribution is the extreme value Type III (EV3), or Weibull, distribution. The GEV distribution is useful in applications where an extreme value distribution is being considered but its type is not known a priori.

Gumbel [12, 13, 14] has fitted the extreme value Type I distribution to long records of river flows from many countries. The cumulative density function (CDF) of the Gumbel distribution is the following double exponential function:

13-16

$$F(x) = e^{-e^{-y}} \tag{6-19}$$

in which $y = (x - u)/\alpha$ is the Gumbel (reduced) variate.

The mean $\bar{y}_n$ and standard deviation $\sigma_n$ of the Gumbel variate are functions of record length $n$. Values of $\bar{y}_n$ and $\sigma_n$ as a function of $n$ are given in Table A-8 (Appendix A). When the record length approaches $\infty$, the mean $\bar{y}_n$ approaches the value of the Euler constant (0.5772) [25], and the standard deviation $\sigma_n$ approaches the value $\pi/\sqrt{6}$. The skew coefficient of the Gumbel distribution is 1.14.

/26

The extreme value Type II distribution is also known as the log Gumbel. Its cumulative density function is

$$F(x) = e^{-y^{1/k}} \tag{6-20}$$

for $k < 0$.

The extreme value Type III distribution has the same CDF as the Type II, but in this case $k > 0$. As $k$ approaches 0, the EV2 and EV3 distributions converge to the EV1 distribution.

9 11 17 19

## 6.2 FLOOD FREQUENCY ANALYSIS

Flood frequency analysis refers to the application of frequency analysis to study the occurrence of floods. Historically, many probability distributions have been used for this purpose. The normal distribution was first used by Horton [18] in 1913, and /19
shortly thereafter by Fuller [10]. Hazen [16] used the lognormal distribution to reduce /11 /17
skewness, whereas Foster [8] preferred to use the skewed Pearson distributions. /9

The logarithmic version of the Pearson Type III distribution, i.e., the log Pearson III, has been endorsed by the U.S. Interagency Advisory Committee on Water Data for general use in the United States [27]. The Gumbel distribution (extreme /28
value Type I, or EV1) is also widely used in the United States and throughout the world. The log Pearson III and Gumbel methods are described in this section.

28

### Selection of Data Series

The complete record of streamflows at a given gaging station is called the *complete duration series.* To perform a flood frequency analysis, it is necessary to select a *flood series,* i.e., a sample of flood events extracted from the complete duration series.

There are two types of flood series: (1) the partial duration series and (2) the extreme value series. The partial duration (or peaks-over-a-threshold (POT) [22]) series consists of floods whose magnitude is greater than a certain base value. When the /23
base value is such that the number of events in the series is equal to the number of years of record, the series is called an *annual exceedence* series.

In the extreme value series, every year of record contributes one value to the extreme value series, either the maximum value (as in the case of flood frequency analysis) or the minimum value (as in the case of low-flow frequency analysis). The former is the *annual maxima* series; the latter is the *annual minima* series.

The annual exceedence series takes into account all extreme events above a certain base value, regardless of when they occurred. However, the annual maxima series considers only one extreme event per yearly period. The difference between the two series is likely to be more marked for short records in which the second largest annual events may strongly influence the character of the annual exceedence series. In practice, the annual exceedence series is used for frequency analyses involving short return periods, ranging from 2 to 10 y. For longer return periods the difference between annual exceedence and annual maxima series is small. The annual maxima series is used for return periods ranging from 10 to 100 y and more.

## Return Period, Frequency, and Risk

The time elapsed between successive peak flows exceeding a certain flow $Q$ is a random variable whose mean value is called the *return period* $T$ (or recurrence interval) of the flow $Q$. The relationship between probability and return period is the following:

$$P(Q) = \frac{1}{T} \tag{6-21}$$

in which $P(Q)$ is the *probability of exceedence* of $Q$, or frequency. The terms frequency and return period are often used interchangeably, although strictly speaking, frequency is the reciprocal of return period. A frequency of $1/T$, or one in $T$ years, corresponds to a return period of $T$ years.

The *probability of nonexceedence* $P(\bar{Q})$ is the *complementary probability* of the probability of exceedence $P(Q)$, defined as

$$P(\bar{Q}) = 1 - P(Q) = 1 - \frac{1}{T} \tag{6-22}$$

The probability of nonexceedence in $n$ successive years is

$$P(\bar{Q}) = \left(1 - \frac{1}{T}\right)^n \tag{6-23}$$

Therefore, the probability, or *risk*, that $Q$ will occur at least once in $n$ successive years is

$$R = 1 - P(\bar{Q}) = 1 - \left(1 - \frac{1}{T}\right)^n \tag{6-24}$$

## Plotting Positions

Frequency distributions are plotted using probability papers. One of the scales on a probability paper is a probability scale; the other is either an arithmetic or logarithmic scale. Normal and extreme value probability distributions are most often used in probability papers.

An *arithmetic probability* paper has a normal probability scale and an arithmetic scale. This type of paper is used for plotting normal and Pearson distributions. A *log probability* paper has a normal probability scale and a logarithmic scale and is used for plotting lognormal and log Pearson distributions. An *extreme value* probability paper has an extreme value scale and an arithmetic scale and is used for plotting extreme value distributions.

Data fitting a normal distribution plot as a straight line on arithmetic probability paper. Likewise, data fitting a lognormal distribution plot as a straight line on log probability paper, and data fitting the Gumbel distribution plot as a straight line on extreme value probability paper.

For plotting purposes, the probability of an individual event can be obtained directly from the flood series. For a series of $n$ annual maxima, the following ratio holds: