Chapter **11**

Concepts of Probability in Hydrology

11.1 INTRODUCTION

There are many hydrologic phenomena in which the variable of interest cannot be uniquely specified as a function of known related variables and conditions. For example, this is the case for the highest discharge that a certain river will attain in the next five years. The previous example is clearly the case of a random variable where the outcome can never be uniquely predicted with the help of physical laws, no matter how much information we gather.

The random aspect of a hydrologic phenomenon may also arise due to our inability to understand all the details of a causal relationship between an input (i.e., rainfall) and an output (i.e., discharge). This lack of knowledge may be inherent in the physical description of the processes or in limitations on availability of data. Summarizing, uncertainty is introduced into hydrologic problems through

1. inherent unexplainable variability of nature;
2. lack of understanding of all causes and effects in physical systems; and
3. lack of sufficient data.

The future of a random variable is not subject to precise prediction and must be described within the domain of the set of possible values it may take (its sample space). The description of the random variable is then accomplished through the concept of probability distributions. The hydrologist or

analyst must consider the possibility of occurrence of particular events and then determine the likelihood of their occurrence.

The collection of all possible outcomes of an experiment or random phenomenon is called its sample space. This space may be discrete or continuous and consists of a set of points—sample points—each of which is associated with one and only one distinguishable outcome. An "event" is a collection of sample points, and it may be simple if the event consists only of one sample point and compound if consisting of two or more sample points.

11.2 REVIEW OF PROBABILITY

Some parts of this section have been adapted from notes by I. Rodriguez-Iturbe.

The classical and simplest interpretation of probability is one of frequency. If A is a given event that may occur from an experiment (i.e., a given streamflow in nature), then the probability of A, $P(A)$, is given by

$$P(A) = \lim_{m \to \infty} \frac{n_A}{m}, \tag{11.1}$$

where n_A is the number of times that event A occurs in m repeated experiments. The true probability measure will only be obtained as the number of experiments m goes to infinity.

Mathematically the probability $P(A)$ must satisfy a series of axioms:

1. The probability of an event is a number between 0 and 1.
2. The probability of a *certain* event is 1. A certain event is one that covers all possible outcomes of an experiment. For example, the probability of a streamflow being greater than or equal to 0, but less than infinity, is 1.
3. The probability of an event that is the sum of two mutually exclusive events is the sum of the probabilities of those two events. Mutually exclusive events are those that by definition preclude the occurrence of each other. For example, a streamflow cannot be less than 300 $m^3 s^{-1}$ and greater than 500 $m^3 s^{-1}$ at the same time. Those are mutually exclusive events. A set of exhaustive events covers all possible outcomes of an experiment.

Union and Intersection

The above axioms allow the definition of some important probabilistic concepts. For example, if two events A and B are not exclusive, they have overlapping components; the set of common points is called the intersection of A and B, $A \cap B$. The union of two events A and B is the collection of all sample points occurring in either A or B. It is represented by $A \cup B$.

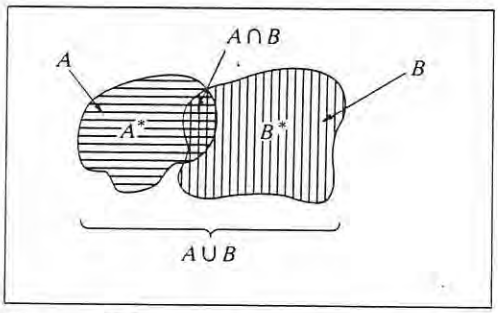


FIGURE 11.1 Illustration of the concept of intersection and union of two events A and B.

Figure 11.1 illustrates an intersection and a union. The probability of event A may be expressed as

$$P[A] = P[A \cap B] + P[A^*], \tag{11.2}$$

since $A \cap B$ and A^* are mutually exclusive (see Fig. 11.1). Similarly,

$$P[B] = P[A \cap B] + P[B^*] \tag{11.3}$$

and

$$P[A \cup B] = P[A^*] + P[B^*] + P[A \cap B]. \tag{11.4}$$

Combining the above leads to

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]. \tag{11.5}$$

Conditional Probability and Probabilistic Independence

The conditional probability of the event A given that the event B has occurred, denoted $P[A\,|\,B]$, is defined as the ratio

$$P[A\,|\,B] = \frac{P[A \cap B]}{P[B]}. \tag{11.6}$$

The condition that "B has occurred" restricts the sample space to the set of sample points in B but should not change the relative likelihood of the simple events in B; we then renormalize the probability measure of those points in B that are also in A, $P[A \cap B]$, dividing by $P[B]$ in order to account

for the reduction in the sample space. When we are considering many events, Eq. (11.6) can be expanded to

$$P[A \cap B \cap C \ldots \cap N] = P[A \mid BC \ldots N] \cdot P[B \mid C \ldots N]P[C \mid D \ldots N] \ldots P[N]. \tag{11.7}$$

If the occurrence of one event A does not alter the probability of occurrence of another event B, both events are said to be statistically independent,

$$P[A \mid B] = P[A], \tag{11.8}$$

which implies

$$P[A \cap B] = P[A]P[B] \tag{11.9}$$

and

$$P[B \mid A] = P[B]. \tag{11.10}$$

N events are independent if it holds that

$$P[A \cap B \cap C \ldots \cap N] = P[A]P[B] \ldots P[N]. \tag{11.11}$$

The hydrologist usually relies on his/her knowledge of the physical situation at hand in order to assume independence or dependence among events. Thus, high flows in a river from year to year may be assumed independent with more confidence than low flows, which are very much influenced by the carryover between years.

Total Probability Theorem

This theorem concerns the probability of a compound event A in a random experiment. Given a set of mutually exclusive, collectively exhaustive events $B_1, B_2, \ldots, B_n$, it is always possible to express the probability of any event A as

$$P[A] = P[A \cap B_1] + P[A \cap B_2] + \cdots + P[A \cap B_n]. \tag{11.12}$$

Every term of Eq. (11.12) can be expressed in the form of a conditional probability to yield

$$P[A] = \sum_{i=1}^{n} P[A \mid B_i]P[B_i], \tag{11.13}$$

which is the most common version of the total probability theorem.

Bayes Theorem

The Bayes theorem is fundamental to engineering and hydrologic analysis and is very important when considering the conditional probability of an event B_j given another event A. We know that

$$P[B_j|A] = \frac{P[B_j \cap A]}{P[A]} = \frac{P[A \cap B_j]}{P[A]}, \tag{11.14}$$

but we have

$$P[A \cap B_j] = P[A|B_j]P[B_j]$$

and

$$P[A] = \sum_{i=1}^{n} P[A|B_i]P[B_i],$$

which substituted in Eq. (11.14) yields

$$P[B_j|A] = \frac{P[A|B_j]P[B_j]}{\sum_{i=1}^{n} P[A|B_i]P[B_i]}. \tag{11.15}$$

The importance of Eq. (11.15) is that it allows the hydrologist to express his/her experience and judgment—which is valuable information—in the form of $P[B_j]$ or probabilities of a certain state of nature, before any sample has been taken. The information obtained from sample A is incorporated through conditional probabilities of the sample given a certain state of nature. In this manner Eq. (11.15) can also be expressed as

$$P[\text{state}|\text{sample}] = \frac{P[\text{sample}|\text{state}]\,P[\text{state}]}{\sum_{\text{all states}} P[\text{sample}|\text{state}]\,P[\text{state}]}, \tag{11.16}$$

where the $P[\text{state}]$ in the right-hand side of Eq. (11.16) represents prior probabilities and the left-hand side of the equation is the posterior probability, which incorporates both the information available before the sample was taken and the information yielded by the sample about the state of nature.

Random Variables, Ensembles, and Distributions

A random variable may be defined as a numerical variable not subject to precise prediction. It must be described in the domain of all its possible values through the aid of probability distributions.

Discrete Random Variables

A random variable is discrete when the value it can take is restricted to countable numbers. An example may be the number of rainy days during one year in Boston.

The function $P_X(x_i)$, which gives the probability of the discrete random variable X taking any possible value x_i, is called the probability mass function (pmf) of X,

$$P_X(x_i) = P[X = x_i].$$

$P_X(x_i)$ is a nonnegative function in accordance with the definition of probability and

$$\sum_{\text{all } i} P_X(x_i) = 1.$$

Clearly, the probability of X being between any two values x_j and $x_k (x_j < x_k)$ is

$$P[x_j \leq X \leq x_k] = \sum_{i=j}^{i=k} P_X(x_i). \tag{11.17}$$

Any random variable can also be described through its cumulative distribution function, which simply gives the probability of the event that the random variable takes a value equal to or less than the argument

$$F_X(x) = P[X \leq x].$$

For discrete random variables we have

$$F_X(x) = \sum_{\text{all } x_i \leq x} P_X(x_i). \tag{11.18}$$

Continuous Random Variables

When the range of variation of a random variable is continuous, the variable is called a continuous random variable. Unlike the discrete random variable, the continuous one is free to take any value on the real line, although this does not mean it must take on values over the entire domain of real numbers.

Separating the real line into a number of intervals of infinitesimal length dx, the probability that a continuous random variable X falls between x and $x + dx$ is given by $f_X(x)\,dx$, where the function $f_X(x)$ is called the probability density function (pdf) of X.

The occurrence of x in different intervals dx constitutes mutually exclusive events and thus, according to the theorem of total probability, we can

compute the probability that X takes a value in an interval of finite length,

$$P[x_1 \leq X \leq x_2] = \int_{x_1}^{x_2} f_X(x)\,dx. \tag{11.19}$$

The value $f_X(x)$ is not itself a probability; it is, rather, a measure of the probability density. The meaning of Eq. (11.19) is illustrated in Figure 11.2.

From the axioms of probability theory, we have

$$f_X(x) \geq 0$$

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

Similarly to the case of discrete random variables, we define the cumulative probability distribution or cumulative density function (cdf) of a continuous random variable by

$$F_X(x) = P[X \leq x] = \int_{-\infty}^{x} f_X(u)\,du \tag{11.20}$$

and we have the relationship

$$\frac{dF_X(x)}{dx} = f_X(x). \tag{11.21}$$

In hydrology there are many cases where the distribution of the random variable is composed of two parts: a discontinuous part, or probability mass; and a continuous part, or probability density. These are the so-called mixed distributions. An example may be the distribution of daily flows for an ephemeral

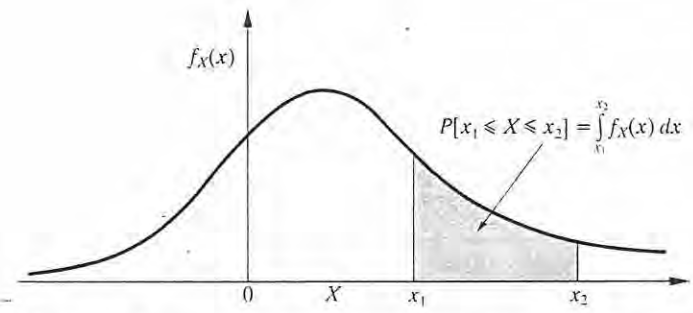


FIGURE 11.2 A probability density function and the probability of a random variable X taking values between x_1 and x_2.

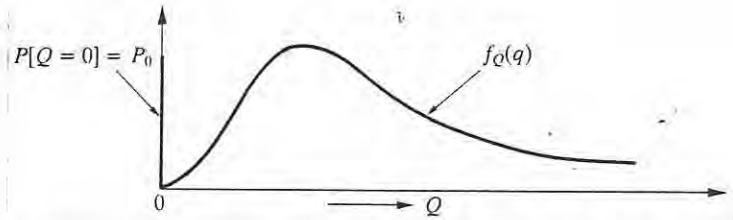


FIGURE 11.3 Example of a mixed distribution with probability P_0 of random variable Q (i.e., discharge) taking a value of 0.

stream. The stream may be dry during certain times of the year and thus we have a probability distribution that looks like Figure 11.3.

Moments and Expectation

Simple numbers are sometimes used to describe the dominant features of the behavior of a random variable, in other words, to describe the general shape of a probability density function. "These numbers usually take the form of weighted averages of certain functions of the random variable. The weights used are the pmf or pdf of the variable, and the average is called the expectation of the function" (Benjamin and Cornell [1970]).

We define the mean or the expected value $E[X]$ of a random variable X as

$$\mu_X = E[X] = \sum_{\text{all } x_i} x_i P_X(x_i) \quad \text{Discrete case}$$
$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx \quad \text{Continuous case} \tag{11.22}$$

The mean is a measure of central tendency.

In many problems we are most interested in the dispersion that the random variable X can have around its expected value. The most useful measure of dispersion is the variance defined as

$$\sigma_X^2 = \text{Var}[X] = \sum_{\text{all } x_i} (x_i - \mu_X)^2 P_X(x_i) \quad \text{Discrete case}$$
$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx. \quad \text{Continuous case} \tag{11.23}$$

The standard deviation is

$$\sigma_X = \sqrt{\sigma_X^2} = \left[\int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\right]^{1/2}.$$

From the definition of expectation we can write

$$\begin{aligned}\text{Var}[X] &= E[(X - \mu_X)^2] \\ &= E[X^2 - 2\mu_X X + \mu_X^2] \\ &= E[X^2] - 2\mu_X E[X] + \mu_X^2 \\ &= E[X^2] - E^2[X].\end{aligned}$$

The third moment around the mean is called the skewness,

$$\begin{aligned}G &= \sum_{\text{all } x_i} (x_i - \mu_X)^3 P_X(x_i) \quad \text{Discrete case} \\ G &= \int_{-\infty}^{\infty} (x - \mu_X)^3 f_X(x)\, dx \quad \text{Continuous case}\end{aligned} \tag{11.24}$$

The commonly used skewness coefficient is defined as

$$\gamma = \frac{G}{\sigma^3} \tag{11.25}$$

and takes negative and positive values. Streamflows usually have positive skewness, implying that their probability density functions have tails that extend to the right (high streamflow values). This is shown in Figure 11.4(a). Figure 11.4(b) shows a negatively skewed pdf. A symmetrical probability density function would have a skewness value of 0.

The fourth moment around the mean is called the kurtosis. It is defined analogously to the skewness.

Joint and Conditional Probability Distribution Functions

The joint probability mass function of two discrete random variables is defined as

$$P_{X,Y}(x_i, y_j) = P[X = x_i;\ Y = y_j].$$

The joint cumulative distribution function is then defined as

$$P[X \le x;\ Y \le y] = F_{X,Y}(x, y) = \sum_{x_i \le x} \sum_{y_j \le y} P_{X,Y}(x_i, y_j), \tag{11.26}$$

which has the property

$$\sum_{\text{all } x_i} \sum_{\text{all } y_j} P_{X,Y}(x_i, y_j) = 1.$$

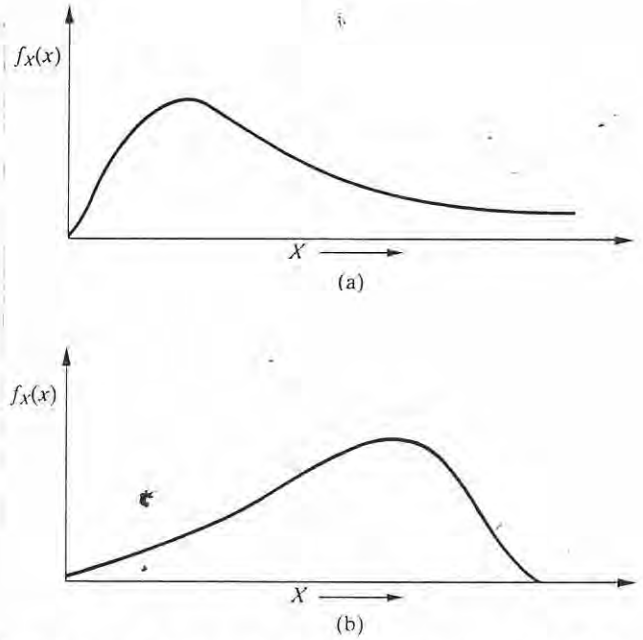


FIGURE 11.4 (a) Probability density function with positive skew coefficient. (b) Probability density function with negative skew coefficient.

The behavior of the random variable X, irrespective of the other random variable Y, is described by the marginal pmf

$$P_X(x_i) = P[X = x_i] = \sum_{\text{all } y_j} P_{X,Y}(x_i, y_j).$$

A different type of distribution is the one that describes the behavior of the random variable X given the random variable Y. It is called the conditional pmf of X given Y:

$$\begin{aligned} P_{X|Y}(x_i, y_j) &= P[X = x_i \mid Y = y_j] \\ &= \frac{P[X = x_i;\ Y = y_j]}{P[Y = y_j]} \\ &= \frac{P_{X,Y}(x_i, y_j)}{P_Y(y_j)}. \end{aligned} \tag{11.27}$$

"It is the relationship between the conditional distribution and the marginal distribution that determines how much an observation of one variable helps in the prediction of the other" (Benjamin and Cornell [1970]).

As for discrete random variables, we can similarly define joint probability functions for continuous random variables. Thus the probability that X lies in the interval $[x, x + dx]$ and Y lies in the interval $[y, y + dy]$ is given by $f_{X,Y}(x,y)\,dx\,dy$, where the function $f_{X,Y}(x,y)$ is called the joint pdf of X and Y.

We then have

$$P[x_1 \le X \le x_2;\ y_1 \le Y \le y_2] = \int_{x_1}^{x_2}\int_{y_1}^{y_2} f_{X,Y}(x,y)\,dx\,dy. \tag{11.28}$$

Clearly, $f_{X,Y}(x,y)$ must satisfy

$$f_{X,Y}(x,y) \ge 0$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1.$$

The joint cumulative distribution function is now

$$F_{X,Y}(x,y) = P[X \le x;\ Y \le y] = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,du\,dv, \tag{11.29}$$

with the property that

$$\frac{\partial^2}{\partial x\,\partial y}F_{X,Y}(x,y) = f_{X,Y}(x,y). \tag{11.30}$$

The behavior of X, irrespective of Y, is given by the marginal pdf of X:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy. \tag{11.31}$$

11.3 MODELS OF PROBABILITY

There are innumerable functions that satisfy the probability axioms and hence are adequate models of probability. Next, we will review some of the simplest and most commonly used models to represent hydrologic variables.

11.3.1 Models of Discrete Random Variables

Bernoulli Trials

Imagine a river discharge of a given magnitude Q^*. From many years of observations it is determined that in any one year the probability of a flood larger or equal to Q^* is P. We can reasonably assume that the flood of any one

year is independent of that of another year. We can define a discrete random variable X that takes a value of 1 if a flood greater than or equal to Q^* occurs in a year; the variable X takes a value of 0 otherwise. The probability mass function of X is then

$$P_X(x) = \begin{cases} P & \text{if } x = 1 \\ 1 - P & \text{if } x = 0 \end{cases} \tag{11.32}$$

The mean or expected value of X is

$$\mu_X = E[X] = 1P + 0(1 - P) = P. \tag{11.33}$$

We have defined expectation as the operation of taking the mean of a function.

The variance of X is

$$E[(X - \mu_X)^2] = (1 - P)P. \tag{11.34}$$

Binomial Distribution

The Bernoulli trial has but two outcomes and is repeated only once. Nevertheless, every year we have the same probability P that a flood greater or equal to Q^* will occur. In other words, the independent Bernoulli trial is repeated year after year. A valid question is then: What is the probability that our flood will occur two times in the next three years? In three trials (three years) two floods can occur in three different ways: in the first and second years; in the second and third; or in the first and third. Since every year is independent of the others we can find the joint probability of the above three-year sequences (after Eq. 11.11):

$$P[1, 0, 1] = P(1 - P)P = P^2(1 - P)$$

$$P[0, 1, 1] = (1 - P)PP = P^2(1 - P)$$

$$P[1, 0, 1] = P(1 - P)P = P^2(1 - P)$$

where 1s are used for indicating a flood, and 0s are years of no flood. Since we have three independent ways of achieving our two years of floods, the desired probability is

$$P[2 \text{ floods in 3 years}] = 3P^2(1 - P).$$

The above result can be generalized. The probability of k floods in n years is

$$P[K = k] = B(n, P) = \binom{n}{k} P^k(1 - P)^{n-k}, \tag{11.35}$$

where K is the random variable representing the number of floods in n years and

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

represents the number of ways that k events (i.e., floods) can occur in n trials (i.e., years).

The above is the binomial distribution. Its first two moments are

$$\mu_K = E[K] = nP \tag{11.36}$$

$$\text{Var}[K] = nP(1-P). \tag{11.37}$$

The results should be clear in the context of the independence of the n Bernoulli trials that lead to the binomial distribution.

If the probability of discharge Q^* being exceeded in any one year is 0.02, then the mean number of times that the flood will be exceeded in 50 years is 1, with a variance of 0.98. Figure 11.5 shows various forms of the binomial for various values of parameters P and n.

Geometric Distribution

Another relevant question in hydrology is: How many years will pass before discharge Q^* is equalled or exceeded? Using the binomial probability concept,

$$P[\text{flood in the } n\text{th year}] = (1-P)^{n-1}P, \tag{11.38}$$

or $n-1$ consecutive failures followed by a flood in the nth year.

The above is called the geometric distribution and represents the probability mass function that a random variable N representing the number of years to the first flood is n. Figure 11.6 shows the geometric distribution.

We can now ask: What is the probability that the next flood will occur in n or less years? The answer is given by the geometric cumulative mass function. A simple way to obtain it is to ask an equivalent question: What is the probability that at least one flood will occur in the next n years? This is the complement of the probability of no floods in n years, which is $(1-P)^n$. Hence,

$$P[N \le n] = 1 - (1-P)^n \qquad n = 1, 2, \ldots. \tag{11.39}$$

The mean of the geometric distribution is

$$\mu_N = 1/P. \tag{11.40}$$

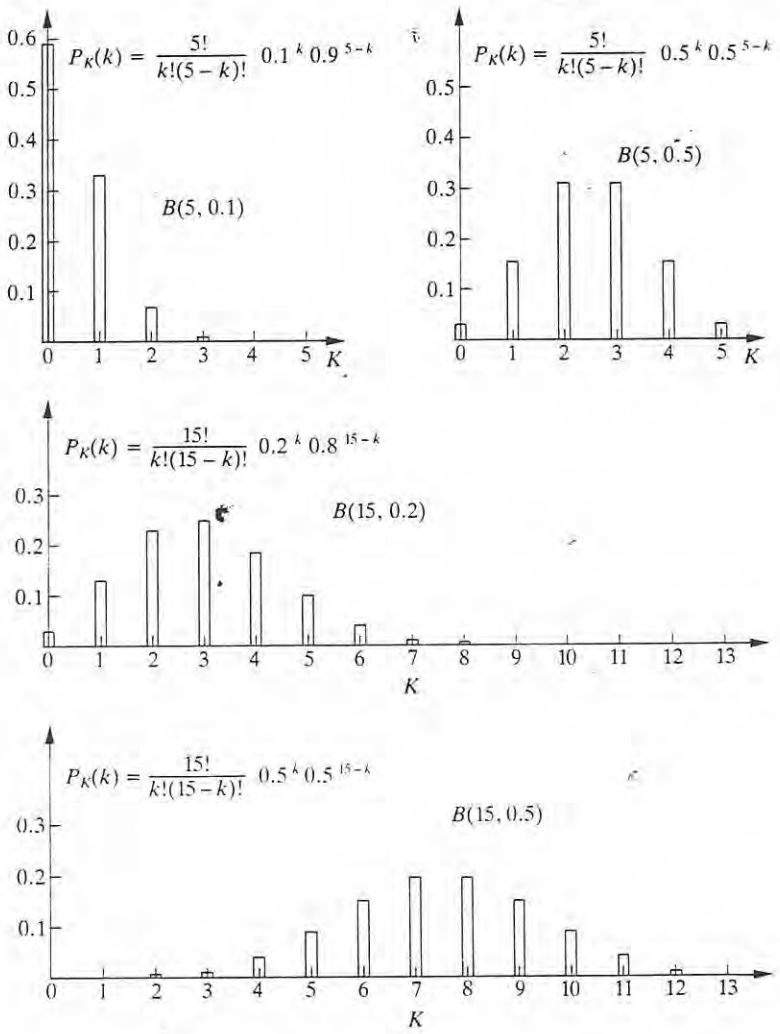


FIGURE 11.5 Binomial distribution $B(n, P)$.

The variance is

$$\sigma_N^2 = (1 - P)/P^2. \tag{11.41}$$

Note that since P is the probability of exceedance in any one year, the mean of the geometric distribution is the recurrence interval or the average number of years that will pass before a flood of magnitude Q^* or greater occurs.

Equation (11.39) is commonly called the risk of flood, since it answers the question: What is the risk of having at least one event (i.e., flood) of recur-

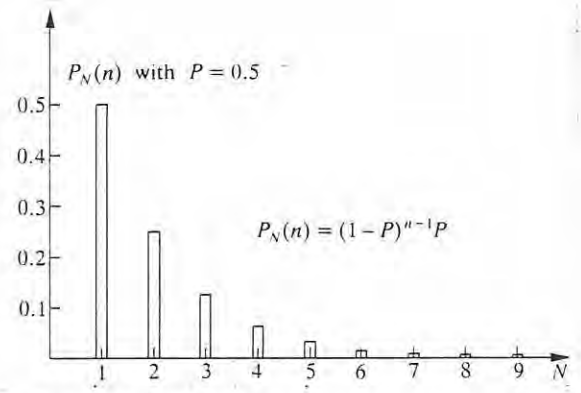


FIGURE 11.6 Geometric distribution $G(P)$.

rence $1/P$ in n years? Table 11.1 evaluates risk for several values of $T = 1/P$ and n, where P is the probability of exceedance in a year.

EXAMPLE 11.1

Uses of the Binomial Distribution

The magnitude of the T-year flood has been defined as that which is exceeded with probability $1/T$ in any given year. If we assume that successive annual floods are independent, several interesting questions can be answered.

A. What is the probability that exactly one flood equal to or in excess of the 50-year flood will occur in a 50-year period? The answer to this type of

TABLE 11.1 Risk of Event Occurring in Specified Number of Years

NUMBER OF YEARS	RECURRENCE INTERVAL			
	10 yr (%)	50 yr (%)	100 yr (%)	500 yr (%)
1	10	2	1	0.2
2	19	4	2	0.4
5	41	10	5	1
10	65	18	10	2
20	88	33	18	4
30	96	45	26	6
50	99	64	40	10
75	99.9	78	53	14
100	99.99	87	63	18
200	99.999	98	87	33
500	99.999	99.99	99	63

question is given by the binomial distribution, Eq. (11.35). The probability of one success (flood) in 50 trials (years) is

$$P[K = 1] = \binom{50}{1}\left(\frac{1}{50}\right)^1\left(1 - \frac{1}{50}\right)^{49} = 0.37.$$

B. What is the probability that exactly three floods will be equal to or will exceed the 50-year flood in 50 years? Again, the binomial distribution gives the answer.

$$P[K = 3] = \binom{50}{3}\left(\frac{1}{50}\right)^3\left(1 - \frac{1}{50}\right)^{47} = 0.06.$$

Note that the probability of three floods of 50-year recurrence in 50 years is much less than the probability of one such flood occurring.

C. What is the probability that one or more floods will equal or exceed the 50-year flood in 50 years? The key to this question is in the words "one or more." Exploiting the properties of mutually exclusive and collectively exhaustive events we can say

$$P[\text{one or more floods in 50 years}] = 1 - P[\text{no floods in 50 years}]$$

or

$$P[\text{one or more floods in 50 years}] = 1 - \binom{50}{0}\left(\frac{1}{50}\right)^0\left(1 - \frac{1}{50}\right)^{50} = 0.64.$$

Note that the probability of one or more floods is nearly twice the probability of a single flood, which was obtained in part A above.

D. If an agency designs each of 20 independent flood-control systems (i.e., systems in widely scattered locations with independent hydrology) for a particular 500-year flood, what is the distribution of the number of systems that will fail, because of the occurrence of floods with 500-year return periods or larger, at least once within the first 50 years after their construction? If each system is independent, the answer to this question is given by a binomial distribution. Define a random variable K as the number of systems that fail. Then

$$P[K = k] = \binom{20}{k} P_1^k (1 - P_1)^{20-k},$$

where P_1 is the probability that any one system fails at least once in 50 years given that it has a probability of failure of 1/500 in any one year. P_1 is obtained as in part C above.

$$\begin{aligned} P_1 &= P[\text{one or more failures of any one system}] \\ &= 1 - P[\text{no failure by any one system}] \\ &= 1 - \binom{50}{0}\left(\frac{1}{500}\right)^0\left(1 - \frac{1}{500}\right)^{50} \\ &= 1 - \left(\frac{499}{500}\right)^{50} = 0.095. \end{aligned}$$

P_1 can be used in the previous equation to find the distribution of the number of systems that fail:

$$P[K = k] = \binom{20}{k}(0.095)^k(1 - 0.095)^{20-k}.$$

For example,

$$P[K = 0] = 0.136,$$
$$P[K = 1] = 0.285. \blacklozenge$$

11.3.2 Models of Continuous Random Variables

Throughout the previous section we have assumed that a probability P of exceeding a given event (i.e., flood) in a given unit time period is known. For example, P could be the probability that a flood of magnitude Q^* is equaled or exceeded in a year or the probability that total rainfall depth in any given day exceeds a magnitude D^*. Both discharge and rainfall depth are continuous variables that, during any time period, can take values between 0 and infinity. The probability of them taking any value within a range must be described by a probability density function. Only through their pdf can we define P.

There are innumerable functions that are valid probability density functions. Following is a very limited collection of some of the most common, particularly in hydrology.

Gaussian or Normal Distribution

The Gaussian probability density curve is probably the most common model of probability. It arises from the argument of the central limit theorem. Briefly, this theorem states that the normal distribution is asymptotically the model for a sum of a large (infinite) number of identically distributed random

variables. Given that many natural and man-made processes result from additive mechanisms, the commonality of this distribution is not surprising.

The Gaussian pdf is shown in Figure 11.7. It is characteristically bell-shaped, symmetrical, and extends from minus infinity to infinity. Since natural processes, like river discharges, are rarely defined as negative, the Gaussian pdf is obviously limited in its relevance to hydrology. It can nevertheless be useful. The curves of Figure 11.7 are described by a two-parameter function,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \qquad -\infty < x \leq \infty. \tag{11.42}$$

The two parameters are the mean μ, which centers the distribution around a preferred value and σ, the standard deviation, which tells us how dispersed are occurrences of the random variable around its mean. The cumulative Gaussian distribution is commonly tabulated as in Table 11.2.

The Gaussian pdf, like any of the other functions we will see in this section, could be used to analyze data sets. The following example defines two common types of streamflow data sets and analyzes them using the Gaussian pdf.

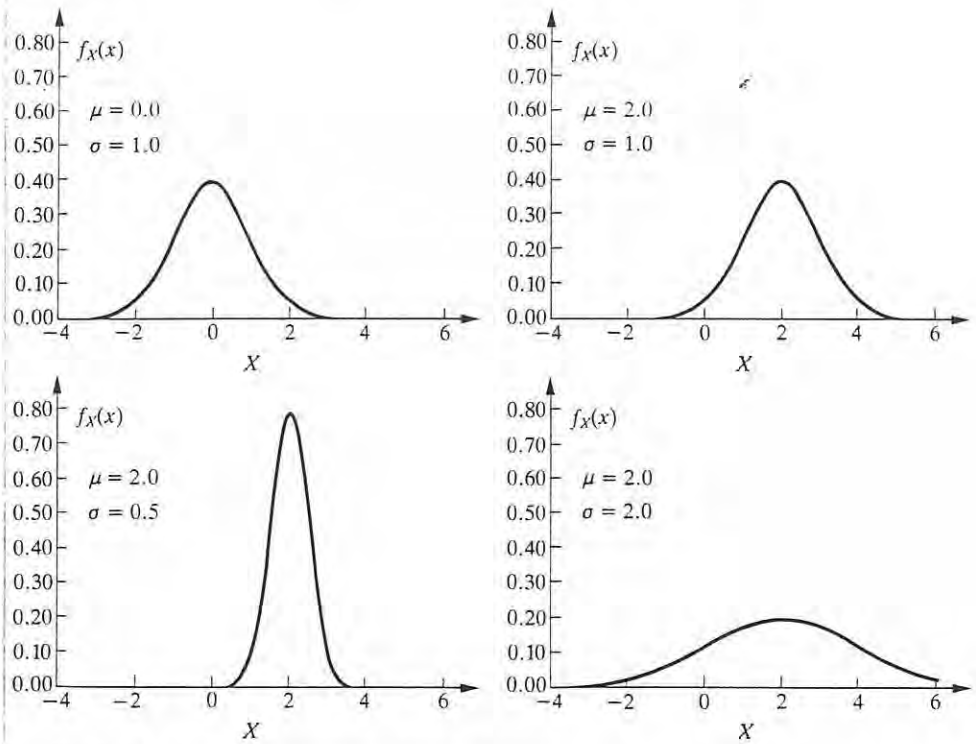


FIGURE 11.7 Gaussian probability density function.

TABLE 11.2 Values of the Standardized Normal Distribution

THE CUMULATIVE DISTRIBUTION FUNCTION, $F_U(u) = \int_{-\infty}^{u} f_U(u)\,du$

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725									
2.1	0.98214									
2.2	0.98610									
2.3	0.98928									
2.4	0.99180									
2.5	0.99379									
3.0	0.99865									
3.5	0.999767									
4.0	0.9999683									
4.5	0.9999966									
5.0	0.99999971									
5.5	0.999999981									

u	2.32	3.09	3.72	4.27	4.75	5.20	5.61	6.00	6.36	6.71
$1 - F_U(u)$	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}	10^{-11}

Sources: A. Hald, *Statistical Tables and Formulas*. Copyright © 1952 by Wiley. Reprinted by permission of John Wiley & Sons, Inc. National Bureau of Standards [1953].

EXAMPLE 11.2

Annual Exceedance and Annual Maxima Series

Hydrologists are commonly interested in extremes, particularly high (floods) and low (droughts) streamflows. Assume you have a daily record of streamflow over N years. That is called the complete duration series because it includes all available information. If you are interested in the very high flows (floods), it is important to study separately that portion of the record that includes the desired extremes. To do that we commonly form a partial duration series. These are series that include only some of the most extreme events of the complete set, regardless of chronological order. Hence, it is assumed that extremes occur independently of each other. There are two types of partial duration series. An annual exceedance series is composed of the N highest (lowest if you are studying droughts) observed values of the process, where N is the total number of years of observation. An annual maxima (or minima) extreme value series is composed of the largest value in each year of observation; therefore it also has N points. Figure 11.8 illustrates these two types of partial duration series.

Although similar, annual exceedance and maxima series are clearly not the same; the difference being mainly at the lower-valued end, where the annual exceedance series will tend to contain larger values.

If, for example, the annual maxima series of a given river is assumed to follow a Gaussian probability distribution with known parameters, it would then be possible to find the annual flood of any desired recurrence. Imagine that indeed you have such a flood series, with mean of 300 $m^3 s^{-1}$ and standard deviation of 100 $m^3 s^{-1}$ and you desire to find the magnitude of the 100-year flood. The object is then to solve the following equation for Q_{100}.

$$P[Q > Q_{100}] = 1 - F(Q_{100}) = \frac{1}{100} = 0.01,$$

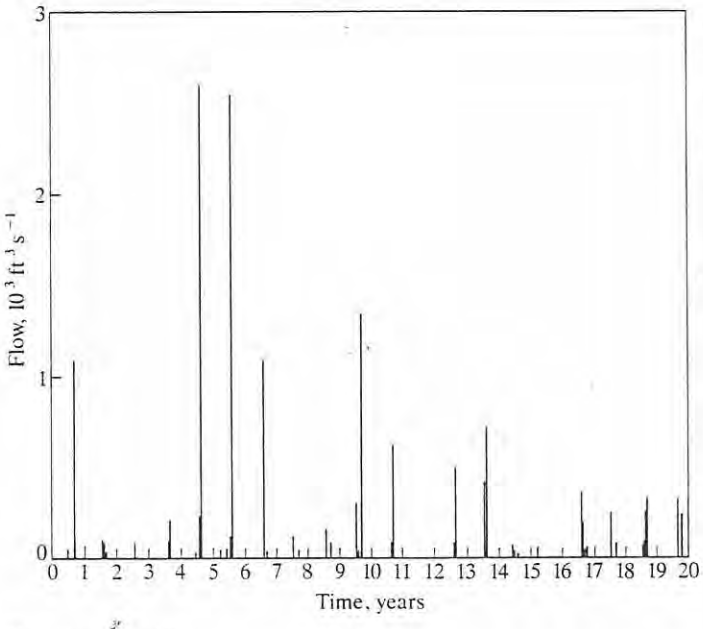
where $F(Q)$ is in this case the cumulative density function of the Gaussian distribution. The above is the same as

$$\int_{-\infty}^{Q_{100}} f_Q(q)\,dq = F(Q_{100}) = 0.99.$$

From Table 11.2, for $F_U(U_{100}) = 0.99$, we find that $U_{100} = 2.32$. The variable U in such tables is a standardized variate of mean 0 and variance 1 (standard normal deviate); therefore

$$U_{100} = \frac{Q_{100} - \mu}{\sigma} = 2.32$$

(a) I. Original flow record



II. Annual maxima

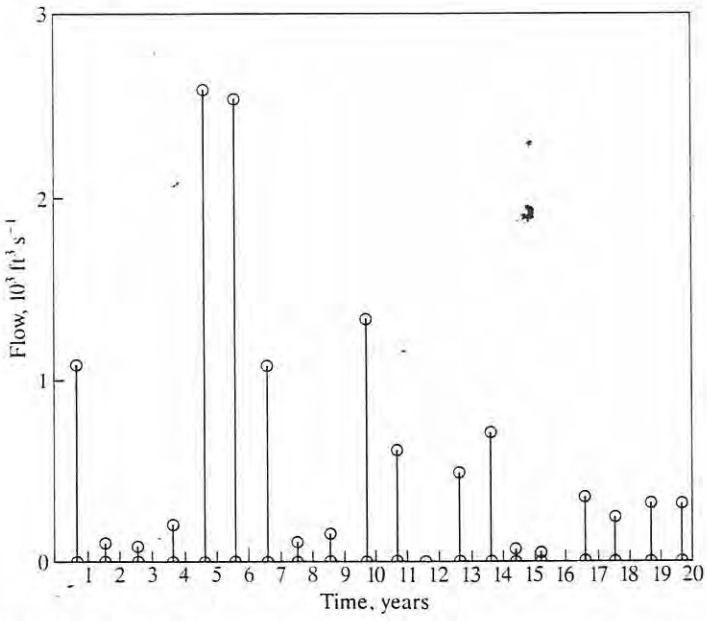
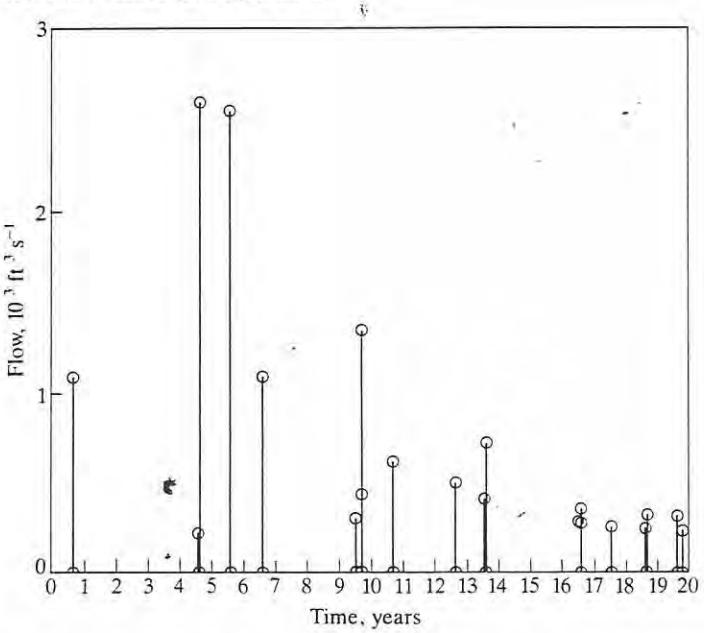


FIGURE 11.8 (a) Definition of annual maxima and annual exceedance series. (Continued on next page.) (b) Difference between the ordered annual maxima and annual exceedance series.

III. Exceedances (20 largest flows)



(b) Annual exceedance and maximum values

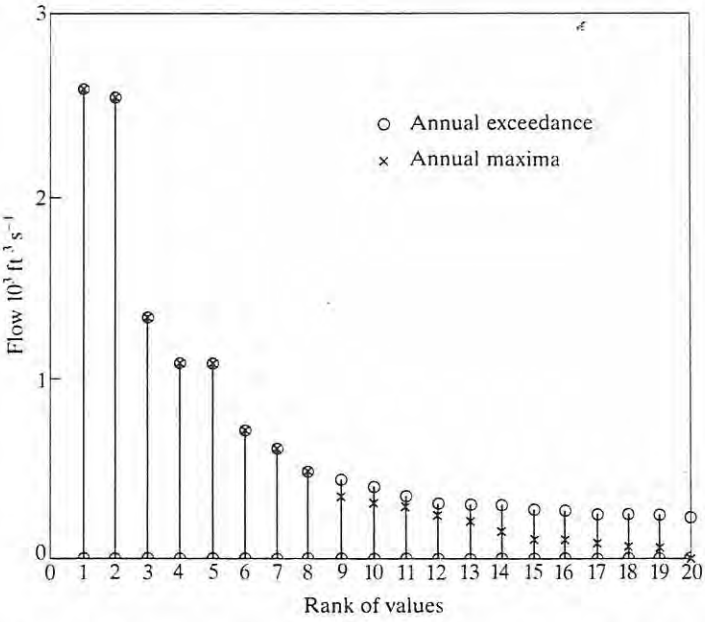


FIGURE 11.8 (*Continued*)

or

$$Q_{100} = \sigma U_{100} + \mu .$$

From the given values of σ and μ, we get $Q_{100} = 532\ \mathrm{m^3 s^{-1}}$.

Note that the desired discharge was the result of an equation of the form

$$Q_T = K_T \sigma + \mu . \tag{11.43}$$

The coefficient K_T is called the frequency factor and is dependent on the probability model assumed. For the normal distribution, K_T is simply the standard normal deviate. ◆

Log-Normal Distribution

Since processes like streamflow or rainfall do not take negative values, the use of the Gaussian pdf is an obvious approximation. If the mean and variance of the modeled process are such that significant probabilities of negative values exist, then the Gaussian approximation would be very bad.

One distribution which is only defined for positive values is the log-normal distribution. It arises naturally from processes that are multiplicative in nature. Let

$$Y = Y_1 \times Y_2 \times \cdots \times Y_n ,$$

where Y_i are independent, identically distributed random variables. Then

$$X = \ln Y = \ln Y_1 + \ln Y_2 + \cdots + \ln Y_n . \tag{11.44}$$

If the $\ln Y_i$ are independent and identically distributed, then for large n the central limit theorem would imply that $X = \ln Y$ is Gaussian or normally distributed. In such cases Y is said to be log-normally distributed.

The log-normal pdf takes the form

$$f_Y(y) = \frac{1}{y\sigma_X\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln y - m_X}{\sigma_X}\right)^2\right] \qquad y \geq 0 , \tag{11.45}$$

where m_X and σ_X are the mean and variance of the transformed variable, $X = \ln Y$. Figure 11.9 illustrates the above equation.

Note that Eq. (11.45) is parameterized in terms of moments of the transformed variable X. It can be shown that the median (value with a 0.5 probability of being exceeded) of Y, $\breve{m}_Y$, is related to the mean of $\ln Y$, m_X, by

$$\ln \breve{m}_Y = \breve{m}_X = m_X . \tag{11.46}$$

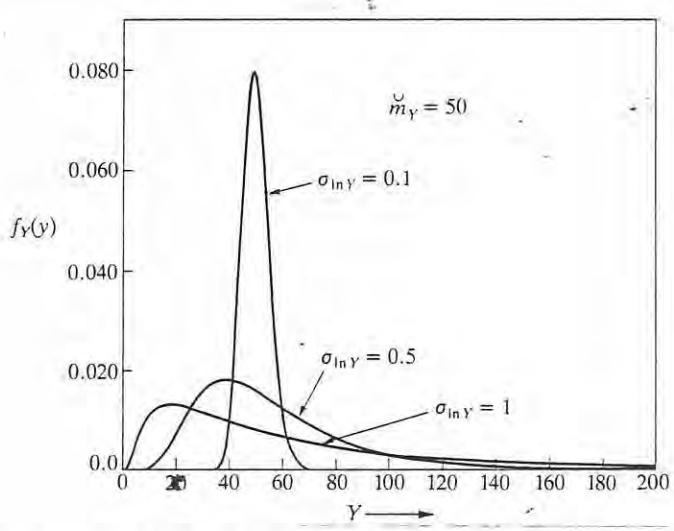


FIGURE 11.9 Log-normal distributions showing influence of $\sigma_{\ln y}$.

Hence, Eq. (11.45) can be expressed as

$$f_Y(y) = \frac{1}{y\sigma_X\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_X}\ln\left(\frac{y}{\breve{m}_Y}\right)\right]^2\right\} \qquad y \geq 0. \tag{11.47}$$

Since the logarithm is a one-to-one monotonically increasing transformation, then it should be clear that

$$F_Y(y) = F_X(\ln y), \tag{11.48}$$

where $F_X(x)$ is the cumulative Gaussian distribution in this case.

In essence, to use the log-normal distribution, the hydrologist should simply take the logarithms of the variable of interest (i.e., peak discharge) and define a standard normal deviate

$$U = \frac{\ln Y - m_{\ln Y}}{\sigma_{\ln Y}} = \frac{\ln(Y/\breve{m}_Y)}{\sigma_{\ln Y}},$$

which can be treated as illustrated in the previous section. An exponentiation of the logarithms is required to revert back to original variables after all operations are finished.

There are some useful relations between moments of log transformed and untransformed variables. The median of the untransformed log-normal variable Y is related to its mean by

$$\breve{m}_Y = m_Y \exp\left(-\frac{1}{2}\sigma^2_{\ln Y}\right). \tag{11.49}$$

The variance of Y is given by

$$\sigma^2_Y = m^2_Y(\exp \sigma^2_X - 1). \tag{11.50}$$

Also,

$$m_X = \ln m_Y - \frac{1}{2}\sigma^2_X. \tag{11.51}$$

EXAMPLE 11.3

Frequency Analysis with Log-Normal Distribution

Let us now assume that the annual maxima streamflows of Example 11.2 are log-normally distributed. What is the 100-year flood in that case?

From Eq. (11.50), we have

$$\sigma^2_X = \ln\left(\frac{\sigma^2_Y}{m^2_Y} + 1\right).$$

Using $m_Y = 300\ \text{m}^3\text{s}^{-1}$ and $\sigma_Y = 100\ \text{m}^3\text{s}^{-1}$ as given in Example 11.2, we get $\sigma^2_X = 0.105$. Using this result in Eq. (11.51),

$$m_X = \ln(300) - \frac{1}{2}(0.105)$$
$$= 5.65.$$

In Example 11.2 we saw that the standard normal deviate with a probability of exceedance of 0.01 (100-year recurrence) is 2.32. Using Eq (11.48),

$$\ln Y_{100} = \sqrt{0.105}\ 2.32 + 5.65$$

or

$$Y_{100} = e^{6.4} = 602\ \text{m}^3\text{s}^{-1}.$$

Note that this answer is different from that obtained under the normal-distribution assumption. Different models will yield different results, some-

times very different! This highlights the importance of model selection. This is a subject of tremendous controversy in hydrology as well as in statistics and probability. It is a difficult issue and one that remains largely unresolved.

The log-normal distribution is very common in hydrology. Annual maxima of discharges, storm depths, and hydraulic conductivities are a few of the hydrologic variables that have been modeled as log-normally distributed. ◆

Extreme-Value Distributions—
The Gumbel Distribution

Mathematically the annual maxima series can be expressed as

$$Y = \max\{X_1, X_2, \ldots, X_m\}, \tag{11.52}$$

where X_i; $i = 1, \ldots, m$ is the set of flows from which the maximum is chosen. If the X_is are independent and identically distributed, then the cumulative probability function of Y is

$$F_Y(y) = P[Y \leq y] = P[X_1 < y]P[X_2 < y] \ldots P[X_m \leq y] = F_X^m(y), \tag{11.53}$$

where $F_X(x)$ is the cdf (cumulative density function) of the Xs. The pdf of Y is then (taking derivatives of Eq. 11.53)

$$f_Y(y) = mF_X^{m-1}(y)f_X(y). \tag{11.54}$$

A powerful asymptotic result is that as m becomes large, and if $F_X(x)$ is such that it goes as $1 - e^{-g(x)}$ for large values of x, then,

$$\begin{aligned} F_Y(y) &= \exp[-e^{-\alpha(y-u)}] \qquad -\infty \leq Y \leq \infty \\ f_Y(y) &= \alpha \exp[-\alpha(y - u) - e^{-\alpha(y-u)}], \end{aligned} \tag{11.55}$$

The above is the Type I extreme, large-value, distribution. In hydrology it is also called the Gumbel distribution. The Gumbel is a two-parameter distribution. The parameter u is the mode (most probable) value of the distribution and α is a measure of dispersion.

The two parameters are related by

$$\mu = u + \frac{0.577}{\alpha} \tag{11.56}$$

$$\sigma^2 = \frac{1.645}{\alpha^2}. \tag{11.57}$$

The distribution has a skewness of 0.577. The previous equations can be used with sample estimates of μ and σ^2 to find parameters u and α. This exercise is

called the method of moments. The general topic of parameter estimation will be briefly treated later.

Figure 11.10 shows the Gumbel distribution and Table 11.3 gives the cdf for a standardized variable of the form $\alpha(y - u)$. The frequency factor of Eq. (11.43) for the Gumbel distribution is given by Chow [1964] as

$$K_T = -\frac{\sqrt{6}}{\pi}\{0.577 + \ln[\ln T - \ln(T - 1)]\}, \tag{11.58}$$

where T is the desired recurrence.

The theoretical underpinning of the Gumbel distribution has made it very popular in the analysis of floods. Nevertheless, it should be viewed as another alternative whose performance must be judged relative to its ability to reproduce observed data.

There are other types of asymptotic extreme-value distributions, including several applicable to small values (droughts). The reader is referred to Gumbel [1958] for some of the original work on extremes and Benjamin and Cornell [1970] for a good summary of extreme-value functions.

EXAMPLE 11.4

Frequency Analysis with Gumbel Distribution

The Gumbel distribution can be used to find the 100-year flood of Example 11.2. One way would be to use Eq. (11.58) to find the approximate

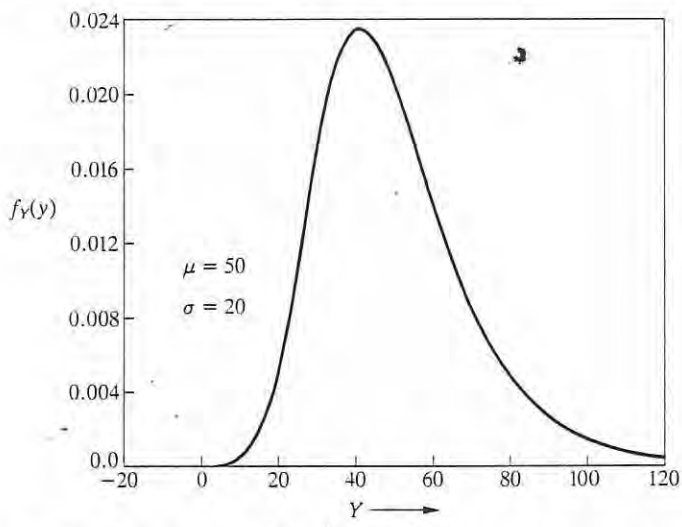


FIGURE 11.10 Type I extreme-value distribution for largest value: $\mu = 50$, $\sigma = 20$.

TABLE 11.3 Values of the Standardized Type I Extreme-Value Distribution (Largest Value) $F_W(w) = e^{-e^{-w}}$

w	cdf	pdf	w	cdf	pdf	w	cdf	pdf
−3.0		0.00000 00	−0.50	0.19229 56	0.31704 19	3.5	0.97025 40	0.02929 91
−2.9	0.00000 00	0.00000 02	−0.45	0.20839 66	0.32638 10	3.6	0.97304 62	0.02658 72
−2.8	0.00000 01	0.00000 12	−0.40	0.22496 18	0.33560 30	3.7	0.97557 96	0.02411 98
−2.7	0.00000 03	0.00000 51	−0.35	0.24193 95	0.34332 85	3.8	0.97787 76	0.02187 59
−2.6	0.00000 14	0.00001 91	−0.30	0.25927 69	0.34998 72	3.9	0.97996 16	0.01983 63
−2.5	0.00000 51	0.00006 24	−0.25	0.27692 03	0.35557 27	4.0	0.98185 11	0.01798 32
−2.40	0.00001 63	0.00017 99	−0.20	0.29481 63	0.36008 95	4.2	0.98511 63	0.01477 24
−2.35	0.00002 79	0.00029 29	−0.15	0.31291 17	0.36355 15	4.4	0.98779 77	0.01212 75
−2.30	0.00004 66	0.00046 47	−0.10	0.33115 43	0.36598 21	4.6	0.98999 85	0.00995 13
			−0.05	0.34949 32	0.36741 21	4.8	0.99180 40	0.00816 23
−2.25	0.00007 58	0.00071 89	0.0	0.36787 94	0.36787 94	5.0	0.99328 47	0.00669 27
−2.20	0.00012 04	0.00108 63	0.1	0.40460 77	0.36610 42	5.2	0.99449 86	0.00548 62
−2.15	0.00018 69	0.00160 46	0.2	0.44099 10	0.36105 29	5.4	0.99549 36	0.00449 62
−2.10	0.00028 41	0.00232 06	0.3	0.47672 37	0.35316 56	5.6	0.99630 90	0.00368 42
−2.05	0.00042 31	0.00328 66	0.4	0.51154 48	0.34289 88	5.8	0.99697 70	0.00301 84
−2.00	0.00061 80	0.00456 63	0.5	0.54523 92	0.33070 43	6.0	0.99752 43	0.00247 26
−1.95	0.00088 61	0.00622 81	0.6	0.57763 58	0.31701 33	6.2	0.99797 26	0.00202 53
−1.90	0.00124 84	0.00834 67	0.7	0.60860 53	0.30222 45	6.4	0.99833 98	0.00165 88
−1.85	0.00172 97	0.01100 04	0.8	0.63805 62	0.28669 71	6.6	0.99864 06	0.00135 85
−1.80	0.00235 87	0.01426 93	0.9	0.66593 07	0.27074 72	6.8	0.99888 68	0.00111 25
−1.75	0.00316 82	0.01823 15	1.0	0.69220 06	0.25464 64	7.0	0.99908 85	0.00091 11
−1.70	0.00419 46	0.02296 12	1.1	0.71686 26	0.23862 28	7.2	0.99925 37	0.00074 60
−1.65	0.00547 82	0.02852 48	1.2	0.73993 41	0.22286 39	7.4	0.99938 89	0.00061 09
−1.60	0.00706 20	0.03497 81	1.3	0.76144 92	0.20751 91	7.6	0.99949 97	0.00050 02
−1.55	0.00899 15	0.04236 34	1.4	0.78145 56	0.19270 46	7.8	0.99959 03	0.00040 96
−1.50	0.01131 43	0.05070 71	1.5	0.80001 07	0.17850 65	8.0	0.99966 46	0.00033 54
−1.45	0.01407 84	0.06001 78	1.6	0.81717 95	0.16498 57	8.5	0.99979 66	0.00020 34
−1.40	0.01733 20	0.07028 48	1.7	0.83303 17	0.15218 12	9.0	0.99987 66	0.00012 34
−1.35	0.02112 23	0.08147 77	1.8	0.84764 03	0.14011 40	9.5	0.99992 51	0.00007 48
−1.30	0.02549 44	0.09354 65	1.9	0.86107 93	0.12879 04	10.0	0.99995 46	0.00004 54
−1.25	0.03049 04	0.10642 20	2.0	0.87342 30	0.11820 50	10.5	0.99997 25	0.00002 75
−1.20	0.03614 86	0.12001 76	2.1	0.88474 45	0.10834 26	11.0	0.99998 33	0.00001 67
−1.15	0.04250 25	0.13423 10	2.2	0.89511 49	0.09918 16	11.5	0.99998 99	0.00001 01
−1.10	0.04958 01	0.14894 68	2.3	0.90460 32	0.09069 45	12.0	0.99999 39	0.00000 61
−1.05	0.05740 34	0.16403 90	2.4	0.91327 53	0.08285 05			
−1.00	0.06598 80	0.17937 41	2.5	0.92119 37	0.07561 62	12.5	0.99999 63	0.00000 37
−0.95	0.07534 26	0.19481 41	2.6	0.92841 77	0.06895 69	13.0	0.99999 77	0.00000 23
−0.90	0.08546 89	0.21021 95	2.7	0.93500 30	0.06283 74	13.5	0.99999 86	0.00000 14
−0.85	0.09636 17	0.22545 23	2.8	0.94100 20	0.05722 24	14.0	0.99999 92	0.00000 08
−0.80	0.10800 90	0.24037 84	2.9	0.94646 32	0.05207 75	14.5	0.99999 95	0.00000 05
−0.75	0.12039 23	0.25487 04	3.0	0.95143 20	0.04736 90	15.0	0.99999 97	0.00000 03
−0.70	0.13348 68	0.26880 94	3.1	0.95595 04	0.04306 48	15.5	0.99999 98	0.00000 02
−0.65	0.14726 22	0.28208 67	3.2	0.96005 74	0.03913 41	16.0	0.99999 99	0.00000 01
−0.60	0.16168 28	0.29460 53	3.3	0.96378 87	0.03554 76	16.5	0.99999 99	0.00000 01
−0.55	0.17670 86	0.30628 08	3.4	0.96717 75	0.03227 79	17.0	1.00000 00	0.00000 00

Source: National Bureau of Standards [1953].

frequency factor for the 100-year event.

$$K_{100} = -\frac{\sqrt{6}}{\pi}\{0.577 + \ln[\ln(100) - \ln(99)]\}$$
$$= 3.14.$$

The 100-year flood is then (from Eq. 11.43)

$$Q_{100} = \mu + K_{100}\sigma$$
$$= 300 + 3.14 \times 100 = 614 \text{ m}^3\text{s}^{-1},$$

which is the largest of the estimates we have obtained for the 100-year flood in this example.

Rather than using Eq. (11.58), we could have used Eqs. (11.56) and (11.57) to estimate Gumbel distribution parameters and Table 11.3 to get the 100-year flood.

From Eq. (11.57)

$$\alpha = \left(\frac{1.645}{100^2}\right)^{1/2} = 0.0128.$$

From Eq. (11.56)

$$u = 300 - \frac{0.577}{0.013} = 255.0.$$

From Table 11.3 the standardized variable with probability of exceedance of 0.01 (or cdf = 0.99) is about 4.6 or

$$\alpha(Q_{100} - u) = 4.6.$$

Therefore,

$$Q_{100} = \frac{4.6}{\alpha} + u = 614 \text{ m}^3\text{s}^{-1}. \blacklozenge$$

The Log-Pearson Type III Distribution

Motivated by a desire to minimize coordination efforts among U.S. government agencies making flood frequency estimates and to provide a consistent approach for defining flood risk, losses, and insurance provisions, the U.S. Water Resources Council [1967] recommended the use of the log-Pearson Type III distribution to model the annual maxima streamflow series. (*Note:* the Water Resources Council is no longer in existence. Its duties in establishing guidelines for flood frequency determinations have been taken over by

the Interagency Advisory Committee on Water Data, coordinated by the U.S. Geological Survey.) The selection was based on a comparison with five other probabilistic models and a nonparametric procedure of estimating flood frequency. The guidelines on how to use the log-Pearson Type III distribution to compute flood frequencies have been revised several times (Water Resources Council [1967, 1976, 1977]; U.S. Geological Survey [1982]). These guidelines specify recommended procedures on how to fit parameters to the distribution and handle several other statistical questions. Some of these will be discussed below at some length since the concepts are general in nature.

The Pearson Type III distribution is

$$f_X(x) = P_0\left(1 + \frac{x}{\alpha}\right)^{\alpha/\delta} e^{-x/\delta}, \tag{11.59}$$

where the most probable point (highest point) or mode is at $X = 0$. The difference between the mean and the mode is δ. The distribution extends from $X = -\alpha$ to infinity. P_0 is the value of the distribution at the mode. The above distribution has three parameters to be estimated. A variable Y is log-Pearson Type III distributed if its logarithm, $X = \log Y$, has Eq. (11.59) as a model.

The Pearson Type III distribution can be shown to be equivalent to a three-parameter gamma distribution, which takes the form (Wallis et al. [1974])

$$f_X(x) = \frac{1}{a\Gamma(b)}\left(\frac{y - m}{a}\right)^{b-1} \exp\left[-\left(\frac{y - m}{a}\right)\right], \tag{11.60}$$

where the first four moments are related to the parameters m, a, and b by

Mean

$$\mu = m + ab$$

Standard deviation

$$\sigma = a(b)^{1/2}$$

Skewness coefficient

$$\gamma = \frac{2}{b^{1/2}}$$

Kurtosis

$$\lambda = 3 + \frac{6}{b}$$

The Water Resources Council recommended that the base 10 logarithms of the annual maxima flood series, $X = \log_{10} Y$, be fitted to a Pearson Type III distribution.

Fitting distributions to data is a major and difficult problem. Many possible procedures exist. Two of the most common are the method of moments and maximum-likelihood techniques. The many issues of statistical inference, as important as they are, are beyond the scope of this book. Nevertheless, the easiest procedure to understand is the method of moments, which we have already utilized in Examples 11.1 through 11.3. In that procedure, parameters are obtained by relating them to the sample moments (i.e., estimates of the true moments obtained from finite-sized records).

The Water Resources Council recommended the method of moments to fit the Pearson Type III distribution to the base 10 logarithms of the annual floods. The procedure is as follows. First, find the logarithm of all records and define a transformed series

$$x_i = \log_{10} y_i . \tag{11.61}$$

Then, compute the first three sample moments using

Mean

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{11.62}$$

Standard deviation

$$S = \left[\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{X})^2 \right]^{0.5} = \left\{ \frac{1}{N-1} \left[\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum x_i \right)^2}{N} \right] \right\}^{0.5} \tag{11.63}$$

Skewness coefficient

$$G = \frac{N \sum_{i=1}^{N} (x_i - \overline{X})^3}{(N-1)(N-2)S^3} = \frac{N^2 \left(\sum_{i=1}^{N} x^3 \right) - 3N \left(\sum_{i=1}^{N} x_i \right) \left(\sum_{i=1}^{N} x_i^2 \right) + 2 \left(\sum_{i=1}^{N} x_i \right)^3}{N(N-1)(N-2)S^2} \tag{11.64}$$

where N is the number of data points available. A reasonable number of data points should be used. More than 25 years (points) of data is a good guideline.

The above estimates of the first three moments are based on a finite number of data points and hence have errors of estimation. Wallis et al. [1974] give the distribution functions and moments for the three statistics, based on small samples (N = 10 through 90 data points) from several parent distributions including the Pearson Type III. Generally, the smaller N is, the greater is the variance of the estimate and the worse is its bias.

The skewness coefficient estimate is particularly sensitive to large values when a small number of data points are available. For values of N less than 100 the Water Resources Council recommended the use of the so-called generalized skew. The generalized skew is a regional estimate of skew based on data from surrounding stations. The hypothesis is that regions of similar climatologic, topographic, and hydrologic characteristics should exhibit similar relations between moments of flood flows and variables such as area, basin shape, slope, channel length, annual precipitation, etc. It is recommended that skew values from surrounding stations (i.e., 40) with at least 25 years of record be related via regression to some of the possible explanatory variables. Efforts should also be made to determine regional parameter patterns, if any. Skew coefficients estimated this way are called generalized skews. The guidelines also provide estimates of generalized skews in the form of the map in Figure 11.11, which should be used except when the analyst feels that he or she has a more accurate procedure to find the generalized skew.

The skew coefficient of the logarithms actually used in fitting the log-Pearson Type III distribution is a weighted average of that computed for the station at hand and the generalized skew,

$$G_w = \frac{\text{MSE}_{\overline{G}} \cdot G + \text{MSE}_G \cdot \overline{G}}{\text{MSE}_{\overline{G}} + \text{MSE}_G}. \tag{11.65}$$

The weighting procedure is adapted from Tasker [1978]. In the above $\text{MSE}_{\overline{G}}$ is the mean square error of the generalized skew and MSE_G is the same for the station skew. The former is a function of the regression equations used. If the generalized skew is obtained from Figure 11.11, the $\text{MSE}_{\overline{G}}$ is taken as the estimated error of that map, or 0.302. The station mean square error is obtained from results of bias and variance of skew coefficients obtained from Pearson Type III distributed variables and given by Wallis et al. [1974]. An approximation to their numerical results is

$$\text{MSE}_G \simeq 10^{\{A - B[\log_{10}(N/10)]\}}, \tag{11.66}$$

where

$$\begin{aligned} A &= -0.33 + 0.08|G| && \text{if } |G| \le 0.90 \\ &= -0.52 + 0.30|G| && \text{if } |G| > 0.90 \\ B &= 0.94 - 0.26|G| && \text{if } |G| \le 1.5 \\ &= 0.55 && \text{if } |G| > 1.5 \end{aligned}$$

in which $|G|$ is the absolute value of the station skew (used as an estimate of the population skew) and N is the record length in years.

Table 11.4 gives mean square error values for station skew as a function of N and G, according to Eq. (11.66).

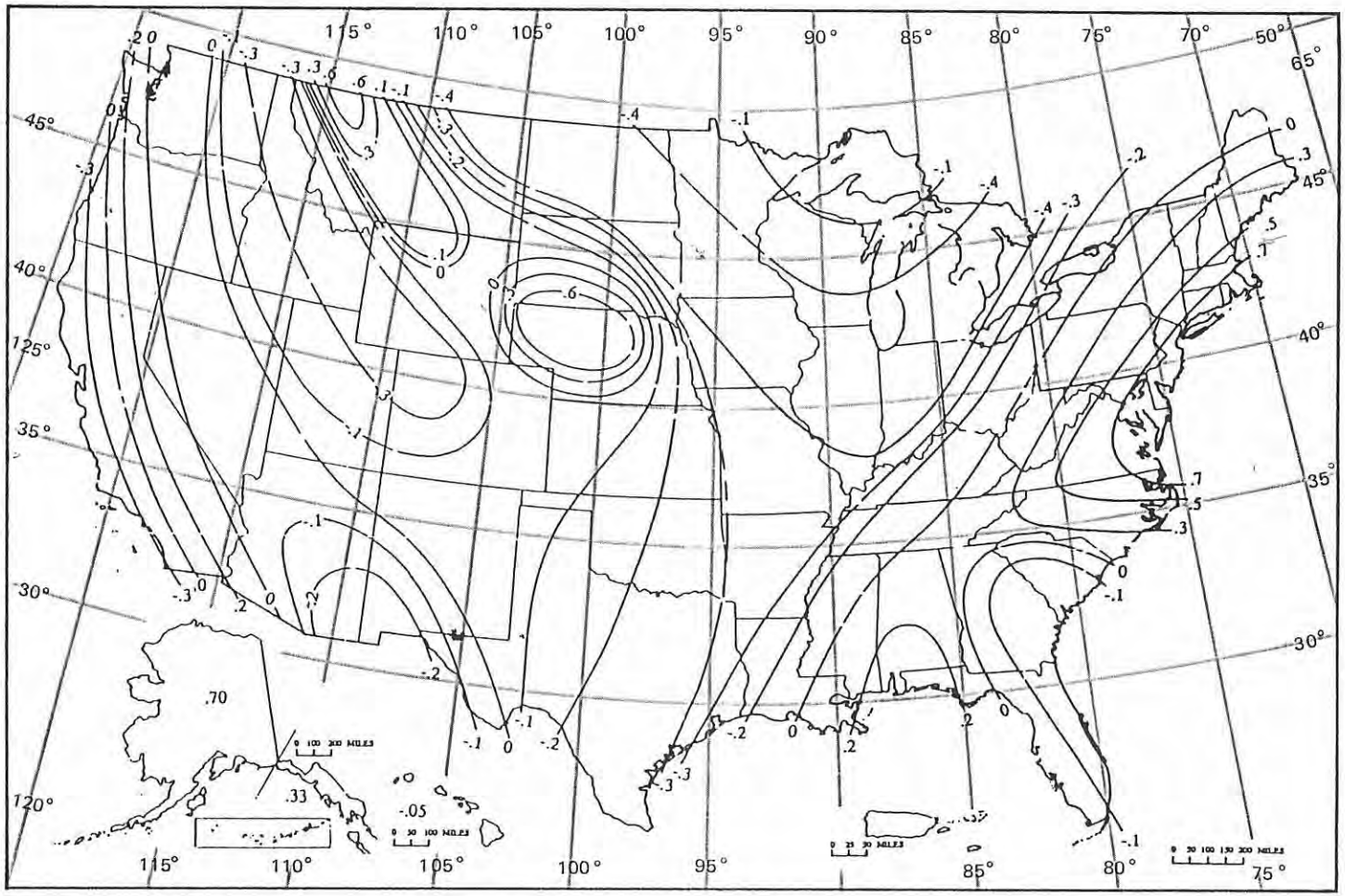


FIGURE 11.11 Generalized skew coefficients of annual maximum streamflow. Source: U.S. Geological Survey [1982].

TABLE 11.4 Summary of Mean Square Error of Station Skew as a Function of Record Length and Station Skew

STATION SKEW (G)	RECORD LENGTH, YEARS (N)									
	10	20	30	40	50	60	70	80	90	100
0.0	0.468	0.244	0.167	0.127	0.103	0.087	0.075	0.066	0.059	0.054
0.1	0.476	0.253	0.175	0.134	0.109	0.093	0.080	0.071	0.064	0.058
0.2	0.485	0.262	0.183	0.142	0.116	0.099	0.086	0.077	0.069	0.063
0.3	0.494	0.272	0.192	0.150	0.123	0.105	0.092	0.082	0.074	0.068
0.4	0.504	0.282	0.201	0.158	0.131	0.113	0.099	0.089	0.080	0.073
0.5	0.513	0.293	0.211	0.167	0.139	0.120	0.106	0.095	0.087	0.079
0.6	0.522	0.303	0.221	0.176	0.148	0.128	0.114	0.102	0.093	0.086
0.7	0.532	0.315	0.231	0.186	0.157	0.137	0.122	0.110	0.101	0.093
0.8	0.542	0.326	0.243	0.196	0.167	0.146	0.130	0.118	0.109	0.100
0.9	0.562	0.345	0.259	0.211	0.181	0.159	0.142	0.130	0.119	0.111
1.0	0.603	0.376	0.285	0.235	0.202	0.178	0.160	0.147	0.135	0.126
1.1	0.646	0.410	0.315	0.261	0.225	0.200	0.181	0.166	0.153	0.143
1.2	0.692	0.448	0.347	0.290	0.252	0.225	0.204	0.187	0.174	0.163
1.3	0.741	0.488	0.383	0.322	0.281	0.252	0.230	0.212	0.197	0.185
1.4	0.794	0.533	0.422	0.357	0.314	0.283	0.259	0.240	0.224	0.211
1.5	0.851	0.581	0.465	0.397	0.351	0.318	0.292	0.271	0.254	0.240
1.6	0.912	0.623	0.498	0.425	0.376	0.340	0.313	0.291	0.272	0.257
1.7	0.977	0.667	0.534	0.456	0.403	0.365	0.335	0.311	0.292	0.275
1.8	1.047	0.715	0.572	0.489	0.432	0.391	0.359	0.334	0.313	0.295
1.9	1.122	0.766	0.613	0.523	0.463	0.419	0.385	0.358	0.335	0.316
2.0	1.202	0.821	0.657	0.561	0.496	0.449	0.412	0.383	0.359	0.339
2.1	1.288	0.880	0.704	0.601	0.532	0.481	0.442	0.410	0.385	0.363
2.2	1.380	0.943	0.754	0.644	0.570	0.515	0.473	0.440	0.412	0.389
2.3	1.479	1.010	0.808	0.690	0.610	0.552	0.507	0.471	0.442	0.417
2.4	1.585	1.083	0.866	0.739	0.654	0.592	0.543	0.505	0.473	0.447
2.5	1.698	1.160	0.928	0.792	0.701	0.634	0.582	0.541	0.507	0.479
2.6	1.820	1.243	0.994	0.849	0.751	0.679	0.624	0.580	0.543	0.513
2.7	1.950	1.332	1.066	0.910	0.805	0.728	0.669	0.621	0.582	0.550
2.8	2.089	1.427	1.142	0.975	0.862	0.780	0.716	0.666	0.624	0.589
2.9	2.239	1.529	1.223	1.044	0.924	0.836	0.768	0.713	0.669	0.631
3.0	2.399	1.638	1.311	1.119	0.990	0.895	0.823	0.764	0.716	0.676

Source: U.S. Geological Survey [1982].

Once the first three sample moments are available, the logarithm of the base 10 discharge of a given recurrence T is

$$X_T = \overline{X} + K_T S, \tag{11.67}$$

where the frequency factor K_T is given in Table 11.5 as a function of the skewness coefficient G_w. Approximate values for K_T can be obtained, when the skewness is between -1.0 and 1.0, using

$$K_T = \frac{2}{G_w}\left\{\left[\left(K_T^n - \frac{G_w}{6}\right)\frac{G_w}{6} + 1\right]^3 - 1\right\}, \tag{11.68}$$

where K_T^n is the standard normal deviate corresponding to recurrence T.

The Water Resources Council [1977] and the U.S. Geological Survey [1982] deal with many other aspects of how to process data and use the log-Pearson Type III distribution. They describe how to handle incomplete records, zero flood years, mix historic and systematic data records (this will be touched on in a more general way later), estimate confidence limits, adjust for limited data, etc. The reader is urged to study the details in the above references. A very brief discussion of the important topic of confidence limits follows.

We have seen that moments estimates are themselves random variables because of the limited size of the data sample used to estimate them. Since these moments are used in the fitting of the distribution, the estimates of exceedance probability for a given discharge or the discharge for a given exceedance probability are themselves uncertain. It is theoretically possible to put bounds, confidence limits, within which we can state that the true answer lies with a given level of certainty. The confidence limits are dependent on the distributions being used and, for nonnormal models, are different for the exceedance probability (for a given discharge) and for the discharge (for a given probability).

A two-sided confidence interval on the logarithm of the discharge with T years recurrence X_T^* is given by

$$P[L_{T,c}(X) \le X_T^* \le U_{T,c}(X)] = 2c - 1, \tag{11.69}$$

where c is the "confidence level" such that

$$P[X_T^* > U_{T,c}(X)] = 1 - c \quad \text{and} \quad P[X_T^* \le L_{T,c}(X)] = 1 - c.$$

$L_{T,c}(X)$ is a lower limit on X_T^* with a probability of exceedance of c. $U_{T,c}(X)$ is an upper limit on X_T^* with a probability of exceedance of $1 - c$. These limits are given by the U.S. Geological Survey [1982] as

$$U_{T,c}(X) = \overline{X} + K_{T,c}^U S \tag{11.70}$$

$$L_{T,c}(\overline{X}) = \overline{X} + K_{T,c}^L S, \tag{11.71}$$

TABLE 11.5 Frequency Factors for the Pearson Type III Distribution

SKEW COEFFICIENT G_w	RETURN PERIOD IN YEARS						
	2	5	10	25	50	100	200
	EXCEEDANCE PROBABILITY						
	0.50	0.20	0.10	0.04	0.02	0.01	0.005
3.0	−0.396	0.420	1.180	2.278	3.152	4.051	4.970
2.9	−0.390	0.440	1.195	2.277	3.134	4.013	4.909
2.8	−0.384	0.460	1.210	2.275	3.114	3.973	4.847
2.7	−0.376	0.479	1.224	2.272	3.093	3.932	4.783
2.6	−0.368	0.499	1.238	2.267	3.071	3.889	4.718
2.5	−0.360	0.518	1.250	2.262	3.048	3.845	4.652
2.4	−0.351	0.537	1.262	2.256	3.023	3.800	4.584
2.3	−0.341	0.555	1.274	2.248	2.997	3.753	4.515
2.2	−0.330	0.574	1.284	2.240	2.970	3.705	4.444
2.1	−0.319	0.592	1.294	2.230	2.942	3.656	4.372
2.0	−0.307	0.609	1.302	2.219	2.912	3.605	4.298
1.9	−0.294	0.627	1.310	2.207	2.881	3.553	4.223
1.8	−0.282	0.643	1.318	2.193	2.848	3.499	4.147
1.7	−0.268	0.660	1.324	2.179	2.815	3.444	4.069
1.6	−0.254	0.675	1.329	2.163	2.780	3.388	3.990
1.5	−0.240	0.690	1.333	2.146	2.743	3.330	3.910
1.4	−0.225	0.705	1.337	2.128	2.706	3.271	3.828
1.3	−0.210	0.719	1.339	2.108	2.666	3.211	3.745
1.2	−0.195	0.732	1.340	2.087	2.626	3.149	3.661
1.1	−0.180	0.745	1.341	2.066	2.585	3.087	3.575
1.0	−0.164	0.758	1.340	2.043	2.542	3.022	3.489
0.9	−0.148	0.769	1.339	2.018	2.498	2.957	3.401
0.8	−0.132	0.780	1.336	1.993	2.453	2.891	3.312
0.7	−0.116	0.790	1.333	1.967	2.407	2.824	3.223
0.6	−0.099	0.800	1.328	1.939	2.359	2.755	3.132
0.5	−0.083	0.808	1.323	1.910	2.311	2.686	3.041
0.4	−0.066	0.816	1.317	1.880	2.261	2.615	2.949
0.3	−0.050	0.824	1.309	1.849	2.211	2.544	2.856
0.2	−0.033	0.830	1.301	1.818	2.159	2.472	2.763
0.1	−0.017	0.836	1.292	1.785	2.107	2.400	2.670
0.0	0	0.842	1.282	1.751	2.054	2.326	2.576
− 0.1	0.017	0.846	1.270	1.716	2.000	2.252	2.482
− 0.2	0.033	0.850	1.258	1.680	1.945	2.178	2.388
− 0.3	0.050	0.853	1.245	1.643	1.890	2.104	2.294
− 0.4	0.066	0.855	1.231	1.606	1.834	2.029	2.201
− 0.5	0.083	0.856	1.216	1.567	1.777	1.955	2.108
− 0.6	0.099	0.857	1.200	1.528	1.720	1.880	2.016
− 0.7	0.116	0.857	1.183	1.488	1.663	1.806	1.926
− 0.8	0.132	0.856	1.166	1.448	1.606	1.733	1.837
− 0.9	0.148	0.854	1.147	1.407	1.549	1.660	1.749

(*continued*)

TABLE 11.5 (*Continued*)

SKEW COEFFICIENT G_w	RETURN PERIOD IN YEARS						
	2	5	10	25	50	100	200
	EXCEEDANCE PROBABILITY						
	0.50	0.20	0.10	0.04	0.02	0.01	0.005
− 1.0	0.164	0.852	1.128	1.366	1.492	1.588	1.664
− 1.1	0.180	0.848	1.107	1.324	1.435	1.518	1.581
− 1.2	0.195	0.844	1.086	1.282	1.379	1.449	1.501
− 1.3	0.210	0.838	1.064	1.240	1.324	1.383	1.424
− 1.4	0.225	0.832	1.041	1.198	1.270	1.318	1.351
− 1.5	0.240	0.825	1.018	1.157	1.217	1.256	1.282
− 1.6	0.254	0.817	0.994	1.116	1.166	1.197	1.216
− 1.7	0.268	0.808	0.970	1.075	1.116	1.140	1.155
− 1.8	0.282	0.799	0.945	1.035	1.069	1.087	1.097
− 1.9	0.294	0.788	0.920	0.996	1.023	1.037	1.044
− 2.0	0.307	0.777	0.895	0.959	0.980	0.990	0.995
− 2.1	0.319	0.765	0.869	0.923	0.939	0.946	0.949
− 2.2	0.330	0.752	0.844	0.888	0.900	0.905	0.907
− 2.3	0.341	0.739	0.819	0.855	0.864	0.867	0.869
− 2.4	0.351	0.725	0.795	0.823	0.830	0.832	0.833
− 2.5	0.360	0.711	0.771	0.793	0.798	0.799	0.800
− 2.6	0.368	0.696	0.747	0.764	0.768	0.769	0.769
− 2.7	0.376	0.681	0.724	0.738	0.740	0.740	0.741
− 2.8	0.384	0.666	0.702	0.712	0.714	0.714	0.714
− 2.9	0.390	0.651	0.681	0.683	0.689	0.690	0.690
− 3.0	0.396	0.636	0.666	0.666	0.666	0.667	0.667

Source: U.S. Geological Survey [1982].

where $K_{T,c}^U$ and $K_{T,c}^L$ are new frequency factors (as in Eq. 11.67), approximated as

$$K_{T,c}^U = \frac{K_T + \sqrt{K_T^2 - ab}}{a} \tag{11.72}$$

$$K_{T,c}^L = \frac{K_T - \sqrt{K_T^2 - ab}}{a} \tag{11.73}$$

in which

$$a = 1 - \frac{U_c^2}{2(N-1)}$$

$$b = K_T^2 - \frac{U_c^2}{N}.$$

In Eqs. (11.72) and (11.73), K_T is the frequency factor with recurrence T (probability of exceedance $P = 1/T$) appearing in Eq. (11.67). It is a function of skewness G_w. U_c is a standard normal deviate with exceedance probability $1 - c$. The length of the record is N.

EXAMPLE 11.5

Calculation of Confidence Limits with the Log-Pearson Type III Distribution

Imagine that from 50 years of records (i.e., 50 data points), we had computed a mean of the base 10 logarithms of the data to be $\overline{X} = 1.5$, the standard deviation of the logarithms to be $S = 1.0$, and the skewness to be $G_w = 0.5$. The logarithm of the 100-year flood would be given by Eq. (11.67) with K_T from Table 11.5. For an exceedance probability of 0.01 and $G_w = 0.5$ the table gives $K_{100} = 2.686$ or

$$X_{100} = 1.5 + 2.686(1.0) = 4.2$$

or $Y = 10^{4.2} = 15{,}850$ in units of volume per unit time.

With 50 years of data we can compute the range in which the true X_{100} must lie, with probability of 0.9. That would imply that there must be a 0.05 probability that it is larger than the upper limit and a 0.05 probability that it is smaller than the lower limit, $c = 0.95$. (See Eq. 11.69.) We can use Eqs. (11.70) to (11.73) to answer the question.

From Table 11.2, the standard normal deviate with cumulative probability of 0.95 ($c = 0.95$) is about 1.645. Hence,

$$a = 1 - \frac{(1.645)^2}{2(49)} = 0.972$$

$$b = (2.686)^2 - \frac{(1.645)^2}{50} = 7.160.$$

From Eq. (11.72),

$$K^U_{T,c} = \frac{2.686 + \sqrt{(2.686)^2 - (0.972)(7.160)}}{0.972}$$

$$= 3.283.$$

Similarly, Eq. (11.73) yields

$$K^L_{T,c} = 2.244.$$

Using Eqs. (11.70) and (11.71)

$$U_{T,c}(X) = 1.5 + 3.283(1) = 4.783$$

$$L_{T,c}(X) = 1.5 + 2.244(1) = 3.744.$$

The 90% confidence limits in discharge units are obtained by raising 10 to the above powers. The results are {5546, 60674}. ◆

11.4 NONPARAMETRIC ESTIMATES OF EXCEEDANCE PROBABILITY

Exceedance probabilities can be estimated using the concept of plotting positions. If we are interested in floods, the N values of the annual maxima series are ordered in descending magnitude. The order or position is called m, giving the value $m = 1$ to the largest value in the set and $m = N$ to the smallest. The plotting position, recurrence ($T = 1/P$), is obtained by using one of the many formulas available, some of which are shown in Table 11.6 (Chow [1964]). The third equation in the table seems to have the best theoretical justification (Chow [1953], Thomas [1948]). Using that equation then,

$$T = \frac{N+1}{m} \quad \text{or} \quad P = \frac{1}{T} = \frac{m}{N+1}. \tag{11.74}$$

Most formulas give essentially the same value for middle values of m but differ in the extremes.

Thomas [1948] has shown that the formula

$$T = \frac{N+1}{m}$$

is in fact the mean recurrence (or probability) obtained by taking the expected value of the distribution of T for a given value of N and m.

TABLE 11.6 Plotting-Position Formulas

FORMULA* FOR T OR $1/P(X \geq x)$	
$T = \frac{N}{m}$	$T = \frac{N+0.4}{m-0.3}$
$T = \frac{2N}{2m-1}$	$T = \frac{N+1/3}{m-3/8}$
$T = \frac{N+1}{m}$	$T = \frac{3N+1}{3m-1}$
$T = \frac{1}{1-0.5^{1/N}}$	$T = \frac{N+0.12}{m-0.44}$

*N = total number of items and m = order number of the items arranged in descending magnitude; thus $m = 1$ for the largest item.

Source: Adapted from Chow [1964].

The distribution of $\overline{P}^c$ (probability of $X \leq x$) is an incomplete beta function (Thomas [1948]), so

$$\text{Prob}[\overline{P}^c \leq P_\circ] = \theta = \binom{N}{m} m \int_0^{P_\circ} P^{N-m}(1-P)^{m-1}\,dP. \tag{11.75}$$

For the largest flood $m = 1$, so

$$\theta = N\int_0^{P_\circ} P^{N-1}\,dP = P_\circ^N. \tag{11.76}$$

For example, assume that in 25 years of data, the largest flow is 4720 $ft^3 s^{-1}$. What is the mean recurrence? What is the probability that in fact the recurrence is between 20 and 100 years? The mean recurrence is

$$\frac{N+1}{1} = 26 \text{ years.}$$

The probability of exceedance corresponding to the two recurrence limits are $1/100 = 0.01$ and $1/20 = 0.05$.

Therefore, $P[Q \leq 4720] = 0.99$ or 0.95. Using Eq. (11.76) for θ

$$\theta_{0.99} - \theta_{0.95} = 0.99^{25} - 0.95^{25} = 0.5004.$$

So there is only a 50% chance that the actual recurrence is within these limits and 50% that it is outside them.

Using a similar approach we can compute the 50% confidence limits on the recurrence of the five largest floods in 25 years of data:

m (RANK)	LOWER LIMIT (YEARS)	UPPER LIMIT (YEARS)
1	18	87
2	10	26
3	6.6	14
4	5.1	9.8
5	4.2	7.3

Note that confidence limits are narrower for smaller floods, relative to N. To be 95% confident of the largest ($m = 1$) of 25 years, you must state the recurrence may be between 7.3 and 987 years!

Thomas [1948] weighted the binomial distribution by possible values of the probability of exceedance P of the annual flood to obtain a nonparametric statement of risk. (See Eqs. 11.35 and 11.39 for the definition of risk.)

$$\phi_k = \frac{m\binom{t}{k}\binom{N}{m}}{(m+k)\binom{t+N}{m+k}}, \tag{11.77}$$

where ϕ_k is the probability that in t future years the mth of N past floods will be exceeded in exactly k years. As before the notation $\binom{N}{m}$ is the number of combinations of m in N,

$$\binom{N}{m} = \frac{N!}{m!(N-m)!}.$$

When $k = 0$,

$$\phi_0 = \frac{\binom{N}{m}}{\binom{t+N}{m}},$$

if $m = 1$,

$$\phi_0 = \frac{\binom{N}{1}}{\binom{t+N}{1}} = \frac{N}{t+N},$$

so if $t = N$ the probability of no exceedance of the largest flood in past N years in a future period of N years is 50%.

An illustrative example is to design a cofferdam to protect against the sixth largest flow in the past 25 years during the five-year construction period of a bigger dam. The probability of no exceedance ($k = 0$) in the next five years is 0.298. So the probability of exceedance at least once in five years is $1 - 0.298 = 0.702$.

The probability of the design flow being exceeded more than twice can be obtained as one minus the probability of zero, one, or two exceedances:

$$1 - \phi_0 - \phi_1 - \phi_2 = 1 - 0.898 = 0.102.$$

*11.5 NOVEL APPROACHES AND FUTURE DIRECTIONS

11.5.1 Derived Distributions

Many times, data may be lacking or completely nonexistent. It may be possible, though, to deterministically relate the variable of interest to a better-defined random variable. Typical examples are soil moisture, basin total yield, and discharge volumes, all related to accessible rainfall data. The derived distribution approach augments information by introducing the engineer's/scientist's knowledge of the physical processes at hand into the probabilistic analysis.

In 1972, Eagleson utilized the concept to derive the distribution of flood peaks as a function of rainfall probabilistic properties and basin characteristics. Eagleson [1978] followed the same ideas in obtaining the probability density function of annual basin yield (surface plus groundwater) as a function of climate and basin-soil properties. Howard [1976] uses derived distributions in obtaining a procedure for the design of combined storage and water-treatment facilities.

The mechanics of derived distributions are well established in probability theory (Benjamin and Cornell [1970]). The conceptual framework is the following. Assume that a variable Q is functionally related to a vector of parameters $\boldsymbol{\theta}$ by

$$q = Q(\boldsymbol{\theta}). \tag{11.78}$$

The elements of the vector $\boldsymbol{\theta}$ are random variables with a given joint probability density function $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and corresponding cumulative distribution $F_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.

Due to the randomness of $\boldsymbol{\theta}$, the variable Q is also a random variable with cumulative distribution,

$$F_Q(q) = \int_{R_q} f_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{11.79}$$

where R_q is the region within the possible values of $\boldsymbol{\theta}$ for which $Q \leq q$.

Chan and Bras [1979] concentrated on obtaining the theoretical distribution of the volume of the hydrograph above a given threshold. The variable in question is illustrated in Figure 11.12. The volume above a threshold discharge is required for the design of flood storage devices in urban areas. For example, the threshold discharge shown in the figure may be the maximum capacity of treatment of water from a combined sewer system. The exceedance volume must be stored or spilled, possibly contaminating receiving water bodies.

Since hydrograph volume data is scarce or nonexistent, the approach taken was to relate the available probabilistic description of the rainfall process to a deterministic model of flood volumes in order to derive the distribution of the latter. The typical behavior of the cumulative density function of volume derived by Chan and Bras [1979] is shown in Figure 11.13. There, only the parameter corresponding to the length of overland flow L is varied while all others remain constant.

The important features are that the probabilistic distribution is mixed. There is a finite probability of attaining zero volume above a given threshold and a continuous probability density for volumes greater than zero. As length of overland flow decreases, the peak discharges relative to qth decrease, leading to an increasing probability of zero volume above a fixed threshold discharge. Any parameter influencing peak discharge in a similar manner will have the same effect on the cumulative distribution function. For example,

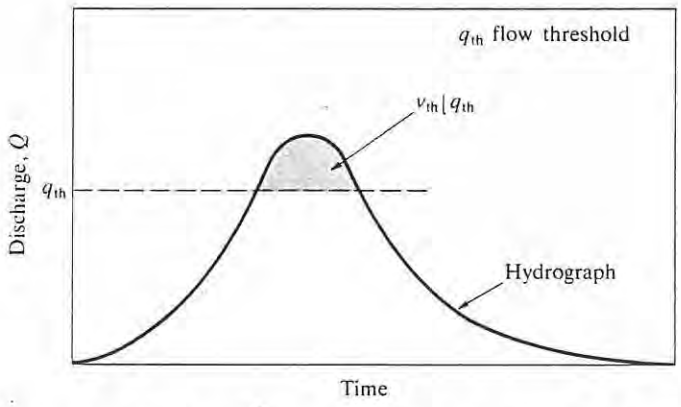


FIGURE 11.12 Volume above a given threshold discharge. Source: S.-O. Chan and R. L. Bras, "Urban Storm Water Management: Distribution of Flood Volumes," *Water Resources Res.* 15(2):371–382, 1979. Copyright by the American Geophysical Union.

increasing qth has the same qualitative effect on the distribution as decreasing length of overland flow.

Hebson and Wood [1982] and Diaz-Granados et al. [1984] used the derived distribution method to obtain the distribution of annual maxima from simple assumptions for the distributions for rainfall intensity and duration. They related rainfall properties to floods via river basin geomorphology (tree network) and models of runoff production (see Chapter 12).

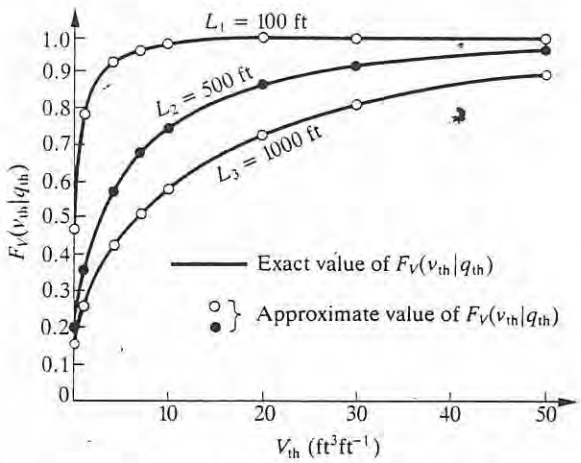


FIGURE 11.13 Cumulative density function of volume above a given threshold discharge. Source: S.-O. Chan and R. L. Bras, "Urban Storm Water Management: Distribution of Flood Volumes," *Water Resources Res.* 15(2):371–382, 1979. Copyright by the American Geophysical Union.

11.5.2 Regional Analysis

It should be intuitively obvious that climatic, geologic, and geomorphologic homogeneity of a region should allow us to transfer information from one basin to another in the region so as to augment data available for statistical inference. There has been a long history of research on how to achieve this objective. For example, investigators have developed regression equations to relate statistical moments such as the mean and variance to basin properties such as area, annual precipitation, slopes, etc. This information is then combined with existing on-site data to obtain improved probabilistic models of extremes at the location of interest.

Regional analysis is by now a fairly well accepted procedure. The Water Resources Council [1977] included it as part of their recommendation for estimating parameters of the log-Pearson Type III distribution of annual streamflow maxima as discussed in Section 11.3.2.

Most recently, Hosking and Wallis [1986a, 1988] reported very encouraging results using one of the simplest of regionalizing procedures, the flood index method. In the flood index method, data from several basins in a region are scaled by dividing by a characteristic quantity, commonly taken as the mean annual flow. The scaled data sets are then lumped and fitted with a common distribution, which becomes the regional distribution. The flow of a given recurrence for any basin in the region is then obtained from the regional distribution multiplied by the sample mean annual flow in that basin. This procedure is summarized in Table 11.7.

TABLE 11.7 Regional Flood Frequency Analysis Steps

1. Define $Q_i(F)$ as the annual maximum flood at site i with cumulative probability F (growth curve).
2. Assume $Q_i(F) = \mu_i q(F)$, where μ_i is a site mean annual flow and $q(F)$ is a regional growth curve.
3. Estimate μ_i by $\hat{\mu}_i = \overline{Q}_i$ (sample mean annual flow).
4. Fit distribution to

   $$q_{ij} \equiv Q_{ij} / \overline{Q}_{ij} \qquad j = 1, \ldots n_i; \qquad i = 1, \ldots N,$$

   where n_i is the number of data points at site i.
5. Let $\hat{q}(F)$ be the estimated inverse cumulative distribution function of the scaled streamflows.
6. The quantile estimator for site i is $\hat{Q}_i(F) = \hat{\mu}_i \hat{q}(F)$.

Source: Adapted from J. R. M. Hosking and J. R. Wallis, "Regional Flood Frequency Analysis Using the Log Normal Distribution," *EOS* 67(44), 1986.

Hosking and Wallis [1986a] suggest that if a three-parameter log-normal distribution is fitted to the regional flows (using a method called probability weighted moments), the results of the index method are surprisingly good and accurate. Using numerical experiments, they show consistently low mean square errors and bias of estimation of floods of various probabilities of exceedance.

Figure 11.14 is one example where real annual maxima are assumed to be distributed with an extreme-value distribution, Type I. Root mean square errors and bias of estimating the 1000-year flood are shown for 21 sites. The coefficients of variation (σ/μ) were 0.5, the skewness 1.14, and the number of data points in each site varied from 10 to 30. Sites were uncorrelated. Tested were assumptions for generalized extreme value distributions at each site (GEV/AS), log-normal distribution at each site (LN3/AS), generalized extreme value as a regional distribution (GEV/R), a Wakeby distribution for the region (WAK/R), and a log-normal distribution for the region (LN3/R). It is surprising how well the latter technique performs even though the assumed distribution, log-normal, is different than the true underlying distribution, which was extreme-value, Type I. Figure 11.15 is even more striking. Discharges of various recurrences Q_T are computed under the assumption that the underly-

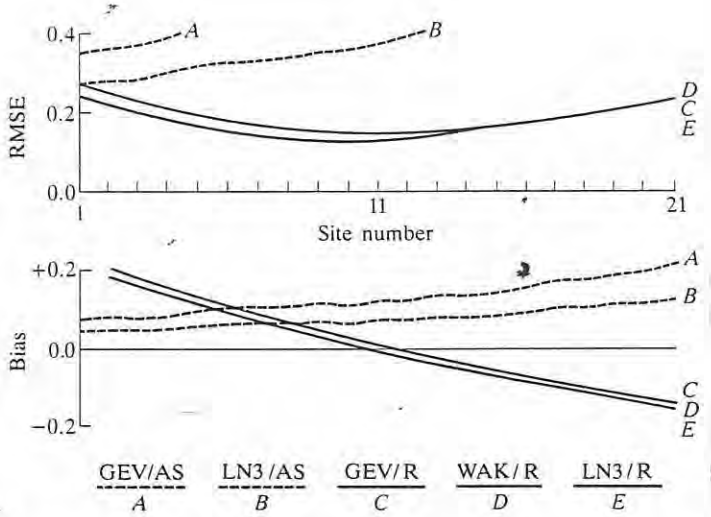


FIGURE 11.14 Estimating the 1000-year flood in 21 sites by different on-site and regional procedures when the population follows an extreme-value distribution, Type I. RMSE denotes root mean square error, GEV generalized extreme value, AS at site, LN3 log-normal distribution, R regional distribution, and WAK Wakeby distribution. Source: J. R. M. Hosking and J. R. Wallis, "Regional Flood Frequency Analysis Using the Log Normal Distribution," *EOS* 67(44), 1986. Copyright by the American Geophysical Union.

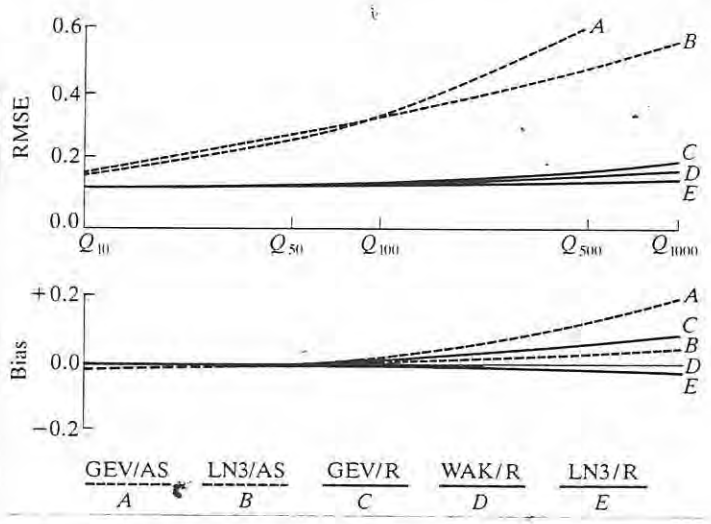


FIGURE 11.15 Estimation of different quantiles when population is homogeneous and log-normal. Different on-site and regional methods are compared. Abbreviations are explained in the legend to Figure 11.14. Source: J. R. M. Hosking and J. R. Wallis, "Regional Flood Frequency Analysis Using the Log Normal Distribution," *EOS* 67(44), 1986. Copyright by the American Geophysical Union.

ing true distribution is log-normal with skewness of 2.5 and all other characteristics as above. The regional procedure performs better than attempts to fit individual distributions at every site.

Lettenmaier [1988] gives a very good comparison of regionalizing procedures. The interested reader is urged to study that reference. He concludes that "regionalization is the most viable way of improving flood quantile estimation." Particularly when the flexibility of three-parameter distributions is required, Lettenmaier states that "the reduction in the variability of flood quantile estimators achieved by proper regionalization is so large that at site estimators should not be seriously considered." A quantile is the value of discharge with a given probability of exceedance.

11.5.3 Paleohydrology and the Value of Historical Information

The Water Resources Council also recognized the value of using historical (vs. systematic record) data. The historical record can consist of written or unwritten accounts of past floods or fragmented records of past civilizations. Information is also available in the geology or related time series like varves, tree rings, etc.

Recently there have been considerable advances in methodology and evaluation of methods to include historical information. Hosking and Wallis [1986b, 1986c] address the information of both paleohydrology and historical data in relation to systematic records. In their work, paleologic information is measured in thousands of years and historical information in hundreds of years. Otherwise there is no significant difference in treatment. Their approach is a [Monte Carlo] simulation where they assume that they have only one historical or paleological maximum event in a period of m years. They process the historical data using the incomplete data likelihood (a maximum-likelihood procedure) and account for errors in estimating the magnitude of the historical maximum event. Typical of their results are Figures 11.16 and 11.17. Their conclusions can be summarized as

1. Historical and paleohydrologic information can be valuable in estimating three-parameter distributions at a single site. It is much less effective when dealing with two-parameter distributions.
2. The effectiveness of historical information reduces with increased sample length.
3. The effectiveness of historical information reduces with increased error in estimating the magnitude of the event. Yet improvements are sometimes observed even in cases when the paleologic maximum event is subject to an error of ±50%.

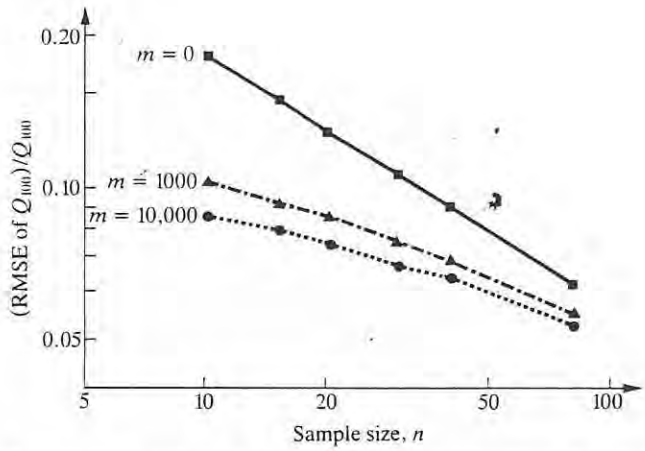


FIGURE 11.16 (Root mean square error of $Q_{100})/Q_{100}$ as a function of gaged record length n and historic period m of paleologic maximum event. Single-site analysis, parent distribution is an extreme-value, Type I (EVI) with coefficient of variation of 0.4 and fitted distribution EVI. Source: J. R. M. Hosking and J. R. Wallis, "Paleoflood Hydrology and Flood Frequency Analysis," *Water Resources Res.* 22(4):543–550, 1986. Copyright by the American Geophysical Union.

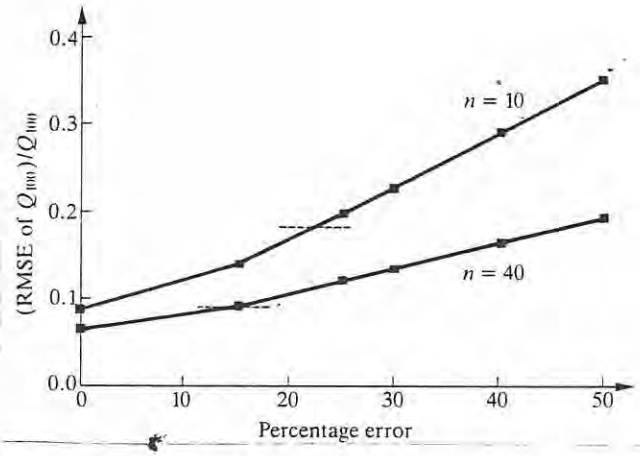


FIGURE 11.17 (Root mean square error of Q_{100})/Q_{100} as a function of measurement error of 10,000-year paleologic maximum event. Dashed lines denote (root mean square error of Q_{100})/Q_{100} when there are no paleologic data; n is length of gaged record. Single-site analysis, parent distribution EVI with coefficient of variation of 0.4 and fitted distribution EVI. Source: J. R. M. Hosking and J. R. Wallis, "Paleoflood Hydrology and Flood Frequency Analysis," *Water Resources Res.* 22(4):543–550, 1986. Copyright by the American Geophysical Union.

4. In a regional analysis using a large number of sites the inclusion of a realistic amount of historical information is unlikely to be useful in practice. Paleologic information in small regions with short records is worthwhile and improves flood estimates even at sites where no paleologic event has been observed, but this can exacerbate biases in estimates.

Stedinger and Cohn [1986] interpreted paleologic or historical records as parts of a censored data set that would include the systematic information. The records are censored in the sense that historical or paleologic information corresponds to events that exceed a given threshold discharge of the given probability of exceedance. Censored data can be of two types, in one there is knowledge of the magnitudes as well as the occurrence of the events, in the other no magnitude information is known. Stedinger and Cohn use likelihood functions corresponding to the two types of data records.

A simulation experiment was again used to evaluate procedures. Reality was assumed to follow a two-parameter, log-normal distribution, and samples were fitted with the same distribution. Their results are well represented in Figures 11.18 and 11.19. The figures give the equivalent number of systematic years of data achieved by using a given historical record length (over and above 20 years of systematic record) for two threshold levels (90% for 10-year recurrence and 99% for 100-year recurrence). It is concluded that

1. Historical information can be very effective in augmenting systematic records.

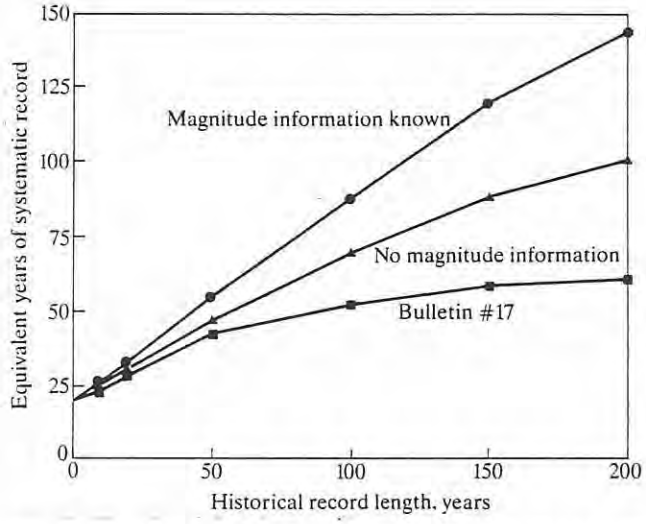


FIGURE 11.18 Effective record length of the two maximum-likelihood estimators and the Bulletin 17 procedures when estimating the 100-year flood. The cases have a 20-year systematic record, a censoring threshold at the 90th percentile of the flow distribution (i.e., 10-year recurrence flood) and between 0 and 200 years of historical information. Source: J. R. Stedinger and T. A. Cohn, "Flood Frequency Analysis with Historical and Paleoflood Information," *Water Resources Res.* 22(5):785–793, 1986. Copyright by the American Geophysical Union.

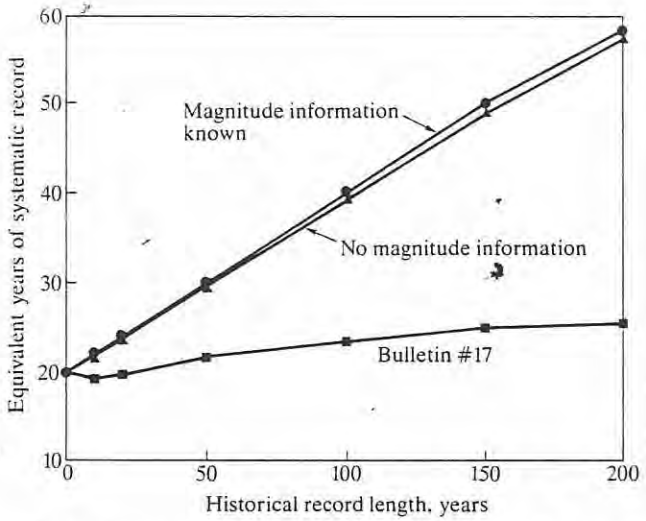


FIGURE 11.19 Effective record length of the two maximum-likelihood estimators and the Bulletin 17 estimator when estimating the 100-year flood. The cases have a 20-year systematic record, a censoring threshold at the 99th percentile of the flow distribution (i.e., 100-year recurrence flood) and between 0 and 200 years of historical information. Source: J. R. Stedinger and T. A. Cohn, "Flood Frequency Analysis with Historical and Paleoflood Information," *Water Resources Res.* 22(5):785–793, 1986. Copyright by the American Geophysical Union.

2. The maximum-likelihood procedures are far more effective than those recommended by the Water Resources Council to handle historical information.
3. Effectiveness increases with higher thresholds.
4. Magnitude information is less important as the censoring threshold increases.
5. The procedures showed robustness, even when reality was assumed to be other than log-normal (i.e., log-Pearson Type III).

Readers interested in plotting position formulas including historical information are referred to Hirsch [1987].

11.6 SUMMARY

Probability and statistics are integral tools of hydrology. As in many natural systems, hydrologic processes are never predictable in exact, deterministic, ways; hence, the need for probabilistic models. The hydrologist also depends on data collection and handling; hence, the need for statistics. You may ask why it took 10 chapters of deterministic conceptualizations to get to the important probability concept. There are two reasons. One is the philosophical belief that a good understanding of physical, deterministic, mechanisms, and their strengths and failings, is the best way to appreciate and ultimately intelligently use probabilistic models. In the end, probabilistic and deterministic thinking must converge and ideally be closely related. After all, they are intended to represent the same phenomena. This convergence of concepts is occurring. Recently, it is most apparent in the study of nonlinear chaotic systems—deterministic phenomena that look stochastic.

The second reason for concentrating on deterministic concepts in this introductory course is that probabilistic thinking is generally harder to accept and understand. The reason for this is not inherent in the mathematics or the ideas but in our system of education, which emphasizes determinism during most formative years. It takes a lot of time and maturity to break the deterministic habit.

This chapter introduced basic ideas of probability theory as applied to random variables. It assumes that the reader has some familiarity with the topics. Most important is the origin of the concept of recurrence, which ties together models of discrete and continuous random variables. The chapter does provide a lot of useful and pragmatic tools for everyday hydrology.

The study of random functions, i.e., random processes in time and space, is not covered, except for a brief hint in the problem set. Most hydrologic variables are indeed random processes. Examples are rainfall in time, daily streamflow, evaporation in time, rainfall distribution in space, and soil hydraulic properties in space. This topic is the next level of study for the serious hy-

drologist. A general, good, random processes book is Parzen [1962]. For hydrologic emphasis at an advanced level see Bras and Rodriguez-Iturbe [1985]. Other useful books are Haan [1977], McCuen and Snyder [1986], and Kottegoda [1980].

Statistics deals with the tools of data handling and parameter estimation in particular. It is barely touched in this chapter. The field is extensive and specialized. The references are innumerable. Some of the references given above present the most common statistical concepts in hydrology.

REFERENCES

Benjamin, J. R., and C. A. Cornell [1970]. *Probability, Statistics and Decision for Civil Engineers*. New York: McGraw-Hill.

Bobée, B. B. [1975]. "The Log-Pearson Type III Distribution and Its Application in Hydrology." *Water Resources Res.* 11(5):681–689.

Bobée, B. B., and R. Robitaille [1977]. "The Use of the Pearson Type III and Log-Pearson Type III Distribution Revisited." *Water Resources Res.* 13(2):427–443.

Bras, R. L., and I. Rodriguez-Iturbe [1985]. *Random Functions and Hydrology*. Reading, Mass.: Addison–Wesley.

Burges, S. J., D. P. Lettenmaier, and C. L. Bates [1975]. "Properties of the Three-Parameter Log-Normal Probability Distribution." *Water Resources Res.* 11(2):229–235.

Chan, S.-O., and R. L. Bras [1979]. "Urban Storm Water Management: Distribution of Flood Volumes." *Water Resources Res.* 15(2):371–382.

Chow, V. T. [1953]. "Frequency Analysis of Hydrologic Data with Special Application to Rainfall Intensities." Urbana, Ill.: University of Illinois Bulletin. 5(31).

Idem. [1964]. *Handbook of Applied Hydrology*. New York: McGraw-Hill.

Diaz-Granados, M. S., J. B. Valdes, and R. L. Bras [1984]. "A Physically Based Flood Frequency Distribution." *Water Resources Res.* 20(7):995–1002.

Eagleson, P. S. [1972]. "Dynamics of Flood Frequency." *Water Resources Res.* 8(4):878–898.

Idem. [1978]. "Climate, Soil, and Vegetation: 1. Introduction to Water Balance Dynamics." *Water Resources Res.* 14(5):705–712.

Gumbel, E. J. [1958]. *Statistics of Extremes*. New York: Columbia University Press.

Haan, C. T. [1977]. *Statistical Methods in Hydrology*. Ames, Iowa: Iowa State University Press.

Hald, A. [1952]. *Statistical Tables and Formulas*. New York: Wiley.

Hebson, C., and E. F. Wood [1982]. "A Derived Flood Frequency Distribution Using Horton Order Ratios." *Water Resources Res.* 18(5):1509–1518.

Hirsch, R. M. [1987]. "Probability Plotting Position Formulas for Flood Records with Historical Information." In: *Analysis of Extraordinary Flood Events*, Special Issue of the *Journal of Hydrology*, 96(1–4):185–199.

Hosking, J. R. M., and J. R. Wallis [1986a]. "Regional Flood Frequency Analysis Using the Log Normal Distribution." *EOS*. 67(44).

Idem. [1986b]. "Paleoflood Hydrology and Flood Frequency Analysis." *Water Resources Res.* 22(4):543–550.

Idem. [1986c]. "The Value of Historical Data in Flood Frequency Analysis." *Water Resources Res.* 22(11):1606–1612.

Idem. [1988]. "The Effect of Intersite Dependence on Regional Flood Frequency Analysis." *Water Resources Res.* 24(4):588–600.

Howard, C. [1976]. "Theory of Storage and Treatment-Plant Overflows." *J. Environ. Engin. Div., A.S.C.E.* 102(EE4):709–722.

Keeney, R. L., and E. F. Wood [1977]. "An Illustrative Example of the Use of Multiattribute Utility Theory for Water Resources Planning." *Water Resources Res.* 13(4):705–712.

Kottegoda, N. T. [1980]. *Stochastic Water Resources Technology.* New York: Wiley.

Landwehr, J. M., N. C. Matalas, and J. R. Wallis [1978]. "Some Comparisons of Flood Statistics in Real and Log Space." *Water Resources Res.* 14(5):902–920.

Lenton, R. L., I. Rodriguez-Iturbe, and J. C. Schaake, Jr. [1974]. "The Estimation of ρ in the First-Order Autoregressive Model: A Bayesian Approach." *Water Resources Res.* 10(2):227–241.

Lettenmaier, D. P. [1988]. "Evaluation and Testing of Flood Frequency Estimation Methods," Proceedings of the International Workshop on Natural Disasters in European Mediterranean Countries, Villa Colombella, Perugia, Italy, June 27–July 1, 1988.

McCuen, R. H., and W. M. Snyder [1986]. *Hydrologic Modeling: Statistical Methods and Applications.* Englewood Cliffs, N.J.: Prentice-Hall.

Moughamian, M. S. [1986]. "Flood Frequency Estimation: A Testing and Analysis of Physically Based Models." Cambridge, Mass.: MIT Department of Civil Engineering. M.S. thesis.

National Bureau of Standards [1953]. "Probability Tables for the Analysis of Extreme-Value Data." *Appl. Math Series* 22. Washington, D.C.

Parzen, E. [1962]. *Stochastic Processes.* San Francisco: Holden Day.

Rao, D. U. [1980]. "Log-Pearson Type III Distribution — Method of Mixed Moments." *J. Hydraul. Div. A.S.C.E.* 106(HY6):999–1019.

Russell, S. O. [1982]. "Flood Probability Estimation." *J. Hydraul. Div. A.S.C.E.* 108(HY1):63–72.

Stedinger, J. R. [1980]. "Fitting Lognormal Distributions to Hydrologic Data." *Water Resources Res.* 16(2):481–490.

Stedinger, J. R., and T. A. Cohn [1986]. "Flood Frequency Analysis with Historical and Paleoflood Information." *Water Resources Res.* 22(5):785–793.

Tasker, G. D. [1978]. "Flood Frequency Analysis with a Generalized Skew Coefficient." *Water Resources Res.* 14(2):373–376.

Thomas, H. R., Jr. [1948]. "Frequency of Minor Floods." *Boston Soc. Civil Eng.* 34:425–442.

U.S. Water Resources Council [1967]. "A Uniform Technique for Determining Flood Flow Frequencies." *Water Resources Council Bull.* 15. Washington, D.C.

Idem. [1976]. "Guidelines for Determining Flood Flow Frequency." *Water Resources Council Bull.* 17. Washington, D.C.

Idem. [1977]. "Guidelines for Determining Flood Flow Frequency." *Water Resources Council Bull.* 17A. Washington, D.C.

U.S. Geological Survey [1982]. "Guidelines for Determining Flood Flow Frequency." *Water Resources Council Bull.* 17B. Washington, D.C.

Wallis, J. R., N. Matalas, and J. R. Slack [1974]. "Just a Moment!" *Water Resources Res.* 10(2):211–219.

Wood, E. F. [1978]. "Analyzing Hydrologic Uncertainty and Its Impact Upon Decision Making in Water Resources." *Adv. Water Resources*. 1(5):299–306.

Wood, E. F., and I. Rodriguez-Iturbe [1975a]. "Bayesian Inference and Decision Making for Extreme Hydrologic Events." *Water Resources Res.* 11(4):533–542.

Idem. [1975b]. "A Bayesian Approach to Analyzing Uncertainty Among Flood Frequency Models." *Water Resources Res.* 11(6):839–843.

PROBLEMS

1. Chapter 4 briefly mentioned intensity–frequency–duration curves in the analysis of precipitation. These are curves that show the relationship between rainfall intensity and recurrence for a given duration. The intensity is higher for storms of larger recurrence intervals (rare, low-probability storms) for a fixed duration. For a fixed recurrence the intensity decreases with duration. In order to obtain intensity–frequency–duration (IFD) curves, the hydrologist groups storms by duration and analyzes each duration group as a realization (i.e., a group of possible values) of a random variable. The recurrence or the probability of exceedance is estimated for each value in the set of equal-duration storms by fitting a distribution to the set or by using nonparametric plotting position concepts as discussed in this chapter.

Given the following data, compute and draw the intensity–frequency–duration curves:

AVERAGE INTENSITY (in. hr^{-1})	DURATION (hr)	AVERAGE INTENSITY (in. hr^{-1})	DURATION (hr)
0.26	12	0.68	6
0.48	6	0.75	6
0.13	24	1.10	6
0.85	3	1.40	6
1.00	3	2.10	3
1.80	3	2.40	2
0.34	24	3.60	1
0.28	24	4.20	1
0.40	24	3.40	1
0.21	24	2.40	1
0.32	12	2.30	1
0.58	6	1.10	3
1.50	2	1.50	2
2.10	1	1.60	2
1.80	1	1.15	3
1.15	2	0.18	24
0.11	24	0.19	24
0.55	12	0.30	24
0.36	12	0.27	24

(*continued*)

AVERAGE INTENSITY (in. hr^{-1})	DURATION (hr)	AVERAGE INTENSITY (in. hr^{-1})	DURATION (hr)
0.34	12	0.48	12
0.64	6	0.60	12
0.52	12	0.42	12
1.00	6	0.40	12
0.93	6	0.25	24
0.78	6	1.60	3
1.25	3	1.40	3
2.10	2	1.30	3
1.70	2	3.20	1
2.00	2	3.05	1
3.00	1	3.40	1
2.60	1	2.30	2
0.90	6	2.00	2
0.85	6	1.85	2
0.80	6	2.70	1
0.44	12	2.90	1
1.70	2	1.50	3
1.40	6	1.45	3
0.23	24	0.24	24
0.67	12	0.47	2
2.80	2		

Fit a function to the intensity–frequency–duration curves.

2. You are building a dam in a U.S. river with 50 years of record. The base 10 logarithms of the annual maxima series have a mean of 0.6, a standard deviation of 0.3, and a skewness coefficient of 0.6. The original data is in cubic meters per second. During construction you want to build a buffer dam to divert the river. You want to design the buffer dam so that the probability that the dam will fail in any of the five years is 0.1. What is the necessary design recurrence and the magnitude of the design flood? (*Hint:* The exceedance probability in any one year is between 0.025 and 0.015.)

3. The annual minimum rate of discharge of a particular river is thought to have the Type III extreme-value distribution for the smallest value:

$$F_Z(z) = 1 - \exp\left[-\left(\frac{z-\varepsilon}{u-\varepsilon}\right)^2\right] \qquad z \geq \varepsilon \geq 0.$$

a) If the observed first and second moments of annual minimum discharge for the river are 350 ft^3 s^{-1} and (160)^2 ft^6 s^{-2}, find the probability that the annual minimum runoff is below 100 ft^3 s^{-1}.

b) What is the probability of having less than two droughts (annual flows less than 100 ft^3 s^{-1}) in 50 years?

c) If 100 $ft^3 s^{-1}$ is the magnitude of the lowest runoff value in a series of 30 years, give a nonparametric answer to the question in part b. (*Hint:* The first and second moments of the Type III distribution are $m_z = \varepsilon + (u - \varepsilon)(\sqrt{\pi}/2)$ and $\sigma_z^2 = (u - \varepsilon)^2 (1 - \pi/4)$.)

4. The 24-hour-duration storm depth at a site follows a Gumbel distribution with mean of 83 mm and a standard deviation of 30 mm. What is the 100-year 24-hour-duration storm at the site? (See Problem 1 and Chapter 4 for a description of intensity–frequency–duration curves.)

5. An impounding reservoir is designed so as to have sufficient capacity to meet the water requirements of a city during a drought year in which the mean rate of streamflow during the filling period is 4.5 $ft^3 s^{-1}$. During 30 years of records, the lowest flow observed during the filling period was 3.8 $ft^3 s^{-1}$ and the second lowest 4.5 $ft^3 s^{-1}$. Find the probability that during the next 20 years a flow less than 4.5 $ft^3 s^{-1}$ will occur two or more times. (*Hint:* In nonparametric analysis, floods and droughts are treated the same way.)

6. In 1958, the 50-year flood was estimated to be of a particular size. In the next 10 years, two floods were observed in excess of that size. If the original estimate was correct, what is the probability of such an observation?

7. Exceedance series—flows greater than some arbitrary base value—have been treated as random variables with a shifted exponential distribution. For 23 years of records during which 56 flows in excess of 5500 $ft^3 s^{-1}$ were observed, the following density function was found adequate:

$$f_X(x) = \left(\frac{1}{6500}\right) e^{-(x-5500)/6500} \quad x \geq 5500,$$

where X is the peak-flow magnitude given a flood in excess of 5500 occurred.

a) Sketch the density function.

b) Compute the probability that a flow in excess of 20,000 $ft^3 s^{-1}$ will be observed given that the flow exceeds the base value of 5500 $ft^3 s^{-1}$.

c) Compute the flow above 5500 $ft^3 s^{-1}$ such that the probability of being exceeded is 1/50. Is this the 50-year flow?

8. A temporary cofferdam is to be built to protect the five-year construction activity for a major cross-valley dam. If the cofferdam is designed to withstand the 20-year flood, what is the risk that the structure will be overtopped (a) in the first year; (b) in the third year exactly; (c) at least once in the five-year construction period; and (d) not at all during the five-year period?

9. Complete the following mathematical statements about the properties of a probability density function by matching the statements on the left with the correct item number on the right. Assume x is a series of annual occurrences from a normal distribution.

a) $\int_{-\infty}^{\infty} F(x)\,dx =$	1. $P[m_1 \geq x \cup x \geq m_2]$
b) $\int_{-\infty}^{m_1} f(x)\,dx =$	2. unity
c) $\int_{m_1}^{m_2} f(x)\,dx =$	3. median
d) $\int_{-\infty}^{\square} f(x)\,dx = 0.5$	4. $P[x \leq m_1]$
e) $1 - \int_{m_1}^{m_2} f(x)\,dx =$	5. standard deviation
	6. $p(m_1 \leq x \leq m_2)$

10. A temporary flood wall has been constructed to protect several homes in the flood plain. The wall was built to withstand any discharge up to the 20-year flood magnitude. The wall will be removed at the end of the three-year period after all the homes have been relocated. Determine the probability that (a) the wall will be overtopped in any year, (b) the wall will not be overtopped during the relocation operation, (c) the wall will be overtopped at least once before all of the homes are relocated, (d) the wall will be overtopped exactly once before all the homes are relocated, and (e) the wall will be adequate for the first two years and overtopped in the third year.

11. A cofferdam designed to withstand flows up to and including the 1944 flow of 1870 $ft^3 s^{-1}$ is planned for the Middle Branch Westfield River. Investigate the degree of protection afforded by the dam of this size during a five-year program of channel improvement and dam construction. (The 1944 flood was the sixth largest during a 25-year record from 1921 to 1945.)

12. The annual maximum rate of discharge of a particular river is thought to have the Type I extreme-value distribution, i.e., Gumbel, with mean 10,000 $ft^3 s^{-1}$ and standard deviation 3000 $ft^3 s^{-1}$.

a) Compute P[annual maximum discharge $\geq$ 15,000 $ft^3 s^{-1}$].

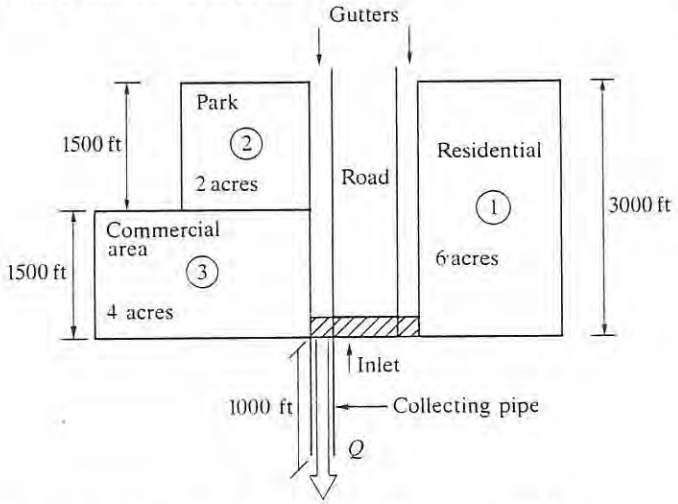
b) Find an expression for the cumulative density function of the river's maximum discharge during the 20-year lifetime of an anticipated flood-control project. Assume that the individual annual maxima are mutually independent random variables.

c) What is the probability that the maximum of 20 years will exceed 15,000 $ft^3 s^{-1}$? *Hint:* The Gumbel cumulative density function is

$$F_Y(y) = \exp[-e^{-\alpha(y-u)}] \quad -\infty \leq y \leq \infty$$

$$m_Y = u + \frac{0.577}{\alpha} \qquad \sigma_Y^2 = \frac{1.645}{\alpha^2}.$$

13. Using the rainfall data and the sketch of an urban development given below, what is the discharge corresponding to a recurrence interval of about seven years? In the analysis you can ignore the road as an area contributing to runoff. See Problem 1 and Chapter 4 for a discussion of intensity–frequency–duration curves.



$t_{c_1} = 20$ min C (Residential) $= 0.7$

$t_{c_2} = 15$ min C (Park) $= 0.5$

$t_{c_3} = 25$ min C (Commercial area) $= 0.9$

Velocity in gutter $= 100\ \text{ft min}^{-1}$

Velocity in pipes $= 200\ \text{ft min}^{-1}$

RAINFALL DATA

Intensity (in. hr^{-1})	Duration (min)	Intensity (in. hr^{-1})	Duration (min)
1.19	45	0.28	55
0.65	50	0.40	55
1.12	50	1.05	55
0.47	55	0.37	50
0.61	55	0.30	50
0.32	55	0.32	45
0.31	50	0.29	45
0.36	45	0.69	45
0.45	45	0.40	45
0.53	45	0.34	50
0.42	50	0.27	50
0.50	50	0.29	55
0.24	55	0.26	55
0.35	55	0.33	45
0.27	45	0.26	50

14. In general, does the discharge of a given recurrence correspond to a rainfall event of the same recurrence? Explain and discuss.

15. Show that the mean of the binomial distribution is $E[K] = nP$ and the variance, $\text{Var}[K] = nP(1 - P)$.

16. Show that the mean of the geometric distribution is $1/P$ and the variance is $(1 - P)/P^2$.

17. A very useful model is the Poisson distribution, which is used to represent the number of "events," for example rainstorms, arrivals in a given time period t. The Poisson distribution is given by

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

where λ is rate of arrival of events in time period t (i.e., number of events per unit time). It has units of inverse time. The mean of the Poisson distribution is $\mu = \lambda t$. The variance is also $\sigma^2 = \lambda t$.

The Poisson is hence a one-parameter distribution; the parameter is $\nu = \lambda t$, or for a fixed time period the parameter is simply ν. In introducing time, the Poisson distribution is a small window into the concept of random processes. A random process is a random function of an argument, commonly time. The Poisson represents the process $X(t)$ of the number of events occurring in the time interval $\{0, t\}$. For each fixed value of t, the process $X(t)$ is a random variable, Poisson distributed, with parameter λt.

The Poisson assumption requires that the process in question satisfy the following conditions: (1) The probability of an event occurring in an arbitrary and small time interval Δt is always $\lambda \Delta t$. This is called stationarity, i.e., the distribution does not change with time. (2) Only one event can occur in a short interval Δt, with probability $\lambda \Delta t$. (3) The occurrence of an event in an interval Δt is independent of the occurrence of an event in any other Δt. This is independence as defined at the beginning of this chapter.

Having introduced the Poisson distribution, here are some useful questions. Assume that daily rainfall in Boston follows a Poisson process, hence only one storm per day is allowed per condition 2 above. Storms arrive in Boston every three days in the average summer month.

a) Draw the Poisson distribution for the 30 days of the month of June.

b) What is the probability that four or less storms occur in June?

c) Call the random variable T the time to the first storm arrival in June. Derive the distribution of T. (*Hint:* The time T is exponentially distributed as $f_T(t) = \lambda e^{-\lambda t}$.)

d) What is the distribution of the time between storm arrivals? (*Hint:* Use the independence property of Poisson events and part c.)

e) What are the mean and variance of the exponential distribution?

f) Show that the conditional distribution of the time T to the next event given that no event has occurred up to time t_0 is exponential.

$$f_{T|T>t_0}(\tau) = \lambda e^{-\lambda\tau} \qquad t \geq 0,$$

where $\tau = t - t_0$.

g) What is the distribution of the time to the kth storm? (*Hint:* The time to the kth storm Γ_k is the sum of the times between storms occuring before that time. That is, $\Gamma_k = T_1 + T_2 + \cdots + T_k$. Each of the T_i are independent and exponentially distributed variables with common parameter λ. The answer is the gamma distribution

$$f_\Gamma(t) = \frac{\lambda(\lambda t)^{k-1} e^{-\lambda x}}{(k-1)!} \qquad t \geq 0.$$

This is analogous to the derivation of the Nash model of the instantaneous unit hydrograph. It involves a convolution of distributions.

18. Assume that coastal flooding in an estuary is due to the combination of wind velocity and river discharge in the estuary. The annual maximum wind velocity is taken to obey an exponential distribution:

$$f_W(w) = \lambda_1 e^{-\lambda_1 w}.$$

The annual maximum discharge is also assumed exponential:

$$f_Q(q) = \lambda_2 e^{-\lambda_2 q}.$$

The level of flooding H is proportional to the sum of W and Q. If winds and river discharges are independent, what is the distribution of H? What is the recurrence of the coastal flooding resulting from the 50-year discharge and the 50-year wind?

19. For a river basin near you, find 20 or more years of streamflow records and fit the log-Pearson Type III distribution. Estimate the 50-year flood. Draw confidence limits on your distribution. In the United States, streamflow records are available from the U.S. Geological Survey or its publications. The U.S. Army Corps of Engineers and state agencies also have streamflow records.

20. Simulation is a powerful tool in hydrology and other earth sciences. Monte Carlo simulation refers to the generation of sequences of numbers following a particular probabilistic distribution. Given a probabilistic model, simulation allows the study of possible sequences (and effects) of hydrologic processes like rainfall and discharges. Simulation exercises are a day-to-day tool in practice and research.

If a random variable obeys a cumulative density function

$$P[X \leq x] = F_X(x). \tag{1}$$

The goal of Monte Carlo simulation is to invert the above equation to obtain values of X. Theoretically, we want to solve

$$x = F^{-1}(P), \tag{2}$$

where P is the probability of X being less than or equal to x. Hence, P is uniformly distributed between 0 and 1. Given Eq. (2), all that needs to be done is to introduce values uniformly distributed between 0 and 1 for P, and solve for x. Uniformly distributed values between 0 and 1 are available from most computers or scientific calculators. Alternatively, look at your local telephone directory and add a period before the last three digits of telephone numbers randomly selected!

Analytical solutions of Eq. (2) are sometimes possible. Most times numerical solutions are required.

The concept of Monte Carlo simulation can be used to create models of hydrologic phenomena. For example, assume that storms can be represented as instantaneous pulses of depth D arriving as Poisson events on the average once every 10 days. The depth D is exponentially distributed.

$$f_D(d) = \alpha e^{-\alpha d}.$$

Generate a sequence of five years of storms, showing time of arrival and depth of each event. Assume that the mean depth is 2 mm.

21. In Chapter 4 we saw that large-duration storms have generally small intensities. Let a storm be represented by rectangular pulses of duration t and intensity i (i.e., depth = it). Assume the conditional distribution of intensity i on duration t is of the form

$$f_{I|T}(i|t) = \alpha t e^{-\alpha t i}$$

and that the storm duration follows

$$g_T(t) = \beta e^{-\beta t}.$$

What is the joint distribution of I and T? If $\beta = 0.25\ \text{hr}^{-1}$ and $\alpha = 2\ \text{cm}^{-1}$, what is the probability of obtaining a storm of mean intensity less than or equal to 0.5 cm hr^{-1} and of duration less than or equal to 3 hours? What is the probability of obtaining a storm of total depth less than or equal to 2 cm?

22. A hydrologist is trying to determine the hydraulic conductivity of a formation that he knows must have one of four geologic origins. From his or her hydrogeologic knowledge he or she can assign the following probabilities to each possibility.

GEOLOGIC ORIGIN OR STATE	ASSOCIATED HYDRAULIC CONDUCTIVITY (cm s^{-1})	PROBABILITY
1	10^{-4}	0.1
2	5×10^{-4}	0.2
3	10^{-3}	0.5
4	10^{-2}	0.2

To determine hydraulic conductivity a well test is performed. Such procedures are not completely accurate. From experience the hydrologist can assign the probability that the test indicates geologic state i given that the true state is j. This probability table is

WELL TEST RESULT	TRUE STATE 1	2	3	4
1	0.5	0.2	0.0	0.0
2	0.3	0.6	0.1	0.0
3	0.1	0.1	0.7	0.2
4	0.1	0.1	0.2	0.8

What is the probability mass function of hydraulic conductivity if the well test indicates that the formation is of type 2? If another test is repeated and the test indicates type 1, what is the new probability mass function of hydraulic conductivity?

23. Using concepts and introductory ideas presented in this chapter to use historical information for frequency analysis, how many years of historical record would you need to augment a 20-year systematic record to an equivalent of 40 years if you are estimating the 100-year flood? Assume your historical record contains (a) only the events (no magnitude) larger than the 10-year flood and (b) only the events (no magnitude) larger than the 100-year flood.

24. Probability paper is one on which a graph of a particular cumulative density function (cdf) plots as a straight line. This is achieved by distorting the ordinate scale in a plot of $F_X(x)$ versus x. The goal is to obtain a relationship of the form

$$y = ax + b,$$

where y is a function of $F_X(x)$. What would y (the necessary scale transformation) be for the case of the exponential distribution

$$F_X(x) = 1 - e^{-\lambda x}?$$

Construct probability paper for the normal distribution. (*Hint:* Do it graphically, if necessary, by first drawing the cdf of the normal on arithmetic scales and seeing how it needs to be distorted to produce a straight line.)